Universidad de Valladolid

Facultad de Ciencias

Trabajo de Fin de Grado

Junio 2021

Grado en Matemáticas

Aprendizaje supervisado basado en Procesos Gaussianos

Autor: Elsa López Pérez

Tutor: Eustasio del Barrio Tellado

Índice general

Índice de figuras

CAPÍTULO 1

INTRODUCCIÓN.

Los procesos Gaussianos (GP) constituyen un método de aprendizaje supervisado no paramétrico basado en la inferencia Bayesiana. Pueden emplearse tanto para problemas de regresión como para problemas de clasificación. A diferencia de otros algoritmos de aprendizaje supervisado que buscan valores exactos para ciertos parámetros, la aproximación Bayesiana infiere una distribución de probabilidad sobre todos los posibles valores de éstos. Para ello, se especifica una distribución a priori sobre los parámetros y, mediante el teorema de Bayes, se obtiene una distribución a posteriori que incorpora información tanto de la distribución a priori como de los datos de entrenamiento proporcionados. Por tratarse de un modelo no paramétrico, la regresión basada en modelos Gaussianos calcula la distribución de probabilidad sobre todas las funciones que ajustan los datos. Esto tiene la ventaja de que es posible determinar intervalos de confianza para las distintas funciones obtenidas. La información a priori del proceso viene determinada por las funciones kernel o funciones núcleo. Es esta función la que determina las propiedades de regularidad del proceso.

Pese a ser modelos muy flexibles, el principal inconveniente de los métodos de aprendizaje basados en procesos Gaussianos es el elevado coste computacional de éstos, ya que es necesario retener todos los datos de la muestra de entrenamiento para realizar nuevas predicciones. En el caso de los modelos paramétricos, una vez determinados los valores de los parámetros podemos "deshacernos"de los datos y realizar predicciones únicamente con los valores de éstos.

En el capítulo 2 de este trabajo se incluyen conceptos básicos en relación con la distribución normal, fundamental en el uso de procesos Gaussianos, así como unas nociones básicas sobre procesos estocásticos y modelos basados en la inferencia Bayesiana. Además, se introducen las ideas fundamentales de la regresión lineal, junto con un método de regresión conocido como regresión Ridge. Se estudiarán a continuación las funciones núcleo en el capítulo 3, que constituyen el elemento clave para los modelos de aprendizaje basados en procesos Gaussianos, incluyendo ejemplos comunes de éstas. Veremos que son las funciones núcleo las que determinan las propiedades de continuidad y regularidad de los procesos Gaussianos. Los siguientes dos capítulos, 4 y 5, están destinados al estudio de los problemas de aprendizaje basados en métodos Gaussianos, tanto de regresión como de clasificación, respectivamente. Se introduce posteriormente el concepto de PAC en

el capítulo 6, aprendizaje probablemente aproximadamente correcto, a través del cual se pueden estimar cotas de error de los algoritmos que emplean procesos Gaussianos. Por último, en el capítulo 7 se realizarán simulaciones de aprendizaje en MATLAB empleando métodos Gaussianos para estudiar la eficacia y el coste computacional de éstos. La primera simulación trata de predecir la concentración de CO_2 en Mauna Loa, un volcán situado en Hawai. Se trata de un ejemplo que permite una visualización clara del problema de regresión. A continuación, se muestra un ejemplo sencillo de modelo Gaussiano para un problema en tres dimensiones. Por último, se plantea un problema real de regresión. Dicho problema consiste en predecir los torques asociados a las diferentes articulaciones de un brazo robótico. En este caso, será necesario solventar uno de los principales problemas de los modelos de regresión basados en procesos Gaussianos: el elevado coste computacional derivado de muestras de entrenamiento de gran tamaño, debido a la necesidad de retener todas éstas para cada uno de los cálculos.

CAPÍTULO 2

RESULTADOS PREVIOS.

El objetivo de este capítulo es introducir conceptos básicos empleados en el aprendizaje estadístico. En la sección 2.1 se incluyen nociones básicas sobre la distribución normal multivariante. Posteriormente, en la sección 2.2 se presentan algunos conceptos básicos en relación con los procesos estocásticos. A continuación, en la sección 2.3 se presenta la forma general de definir la probabilidad condicionada, de manera que podamos tratar indistintamente modelos con o sin densidad asociada. En la sección 2.4 se introducen las ideas fundamentales de la inferencia Bayesiana en problemas de regresión, en contraposición con el modelo clásico. A continuación, se introduce un concepto muy empleado en probabilidad e informática, la divergencia de Kullback Leibler, 2.5, que es una medida no simétrica de la similitud o diferencia entre dos funciones de distribución de probabilidad. Éste será importante en el capítulo 6, en el que trataremos el aprendizaje PAC (aprendizaje probablemente aproximadamente correcto). Por último, en las secciones 2.6 y 2.7 se incluyen los conceptos básicos de la regresión lineal y los fundamentos de la regresión Ridge, respectivamente.

2.1. Distribución normal.

En este trabajo nos centraremos en un tipo concreto de procesos estocásticos: los procesos Gaussianos. Con ese fin, introducimos algunos conceptos básicos relacionados con la distribución normal. Este tipo de distribución es la distribución de probabilidad continua por antonomasia. Se trata de la campana de Gauss que fue introducida por el propio Gauss en 1797 como modelo para explicar la distribución de errores en observaciones astronómicas. En este experimento, mostró que realizando un número grande de observaciones de una constante astronómica desconocida se obtenían valores cuya distribución muestral parecía aproximarse a la distribución normal.

Recordamos que una variable aleatoria X sigue una distribución normal de parámetros μ y σ^2, $\mu \in \mathbb{R}$, $\sigma > 0$, y se escribe $X \sim \mathcal{N}(\mu, \sigma^2)$, si la distribución de probabilidad de X viene dada por la función de densidad

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty < x < \infty. \tag{2.1}$$

Es posible extender esta definición al caso de una distribución normal multivariante. Para

ello, es necesario introducir previamente algunos conceptos, como es el de matriz semidefinida positiva. Así pues, diremos que una matriz $A \in \mathcal{M}_{n\times n}$ con entradas A_{ij} es simétrica semidefinida positiva si $A_{ij} = A_{ji}$ para todos i, j y se verifica que $\mathbf{v}^T A\mathbf{v} \geq 0$ para todo $\mathbf{v} \in \mathbb{R}^n$,

$$\mathbf{v}^T A\mathbf{v} = \sum_{i=1}^{n}\sum_{j=1}^{n} vi A_{ij} v_j \geq 0. \tag{2.2}$$

Supongamos que tenemos un vector aleatorio n-dimensional $\mathbf{X} = [X_1, ..., X_n]^T$. Entonces el vector de medias de $\mathbf{X}$ viene dado por $\mu = [E(X_1), ..., E(X_n)]$ y la matriz de covarianzas es la matriz $\Sigma = [\sigma_{i,j}]_{1\leq i,1\leq j}$ con $\sigma_{i,j} = \mathrm{Cov}(X_i, X_j)$. A partir de las propiedades de linealidad de la esperanza y la covarianza, es posible obtener el vector de medias y la matriz de covarianzas de una trasformación lineal de un vector aleatorio.

Proposición 2.1.1. *Si $\mathbf{X} = [X_1, ..., X_n]^T$ es un vector aleatorio con vector de medias μ y matriz de covarianzas $\Sigma \in \mathcal{M}_{m\times n}(\mathbb{R})$ entonces:*

a) $A\mathbf{X}$ tiene vector de medias $A\mu$ y matriz de covarianzas $A\Sigma A^T$. En particular, si $\mathbf{a} \in \mathbb{R}^n$, entonces $\mathbf{a}^T\mathbf{X}$ tiene media $\mathbf{a}^T\mu$ y varianza $\mathbf{a}^T\Sigma\mathbf{a}$.

b) Σ es una matriz semidefinida positiva.

Por otro lado, una propiedad importante de la distribución de un vector aleatorio es que está totalmente determinada por la distribución de sus proyecciones unidimensionales $\mathbf{a}^T\mathbf{X}, \mathbf{a} \in \mathbb{R}^n$. (Véase el Teorema de Cramér-Wold, [10, Teorema 29.4.]). Esto nos permite dar la siguiente definición:

Definición 2.1.1. *Diremos que $\mathbf{X} = [X_1, ..., X_n]^T$ sigue una distribución normal si $\mathbf{a}^T\mathbf{X}$ sigue una distribución normal para cada $\mathbf{a} \in \mathbb{R}^n$.*

De acuerdo con esta definición, si un vector aleatorio $\mathbf{X} = [X_1, ..., X_n]^T$ tiene una distribución normal, con vector de medias μ y matriz de covarianzas Σ, entonces dado $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{a}^T\mathbf{X}$ seguirá una distribución normal de media $\mathbf{a}^T\mu$ y varianza $\mathbf{a}^T\Sigma\mathbf{a}$. Otro vector aleatorio normal tendrá la misma distribución que $\mathbf{X}$ si y sólo si las medias y las varianzas de sus proyecciones coinciden con las de $\mathbf{X}$, y por tanto, si y sólo si su vector de medias es μ y su matriz de covarianzas es Σ. Así pues, al igual que ocurre en el caso unidimensional, el vector de medias y la matriz de covarianzas determinan por completo la distribución normal. Denotamos entonces por $\mathbf{X} \sim \mathcal{N}_n(\mu, \Sigma)$ para indicar que $\mathbf{X}$ sigue una distribución normal multivariante n-dimensional. El siguiente resultado nos permitirá introducir la función de densidad asociada a una distribución normal multivariante.

Proposición 2.1.2. *Sea $\mathbf{X} = [X_1, ..., X_n]^T$ un vector aleatorio n-dimensional. Entonces:*

- *$\mathbf{X} \sim \mathcal{N}_n(\mu, K)$ si y sólo si $\mathbf{a}^T\mathbf{X} \sim \mathcal{N}(\mathbf{a}^T\mu, \mathbf{a}^T\Sigma\mathbf{a})$.*
- *si $\mathbf{X} \sim \mathcal{N}_n(\mu, K)$, $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ y $\mathbf{b} \in \mathbb{R}^m$, entonces $A\mathbf{X}+\mathbf{b} \sim \mathcal{N}_m(A\mu+\mathbf{b}, A\Sigma A^T)$.*

Como consecuencia de este resultado, tenemos que dado $Z = (Z_1, ..., Z_n)$ un vector aleatorio, con $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$, entonces, $Z \sim \mathcal{N}(0, I_n)$. Por consiguiente, si definimos $\Sigma = AA$, con A una matriz simétrica, entonces tenemos que la variable X dada por $X = AZ+\mu$ sigue una distribución normal n-dimensional de media μ y varianza Σ; es decir, $X \sim \mathcal{N}(\mu, \Sigma)$. Por consiguiente, la función de densidad asociada tendrá la siguiente expresión:

$$f_X(x_1, ..., x_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right). \tag{2.3}$$

La distribución normal multivariante es importante porque aparece de manera natural en muchas ocasiones. En concreto, el teorema central del límite establece que bajo ciertas condiciones, la suma de m variables aleatorias se aproxima a una distribución normal cuando m toma un valor suficientemente grande.

Un concepto importante que aparecerá en los problemas de aprendizaje basados en la inferencia Bayesiana es el de la probabilidad condicionada, que será estudiado en más detalle en la sección 2.3. En ocasiones, dada la distribución conjunta de dos vectores aleatorios $\mathbf{a}$ y $\mathbf{b}$ estamos interesados en conocer la distribución de $\mathbf{a}$ condicionada por $\mathbf{b}$, o viceversa. Adaptando la notación al caso de los problemas de aprendizaje, supongamos que tenemos un vector $\mathbf{f}$ que contiene los valores de las observaciones y un vector $\mathbf{x}_*$ que contiene datos de prueba . Queremos entonces predecir el valor de $\mathbf{f}^* = f(\mathbf{x}^*)$, basándonos en las observaciones $\mathbf{f}$. Es decir, buscamos $p(\mathbf{f}^*|\mathbf{f})$. El siguiente resultado nos muestra cómo obtener la distribución de probabilidad condicionada a partir de la distribución de probabilidad conjunta [34, Apéndice A.2].

Proposición 2.1.3. *Sean $\mathbf{a} \in \mathbb{R}^n$ y $\mathbf{b} \in \mathbb{R}^m$ vectores aleatorios con una distribución Gaussiana, tales que*

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix} \right),$$

donde $\mu_a \in \mathbb{R}^n$, $\mu_b \in \mathbb{R}^m$ son los vectores de media, $A \in \mathbb{R}^{n\times n}$ y $B \in \mathbb{R}^{m\times m}$ son las matrices de covarianza, siendo A definida positiva, y $C \in \mathbb{R}^{n\times m}$. Entonces, la distribución de $\mathbf{b}$ condicionada por $\mathbf{a}$ es

$$\mathbf{b}|\mathbf{a} \sim \mathcal{N}(\mu_b + C^T A^{-1}(\mathbf{a} - \mu_a), B - C^T A^{-1} C).$$

2.2. Procesos estocásticos.

En esta sección introducimos algunos conceptos básicos relacionados con los procesos estocásticos. En la literatura es habitual denotar por T al conjunto de índices para enfatizar que se trata de un espacio temporal. De esta manera, el proceso estocástico en sí suele denotarse por la letra X. Este será el convenio de notación adoptado en esta sección. No obstante, puesto que el trabajo está basado en métodos de regresión, en secciones posteriores, y por motivos prácticos, denotaremos por X al conjunto de índices y por f al proceso Gaussiano. La siguiente definición introduce el concepto de proceso estocástico [30, Definición 3.1].

Definición 2.2.1 (Proceso estocástico)**.** *Sea T un conjunto de índices y $(\Omega', \mathcal{F}')$ un espacio medible, donde generalmente $\Omega' = \mathbb{R}$ y $\mathcal{F}' = \beta$. Un proceso estocástico sobre T es una colección de variables aleatorias $\{X_t, t \in T\}$ definidas en un espacio probabilístico $(\Omega, \mathcal{F}, P)$ con valores en $(\Omega', \mathcal{F}')$. Equivalentemente, es una colección de aplicaciones medibles X_t desde un espacio probabilístico $(\Omega, \mathcal{F}, P)$ a $(\Omega', \mathcal{F}')$.*

Ω recibe el nombre de espacio muestral mientras que Ω' es el espacio de estados. Para $w \in \Omega$ fijo, la aplicación $t \in T \to X_t(w)$ recibe el nombre de trayectoria de w y es un elemento de Ω'^T. Puesto que el conjunto de índices T suele hacer referencia al tiempo, con frecuencia T recibe el nombre de espacio temporal del proceso.

Observación 2.2.1.

- El término proceso estocástico se emplea para describir modelos matemáticos que representan el estado de un sistema dependiente de un parámetro, generalmente el tiempo, y del azar. Un modelo de esta forma es una aplicación $(t, w) \to X(t, w)$ definida en $T \times \Omega$ y con valores en Ω' que describe los estados del sistema. En un instante fijo $t \in T$ el sistema depende únicamente del azar, y queda descrito por X_t, que es una variable aleatoria que recibe el nombre de estado del sistema en el instante t.
- Normalmente T es un subconjunto de $\mathbb{R}$. Puede tratarse de un intervalo de $\mathbb{R}$, en cuyo caso el soporte será continuo, o puede ser un intervalo de $\mathbb{Z}$, siendo entonces un soporte discreto.

Definición 2.2.2 (Distribución finito dimensional)**.** *[30, Definición 3.2] Sea $X = (X_t)_{t\in T}$ un proceso estocástico. Las distribuciones conjuntas de las subfamilias finitas de $(X_t)_{t\in T}$ reciben el nombre de distribuciones finito dimensionales. Así pues, si $t_1, ..., t_n \in T$, la distribución de probabilidad*

$$P_{t_1,...,t_n}(C) = P[(X_{t1}, ..., X_{tn}) \in C],$$

con $C \in \mathcal{F'}^n$ es una distribución finito dimensional.

La familia de las distribuciones finito dimensionales de un proceso estocástico constituye uno de los aspectos más importantes del mismo, ya que esta familia determina la distribución del proceso de forma unívoca, como veremos más adelante mediante el Teorema de extensión de Kolmogorov [Anexo A].

Las siguientes propiedades de las distribuciones finito dimensionales nos serán útiles en lo que sigue:

- Si σ es una permutación en $\{1, ..., n\}$ y $C_1, ..., C_n \in \mathcal{F}'$ entonces los sucesos

  $$\{(X_{t_1}, ..., X_{t_n}) \in C_1 \times ... \times C_n\} \text{ y } \left\{(X_{t_{\sigma(1)}}, ..., X_{t_{\sigma(n)}}) \in C_{\sigma(1)} \times ... \times C_{\sigma(n)}\right\}$$

  son iguales y

  $$P_{(t_1,...,t_n)}(C_1 \times ... \times C_n) = P_{(t_{\sigma(1)},...,t_{\sigma(n))}}(C_{\sigma(1)} \times ... \times C_{\sigma(n)}).$$

- $P_{(t_1,...,t_{n-1})}(C_1 \times ... \times C_{n-1}) = P_{(t_1,...,t_n)}(C_1 \times ... \times C_{n-1} \times \Omega')$.

La primera condición nos permite considerar únicamente las distribuciones finito dimensionales de la forma $P_{(t_1,...,t_n)}$ tales que $t_1 < ... < t_n$, ya que con ellas quedan determinadas todas las demás. Para aligerar la notación, denotamos por P_V a $P_{(t_1,...,t_n)}$, siendo por tanto $V = \{t_1, ..., t_n\}$.

Resumiendo, hemos definido un proceso estocástico como una familia $(X_t)_{t\in T}$ de variables aleatorias en $(\Omega, \mathcal{F}, P)$. Hemos visto como podemos definir un proceso estocástico como una aplicación $X : (t, w) \in T \times \Omega \to X(t, w) \in \mathbb{R}$ donde, para cada $t \in T$, $X(t, \cdot)$ es una variable aleatoria en Ω. Podemos verlo también de otra forma: consideramos una aplicación X que asocia a cada $w \in \Omega$ fijo una aplicación $t \in T \to X_t(w)$. De esta forma, X es un aplicación de Ω en el conjunto de las aplicaciones de T en $\mathbb{R}$, $\mathbb{R}^T$. Observamos además

que una aplicación $Y : (\Omega, F) \to (\mathbb{R}^T, \mathcal{R}^T)$ es una variable aleatoria si y sólo si $\pi_t(Y)$ lo es para cada $t \in T$, donde π_t denota la aplicación proyección en $\mathbb{R}^T$. Así pues, podemos ver un proceso estocástico como una variable aleatoria de $(\Omega, \mathcal{F}, P)$ en $(\mathbb{R}^T, \mathcal{R}^T)$. Es por ello por lo que en ocasiones los procesos estocásticos reciben el nombre de funciones aleatorias.

La siguiente definición introduce el concepto de σ-álgebra engendrada por cilindros medibles, a partir de la cual podremos definir adecuadamente un proceso mediante la ley de éste.

Definición 2.2.3. *Sea T un conjunto no vacío y para cada $t \in T$ suponemos que (Ω_t, F_t) es un espacio medible. Sea $\mathbb{R}^T = \prod_{t\in T} \Omega_t$. Llamaremos cilindro medible n-dimensional en Ω a un subconjunto de Ω de la forma*

$$c(B) = \{w \in \Omega : (w_1, ..., w_n) \in B\},$$

donde $B \in \prod_{i=1}^n \mathcal{F}_{ti}$. Denotaremos por $\prod_{t\in T} \mathcal{F}_t$ a la σ-álgebra en Ω engendrada por cilindros medibles en Ω.

Observamos que de acuerdo con esta definición, la familia de los cilindros medibles en Ω es un álgebra en Ω que genera las σ-álgebra producto. Por otro lado, si todos los espacios medibles $(\Omega_t, \mathcal{F}_t)$ coinciden con un cierto espacio medible (Ω, F), el espacio medible lo denotaremos por (Ω^T, F^T). De esta forma, si definimos $X : \Omega \to \{\text{Funciones de } T \text{ en } \Omega'\}$, entonces X_t será medible si y sólo si F^T es medible, por lo que tendrá sentido hablar de la ley del proceso.

El objetivo es construir en $(\mathbb{R}^T, \mathcal{R}^T)$ una probabilidad a partir de probabilidades $P_{(t_1,...,t_n)}$ en $\mathcal{R}^n$ definidas para cada colección creciente de índices $t_1 < ... < t_n$ y cada $n \in \mathbb{N}$, supuesto que estas probabilidades satisfacen una cierta condición de consistencia. En virtud del Teorema de extensión de Kolmogorov [Anexo A], es posible construir un proceso estocástico a partir de unas distribuciones finito dimensionales dadas de antemano, suponiendo que verifican una condición de consistencia. Las definiciones siguientes precisan hasta qué punto un proceso estocástico queda determinado por sus distribuciones finito dimensionales.

Definición 2.2.4.

a) Consideremos dos procesos estocásticos reales sobre el mismo espacio temporal $(\Omega, \mathcal{F}, P, (X_t)_{t\in T})$ y $(\Omega', \mathcal{F}', P', (X'_t)_{t\in T})$. Diremos que dichos procesos son equivalentes si

$$P(X_{t_1} \in F_1, ..., X_{t_n} \in F_n) = P'(X'_{t_1} \in F_1, ..., X'_{t_n} \in F_n)$$

para cada subconjunto finito $\{t_1, ..., t_n\}$ de T y cada familia finita $F_1, ..., F_n$ en $\mathcal{R}$. Esto es equivalente a decir que ambos procesos tienen la misma distribución, y, por tanto, tienen las mismas distribuciones finito dimensionales.

b) Sean $(X_t)_{t\in T}$ e $(Y_t)_{t\in T}$ dos procesos estocásticos reales en el mismo espacio probabilístico (Ω, F, P) y sobre el mismo espacio temporal T. Diremos que (Y_t) es una modificación de (X_t) si $X_t \overset{P}{=} Y_t$ c.s. para cada $t \in T$. Diremos que dichos procesos son P-indistinguibles si existe $F \in \mathcal{F}$ tal que $P(F) = 0$ y $X_t(w) = Y_t(w)$ para cada $w \in F^c$ y cada $t \in T$.

Por último, hemos visto que las distribuciones finito dimensionales de un proceso estocástico nos permiten, en virtud del Teorema de extensión de Kolmogorov, determinar de manera unívoca el proceso. No obstante, dichas distribuciones no son suficientes para abordar

cuestiones relacionadas con propiedades de regularidad de las trayectorias, como puede ser la continuidad de éstas cuando T es un intervalo de la recta real. Esto quiere decir que podemos tener dos procesos estocásticos, uno modificación del otro, tales que tengan las mismas distribuciones finito dimensionales, y sin embargo las trayectorias de uno pueden ser continuas siendo las del otro discontinuas.

Con los ideas introducidas hasta ahora es posible definir el concepto de proceso Gaussiano [30, Definición 3.4], que será la base de este trabajo:

Definición 2.2.5 (Proceso Gaussiano)**.** *Un proceso estocástico $(X_t)_{t\in T}$ es Gaussiano si para todo $n \in \mathbb{N}$ finito y todos $t_1, ..., t_n \in T$ el vector aleatorio $(X_{t_1}, ..., X_{t_k})$ tiene una distribución Gaussiana con valores en $\Omega' = \mathbb{R}$. Es decir, si todas las distribuciones finito dimensionales son Gaussianas.*

De acuerdo con los cambios en la notación mencionados al principio de esta sección, diremos que f es un proceso Gaussiano con función de media m, $m(x) = E[f(x)]$, y función de covarianza k, $k(x, x') = \text{cov}(f(x), f(x'))$, que denotamos por $f \sim \mathcal{GP}(m, k)$, si sus distribuciones finito dimensionales $f_X = (f(x_1), ..., f(x_n)) \in \mathbb{R}^n$ son Gaussianas; es decir, tienen una distribución normal multivariante $\mathcal{N}(m_X, k_{XX})$ con matriz de covarianza $k_{XX} = (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$ y vector de medias $m_X = (m(x_1), ..., m(x_n))^T$.

Así pues, dado un proceso Gaussiano f existen funciones $m : \mathcal{X} \to \mathbb{R}$ y $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ que son respectivamente la media y la función de covarianzas asociadas a dicho proceso. A partir del Teorema de extensión de Kolmogorov y la equivalencia entre las funciones núcleo y las funciones de covarianza que introduciremos en el capítulo 3, Teorema 3.2.1, es posible mostrar que el recíproco también es cierto. Por consiguiente, para toda función de media m y toda función semidefinida positiva k, existe un proceso Gaussiano asociado que queda unívocamente determinado por dichas funciones. La elección de éstas otorga al proceso ciertas propiedades, como pueden ser la diferenciabilidad de las muestras pertenecientes a éste.

2.3. Distribuciones condicionadas.

El manejo de distribuciones condicionadas es fundamental en la aproximación bayesiana a la inferencia. En los cursos elementales de estadística y probabilidad las definiciones de distribuciones condicionadas se suelen hacer de manera separada para distribuciones discretas y distribuciones con densidad conjunta. Esto puede resultar insuficiente por que excluye, entre otros, a modelos en los que parte del vector es discreto mientras que otra parte del vector tiene densidad. Para evitar problemas de este tipo, incluimos una pequeña sección sobre la forma general de definir la probabilidad condicionada. Esto es posible a través del Teorema de Radon Nikodym [15]. Supongamos que tenemos un espacio probabilístico de la forma $(\Omega, \mathcal{F}, P)$, donde $\mathcal{G} \subset \mathcal{F}$. Definimos la siguiente medida:

$$\nu(A) = E[XI_A], \quad \text{donde } A \in \mathcal{G}, x \geq 0. \tag{2.4}$$

De esta forma, tenemos que ν es una medida positiva en el correspondiente espacio de medida $(\Omega, \mathcal{G})$, ya que cumple las siguientes propiedades:

- $\nu(\emptyset) = 0$.
- $\nu(A) \geq 0 \, \forall A \in \mathcal{G}$.

- $\nu(\amalg_{i=1}^{\infty} A_1) = \sum_{i=1}^{\infty} \nu(A_i)$, donde el símbolo $\amalg$ se usa para denotar la unión disjunta.

Además, $\nu(A) = 0$ si $P(A) = 0$. La última propiedad implica que la medida ν es absolutamente continua respecto de P, lo que se suele denotar por $\nu << P$. Esto nos permite emplear el Teorema de Radon Nikodym [15], que garantiza la existencia de una única función $\mathcal{G}$-medible verificando que

$$\nu(A) = \int_A \frac{d\nu}{dP} dP, \quad \forall A \in \mathcal{G}. \tag{2.5}$$

Podemos entonces definir lo siguiente:

$$E[X|\mathcal{G}] = \frac{d\nu}{dP}. \tag{2.6}$$

De esta forma, $Z = E[X|\mathcal{G}]$ es la única función $\mathcal{G}$-medible que verifica:

$$E[XI_A] = E[ZI_A] \quad \forall A \in \mathcal{G}, \tag{2.7}$$

ya que:

$$E[XI_A] = \nu(A) = \int_A \frac{d\nu}{dP} dP = \int_A Z dP = E[ZI_A].$$

En el caso $\mathcal{G} = \sigma(Y)$, las funciones $\mathcal{G}$-medibles son de la forma $\Phi(Y)$, donde $\Phi : \mathbb{R}^k \to \mathbb{R}$ es una función medible. Con esta notación, podemos reescribir lo anterior de la siguiente forma: $E[X|Y] = \Phi(Y)$, donde $\Phi(Y)$ es la única función (c.s.) medible tal que

$$E[XI_{Y\in B}] = E[\Phi(Y)I_{Y\in B}], \forall B \in \mathcal{B}(\mathbb{R}^k).$$

En el caso de trabajar con funciones indicadoras, sea $X = I(Z \in A)$, tenemos que, por definición,

$$E[X|Y] := P(Z \in A|Y), \tag{2.8}$$

y, de nuevo, en virtud del Teorema de Radon Nikodym [15], $P(Z \in A|Y)$ es la única función Y-medible tal que

$$P(Z \in A, Y \in B) = E[P(Z \in A|Y)I_{Y\in B}] \quad \forall B \in \mathcal{B} \in \mathbb{R}^k. \tag{2.9}$$

Por consiguiente, hemos obtenido una forma general de definir la probabilidad condicionada de forma única. Veamos a continuación cómo expresarla en términos de las funciones de densidad. Observamos que si

$$P(Z \in A|Y = y) = \int_A \frac{f_{Z,Y}(z,y)}{f_Y(y)} dz,$$

entonces

$$\begin{aligned} E[P(Z \in A|Y)I_{Y\in B}] &= \int_B \left(\int_A \frac{f_{Z,Y}(z,y)}{f_Y(y)} dz \right) f_Y(y) dy = \int_{B\times A} f_{Z,Y}(z,y) dz dy \\ &= P(Z \in A, Y \in B), \end{aligned} \tag{2.10}$$

obteniendo de nuevo la igualdad dada en (2.9). Podemos entonces concluir que la probabilidad condicionada se define de la siguiente forma:

$$P(Z \in A|Y = y) = \int_A \frac{f_{Z,Y}(z,y)}{f_Y(y)} dz, \tag{2.11}$$

y se trata de un resultado general.

Nos interesa ahora adaptar este resultado al caso discreto y al caso continuo. A través de la regla de multiplicación podemos escribir lo siguiente:

$$f_{Z,Y}(z,y) = f_Z(z)f_{Y|Z=z}(y) = f_Y(y)f_{Z|Y=y}(z), \tag{2.12}$$

A partir de esta igualdad, es posible obtener la expresión del Teorema de Bayes para el caso con densidad:

$$f_{Z|Y=y} = \frac{f_{Z,Y}(z,y)}{f_Y(y)} = \frac{f_Z(z)f_{Y|Z=z}(y)}{\int f_Z(z)f_{Y|Z=z}(y)dz}. \tag{2.13}$$

Esta expresión es válida cuando ambas variables tienen una función de densidad asociada. Veamos ahora cómo podemos adaptarlo al caso en el que una de ellas tiene un soporte discreto, como ocurre en los problemas de clasificación. Supongamos que la variable observada, y, toma valores en un conjunto discreto, sea $y \in \{1, ..., D\}$, y que z tiene una función de densidad asociada, sea $f_Z(z) = p(z) << l_d = \mu_d$. Es decir, la densidad de z es absolutamente continua con respecto de la medida de Lebesgue. Tenemos por tanto que $(z, y) \in \mathbb{R}^d \times \{1, ..., D\}$, por lo que la medida asociada será el producto directo de ambas medidas, $\mu = \mu_d \otimes \mu_c$. Entonces $p(z, y) << \mu$. Es decir,

$$\mu(C) = 0 \Rightarrow P((Z, Y) \in C) = 0.$$

Es fácil comprobar esto para el caso de conjuntos de la forma $C = A \times B$. Supongamos que $\mu(C) = 0$. Entonces, tenemos que

$$\mu(C) = 0 \Rightarrow \mu_d(A)\mu_c(B) = 0 \Rightarrow \begin{cases} \mu_d(A) = 0 \\ \text{ó bien} \\ \mu_c(B) = 0 \end{cases}. \tag{2.14}$$

Puesto que $f_Z(z) = p(z) << l_d = \mu_d$ y $f_Y(y) = p(y) << \mu_c$, esta última por ser un conjunto discreto, tenemos que

$$\begin{cases} P(Z \in A) = 0 \\ \text{o bien} \\ P(Y \in B) = 0 \end{cases} \Rightarrow P((Z, Y) \in C) = 0. \tag{2.15}$$

Buscamos reescribir el Teorema de Bayes para el caso discreto. Por un lado tenemos que

$$P((Z, Y) \in C) = \int_C f_{Z,Y}(z, y)d\mu(z, y) = \int_C f_{Z,Y}(z, y)d\mu_d(z)d\mu_c(y). \tag{2.16}$$

Por otro lado, podemos obtener las densidades marginales de la siguiente forma:

$$\begin{cases} f_Z(z) = \int_{\{1,...,D\}} f_{Z,Y}(z, y)d\mu_c(y) = \int_{\{1,...,D\}} f_Y(y)f_{Z|Y=y}(z)d\mu_c(y) = \sum_{j=1}^{D} f_Y(j)f_{Z|y=j}(z). \\ f_Y(y) = \int_{\mathbb{R}^d} f_{Z,Y}(z, y)d\mu_d(z). \end{cases} \tag{2.17}$$

Con estas expresiones, podemos reescribir (2.13) de la siguiente forma para el caso en el que una de las variables toma valores en un conjunto discreto:

$$f_{Z|Y=y} = \frac{f_{Z,Y}(z,y)}{f_Y(y)} = \frac{f_Z(z)f_{Y|Z=z}(y)}{\sum_{j=1}^{D} f_Y(j)f_{Z|Y=j}(z)}. \tag{2.18}$$

Esto nos muestra que el tratamiento va a ser el mismo independientemente de que trabajemos en un conjunto discreto o continuo. Por tanto, los problemas de regresión y de clasificación van a poder tratarse de forma muy similar.

2.4. Aproximación Bayesiana a la inferencia.

Para el aprendizaje basado en procesos Gaussianos la aproximación clásica a la inferencia resulta insuficiente y es más conveniente recurrir a la aproximación bayesiana. El marco teórico en el que se aplica la inferencia bayesiana es similar al clásico: disponemos de un parámetro poblacional respecto al cual se desea realizar inferencias y se tiene un modelo que determina la probabilidad de observar diferentes valores de la variable observable, bajo diferentes valores de los parámetros. Sin embargo, la diferencia fundamental es que la inferencia bayesiana considera al parámetro como una variable aleatoria. De este modo, en la estadística clásica solo se tiene en cuenta la información de la muestra obtenida, suponiendo, para los desarrollos matemáticos, que se podría tomar una muestra de tamaño infinito de manera hipotética. En el caso bayesiano, sin embargo, además de la muestra también juega un papel importante la información previa o externa que se pose en relación con los fenómenos que se tratan de modelizar. En esencia, la inferencia bayesiana está basada en la distribución de probabilidad del parámetro dados los datos (distribución a posteriori de probabilidad), en lugar de la distribución de los datos dado el parámetro. Por consiguiente, se requiere especificar previamente una distribución a priori de probabilidad que representa el conocimiento acerca del parámetro antes de obtener cualquier información respecto a los datos. Podemos dividir el problema de inferencia bayesiana en varias etapas:

1. Especificar un modelo de probabilidad completo que incluya algún tipo de conocimiento previo sobre los parámetros del modelo dado. Debe seleccionarse una distribución de probabilidad conjunta para todas las cantidades observables y no observables. El modelo debe ser consistente con el conocimiento acerca del problema fundamental y el proceso de recolección de la información.

2. Actualizar el conocimiento sobre los parámetros desconocidos condicionando este modelo de probabilidad a los datos observados. Calcular e interpretar la distribución a posteriori apropiada que se define como la distribución de probabilidad condicionada por las cantidades no observadas de interés, dados los datos observados.

3. Evaluar el ajuste del modelo y las implicaciones de la distribución a posteriori resultante. ¿Es el modelo apropiado a los datos?, ¿son las conclusiones razonables? Si fuese necesario, alterar o ampliar el modelo, y repetir las tres etapas mencionadas.

Por consiguiente, vemos que la inferencia bayesiana pretende obtener información sobre cómo cambia nuestra idea de los parámetros a partir de los datos de entrenamiento. Para ello, es necesario obtener la distribución de probabilidad de los parámetros condicionada por los datos de entrenamiento. Este proceso de aprendizaje inductivo por medio de la regla de Bayes es la base de la inferencia bayesiana, y da lugar a las distribuciones a posteriori dadas por (2.13) y (2.18). Además, de cara a la inferencia resulta conveniente manejar la distribución predicativa dada por $p(x_{n+1}, ..., x_{n+m}|x_1, ..., x_n)$, para cualquier $m \geq 1$ y $n \geq 1$. De acuerdo con la definición de probabilidad condicionada, esto se reduce a lo siguiente:

$$p(x_{n+1}, ..., x_{n+m}|x_1, ..., x_n) = \frac{p(x_1, ..., x_{n+m})}{p(x_1..., x_n)}. \tag{2.19}$$

Cuando el modelo de predicción admite una representación en términos de unos parámetros y de unas distribuciones a priori, podemos emplear el Teorema de Bayes para obtener la distribución a posteriori en términos del modelo paramétrico de $x_1, ..., x_n$ dado θ, y la

densidad a priori de θ.

Con el objetivo de aligerar la notación, las ecuaciones (2.13) y (2.18) se han reescrito con una notación más compacta, que será la que empleemos a lo largo del trabajo:

$$p(z|y) = \frac{p(z,y)}{p(y)} = \frac{p(z)p(y|z)}{\int p(z)p(y|z)dz}. \tag{2.20}$$

$$p(z|y) = \frac{p(z,y)}{p(y)} = \frac{p(z)p(y|z)}{\sum_{j=1}^{D} p(j)p(z|y=j)}. \tag{2.21}$$

2.5. Divergencia de Kullback-Leibler.

La divergencia de Kullback-Leibler tiene su origen en la teoría de la información. El objetivo de la divergencia de Kullback-Leibler es medir la similitud entre dos distribuciones de probabilidad, $\mathcal{P}$ y $\mathcal{Q}$. Generalmente $\mathcal{Q}$ representa los datos, las observaciones o la distribución de probabilidad que hemos medido. Por el contrario, $\mathcal{P}$ se emplea para denotar un modelo teórico, una descripción o aproximación de $\mathcal{Q}$. De esta forma, podemos interpretar la divergencia de Kullback-Leibler como la diferencia o el número de bits que necesitamos para representar muestras dadas por $\mathcal{Q}$ usando un código que optimiza $\mathcal{P}$ en lugar de uno que optimiza $\mathcal{Q}$. En este sentido, está estrechamente relacionada con el concepto de entropía [47, Capítulo 2], denotado típicamente por la letra H. La entropía constituye una medida de la incertidumbre asociada con una variable aleatoria X que toma m valores diferentes x_j y cuya distribución viene dada por f, de modo que $f(x_j) = P(X = x_j) = p_j$.

Definición 2.5.1. *La entropía de Shannon, o simplemente entropía, de una variable aleatoria discreta viene definida por la siguiente expresión:*

$$H(X) = -\sum_{j=1}^{m} p_j \log P(X = x_j) = -E[\log P(X = x_j)] = -E\left[\log \frac{1}{p_j}\right], \tag{2.22}$$

cuando la suma existe.

Dicha función H satisface, entro otras, las siguientes propiedades:

- $H(X) \geq 0$.
- $H(X) = 0$ si y sólo si $\exists x_0 \in X$ tal que $X \overset{c.s}{=} x_0$.
- Si X puede tomar m valores diferentes, siendo m finito, entonces $H(X) \leq \log(m)$, y se da la igualdad si X está uniformemente distribuida.
- $H(X) + H(Y) \geq H(X, Y)$, y se da la igualdad si X, Y son independientes.

Las tres primeras propiedades pueden resumirse diciendo que una distribución uniforme maximiza $H(X)$. Es decir, que la entropía de una variable aleatoria alcanza su valor máximo cuando todos los posibles valores de X son equiprobables, mientras que el valor de $H(X)$ es cero cuando todos los valores tienen probabilidad cero a excepción de uno, que ocurre con probabilidad uno.

La definición 2.5.1 puede generalizarse al caso continuo, pero hay que tener precaución ya que la probabilidad de que una variable aleatoria continua tome un valor concreto es cero. Para solventar dicho problema, es necesario introducir una medida de referencia, a través de la cual es posible definir la entropía de la siguiente forma [47, Capítulo 2]:

Definición 2.5.2. *Dadas dos distribuciones de probabilidad, $\nu << \mu$, la entropía de ν con respecto de μ, o la divergencia de Kullback-Leibler de ν con respecto de μ, se define como*

$$D(\mu||\nu) = -E_\mu\left[\log\frac{d\nu}{d\mu}\right] = E_\mu\left[\log\frac{d\mu}{d\nu}\right]. \tag{2.23}$$

Si ν no es absolutamente continua respecto de μ, entonces $D(\mu||\nu) = \infty$. Es decir, tenemos que

$$D(\mu||\nu) = \begin{cases} \int \frac{d\mu}{d\nu}\log\frac{d\mu}{d\nu}d\nu = \int \log\frac{d\mu}{d\nu}d\mu, & si \quad \nu << \mu. \\ \\ \infty, & en\ caso\ contrario. \end{cases} \tag{2.24}$$

En el contexto del *Machine Learning*, dadas dos distribuciones $\mathcal{Q}$ y $\mathcal{P}$, la divergencia de Kullback-Leibler, $D(\mathcal{Q}||\mathcal{P})$, cuantifica la información que obtendríamos si empleásemos $\mathcal{Q}$ en lugar de $\mathcal{P}$. En términos Bayesianos, es una medida de la información adicional que obtendríamos si empleásemos la distribución a posteriori $\mathcal{Q}$ en lugar de la distribución a priori $\mathcal{P}$. Es decir, es la cantidad de información que perdemos al emplear $\mathcal{P}$ para aproximar $\mathcal{Q}$. Aunque a menudo se emplea el término distancia para referirse a la divergencia de Kullbak-Leibler, no es correcto puesto que ni verifica la desigualdad triangular ni es simétrica. Algunas propiedades importantes de la divergencia de Kullback Leibler son las siguientes:

- $D(\mathcal{Q}||\mathcal{P}) \geq 0$, y se da la igualdad si y sólo si $\mathcal{P} = \mathcal{Q}$ casi siempre.
- $D(\mathcal{Q}||\mathcal{P}) = D(\mathcal{Q}_1||\mathcal{P}_1) + D(\mathcal{Q}_2||\mathcal{P}_2)$ si $\mathcal{P}_1, \mathcal{P}_2$ y $\mathcal{Q}_1, \mathcal{Q}_2$ son independientes, donde $\mathcal{Q} = \mathcal{Q}_1 \otimes \mathcal{Q}_2$ y $\mathcal{P} = \mathcal{P}_1 \otimes \mathcal{P}_2$ siendo $\mathcal{Q}$ y $\mathcal{P}$ distribuciones de probabilidad en $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. Esta igualdad se deduce de la siguiente forma:

$$\begin{aligned} \mathcal{Q}(A \times B) &= \mathcal{Q}_1(A)\mathcal{Q}_2(B) = \int_A \frac{d\mathcal{Q}_1}{d\mathcal{P}_1}d\mathcal{P}_1 \int_B \frac{d\mathcal{Q}_2}{d\mathcal{P}_2}d\mathcal{P}_2 \\ &= \int_{A\times B}\left(\frac{d\mathcal{Q}_1}{d\mathcal{P}_1}(x) \times \frac{d\mathcal{Q}_2}{d\mathcal{P}_2}(y)\right) d\mathcal{P}_1(x) \times d\mathcal{P}_2(y). \end{aligned} \tag{2.25}$$

  Por tanto, tenemos que

$$\frac{d\mathcal{Q}}{d\mathcal{P}}(x,y) = \frac{d\mathcal{Q}_1}{d\mathcal{P}_1}(x) \times \frac{d\mathcal{Q}_2}{d\mathcal{P}_2}(y). \tag{2.26}$$

  Sustituyendo en la expresión de la divergencia de Kullback-Leibler,

$$D(\mathcal{Q}||\mathcal{P}) = \int \log\frac{d\mathcal{Q}}{d\mathcal{P}}d\mathcal{P} = \int\left(\log\frac{d\mathcal{Q}_1}{d\mathcal{P}_1}(x) + \log\frac{d\mathcal{Q}_2}{d\mathcal{P}_2}(y)\right) d\mathcal{P}_1(x)d\mathcal{P}_2(y). \tag{2.27}$$

  Aplicando el Teorema de Fubini, obtenemos la igualdad

$$D(\mathcal{Q}||\mathcal{P}) = D(\mathcal{Q}_1||\mathcal{P}_1) + D(\mathcal{Q}_2||\mathcal{P}_2). \tag{2.28}$$

En el caso particular de los procesos Gaussianos, puesto que tanto la distribución a priori como la distribución a posteriori son Gaussianas, podemos expresar la divergencia de Kullback Leibler de forma particularmente sencilla. Denotando por $f(\mathbf{x})$ a los valores de la función de predicción o hipótesis del modelo, podemos dividir este vector en dos partes: una primera parte que denotamos por $\mathbf{f}$ y que se corresponde con los valores de f en los puntos de entrenamiento, $\mathbf{x}_1, ..., \mathbf{x}_n$, y una segunda parte que representa el valor de f en el

resto de puntos del dominio, $\mathbf{x} \in \mathcal{X}$, que denotaremos por $\mathbf{f}_*$. De esta forma, denotando por q a la distribución a posteriori para nuestro proceso Gaussiano, y por p a la distribución a priori, tenemos que

$$\begin{aligned} q(\mathbf{f}, \mathbf{f}_*|\mathbf{y}) &= q(\mathbf{f}|\mathbf{y})p(\mathbf{f}_*|\mathbf{f}). \\ p(\mathbf{f}, \mathbf{f}_*) &= p(\mathbf{f})p(\mathbf{f}_*|\mathbf{f}). \end{aligned}$$

La divergencia de Kullback-Leibler se escribe entonces como

$$\begin{aligned} KL(q||p) &= \int q(\mathbf{f}, \mathbf{f}_*|\mathbf{y}) \ln \frac{q(\mathbf{f}, \mathbf{f}_*|\mathbf{y})}{p(\mathbf{f}, \mathbf{f}_*)} d\mathbf{f} d\mathbf{f}_* = \int q(\mathbf{f}|\mathbf{y})p(\mathbf{f}_*|\mathbf{f}) \ln \frac{q(\mathbf{f}|\mathbf{y})p(\mathbf{f}_*|\mathbf{f})}{p(\mathbf{f})p(\mathbf{f}_*|\mathbf{f})} d\mathbf{f} d\mathbf{f}_* \\ &= \int q(\mathbf{f}|\mathbf{y}) \ln \frac{q(\mathbf{f}|\mathbf{y})}{p(\mathbf{f})} d\mathbf{f}. \end{aligned} \tag{2.29}$$

De esta forma se puede transformar la integral de dimensión posiblemente infinita en una integral n-dimensional, que es significativamente más manejable. Por otro lado, usando la condición de que ambas distribuciones son Gaussianas, es posible simplificar aún más la expresión. En particular, si tenemos dos distribuciones Gaussianas $\mathcal{N}(\mu_p, \Sigma_p)$ y $\mathcal{N}(\mu_q, \Sigma_q)$, que denotamos por $\mathcal{N}_p$ y $\mathcal{N}_q$ respectivamente, tenemos la siguiente expresión de la divergencia de Kullback-Leibler, donde hemos empleado la definición de la densidad de probabilidad de una normal n-dimensional (2.3):

$$\begin{aligned} KL(q||p) &= E_q[\ln(q)/\ln(p)] = E_q[\ln(q) - \ln(p)] \\ &= E_q\left[\frac{1}{2}\ln\frac{|\Sigma_p|}{|\Sigma_q|} - \frac{1}{2}(\mathbf{x}-\mu_\mathbf{q})^T\Sigma_q^{-1}(\mathbf{x}-\mu_\mathbf{q}) + \frac{1}{2}(\mathbf{x}-\mu_\mathbf{p})^T\Sigma_p^{-1}(\mathbf{x}-\mu_\mathbf{p})\right] \\ &= \frac{1}{2}E_q\left[\ln\frac{|\Sigma_p|}{|\Sigma_q|}\right] - \frac{1}{2}E_q\left[(\mathbf{x}-\mu_\mathbf{q})^T\Sigma_q^{-1}(\mathbf{x}-\mu_\mathbf{q})\right] + \frac{1}{2}E_q\left[(\mathbf{x}-\mu_\mathbf{p})^T\Sigma_p^{-1}(\mathbf{x}-\mu_\mathbf{p})\right] \\ &= \frac{1}{2}\ln\frac{|\Sigma_p|}{|\Sigma_q|} - \frac{1}{2}E_q\left[(\mathbf{x}-\mu_\mathbf{q})^T\Sigma_q^{-1}(\mathbf{x}-\mu_\mathbf{q})\right] + \frac{1}{2}E_q\left[(\mathbf{x}-\mu_\mathbf{p})^T\Sigma_p^{-1}(\mathbf{x}-\mu_\mathbf{p})\right] \end{aligned} \tag{2.30}$$

Teniendo en cuenta que $\left[(\mathbf{x}-\mu_q)^T\Sigma_q^{-1}(\mathbf{x}-\mu_q)\right] \in \mathbb{R}$, podemos escribirlo como $\mathrm{tr}\left(\left[(\mathbf{x}-\mu_q)^T\Sigma_q^{-1}(\mathbf{x}-\mu_q)\right]\right)$, donde tr denota el operador traza. Además, la esperanza y la traza se pueden intercambiar [12, Ecuación 16], y la traza tiene la propiedad conmutativa, aunque el producto de matrices no la cumpla, por lo que podemos escribir el segundo término como

$$\frac{1}{2}E_q\mathrm{tr}\left(\left[(\mathbf{x}-\mu_q)^T(\mathbf{x}-\mu_q)\right]\Sigma_q^{-1}\right) = \frac{1}{2}\mathrm{tr}\left(E_q\left[(\mathbf{x}-\mu_q)^T(\mathbf{x}-\mu_q)\right]\Sigma_q^{-1}\right).$$

Puesto que $E_q\left[(\mathbf{x}-\mu_q)^T(\mathbf{x}-\mu_q)\right] = \Sigma_q$, podemos escribir lo siguiente:

$$\frac{1}{2}\mathrm{tr}\left(E_q\left[(\mathbf{x}-\mu_q)^T(\mathbf{x}-\mu_q)\right]\Sigma_q^{-1}\right) = \frac{1}{2}\mathrm{tr}\left(\Sigma_q\Sigma_q^{-1}\right) = \frac{1}{2}\mathrm{tr}\left(I_n\right) = \frac{n}{2}.$$

Por último, el tercer término puede simplificarse de la siguiente forma [12, Ecuación 380]:

$$E_q\left[(\mathbf{x}-\mu_p)^T\Sigma_p^{-1}(\mathbf{x}-\mu_p)\right] = (\mu_q-\mu_p)^T\Sigma_p^{-1}(\mu_q-\mu_p) + \mathrm{tr}\left(\Sigma_p^{-1}\Sigma_q\right).$$

Así pues, para el caso de distribuciones Gaussianas, la divergencia de Kullback-Leibler adquiere la siguiente forma:

$$KL(q||p) = \frac{1}{2}\left[\ln\frac{|\Sigma_p|}{|\Sigma_q|} - n + (\mu_q-\mu_p)^T\Sigma_p^{-1}(\mu_q-\mu_p) + \mathrm{tr}\left(\Sigma_p^{-1}\Sigma_q\right)\right]. \tag{2.31}$$

2.6. Regresión lineal.

Es habitual en aplicaciones científicas o tecnológicas que uno esté interesado en estudiar la relación que existe entre dos variables, digamos x e y. A la variable x se la conoce como variable explicativa o regresor, mientras que a y nos referimos como variable dependiente o variable respuesta. Frecuentemente asumimos que existe una relación entre ambas dada por $y = f(x)$, donde la función f es desconocida, aunque en algunos casos puede venir especificada por un modelo teórico, por ejemplo, la ley de Hooke. No obstante, hay casos en los que no disponemos a priori de información acerca de dicha función. Ésta debe ser entonces estudiada mediante una serie de experimentos a partir de los cuales se obtienen unos datos de entrenamiento que pueden ser posteriormente analizados para buscar relaciones entre las variables respuesta y_i y las condiciones iniciales $x_i \in \mathcal{X}$. El método de regresión nos permite determinar una función f que mejor se ajusta a los datos observados. Es decir, nos permite encontrar un estimador para la función subyacente, de manera que dado cualquier vector de características $x_* \in \mathcal{X}$ podemos predecir el valor de la respuesta y_* como $f(\mathbf{x}_*)$.

Con frecuencia se asume que existe una relación lineal entre las variables, de modo que f es una función lineal, $f = w_1 x + w_0$. Con el fin de dar mayor flexibilidad al modelo, se introduce un pequeño error aleatorio, de modo que diremos que $y_i = f(x_i) + \epsilon_i, i = 1, ..., n$, siendo n el número de datos observados o datos de entrenamiento. Es habitual asumir que dichos errores son variables aleatorias normales independientes e igualmente distribuidas con media cero y varianza σ^2, esto es,

$$\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, ..., n.$$

Hablamos entonces de un modelo de regresión lineal normal. Los parámetros w_0 y w_1, que reciben el nombre de intercept y pendiente, respectivamente, determinan la función de regresión $f(x) = w_1 x + w_0$. Observamos además que, de acuerdo con las propiedades de la distribución normal, $y_1, ..., y_n$ son variables aleatorias normales independientes con media $E[y_i] = w_1 x_i + w_0$ y varianza $\text{Var}[y_i] = \sigma^2$. Puesto que el valor de la variable y depende de x, generalmente se escribe $E[y|x] = w_1 x + w_0$ con el fin de enfatizar dicha dependencia.

Observamos además que en ocasiones podemos estar interesados en establecer una relación lineal entre una variables respuesta y_i y más de una variable explicativa, digamos $\mathbf{x}_i = [x_{i,1}, ..., x_{i,n}]^T$, obteniendo la siguiente relación:

$$\mathbf{y} = X^T \mathbf{w} + \epsilon. \tag{2.32}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2 I_n). \tag{2.33}$$

donde $\mathbf{w} = [w_0, w_1, ..., w_d]^T$ y $X = [\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_d]^T$. Es decir, X es la matriz cuyas columnas son los datos de entrenamiento, siendo d el número de variables explicativas en este caso.

En lo que sigue, denotamos por $\mathcal{D} = \{(\mathbf{x}_i, y_i), x_i \in \mathcal{X}, y_i \in \mathbb{R}\}_{i=1}^n$ al conjunto de datos de entrenamiento, donde $\mathbf{x}_i$ denota el vector de datos de entrada de dimensión d e y_i son los resultados obtenidos en la observación. Denotamos por X la matriz cuyas columnas son los vectores que constituyen los datos de entrada, por lo que $X \in \mathcal{M}^{d \times n}$. Esta escritura no es la más frecuente, sino que se corresponde con la traspuesta de la matriz de diseño habitual. Por otro lado, agrupamos los datos de salida en un vector $\mathbf{y}$, de modo que $\mathcal{D} = (X, \mathbf{y})$.

Asumimos inicialmente que existe una relación lineal entre los datos, de modo que tenemos el siguiente problema de regresión planteado al comienzo de la sección:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}, \quad y = f(\mathbf{x}) + \epsilon. \tag{2.34}$$

donde $\mathbf{x}$ es el vector de datos de entrada, $\mathbf{w}$ es el vector de pesos (parámetros), y son los datos de salida observados y f es la función de regresión que se desea estimar. Al igual que antes, asumiremos sin pérdida de generalidad que el ruido asociado a las observaciones sigue una distribución normal con media cero y varianza σ_n^2,

$$\epsilon \sim \mathcal{N}(0, \sigma_n^2). \tag{2.35}$$

En el caso exacto en el que $n = d$, podemos encontrar los parámetros $\mathbf{w}$ sin más que resolver el sistema $\mathbf{X}^T\mathbf{w} = \mathbf{y}$. No obstante, si el numero de datos es menor que la dimensión del espacio, hay muchas posibilidades para el vector $\mathbf{w}$ y todos ellos describen los datos de manera correcta, por lo que debemos seleccionar un criterio de preferencia. Por norma general, se favorecerán aquellos vectores $\mathbf{w}$ que tengan menor norma. En el caso en el que haya ruido en las observaciones, no es posible encontrar una solución exacta y será necesario usar una aproximación. En este caso seleccionaremos el vector $\mathbf{w}$ que proporcione un menor error. En situaciones en las que se den ambos problemas la solución será una combinación de ambas. Generalmente se estudian los cuadrados de los errores, dando lugar a la siguiente función de pérdida [42, Capítulo 2]:

$$\mathcal{L}(f, \mathcal{D}) = \mathcal{L}(\mathbf{w}, \mathcal{D}) = \sum_{i=1}^{n}(y_i - f(\mathbf{x}_i)^2 = \sum_{i=1}^{n}\mathcal{L}((\mathbf{x}_i, y_i), f), \tag{2.36}$$

donde $\mathcal{L}((\mathbf{x}_i, y_i), f)$ denota el error al cuadrado o pérdida de la función f en el ejemplo $(\mathbf{x}_i, y_i)$ y $\mathcal{L}(f, \mathcal{D})$ denota el error global de la función f en el conjunto de datos $\mathcal{D}$. El problema se reduce entonces a encontrar el vector de parámetros $\mathbf{w}$ que minimice la función de pérdida global. Usando notación matricial, la función de pérdida se puede escribir como

$$\mathcal{L}(f, \mathcal{D}) = \mathcal{L}(\mathbf{w}, \mathcal{D}) = (\mathbf{y} - X^T\mathbf{w})^T(\mathbf{y} - X^T\mathbf{w}). \tag{2.37}$$

Derivando ahora respecto de los parámetros $\mathbf{w}$ e igualando a cero las derivadas, obtenemos la solución óptima:

$$\frac{\partial\mathcal{L}(\mathbf{w}, \mathcal{D})}{\partial\mathbf{w}} = -2X\mathbf{y} + 2XX^T\mathbf{w} = 0 \Rightarrow XX^T\mathbf{w} = X\mathbf{y}. \tag{2.38}$$

Esta última ecuación recibe el nombre de ecuación normal. Siempre que la inversa de XX^T exista, podemos despejar $\mathbf{w}$ y obtener

$$\mathbf{w} = (XX^T)^{-1}X\mathbf{y}. \tag{2.39}$$

2.7. Regresión Ridge.

Vemos que para que (2.39) tenga sentido, la matriz XX^T debe ser invertible. Esto no siempre se cumple, ya que puede ser que no haya datos suficientes para garantizar que lo sea. También puede ocurrir que haya ruido en las observaciones, de modo que no sea lógico tratar de encontrar una solución exacta al problema de regresión. Para solventar estos inconvenientes se introduce un término de regularización. El más simple consiste, como se ha comentado anteriormente en la sección 2.6, en favorecer aquellos parámetros que tengan menor norma. Eligiendo este criterio, se obtiene la regresión Ridge [42, Capítulo 2].

Definición 2.7.1 (Regresión Ridge)**.** *El problema de regresión Ridge consiste en resolver el siguiente problema de optimización:*

$$\min_{\mathbf{w}} \mathcal{L}_\lambda(\mathbf{w}, \mathcal{D}) = \min_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i),$$

donde λ es un parámetro que controla el grado de regularización.

De nuevo, para obtener el valor óptimo de los parámetros se efectúan las derivadas parciales y se iguala a cero el resultado.

$$\frac{\partial \mathcal{L}(\mathbf{w}, \mathcal{D})}{\partial \mathbf{w}} = -2X\mathbf{y} + 2XX^T\mathbf{w} + 2\lambda\mathbf{w} = 0 \Rightarrow (XX^T + \lambda I_n)\mathbf{w} = X\mathbf{y}. \tag{2.40}$$

Observamos como en este caso la matriz $(XX^T + \lambda I_n)$ es invertible siempre que $\lambda > 0$, por lo que la solución óptima puede escribirse como:

$$\mathbf{w} = (XX^T + \lambda I_n)^{-1} X\mathbf{y}. \tag{2.41}$$

De esta ecuación observamos que necesitamos mantener tantos parámetros como dimensión tenga el espacio, además de invertir una matriz de dimensión $d \times d$. En el caso de trabajar en espacios de elevada dimensión, este procedimiento resulta poco eficiente. Resulta entonces conveniente expresar la solución en la forma dual de manera que no dependa de la dimensión del espacio sino del número de datos de entrenamiento:

$$\mathbf{w} = \lambda^{-1} X(\mathbf{y} - X^T\mathbf{w}) = X\alpha \Rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i, \tag{2.42}$$

con $\alpha = \lambda^{-1}(\mathbf{y} - X^T\mathbf{w})$. Operando, podemos obtener una expresión equivalente para α:

$$\lambda\alpha = (\mathbf{y} - X^T X\alpha) \Rightarrow (X^T X + \lambda I_n)\alpha = \mathbf{y} \Rightarrow \alpha = (K + \lambda I_n)^{-1}\mathbf{y},$$

donde $K = X^T X$, que es una matriz de dimensión $n \times n$, y depende por tanto del número de datos de entrenamiento y no de la dimensión del espacio. La función de regresión que buscamos puede escribirse entonces de la siguiente forma:

$$\hat{f}(\mathbf{x}) = \mathbf{x}^T\mathbf{w} = \langle \mathbf{x}, \mathbf{w} \rangle = \left\langle \sum_{i=1}^{n} \alpha_i \mathbf{x}_i, \mathbf{x} \right\rangle = \sum_{i=1}^{n} \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle = {k_{Xx}}^T (K + \lambda I_n)^{-1}\mathbf{y}, \tag{2.43}$$

con $[k_{Xx}]_i = \langle \mathbf{x}_i, \mathbf{x} \rangle = [k_{Xx}^T]_i$, de manera que basta con conocer los productos internos entre los datos para definirla. Esto nos será útil en secciones posteriores, donde introduciremos el truco del núcleo (en inglés, *kernel trick*), que nos permite proyectar los datos de entrenamiento, que inicialmente no guardan una relación lineal, en un espacio en el que dichas relaciones sean lineales.

CAPÍTULO 3

FUNCIONES NÚCLEO.

En muchas ocasiones las relaciones entre las variables no son lineales, por lo que los métodos de aprendizaje basados en modelos lineales no son válidos. No obstante, es posible obtener nuevos datos de entrenamiento a través de una transformación $\phi : \mathcal{X} \to \mathcal{F}$, con $\mathcal{F}$ un espacio de características adecuado. Dicho espacio será generalmente un espacio de Hilbert dotado de producto interno en el que las variables sí presenten relaciones lineales, pudiendo así aplicar los métodos de aprendizaje propios de este caso en dicho espacio. Además, veremos más adelante que tan sólo necesitamos efectuar los productos internos de los datos transformados, es decir, $k(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$, sin necesidad de calcular explícitamente las imágenes de los datos de entrenamiento por la transformación ϕ. Parece lógico plantearse entonces si es posible construir funciones $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ adecuadas que efectúen productos internos en un cierto espacio de características, en general desconocido. Este método se conoce como truco del núcleo (*kernel trick*) y plantea las siguientes cuestiones:

- ¿Cuándo una función $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ es una función núcleo?
- ¿Cómo se pueden definir funciones núcleo sin construir explícitamente la transformación ϕ?
- Dada una función núcleo, ¿es posible encontrar una transformación y un espacio asociado?
- ¿Cuál es el coste computacional de emplear el truco del núcleo en modelos de aprendizaje?

En la sección 3.1 se presentan los conceptos básicos en relación con las funciones núcleo. En la sección 3.2 se introduce el espacio de Hilbert reproductor del núcleo, que permite definir una función núcleo sin calcular explícitamente la transformación a un determinado espacio de características. A continuación, en la sección 3.3 se proporciona una representación en serie para núcleos continuos en dominios compactos, que puede ser posteriormente usada para definir el espacio de Hilbert reproductor del núcleo. En la sección 3.4 se definen conceptos como la continuidad y la diferenciabilidad en media cuadrática y se proporcionan ejemplos básicos de funciones núcleo. Por último, se presenta una adaptación a la regresión Ridge para el caso de variables que no presentan relaciones lineales. Ésta consiste en proyectar los datos en un espacio de características adecuado en el que poder emplear los métodos presentados en la sección 2.7 del capítulo anterior.

3.1. Introducción a las funciones núcleo.

A continuación se define el concepto de función núcleo [42, Definición 2.8].

Definición 3.1.1 (Función núcleo)**.** *Sea $\mathcal{X}$ un conjunto no vacío. Una función $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ recibe el nombre de función kernel o función núcleo si existe un espacio de Hilbert $\mathcal{F}$ y una transformación $\phi : \mathcal{X} \to \mathcal{F}$ tal que para todos $\mathbf{x}$, $\mathbf{z} \in \mathcal{X}$ se tiene que*

$$k(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle .$$

El espacio $\mathcal{F}$ recibe el nombre de espacio de características asociado a la función k y la transformación ϕ se conoce como transformación característica.

Es importante observar que dada una función núcleo, ni la transformación ϕ ni el espacio $\mathcal{F}$ están unívocamente determinados. Ilustramos esto con el siguiente ejemplo [42, Ejemplo 2.9].

Ejemplo 3.1.1. Sea $\mathcal{X} \subset \mathbb{R}^2$. Consideramos las siguientes transformaciones:

$$\begin{aligned} \phi_1 &: \mathbf{x} = (x_1, x_2) \to \phi_1(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2) \in \mathcal{F}_1 \subset \mathbb{R}^3. \\ \phi_2 &: \mathbf{x} = (x_1, x_2) \to \phi_2(\mathbf{x}) = (x_1^2, x_2^2, x_1x_2, x_2x_1) \in \mathcal{F}_2 \subset \mathbb{R}^4. \end{aligned}$$

Si efectuamos el producto interno de dos elementos cualesquiera obtenemos lo siguiente:

$$\begin{aligned} \langle \phi_1(\mathbf{x}), \phi_1(\mathbf{z}) \rangle &= \left\langle (x_1^2, x_2^2, \sqrt{2}x_1x_2), (z_1^2, z_2^2, \sqrt{2}z_1z_2) \right\rangle = x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\ &= (x_1z_1 + x_2z_2)^2 = \langle \mathbf{x}, \mathbf{z} \rangle^2 . \\ \langle \phi_2(\mathbf{x}), \phi_2(\mathbf{z}) \rangle &= \left\langle (x_1^2, x_2^2, x_1x_2, x_2x_1), (z_1^2, z_2^2, z_1z_2, z_2z_1) \right\rangle = x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\ &= (x_1z_1 + x_2z_2)^2 = \langle \mathbf{x}, \mathbf{z} \rangle^2 . \end{aligned}$$

Como resultado, vemos que la función $k(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle^2$ es una función núcleo con más de un espacio de características y más de una transformación asociadas.

En relación con los procesos Gaussianos, existe una equivalencia entre las funciones núcleo y las funciones de covarianza. Esta relación es la que comentábamos al final de la sección 2.2, gracias a la cual es posible, a través del Teorema de extensión de Kolmogorov, definir de forma unívoca un proceso a partir de una función de media y una función semidefinida positiva, que será una función de covarianza o función núcleo. Para establecer dicha relación, es necesario introducir la siguiente definición sobre las funciones semidefinidas positivas [46, Definición 4.15.]:

Definición 3.1.2. *Dado $\mathcal{X}$ un conjunto no vacío, diremos que una función $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ es semidefinida positiva si para todo $n \in \mathbb{N}$, $(c_1, ..., c_n) \subset \mathbb{R}$ y $(x_1, ..., x_n) \subset \mathcal{X}$,*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_jk(x_i, x_j) \geq 0.$$

Además, diremos que k es definida positiva si para $(x_1, ..., x_n) \subset \mathcal{X}$, todos ellos distintos, la igualdad se da solo si $c_1 = ... = c_n = 0$. Por último, diremos que k es simétrica si verifica que $k(x, z) = k(z, x)$ para todos $x, z \in \mathcal{X}$.

En consecuencia, sea $f \sim \mathcal{GP}(m, k)$ un proceso Gaussiano, que podemos suponer centrado, con $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$. Se tiene entonces que, por definición de la función de covarianza,

$$k(\mathbf{x}, \mathbf{y}) = E[f(\mathbf{x})f(\mathbf{y})].$$

Es fácil comprobar que dicha función de covarianza es una función núcleo en el sentido de la definición 3.1.1. Basta con considerar la siguiente transformación

$$\phi : \mathcal{X} \longrightarrow \mathcal{F} = \mathcal{L}^2(\Omega, \mathcal{A}, P),$$

dada por $\phi(\mathbf{x}) = f(\mathbf{x})$. De esta forma, tenemos que

$$k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle = E[f(\mathbf{x}), f(\mathbf{y})].$$

Así pues, una función de covarianza es automáticamente una función núcleo. El recíproco también es cierto por ser una función núcleo simétrica y semidefinida positiva, ya que el teorema de extensión de Kolmogorov nos permitiría definir un proceso Gaussiano cuya función de covarianza sea dicha función simétrica y semidefinida positiva. No obstante, en la siguiente subsección se discutirá de forma más detallada la equivalencia entre funciones de covarianza y funciones núcleo.

Por otro lado, es posible construir nuevas funciones núcleo a partir de unas dadas mediante combinaciones lineales y transformaciones adecuadas. La siguiente proposición [42, Proposición 3.22] introduce algunas de las operaciones válidas para obtener nuevas funciones núcleo. La demostración de este resultado es trivial, por lo que se ha omitido.

Proposición 3.1.1 (Construcción de funciones núcleo)**.** *Sean $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \mathcal{X} \subset \mathbb{R}^n$ funciones núcleo, $a \in \mathbb{R}^+$,$f(\cdot)$ una función real definida en $\mathcal{X}$, $\phi : \mathcal{X} \to \mathbb{R}^N$ con $k_3 : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ una función núcleo sobre $\mathbb{R}^N \times \mathbb{R}^N$ y $\mathbf{B} \in \mathcal{M}^{n \times n}$ una matriz simétrica semidefinida positiva. Entonces, las siguientes funciones son funciones núcleo:*

- $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) + k_2(\mathbf{x}, \mathbf{z})$.
- $k(\mathbf{x}, \mathbf{z}) = ak_1(\mathbf{x}, \mathbf{z})$.
- $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z})k_2(\mathbf{x}, \mathbf{z})$.
- $k(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z})$.
- $k(\mathbf{x}, \mathbf{z}) = k_3(\phi(\mathbf{x}), \phi(\mathbf{z}))$.
- $k(\mathbf{x}, \mathbf{z}) = \mathbf{x}' B \mathbf{x}$.

3.2. Caracterización de la función núcleo. RKHS.

No siempre resulta sencillo construir un determinado espacio de características, por lo que es conveniente establecer criterios que permitan construir funciones núcleo sin explicitar la transformación empleada. Por otro lado, dada una función $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ y dados $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathcal{X}$ fijos, la matriz $[k_{XX}]_{ij} = k(x_i, x_j)$ recibe el nombre de matriz de Gram asociada. Con esta notación, la definición 3.1.2 es equivalente a que la matriz de Gram asociada sea semidefinida positiva.

Es fácil ver que toda función núcleo es simétrica, ya que el producto interno en $\mathcal{F}$ es simétrico y hemos definido la función núcleo como un producto interno en un cierto espacio de características $\mathcal{F}$. Además, la función núcleo es también semidefinida positiva, ya que

$$\sum_{i,j=1}^{n} v_i v_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i,j=1}^{n} v_i v_j \langle \phi(\mathbf{x}_i, \phi(\mathbf{x}_j \rangle = \left\langle \sum_{i=1}^{n} v_i \phi(\mathbf{x}_i), \sum_{j=1}^{n} v_j \phi(\mathbf{x}_j) \right\rangle = \| \sum_{i=1}^{n} v_i \phi(\mathbf{x}_i) \|^2 \geq 0.$$

El siguiente teorema [46, Teorema 4.16] muestra que estas propiedades no son solo necesarias sino suficientes para definir una función núcleo. La clave consiste en construir un espacio de Hilbert asociado a dicha función núcleo que recibe el nombre de Espacio de Hilbert Reproductor del núcleo, RKHS (del inglés, *Reproducing Krenel Hilbert Space*). Veremos más adelante que este espacio tiene la propiedad de que si dos funciones f y g están próximas, en términos de la distancia definida por el producto interno asociado al espacio, entonces los valores de $f(\mathbf{x})$ serán cercanos a los valores de $g(\mathbf{x})$. Como consecuencia, veremos que los RKHS se caracterizan porque las funciones de evaluación, $\delta : x \to f(x)$ son aplicaciones continuas.

Teorema 3.2.1 (Caracterización del núcleo)**.** *Una función $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, es una función núcleo si y sólo si es simétrica y semidefinida positiva.*

Demostración. La discusión anterior prueba que una función núcleo es simétrica y semidefinida positiva, por lo que es suficiente con probar la otra implicación. Supongamos entonces que k es una función semidefinida positiva, de manera que buscamos construir una transformación ϕ con llegada en un espacio de Hilbert del cual k es la función núcleo.

Sea

$$\mathcal{F}_0 = \left\{ \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \cdot) : n \in \mathbb{N}, \mathbf{x}_i \in \mathcal{X}, \alpha_i \in \mathbb{R}, i = 1, ..., n \right\}.$$

Observamos que los elementos de este espacio son funciones. Puesto que tanto la suma de elementos del espacio como el producto de un elemento de éste por un escalar pertenecen a $\mathcal{F}_0$, tenemos que $\mathcal{F}_0$ es un espacio vectorial. Sean $f, g \in \mathcal{F}_0$ dadas por

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i k(\mathbf{x}_i \mathbf{x}) \text{ y } g(\mathbf{x}) = \sum_{j=1}^{n} \beta_j k(\mathbf{z}_j, \mathbf{x}).$$

Podemos definir el siguiente producto escalar:

$$\langle f, g \rangle = \sum_{i=1}^{l} \sum_{j=1}^{n} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{l} \alpha_i g(\mathbf{x}_i) = \sum_{j=1}^{n} \beta_j f(\mathbf{z}_j), \tag{3.1}$$

donde las dos últimas igualdades muestran que éste es independiente de la representación de f y g. Observamos que se cumplen las propiedades propias de un producto interno, ya que es simétrico, bilineal y $\langle f, f \rangle \geq 0$. Esto último se deduce de la hipótesis de que la función núcleo es definida positiva,

$$\langle f, f \rangle = \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0.$$

Por tratarse de un producto escalar, es fácil ver que verifica la desigualdad de Cauchy-Schwarz,

$$| \langle f, g \rangle |^2 \leq \langle f, f \rangle \langle g, g \rangle \qquad f, g \in \mathcal{F}_0.$$

Por otro lado, sea $f \in \mathcal{F}_0$ tal que $\langle f, f \rangle = 0$. Entonces, para todo $\mathbf{x} \in \mathcal{X}$ se tiene que

$$|f(\mathbf{x})|^2 = |\sum_{i=1}^{l} \alpha_i k(\mathbf{x}_i \mathbf{x})|^2 = |\langle f, k(\mathbf{x}, \cdot) \rangle|^2 \leq \langle k(\mathbf{x}, \cdot), k(\mathbf{x}, \cdot) \rangle \cdot \langle f, f \rangle = 0,$$

por lo que necesariamente $f = 0$. Así pues, queda demostrado que (3.1) es un producto escalar en $\mathcal{F}_0$.

Por último, para poder definir un espacio de Hilbert necesitamos que éste sea completo. Consideramos un vector fijo $\mathbf{x}$ y una sucesión de Cauchy $(f_n)_{n=1}^{\infty}$. Tenemos entonces que

$$(f_n(\mathbf{x}) - f_m(\mathbf{x}))^2 = \langle f_n - f_m, k(\mathbf{x}, \cdot) \rangle^2 \leq \|f_n - f_m\|^2 k(\mathbf{x}, \mathbf{x}),$$

donde hemos utilizado la desigualdad de Cauchy-Schwarz, que sabemos que verifica el producto escalar (3.1). Así pues, $(f_n(\mathbf{x}))_{n=1}^{\infty}$ es una sucesión de Cauchy de números reales. Se deduce entonces que ha de estar acotada, por lo que tiene límite. Definimos entonces la función

$$f(\mathbf{x}) = \lim_{n \to \infty} f_n(\mathbf{x}),$$

e incluimos todas las funciones de este tipo en el espacio $\mathcal{F}_0$, completando así el espacio y obteniendo un espacio de Hilbert $\mathcal{F}$ asociado a la función núcleo k. Podemos entonces definir el producto escalar entre elementos de $\mathcal{F}$ de la siguiente forma:

$$\langle f, g \rangle = \lim_{n \to \infty} \langle f_n, g_n \rangle,$$

donde $(f_n(\mathbf{x}))_{n=1}^{\infty}$ y $(g_n(\mathbf{x}))_{n=1}^{\infty}$ son sucesiones de Cacuhy en $\mathcal{F}_0$, dotado del producto escalar que hemos definido antes, cuyos límites son, respectivamente, f y g, elementos de $\mathcal{F}$.

Además, así construido, el espacio $\mathcal{F}$ está dotado de una propiedad adicional que recibe el nombre de *propiedad reproductiva* y se obtiene directamente de la definición del producto interno (3.1) tomando $g = k(\mathbf{x}, \cdot)$:

$$\langle f, k(\mathbf{x}, \cdot) \rangle = \sum_{i=1}^{l} \alpha_i k(\mathbf{x}_i, \mathbf{x}) = f(\mathbf{x}). \tag{3.2}$$

Hemos construido de esta forma un espacio de características y tan sólo quedaría especificar la transformación correspondiente, definida como sigue:

$$\phi : \mathbf{x} \in \mathcal{X} \to \phi(\mathbf{x}) = k(\mathbf{x}, \cdot) \in \mathcal{F}.$$

□

Dada una función k simétrica y semidefinida positiva nos referiremos al correspondiente espacio $\mathcal{F}_k$ como *espacio de Hilbert reproductor del núcleo (RKHS)* y denotaremos por $\langle \cdot, \cdot \rangle_{\mathcal{F}_k}$ al producto escalar asociado. De la definición del espacio $\mathcal{F}_k$ es fácil observar que las funciones f pertenecientes al RKHS asociado a la función núcleo k heredan las propiedades de éste. Así pues, si el núcleo es s-veces diferenciable, también lo serán las funciones en el espacio de Hilbert asociado. Por otro lado, una propiedad importante del RKHS es que la norma $\|f\|_{\mathcal{F}_k}$ recoge información tanto de la dimensión de la función como de su suavidad. Así pues, cuanto menor sea la norma, más suave será la función y viceversa. Este aspecto será de especial importancia al tratar con la regresión Ridge, ya que impondremos un término de regularización, que es precisamente $\|f\|_{\mathcal{F}_k}$, para evitar problemas

de sobreajuste.

La siguiente definición recoge las características fundamentales del espacio RKHS [46, Definición 4.18].

Definición 3.2.1. *Sea $\mathcal{X}$ un espacio no vacío y sea $\mathcal{F}$ un espacio de Hilbert sobre $\mathcal{X}$, es decir, un espacio de Hilbert formado por funciones $f : \mathcal{X} \to \mathbb{R}$.*

- *Una función $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ es un núcleo reproductor de $\mathcal{F}$ si verifica que $k(\mathbf{x}, \cdot) \in \mathcal{F}$ para todo $\mathbf{x} \in \mathcal{X}$. Además, debe cumplirse la propiedad reproductiva,*

  $$f(\mathbf{x}) = \langle f, k(\mathbf{x}, \cdot)\rangle$$

  para todas $f \in \mathcal{F}$ y todos $\mathbf{x} \in \mathcal{X}$.

- *El espacio $\mathcal{F}$ recibe el nombre de espacio de Hilbert reproductor del núcleo sobre $\mathcal{X}$ si para todos $\mathbf{x} \in \mathcal{X}$ la función delta de Dirac $\delta_x : \mathcal{F} \to \mathbb{R}$ definida por*

  $$\delta_x(f) = f(\mathbf{x})$$

  es continua.

El siguiente lema muestra que un núcleo reproductor es una función núcleo en el sentido de la definición 3.1.1 [46, Lema 4.19].

Lema 3.2.1 (Un núcleo reproductor es un núcleo)**.** *Sea $\mathcal{F}$ un espacio de Hilbert de funciones sobre $\mathcal{X}$ con núcleo reproductor k. Entonces, $\mathcal{F}$ es un RKHS y es también un espacio de características del núcleo k, donde la transformación $\phi : \mathcal{X} \to \mathcal{F}$ está dada por*

$$\phi(\mathbf{x}) = k(\mathbf{x}, \cdot).$$

Dicha transformación recibe el nombre de transformación canónica característica.

Demostración. Por la propiedad reproductiva, toda función de Dirac puede representarse a través del núcleo reproductor:

$$|\delta_x(f)| = |f(\mathbf{x})| = |\langle f, k(\mathbf{x}, \cdot)\rangle| \leq \|k(\mathbf{x}\cdot)\|_{\mathcal{F}}\|f\|_{\mathcal{F}}, \tag{3.3}$$

para todos $\mathbf{x} \in \mathcal{X}$, $f \in \mathcal{F}$. Esto prueba la continuidad de las funciones delta de Dirac. Por otro lado, fijamos $\mathbf{x}' \in \mathcal{X}$ y escribimos $f := k(\mathbf{x}', \cdot)$. Entonces, para todo $\mathbf{x} \in \mathcal{X}$, en virtud de la propiedad reproductiva, se tiene

$$\big\langle \phi(\mathbf{x}), \phi(\mathbf{x}')\big\rangle = \big\langle k(\mathbf{x}, \cdot), k(\mathbf{x}', \cdot)\big\rangle = \langle k(\mathbf{x}, \cdot), f\rangle = f(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}').$$

□

Es posible demostrar que el RKHS determina unívocamente la función núcleo k y viceversa. En este sentido, el Teorema de Moore-Aronszajn [46, Teoremas 4.21 y 4.22] establece que el RKHS asociado a una función núcleo es único, por lo que dos espacios RKHS con la misma función núcleo son isomorfos.

3.3. Representación de Mercer del RKHS.

En esta sección introducimos el Teorema de Mercer, que proporciona un desarrollo en serie para núcleos continuos en dominios compactos. Esta representación en serie puede ser usada posteriormente para describir el RKHS. Definimos el siguiente operador $T_k f : \mathcal{L}_2(\nu) \to \mathcal{L}_2(\nu)$ como

$$(T_k f)(\cdot) := \int_{\mathcal{X}} k(\cdot, \mathbf{x}) f(\mathbf{x}) d\nu(\mathbf{x}), \quad f \in \mathcal{L}_2(\nu), \tag{3.4}$$

donde ν es una medida de Borel finita en $\mathcal{X}$ y $\mathcal{L}_2(\nu)$ es el espacio de Hilbert de las funciones de cuadrado integrable con respecto a la medida ν. En el caso de que la función núcleo k sea invariante por traslación, dicho operador no es más que la convolución de f y k, por lo que si k es suave, $T_k f$ es una versión más suave que f. Además, el operador $T_k f$ es positivo, ya que

$$\int \int_{\mathcal{X} \times \mathcal{X}} k(\mathbf{z}, \mathbf{x}) f(\mathbf{z}) f(\mathbf{x}) d\nu(\mathbf{x}) d\nu(\mathbf{z}) \geq 0 \text{ para todo } f \in \mathcal{L}_2(\nu).$$

Puesto que es también un operador compacto y autoadjunto [46, Teorema 4.27], de acuerdo con el teorema espectral [46, Teorema A.5.13] existe un conjunto numerable de vectores ortonormales $(e_i)_{i \in I}$ y una familia $(\lambda_i)_{i \in I} \subset \mathbb{R}$ que converge a 0 tal que $|\lambda_1| \geq |\lambda_2| \geq ... > 0$ y

$$T_k f = \sum_{i \in I} \lambda_i \langle f, e_i \rangle e_i, \tag{3.5}$$

donde $f \in \mathcal{L}_2(\nu)$. De hecho, $\{\lambda_i : i \in I\}$ es el conjunto de autovalores no nulos del operador $T_k f$ y los e_i son los autovectores asociados. El Teorema de Mercer [46, Teorema 4.49] muestra que el núcleo k puede escribirse en términos de los autovalores y autovectores del operador $T_k f$.

Teorema 3.3.1 (Teorema de Mercer)**.** *Sea $\mathcal{X}$ un espacio métrico compacto, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ una función núcleo continua, ν una medida de Borel cuyo soporte es $\mathcal{X}$ y $(e_i, \lambda_i)_{i \in I}$ dados por (3.5). Entonces*

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i \in I} \lambda_i e_i(\mathbf{x}) e_i(\mathbf{x}'), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \tag{3.6}$$

donde la convergencia es absoluta y uniforme sobre $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$.

Con la notación introducida, es evidente que $\phi : \mathcal{X} \to \mathbb{R}$ dada por $\phi(\mathbf{x}) := \lambda_i e_i(\mathbf{x})_{i \in I}, \mathbf{x} \in \mathcal{X}$ es una transformación característica asociada al núcleo k.

El siguiente teorema introduce la representación de Mercer del RKHS [46, Teorema 4.51].

Teorema 3.3.2 (Representación de Mercer del RKHS)**.** *Con la notación y las hipótesis empleadas en el Teorema 3.3.1, definimos*

$$\mathcal{F} := \left\{ \sum_{i \in I} a_i \sqrt{\lambda_i} e_i : (a_i) \in l_2(I) \right\}.$$

Es más, para $f := \sum_{i \in I} a_i \sqrt{\lambda_i} e_i \in \mathcal{F}$ y $g := \sum_{i \in I} b_i \sqrt{\lambda_i} e_i \in \mathcal{F}$ definimos el producto escalar como

$$\langle f, g \rangle_{\mathcal{F}} = \sum_{i \in I} a_i b_i. \tag{3.7}$$

Entonces, $\mathcal{F}$ con el producto escalar (3.7) es el RKHS asociado a k.

3.4. Propiedades de regularidad de las funciones núcleo.

Tal y como se introdujo en la sección 2.2, la función de covarianza asociada a un proceso Gaussiano, que es una función núcleo, recoge información acerca de la regularidad de las trayectorias de tal proceso. Así pues, la función de covarianza o función núcleo juega un papel importante en los métodos de aprendizaje basados en procesos Gaussianos. En problemas de regresión lineal parece lógico pensar que las observaciones asociadas a dos puntos próximos estarán también próximas, por lo que las trayectorias deberán ser, como mínimo, continuas. Por tanto, los datos de entrenamiento que estén cerca de un cierto dato de prueba deberían ser informativos sobre el valor esperado de la observación de éste. En el caso de los procesos Gaussianos, será la función de covarianza la que proporcione información acerca de la similitud entre datos.

Ya hemos visto que una función de covarianza o función núcleo debe verificar una serie de propiedades, tales como ser simétrica y semidefinida positiva. Por tanto, está claro que una función arbitraria de $\mathbf{x}$ y $\mathbf{x}'$ no será, en general, una función de covarianza. El objetivo de esta sección es proporcionar ejemplos de funciones de covarianza usadas frecuentemente así como estudiar sus propiedades. En la subsección 3.4.1 se definen las funciones de covarianza estacionarias e isotrópicas. Posteriormente, en la subsección 3.4.2 se introducen los conceptos de continuidad y diferenciabilidad en media cuadrática. Por último, en las secciones 3.4.3 y 3.4.4 se dan ejemplos de las funciones de covarianza más empleadas en la literatura y se define el concepto de funciones de base radial, respectivamente.

3.4.1. Funciones de covarianza estacionarias e isotrópicas.

En ocasiones es necesario imponer ciertas restricciones sobre los procesos que deseamos estudiar, de manera que sea posible obtener la distribución de probabilidad que genera unos ciertos datos de entrenamiento a partir de un número finito de estos. Una simplificación habitual consiste en suponer que la distribución de probabilidad es similar en distintos puntos del dominio $\mathcal{X}$. Una forma de definir este concepto es a través de la estacionaridad estricta [44, Sección 2]: para todo $n \in \mathbb{N}$ finito, $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathcal{X}$, $t_1, ..., t_n \in \mathbb{R}$ y $\mathbf{x} \in \mathcal{X}$,

$$P[f(\mathbf{x}_1 + \mathbf{x}) \leq t_1, ..., f(\mathbf{x}_n + \mathbf{x}) \leq t_n] = P[f(\mathbf{x}_1) \leq t_1, ..., f(\mathbf{x}_n) \leq t_n].$$

Es posible definir también la estacionaridad en términos de los dos primeros momentos del proceso f. En este sentido, supongamos que la función de covarianza de f depende únicamente de la diferencia entre dos puntos, $\mathbf{x} - \mathbf{y}$. Tenemos entonces que existe una función K denominada función de autocovarianza, tal que $\text{cov}(f(\mathbf{x}), f(\mathbf{y}))= K(\mathbf{x} - \mathbf{y})$ para todos $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Diremos entonces que un proceso es débilmente estacionario si tiene momentos finitos de orden dos, sus funciones de media son constantes y posee una función de autocovarianza. Notemos además que cualquier proceso que sea estrictamente estacionario es también débilmente estacionario, como se muestra a continuación:

$$\begin{aligned}(f(\mathbf{x}_1 + \mathbf{x}), f(\mathbf{x}_2 + \mathbf{x})) \overset{d}{=} (f(\mathbf{x}_1), f(\mathbf{x}_2)) \Rightarrow E[f(\mathbf{x}_1 + \mathbf{x})f(\mathbf{x}_2 + \mathbf{x})] = E[f(\mathbf{x}_1)f(\mathbf{x}_1)] \\ \Rightarrow k(\mathbf{x}_1 + \mathbf{x}, \mathbf{x}_2 + \mathbf{x}) = k(\mathbf{x}_1, \mathbf{x}_2) \Rightarrow k(\mathbf{x}_1, \mathbf{x}_2) = k(\mathbf{x}_1 - \mathbf{x}_2, 0) = \hat{k}(\mathbf{x}_1 - \mathbf{x}_2). \quad (3.8)\end{aligned}$$

Podemos entender entonces la estacionaridad como invarianza frente a traslaciones.

Por otro lado, es posible definir también invarianza frente a rotaciones [44, Sección 2]. En este sentido, diremos que un proceso es estrictamente isotrópico si sus distribuciones finito

dimensionales son invariantes frente a cualquier tipo de movimiento rígido. Es decir, para cualquier matriz ortogonal H de tamaño $d \times d$ y cualquier $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ se tiene:

$$P[f(H\mathbf{x}_1 + \mathbf{x}) \leq t_1, ..., f(H\mathbf{x}_n + \mathbf{x}) \leq t_n] = P[f(\mathbf{x}_1) \leq t_1, ..., f(\mathbf{x}_n) \leq t_n]$$

para todo $n \in \mathbb{N}$ finito, $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathcal{X}$, $t_1, ..., t_n \in \mathbb{R}$ y $\mathbf{x} \in \mathcal{X}$. Al igual que ocurría con la estacionaridad, podemos definir un proceso débilmente isotrópico como un proceso tal que existe una constante m y una función K en $[0, \infty)$ tales que $m(\mathbf{x}) = m$ y $\text{cov}(f(\mathbf{x}), f(\mathbf{y}))= K(\|\mathbf{x} - \mathbf{y}\|)$ para todos $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. La función K recibe el nombre de función de autocovarianza isotrópica del proceso f. De esta definición se deduce que un proceso isotrópico es estacionario. Por otro lado, se deduce que todo proceso estrictamente isotrópico es débilmente isotrópico. La demostración de esto último es equivalente a la dada por (3.8).

$$\begin{aligned}(f(H\mathbf{x}_1 + \mathbf{x}), f(H\mathbf{x}_2 + \mathbf{x})) \overset{d}{=} (f(\mathbf{x}_1), f(\mathbf{x}_2)) \Rightarrow E[f(H\mathbf{x}_1 + \mathbf{x})f(H\mathbf{x}_2 + \mathbf{x})] = E[f(\mathbf{x}_1)f(\mathbf{x}_1)] \\ \Rightarrow k(H\mathbf{x}_1 + \mathbf{x}, H\mathbf{x}_2 + \mathbf{x}) = k(\mathbf{x}_1, \mathbf{x}_2) \Rightarrow k(\mathbf{x}_1, \mathbf{x}_2) = k(H(\mathbf{x}_1 - \mathbf{x}_2), 0) = \hat{k}(\|\mathbf{x}_1 - \mathbf{x}_2\|)\end{aligned} \quad (3.9)$$

La ventaja que poseen las funciones de covarianza estacionarias es que es posible caracterizarlas mediante su densidad espectral a través del teorema de Bochner [20, Sección 2, Teorema 2].

Teorema 3.4.1 (Teorema de Bochner)**.** *Sea k una función definida en $\mathbb{R}^d$. Entonces k es una función de autocovarianza de un proceso estocástico continuo y débilmente estacionario si y sólo si se puede escribir como*

$$k(\mathbf{x}) = \int_{\mathbb{R}^d} e^{2\pi i \langle \mathbf{x}, \mathbf{u} \rangle} F(d\mathbf{u}), \quad (3.10)$$

donde F es una medida finita y positiva.

Por consiguiente, cuando la medida F tiene una densidad asociada, podemos definir una función f denominada densidad espectral de k. Diremos que k y f son duales de Fourier, y podemos calcularlas analíticamente de la siguiente manera:

$$k(\mathbf{x}) = \int e^{2\pi i \langle \mathbf{x}, \mathbf{u} \rangle} f(\mathbf{u}) d\mathbf{u}. \quad (3.11)$$

$$f(\mathbf{u}) = \int e^{-2\pi i \langle \mathbf{x}, \mathbf{u} \rangle} k(\mathbf{x}) d\mathbf{x}. \quad (3.12)$$

De acuerdo con este resultado, las siguientes afirmaciones son equivalentes:

- $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$ es una función de covarianza.
- $k(\mathbf{x}, \mathbf{y})$ es semidefinida positiva.
- $k(\mathbf{x}) = \int_{\mathbb{R}^d} e^{2\pi i \langle \mathbf{x}, \mathbf{u} \rangle} dF(\mathbf{u})$. Para el caso en el que la medida tenga una densidad asociada, $k(\mathbf{x}) = \int_{\mathbb{R}^d} e^{2\pi i \langle \mathbf{x}, \mathbf{u} \rangle} f(\mathbf{u}) du$, donde $f(\mathbf{u})$ es la densidad espectral asociada a dicha medida.

Por tanto, si podemos encontrar una función $f(\mathbf{u})$ integrable y positiva, tal que $k(\mathbf{x})$ es la transformada de Fourier de ésta, entonces queda demostrado que k es en efecto una función de covarianza.

3.4.2. Continuidad y diferenciabilidad en media cuadrática.

Resulta conveniente establecer una relación entre la suavidad de un proceso y su función de covarianza. Una forma de establecer dicha relación es a través de lo que se conoce como propiedades en media cuadrática del proceso. En este sentido, definimos la continuidad en media cuadrática de la siguiente forma [44, Sección 2.4]:

Definición 3.4.1 (Continuidad en media cuadrática)**.** *Diremos que un proceso $f(\mathbf{x})$ es continuo en media cuadrática en $\mathbf{x}^*$ si*

$$E[|f(\mathbf{x}) - f(\mathbf{x}^*)|]^2 \underset{\mathbf{x}\to\mathbf{x}^*}{\longrightarrow} 0.$$

Si esto es cierto para todo $\mathbf{x}^* \in A \subset \mathbb{R}^d$, diremos que $f(\mathbf{x})$ es continuo en media cuadrática en A.

En particular, para un proceso débilmente estacionario con función de autocovarianza K, $E[|f(\mathbf{x}) - f(\mathbf{x}^*)|]^2 = E[f(\mathbf{x})f(\mathbf{x})] - E[f(\mathbf{x})f(^*\mathbf{x})] - E[f(\mathbf{x}^*)f(\mathbf{x})] + E[f(\mathbf{x}^*)f(\mathbf{x}^*)] = 2[K(0) - K(\mathbf{x} - \mathbf{x}^*)]$, de manera que f es continuo en media cuadrática en $\mathbf{x}^*$ si y sólo si K es continua en el origen. Puesto que un proceso débilmente estacionario es continuo en media cuadrática en todos sus puntos, o no lo es en ninguno, podemos afirmar que un proceso será continuo en media cuadrática si y sólo si su función de autocovarianza es continua en el origen. Si la función K es continua en el origen, entonces lo es en todos su puntos:

$$\begin{aligned} |K(\mathbf{x}) - K(\mathbf{y})| &= |\mathrm{cov}(f(\mathbf{x}) - f(\mathbf{y}), f(0))| \leq [\mathrm{var}(f(\mathbf{x}) - f(\mathbf{y}))\mathrm{var}(f(0))]^{1/2} \\ &= [2(K(0) - K(\mathbf{x},\mathbf{y}))K(0)]^{1/2} \underset{\mathbf{y}\to\mathbf{x}}{\longrightarrow} 0. \end{aligned}$$

Es posible definir de manera similar la diferenciabilidad en media cuadrática a través de un límite en L^2 [44, Sección 2.4].

Definición 3.4.2 (Diferenciabilidad en media cuadrática)**.** *Un proceso f en $\mathbb{R}$ con momentos de orden dos finitos es diferenciable en media cuadrática en x si $\frac{f(x+h_n)-f(x)}{h_n}$ converge en media cuadrática para toda sucesión $(h_n) \underset{n\to\infty}{\longrightarrow} 0$, siendo dicho límite independiente de la sucesión escogida. Si dicho límite existe, lo denotamos por $f'(x)$.*

Para un proceso débilmente estacionario con función de autocovarianza K, definimos

$$f_h(x) = \frac{f(x+h) - f(x)}{h},$$

cuya función de autocovarianza es

$$K_h(x) = E[f_h(x)f_h(x)] = \frac{1}{h^2}\left(2K(x) - K(x+h) - K(x-h)\right).$$

Si K es dos veces diferenciable, entonces

$$\lim_{h\to 0} K_h(x) = -K''(x).$$

Es posible probar que un proceso f es diferenciable en media cuadrática si y sólo si $K''(x)$ existe y es finito, y además, si f es diferenciable en media cuadrática, entonces la función de autocovarianza de f' es $-K''$. Podemos definir derivadas de orden mayor de la siguiente forma: diremos que f es m-veces diferenciable en media cuadrática si es $(m-1)$-veces diferenciable en media cuadrática y K^{m-1} es diferenciable en media cuadrática. Repitiendo

los argumentos presentados, es posible probar que un proceso f es m-veces diferenciable en media cuadrática si y sólo si $K^{2m}(0)$ existe y es finito. De ser así, la función de autocovarianza de $f^{(m)}$ es $(-1)^m K^{(2m)}$.

Buscamos extender la noción de diferenciablidad en media cuadrática a un conjunto $\mathcal{X} \subset \mathbb{R}^d$. Sea $\{f(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ un proceso con función de covarianza $k(\mathbf{x}, \mathbf{y})$. Definimos en primer lugar la derivada direccional del proceso en la dirección $\mathbf{u}$ como

$$D_{\mathbf{u}} f(\mathbf{x}) = \lim_{h \to 0} f_{\mathbf{u},h}(\mathbf{x}), \quad \text{donde} \quad f_{\mathbf{u},h}(\mathbf{x}) = \frac{f(\mathbf{x} + h\mathbf{u}) - f(\mathbf{x})}{h}.$$

En esta definición, h es un escalar y el límite es en L^2. Definimos la función de covarianza de la derivada direccional del proceso como $k(\mathbf{x}, \mathbf{y}) = E[D_{\mathbf{u}} f(\mathbf{x}) D_{\mathbf{u}} f(\mathbf{y})]$. Entonces,

$$K_{\mathbf{u}} = \lim_{h \to 0} \lim_{k \to 0} E[D_{\mathbf{u},h} f(\mathbf{x}) D_{\mathbf{u},k} f(\mathbf{y})]. \tag{3.13}$$

Es decir,

$$f_{\mathbf{u},h}(\mathbf{x}) \xrightarrow[h \to 0]{L^2} D_{\mathbf{u},h} f(\mathbf{x}).$$
$$f_{\mathbf{u},k}(\mathbf{y}) \xrightarrow[k \to 0]{L^2} D_{\mathbf{u},k} f(\mathbf{y}).$$

por lo que

$$\lim_{h \to 0} \lim_{k \to 0} f_{\mathbf{u},h}(\mathbf{x}) f_{\mathbf{u},k}(\mathbf{y}) = D_{\mathbf{u},h} f(\mathbf{x}) D_{\mathbf{u},k} f(\mathbf{y}),$$

entendiendo las derivadas en L^2. Esto implica que

$$\lim_{h \to 0} \lim_{k \to 0} E[f_{\mathbf{u},h}(\mathbf{x}) f_{\mathbf{u},k}(\mathbf{y}) - D_{\mathbf{u},h} f(\mathbf{x}) D_{\mathbf{u},k} f(\mathbf{y})] = 0,$$

ya que la convergencia en L^2 implica convergencia en L^1, por lo que obtenemos (3.13). Podemos entonces extender el concepto de diferenciabilidad en media cuadrática a dominios de mayor dimensión de la siguiente forma [5, Sección 3]:

Definición 3.4.3 (Diferenciabilidad en media cuadrática)**.** *Un proceso $\{f(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ es diferenciable en media cuadrática en $\mathbf{x}_0$ si para cualquier dirección $\mathbf{u}$, existe un proceso $L_{\mathbf{x}_0}(\mathbf{u})$, lineal en $\mathbf{u}$, tal que*

$$f(\mathbf{x}_0 + \mathbf{u}) = f(\mathbf{x}_0) + L_{\mathbf{x}_0}(\mathbf{u}) + R(\mathbf{x}_0, \mathbf{u}), \quad \textit{donde} \quad \frac{R(\mathbf{x}_0, \mathbf{u})}{\|\mathbf{u}\|} \xrightarrow{L^2} 0. \tag{3.14}$$

En otras palabras, existe un proceso lineal $L_{\mathbf{x}_0}(\mathbf{u})$ tal que

$$\lim_{\mathbf{u} \to 0} E\left\{ \frac{f(\mathbf{x}_0 + \mathbf{u}) - f(\mathbf{x_0}) - L_{\mathbf{x}_0(\mathbf{u})}}{\|\mathbf{u}\|} \right\}^2 = 0.$$

Es importante observar que la existencia del límite $\lim_{h \to 0} \frac{f(\mathbf{x}_0 + \mathbf{u}) - f(\mathbf{x}_0)}{h}$ para todo $\mathbf{u}$ no implica que $f(\mathbf{x})$ sea continuo en media cuadrática en $\mathbf{x}_0$. No obstante, si un proceso $f(\mathbf{x})$ es diferenciable en media cuadrática en $\mathcal{X} \subset \mathbb{R}^d$ entonces es continuo en media cuadrática. Escribiendo la dirección $\mathbf{u}$ como $h\mathbf{v}$ donde $\mathbf{v}$ es un vector unitario en la dirección de $\mathbf{u}$ y h es un escalar que indica la magnitud de $\mathbf{u}$, tenemos que

$$f(\mathbf{x}_0 + \mathbf{u}) = f(\mathbf{x}_0 + h\mathbf{v}) = f(\mathbf{x}_0) + L_{\mathbf{x}_0}(h\mathbf{v}) + R(\mathbf{x}_0, h\mathbf{v}) = f(\mathbf{x}_0) + h\left(L_{\mathbf{x}_0}(\mathbf{v}) + \frac{R(\mathbf{x}_0, \mathbf{v})}{h} \right),$$

donde la última igualdad se deduce de la linealidad de $L_{\mathbf{x}_0}(\mathbf{u})$. Usando entonces que f es diferenciable en media cuadrática, se tiene que $f(\mathbf{x}_0+\mathbf{u}) \xrightarrow{L^2} f(\mathbf{x}_0)$ si $h \to 0$. Además, si f es diferenciable en media cuadrática, entonces las derivadas direccionales existen para cualquier dirección $\mathbf{u}$ y $D_{\mathbf{u},h}(\mathbf{x}) = L_{\mathbf{x}}(\mathbf{u})$. Esto se deduce de lo siguiente:

$$f_{\mathbf{u},h}(\mathbf{x}) = L_{\mathbf{x}}(\mathbf{u}) + \frac{R(\mathbf{x}, h\mathbf{u})}{h}.$$

Puesto que $\frac{R(\mathbf{x},h\mathbf{u})}{h} \xrightarrow[h\to 0]{L^2} 0$, tenemos que $f_{\mathbf{u},h}(\mathbf{x}) \xrightarrow[h\to 0]{L^2} L_{\mathbf{x}}(\mathbf{u})$, por lo que por definición de derivada direccional, se tiene que $D_{\mathbf{u},h}(\mathbf{x}) = L_{\mathbf{x}}(\mathbf{u})$.

3.4.3. Ejemplos de funciones de covarianza.

En esta sección introduciremos algunos ejemplos comunes de funciones de covarianza, siguiendo los presentados en [34, Capítulo 4].

- **Función de covarianza de Matern:** La covarianza de Matern entre dos puntos viene dada por la siguiente expresión:

$$k_{\text{Matern}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu} r}{l} \right), \tag{3.15}$$

  donde ν y l son parámetros positivos y K_{ν} es la función de Bessel modificada [1, Sección 9.6].

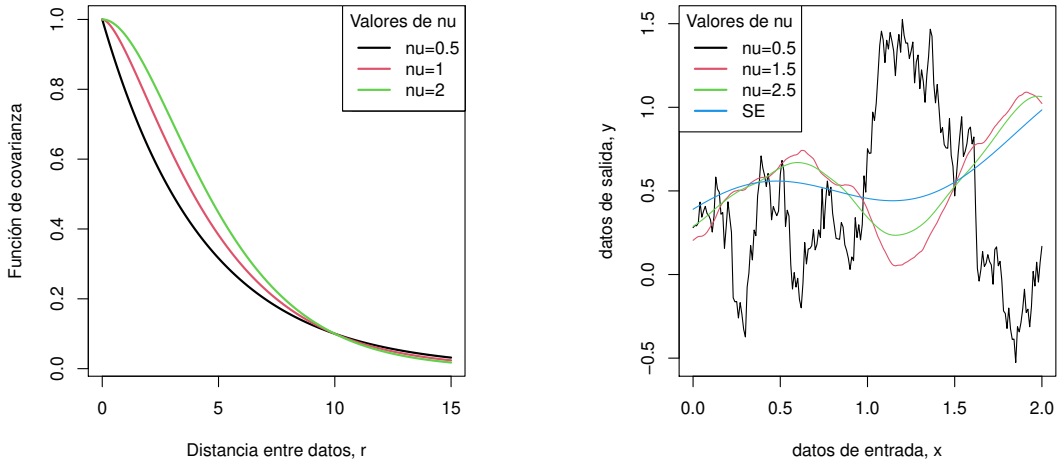


Figura 3.4.1: Función de covarianza de Matern y trayectorias de un proceso Gaussiano con dicha función.

El parámetro mas importante de dicha función de covarianza es μ, ya que determina la suavidad del proceso f. Cuánto mayor sea dicho parámetro, más suave será el proceso asociado. Para valores $\nu = p + \frac{1}{2}$, las funciones de covarianza de Matern son más suaves, por lo que suelen ser las más empleadas. En particular, se usan con mucha

frecuencia los valores $\nu = \frac{3}{2}$ y $\nu = \frac{5}{2}$. Los procesos asociados a funciones de covarianza de este tipo son $\nu - 1$ veces diferenciables, por lo que el parámetro ν controla el grado de suavidad de las muestras de los procesos. Así, para $\nu = 3/2$ tendremos procesos una vez diferenciables, mientras que para $\nu = 5/2$ serán doblemente diferenciables. Por otro lado, la función de covarianza exponencial al cuadrado, que se obtiene cuando $\nu \to \infty$, será infinitas veces diferenciable.

- **Función de covarianza exponencial y proceso de Ornstein-Uhlenbeck:** Si $\nu = \frac{1}{2}$ en la función de covarianza de Matern, obtenemos la función de covarianza exponencial. Un proceso con dicha función de covarianza será continuo en media cuadrática pero no diferenciable.

$$k(r) = \exp\left(-\frac{r}{l}\right). \tag{3.16}$$

- **Función de covarianza γ-exponencial:** Esta familia de funciones de covarianza incluye también la exponencial y la exponencial cuadrada. Su expresión viene dada por

$$k(r) = \exp\left(-\left(\frac{r}{l}\right)^{\gamma}\right) \text{ para } 0 < \gamma \leq 2. \tag{3.17}$$

Este tipo de funciones son menos flexibles que las funciones de covarianza de Matern, ya que el proceso asociado a éstas no es diferenciable en media cuadrática salvo para el caso $\gamma = 2$.

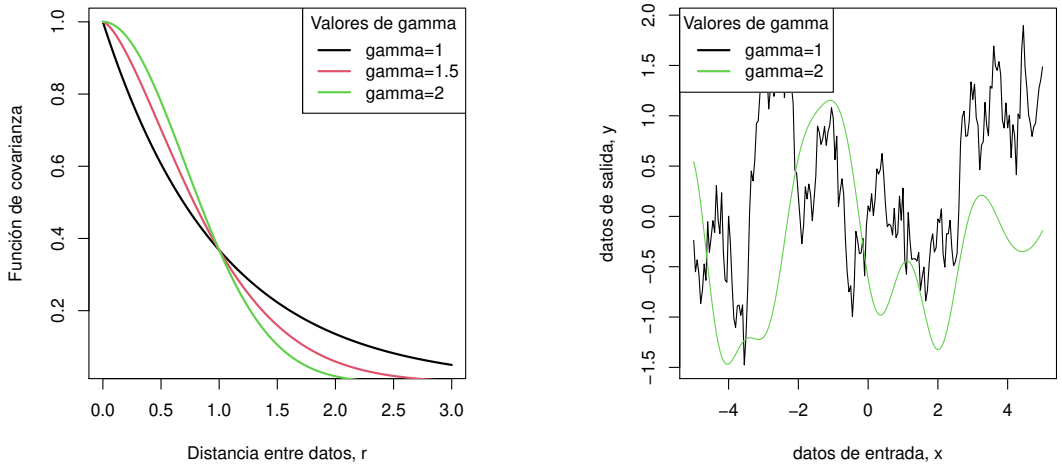


Figura 3.4.2: Función de covarianza γ-exponencial y trayectorias de un proceso Gaussiano asociado.

- **Función de covarianza racional cuadrática:** La expresión de esta función de covarianza viene dada por

$$k_{\mathrm{RQ}}(r) = \left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha}, \tag{3.18}$$

donde $\alpha, l > 0$. Este tipo de funciones de covarianza puede verse como una suma infinita de funciones de covarianza exponenciales al cuadrado con diferentes parámetros de longitud l. A diferencia de lo que ocurría con las funciones de covarianza

de Matern, este tipo de funciones de covarianza son infinitamente diferenciables en media cuadrática para todo α.

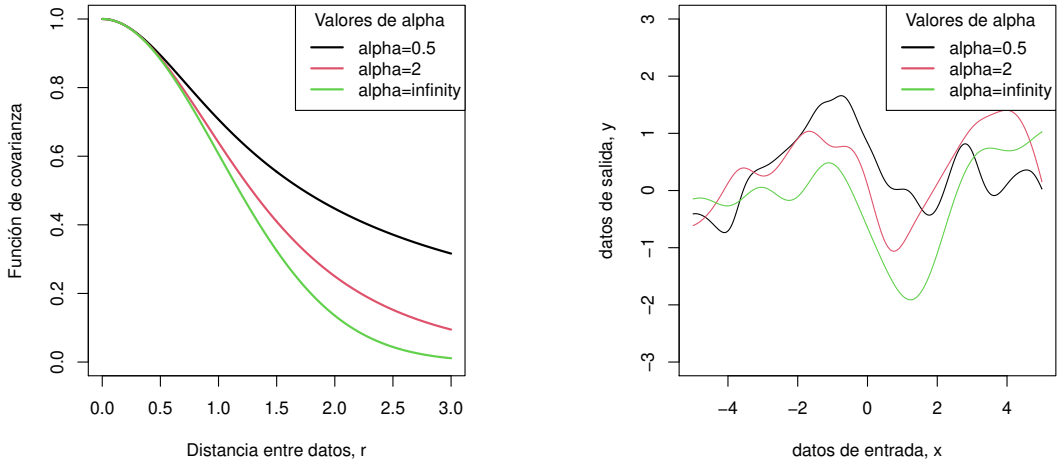


Figura 3.4.3: Función de covarianza racional cuadrática y trayectorias de un proceso Gaussiano asociado.

- **Función de covarianza periódica:** Dicha función de covarianza viene dada por la siguiente expresión

$$k(r) = \exp\left(-\frac{2sin^2\left(\pi|r|/p\right)}{l^2}\right). \tag{3.19}$$

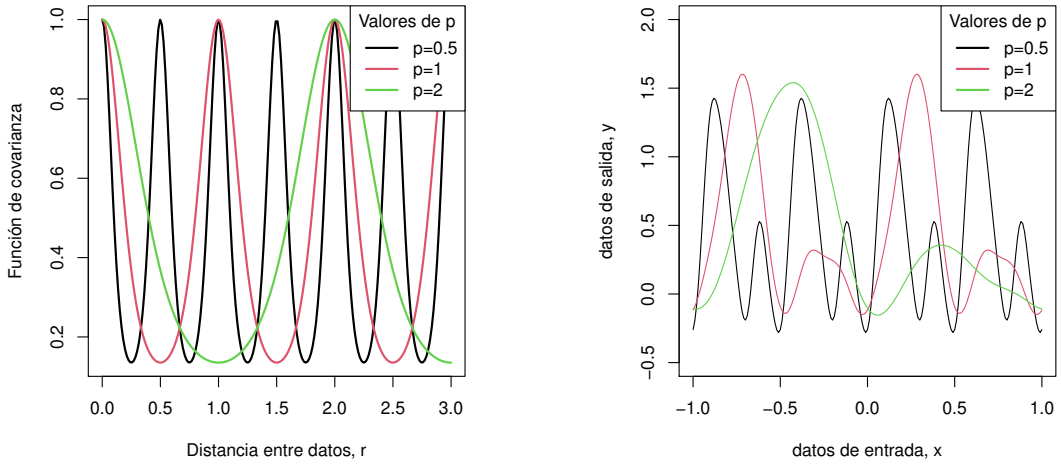


Figura 3.4.4: Función de covarianza periódica y trayectorias de un proceso Gaussiano con dicha función.

- **Función de covarianza exponencial cuadrada:** Hemos visto como esta función núcleo puede considerarse como una función de covarianza de Matern donde el

parámetro $\nu \to \infty$.

$$k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) \to k(r) = \exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{3.20}$$

Esta función núcleo tiene muchas aplicaciones y se usa habitualmente por ser infinitas veces diferenciable. Puesto que los procesos Gaussianos heredan las propiedades de las funciones de covarianza asociadas, un proceso Gaussiano con función de covarianza exponencial cuadrada tendrá derivadas en media cuadrática de todos los órdenes, siendo por tanto un proceso suave.

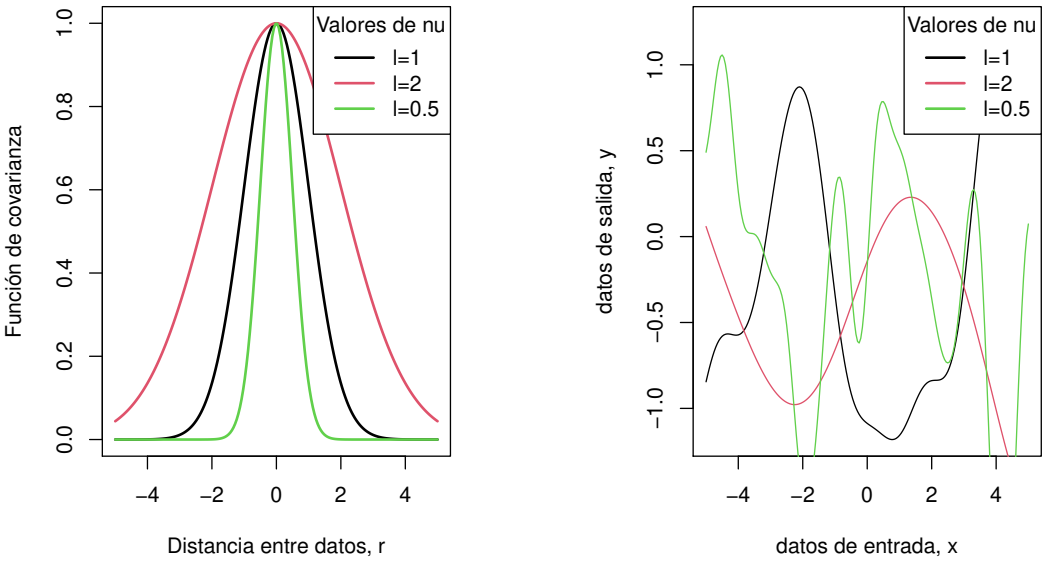


Figura 3.4.5: Función de covarianza exponencial cuadrada y trayectorias de un proceso Gaussiano con dicha función.

Existen varias formas de probar que la función de covarianza exponencial cuadrada es en efecto una función núcleo. Por un lado, podemos buscar una aplicación ϕ tal que $k(\mathbf{x},\mathbf{y}) = \langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle$. Definimos la siguiente transformación:

$$\phi : \mathcal{X} \longrightarrow \mathcal{H} \quad \text{con} \quad \phi(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{2\sigma^2}\right). \tag{3.21}$$

Por tanto, tenemos que

$$k(\mathbf{x},\mathbf{y}) = \int \exp\left(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{2\sigma^2}\right)\exp\left(-\frac{\|\mathbf{y}-\mathbf{z}\|^2}{2\sigma^2}\right)d\mathbf{z} = \sqrt{\pi}\sigma\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{4\sigma^2}\right) \tag{3.22}$$

Basta entonces con redefinir la desviación estándar, de modo que tomando $\sigma/\sqrt{2}$ y

$$\phi(\mathbf{x}) = \frac{2^{1/4}}{\pi^{1/4}\sqrt{\sigma}}\exp\left(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{2(\sigma/\sqrt{2})^2}\right),$$

tenemos que $k(\mathbf{x},\mathbf{y}) = \langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle$.

Es posible también emplear el Teorema de Bochner presentado en 3.4.1. Si encontramos una función de densidad adecuada $f(\mathbf{u})$, positiva e integrable, tal que $k(\mathbf{x})$ sea la transformada de Fourier de $f(\mathbf{u})$, queda probado que k es una función núcleo. Definimos la siguiente función $f(\mathbf{u})$:

$$f(\mathbf{u}) = (2\pi l^2)^{d/2} \exp(-2\pi^2 l^2 \mathbf{u}^2). \tag{3.23}$$

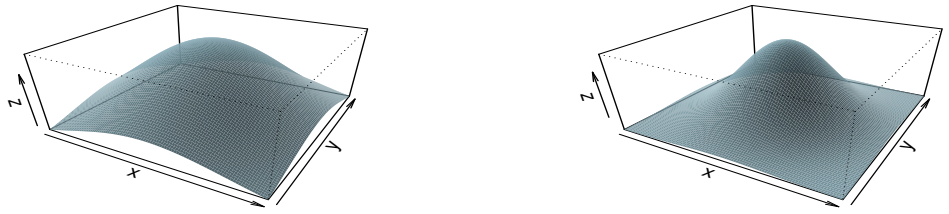
Se puede probar que la transformada de Fourier de dicha función es en efecto la función de covarianza exponencial al cuadrado.

3.4.4. Funciones de base radial.

En términos generales, las funciones de base radial constituyen una forma de aproximar funciones continuas que dependen de dos o más variables mediante combinaciones lineales de términos basados en una función de una única variable [13]. Denotamos a estas funciones por $\phi(r)$, enfatizando que dependen tan solo de la distancia r a un punto central, de modo que son funciones radialmente simétricas. Las funciones de base radial se usan en diversas aplicaciones, como la interpolación de datos, la resolución de ecuaciones diferenciales, resolución de ecuaciones integrales o ecuaciones diferenciales estocásticas. Algunos de las funciones de base radial (RBFs) más usadas son [24]:

- Gaussiana (GA): $\phi(r) = exp(-(\alpha r)^2)$.
- Cuadrática inversa (IQ): $\phi(r) = \frac{1}{(\alpha r)^2+1}$.
- Multicuádrica (MQ): $\phi(r) = \sqrt{(\alpha r)^2 + 1}$.
- Multicuádrica inversa (IMQ): $\phi(r) = \frac{1}{\sqrt{(\alpha r)^2+1}}$.
- Cúbica: $\phi(r) = r^3$.

donde en todos los casos α es un parámetro de forma.



(a) RBF Gaussiana con parámetro $\alpha = 1$ (b) RBF Gaussiana con parámetro $\alpha = 3$

Figura 3.4.6: Función de base radial.

Definición 3.4.4 (RBF [19]). *Una función $\phi : \mathbb{R}^d \to \mathbb{R}$ recibe el nombre de función radial si existe una función de una única variable $\psi : [0, \infty) \to \mathbb{R}$ tal que $\phi(\mathbf{x}) = \psi(r)$, donde $r = \|\mathbf{x}\|$ y $\|.\|$ denota alguna norma en $\mathbb{R}^d$, generalmente la norma Euclídea.*

3.5. Regresión Ridge.

A la vista de los resultados estudiados en este capítulo sobre las funciones núcleo, vamos a volver a analizar el método de regresión Ridge. Veremos que las funciones núcleo nos proporcionan una mayor flexibilidad a la hora de ajustar modelos de regresión. En las secciones 2.6 y 2.7 hemos introducido los conceptos básicos de este método de regresión. No obstante, hemos supuesto que existe una relación lineal entre las variables, de manera que la función buscada satisfacía la relación dada por la ecuación (2.34). No obstante, éste no es generalmente el caso. Una forma de solventar este problema es proyectar las variables desde el espacio de origen hacia un espacio de características en el que sí haya relaciones lineales entre las variables, de manera que los modelos lineales de regresión lineal sigan siendo válidos. Así pues, supongamos que tenemos la siguiente transformación:

$$\phi : \mathbf{x} \in \mathbb{R}^d \to \phi(\mathbf{x}) \in \mathcal{F}.$$

El conjunto de datos de entrenamiento es ahora de la forma $\mathcal{D} = \{(\phi(\mathbf{x}_1), y_1), ...(\phi(\mathbf{x}_n), y_n)\}$. Denotamos por $\Phi(X)$ la matriz cuyas columnas son las correspondientes transformaciones $\phi(\mathbf{x})$ de los n datos de entrenamiento. Así pues, $\Phi(X) \in \mathcal{M}^{N \times n}$. Con esta notación, el modelo de regresión es de la siguiente forma:

$$f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w}, \tag{3.24}$$

donde el vector de parámetros tiene dimensión N en lugar de d. Podemos repetir el procedimiento descrito anteriormente sin más que sustituir X por $\Phi(X)$. De esta forma, la regresión Ridge consistirá en minimizar la siguiente función de pérdida:

$$\min_{\mathbf{w}} \mathcal{L}_\lambda(\mathbf{w}, \mathcal{D}) = \min_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^{n} (y_i - f(\phi(\mathbf{x}_i))^2. \tag{3.25}$$

Sustituyendo $\mathbf{x}$ por $\phi(\mathbf{x})$ en las expresiones anteriores y denotando por $G = \Phi(X)^T \Phi(X)$, llegamos a las siguientes ecuaciones en relación con la función de regresión, que se corresponden con la representación primal y dual respectivamente:

$$\begin{aligned} \hat{f}(\mathbf{x}) &= \phi(\mathbf{x})^T \mathbf{w} = \phi(\mathbf{x})^T (\Phi(X)\Phi(X)^T + \lambda I_N)^{-1} \Phi(X)\mathbf{y} \\ &= \phi(\mathbf{x})^T (G^T + \lambda I_N)^{-1} \Phi(X)\mathbf{y}, \end{aligned} \tag{3.26}$$

$$\begin{aligned} \hat{f}(\mathbf{x}) &= \phi(\mathbf{x})^T \mathbf{w} = \langle \phi(\mathbf{x}), \mathbf{w} \rangle = \left\langle \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i), \phi(\mathbf{x}) \right\rangle = \sum_{i=1}^{n} \alpha_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle \\ &= {k_{Xx}}^T (G + \lambda I_n)^{-1} \mathbf{y}, \end{aligned} \tag{3.27}$$

donde en este caso $[k_{Xx}]_i = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle$. Vemos que la representación primal (3.26) involucra matrices de tamaño $N \times N$, por lo que si el espacio de características es de dimensión elevada, resulta poco eficiente. Por el contrario, la representación dual (3.27) involucra matrices de tamaño $n \times n$ y la información acerca del espacio de características está recogida en forma de productos escalares, por lo que empleando funciones núcleo adecuadas que nos permitan definir dichos producto escalares como función de los datos de entrada no es

necesario definir de forma explícita el espacio $\mathcal{F}$.

Resumiendo, tenemos dos maneras de expresar la solución óptima: Una es mediante la representación primal, en la que interviene la dimensión del espacio y el vector de parámetros se calcula explícitamente. Otra es la representación dual en la que la solución se expresa como combinación lineal de productos escalares calculables a partir de los datos de entrada, de manera que tenemos un sistema de tantas ecuaciones como datos, y la dimensión del espacio no interviene en la resolución de éste. Por tanto, si la dimensión d del espacio es mayor que el número de datos de entrenamiento, será más eficiente resolver la ecuación en la forma dual.

En la Figura 3.5.1 se ha representado la función $f(x) = x\sin(x) + \epsilon$, donde ϵ representa un ruido Gaussiano. Se ha empleado posteriormente el método de regresión Ridge para obtener una aproximación de la función. Vemos que se obtienen un buen acuerdo empleando un núcleo Gaussiano.

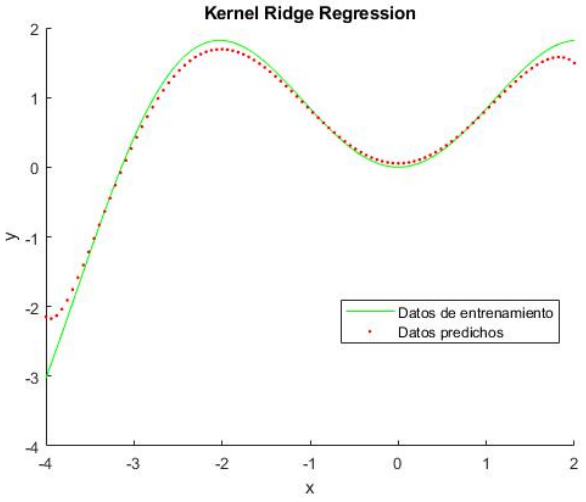
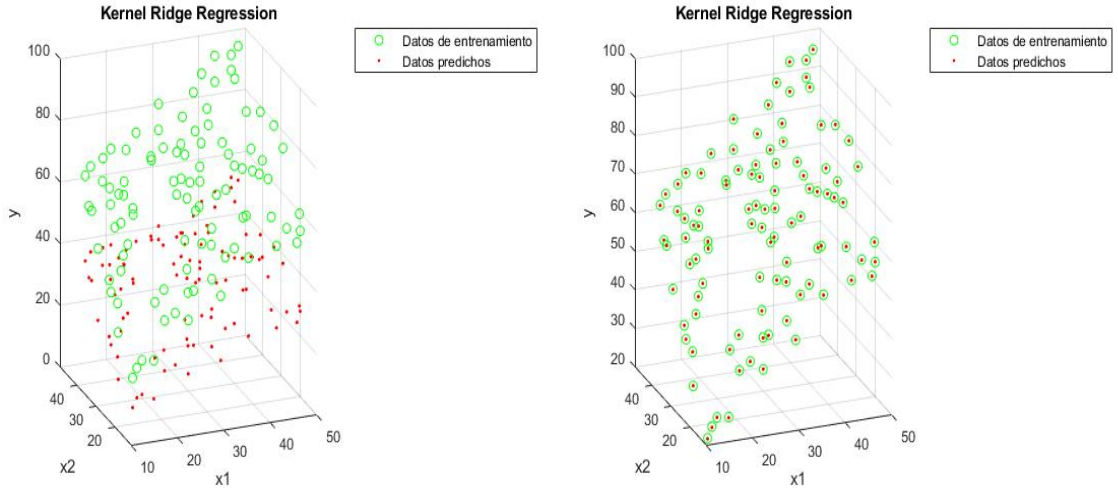


Figura 3.5.1: Regresión Rige con un núcleo Gaussiano.

No obstante, es importante notar que la elección de un núcleo adecuado determina en gran medida la calidad de la aproximación. Ilustramos esto con el ejemplo siguiente. Se han representado unos datos aleatorios en x_1 y x_2, siendo $y = x_1 + x_2 + \epsilon$, con ϵ un error aleatorio. En la figura 3.5.2a se ha empleado un núcleo Gaussiano, obteniendo una muy mala aproximación. En la imagen 3.5.2b, por el contrario, se ha empleado un núcleo lineal. Vemos como en este caso la calidad de la aproximación es notablemente mejor. Es por ello por lo que la determinación del núcleo juega un papel crucial en los problemas de regresión.



(a) Regresión Rige con un núcleo Gaussiano. (b) Regresión Ridge con un núcleo lineal.

Figura 3.5.2: Regresión Ridge para unos mismos datos cambiando la función núcleo empleada.

CAPÍTULO 4

REGRESIÓN BASADA EN PROCESOS GAUSSIANOS.

Hemos visto en la sección 2.4 en qué consiste el tratamiento del problema desde el punto de vista de la inferencia Bayesiana. Con respecto a los modelos de regresión basados en procesos Gaussianos, existen varias formas de interpretarlos. Por un lado, podemos definir un proceso Gaussiano como una distribución sobre funciones, por lo que la inferencia tendrá lugar en un espacio de funciones. Otro punto de vista consiste en utilizar un espacio de pesos en el cual se lleva a cabo la inferencia. En las secciones 4.1 y 4.2 se detallan ambos puntos de vista [34, Capítulo 2]. Es importante darse cuenta de que la regresión basada en procesos Gaussianos es un método no paramétrico, tal como se explica en la sección 4.2, por lo que la fase de entrenamiento no consiste en la optimización de una serie de parámetros sino que, desde el punto de vista numérico, se basa simplemente en la inversión de una matriz. En la sección 4.3 se introducen técnicas para la elección de los hiperparámetros óptimos. Por último, en las secciones 4.4 y 4.5 se introducen algunas consideraciones sobre los procesos Gaussianos y se comparan éstos con la regresión Ridge de capítulos anteriores.

4.1. Espacio de pesos.

Denotamos por $\mathcal{D} = \{(\mathbf{x}_i, y_i), x_i \in \mathcal{X}, y_i \in \mathbb{R}\}_{i=1}^n$ al conjunto de datos de entrenamiento, donde $\mathbf{x}_i$ denota el vector de datos de entrada de dimensión d e y_i es el resultado obtenido en cada observación. Denotamos por X la matriz cuyas columnas son los vectores que constituyen los datos de entrada, por lo que $X \in \mathcal{M}^{d\times n}$, y agrupamos los datos de salida en un vector $\mathbf{y}$, de modo que $\mathcal{D} = (X, \mathbf{y})$. En los problemas de regresión estaremos interesados en encontrar relaciones entre los datos de entrada y de salida.

En un primero momento asumimos que los datos siguen un modelo lineal de modo que tenemos el siguiente problema de regresión:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}, \quad y = f(\mathbf{x}) + \epsilon, \tag{4.1}$$

donde $\mathbf{x}$ es el vector de datos de entrada, $\mathbf{w}$ es el vector de pesos (parámetros), y son los datos de salida observados y f es la función de regresión que se desea estimar. La teoría es mucho más manejable si asumimos que el ruido asociado a las observaciones sigue una distribución normal con media cero y varianza σ_n^2:

$$\epsilon \sim \mathcal{N}(0, \sigma_n^2). \tag{4.2}$$

Definimos la función de verosimilitud como la densidad de las observaciones dados los parámetros. Es fácil ver que dicha densidad es normal de media $X^T\mathbf{w}$ y varianza $\sigma_n^2 I$.

$$\begin{aligned} p(\mathbf{y}|X,\mathbf{w}) &= \prod_{i=1}^{n} p(y_i|\mathbf{x}_i,\mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(y_i - \mathbf{x_i}^T\mathbf{w})^2}{2\sigma_n^2}\right) \\ &= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{|\mathbf{y} - X^T\mathbf{w}|^2}{2\sigma_n^2}\right) = \mathcal{N}(X^T\mathbf{w}, \sigma_n^2 I). \end{aligned} \tag{4.3}$$

Resulta conveniente asumir que los pesos se distribuyen de acuerdo con una normal de media cero y varianza Σ_p, ya que en otro caso, las integrales que aparecerán a lo largo del capítulo no pueden resolverse de forma analítica:

$$\mathbf{w} \sim \mathcal{N}(0, \Sigma_p). \tag{4.4}$$

Con el objetivo de obtener la distribución a posteriori, empleamos la regla de Bayes [34, Apéndice A.3]:

$$p(\mathbf{w}|\mathbf{y}, X) = \frac{p(\mathbf{y}|X,\mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|X)}. \tag{4.5}$$

Puesto que el denominador es una constante que no depende de los parámetros, obtenemos lo siguiente en relación a la distribución a posteriori:

$$\begin{aligned} p(\mathbf{w}|\mathbf{y}, X) &\propto \exp\left(-\frac{1}{2\sigma_n^2}(\mathbf{y} - X^T\mathbf{w})^T(\mathbf{y} - X^T\mathbf{w})\right) \exp\left(-\frac{1}{2}\mathbf{w}^T\Sigma_p^{-1}\mathbf{w}\right) \\ &\propto \exp\left(-\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^T(\frac{1}{\sigma_n^2}XX^T + \sigma_p^{-1})(\mathbf{w} - \bar{\mathbf{w}})\right), \end{aligned} \tag{4.6}$$

donde $\bar{\mathbf{w}} = \sigma_n^{-2}(\sigma_n^{-2}XX^T + \Sigma_p^{-1})^{-1}X\mathbf{y}$, de manera que tenemos una distribución normal de media $\bar{\mathbf{w}}$ y varianza A^{-1}, con $A = (\sigma_n^{-2}XX^T + \Sigma_p^{-1})$,

$$p(\mathbf{w}|\mathbf{y}, X) \sim \mathcal{N}(\bar{\mathbf{w}}, A^{-1}). \tag{4.7}$$

Denotando por f_z los valores predichos para un conjunto de datos de prueba $Z = (\mathbf{z}_1, ..., \mathbf{z}_m)$, podemos obtener la distribución de la función de regresión que buscamos usando las propiedades de linealidad de la distribución normal y recordando que f verifica (2.34):

$$p(f_z|\mathbf{z}, X\mathbf{y}) = \mathcal{N}(\frac{1}{\sigma_n^2}\mathbf{z}^T A^{-1}X\mathbf{y}, \mathbf{z}^T A^{-1}\mathbf{z}). \tag{4.8}$$

De nuevo, hemos supuesto que la relación entre las variables es lineal en el espacio de origen. Cuando esto no sea cierto, es posible que, transformando los datos de entrada desde dicho espacio a un espacio de características adecuado $\mathcal{F}$, dichas variables presenten relaciones lineales. Supongamos que $\mathcal{F} = \mathbb{R}^N$. Usando la transformación presentada en el capítulo 3,

$$\phi : \mathbf{x} \in \mathbb{R}^d \to \phi(\mathbf{x}) \in \mathcal{F},$$

y denotando por S al conjunto de datos de entrenamiento transformados, la distribución a posteriori descrita en (4.8) puede escribirse como

$$p(f_z|\mathbf{z}, X\mathbf{y}) = \mathcal{N}(\frac{1}{\sigma_n^2}\phi(\mathbf{z})^T A^{-1}\Phi(X)\mathbf{y}, \phi(\mathbf{z})^T A^{-1}\phi(\mathbf{z})), \tag{4.9}$$

con $A = (\sigma_n^{-2}\Phi(X)\Phi(X)^T + \Sigma_p^{-1})$. El problema que plantea esta expresión es la dimensión de la matriz A, de tamaño $N \times N$. Si estamos trabajando en un espacio de características de elevada dimensión resulta poco eficiente a nivel computacional. No obstante, es posible obtener una expresión equivalente denotando por $k_{XX} = \Phi(X)^T\Sigma_p\Phi(X)$, $k_{ZX} = \phi(\mathbf{z})^T\Sigma_p\Phi(X)$ y $k_{ZZ} = \phi(\mathbf{z})^T\Sigma_p\phi(\mathbf{z})$. Con esta notación, tenemos que

$$\sigma_n^{-2}\Phi(X)(k_{XX} + \sigma_n^2 I) = \sigma_n^{-2}\Phi(X)(\Phi(X)^T\Sigma_p\Phi(X) + \sigma_n^2 I) = A\Sigma_p\Phi(X).$$

Multiplicando ambos lados por A^{-1} obtenemos la siguiente igualdad:

$$\sigma_n^2 A^{-1}\Phi(X) = \Sigma_p\Phi(X)(k_{XX} + \sigma_n^2 I).$$

Por tanto, la media de la distribución a posteriori puede escribirse de la siguiente forma:

$$\bar{\mathbf{w}} = \bar{m} = \phi(\mathbf{z})^T\Sigma_p\Phi(X)(k_{XX} + \sigma_n^2 I)^{-1}\mathbf{y} = k_{ZX}(k_{XX} + \sigma_n^2 I)^{-1}\mathbf{y}. \tag{4.10}$$

Para obtener una expresión equivalente de la varianza, usamos la identidad matricial de Woodbury [Anexo B] con $A^{-1} = \Sigma_p$, $C^{-1} = \sigma_n^2 I_n$ $U = V = \Phi(X)$. Obtenemos de esta forma una expresión más sencilla:

$$\begin{aligned} \bar{k} &= \phi(\mathbf{z})^T\Sigma_p\phi(\mathbf{z}) - \phi(\mathbf{z})^T\Sigma_p\Phi(X)(k_{XX} + \sigma_n^2 I)^{-1}\Phi(X)^T\Sigma_p\phi(\mathbf{z}) \\ &= k_{ZZ} - k_{ZX}(k_{XX} + \sigma_n^2 I)^{-1}k_{ZX}^T. \end{aligned} \tag{4.11}$$

La ventaja que presenta esta escritura es que las matrices involucradas son de tamaño $n \times n$ en lugar de $N \times N$, lo cual resulta mas eficiente a la hora de calcular sus respectivas inversas si la dimensión del espacio de características es elevada. Por otro lado, observamos que el espacio de características aparece siempre en la forma $\phi(\mathbf{x})^T\Sigma_p\phi(\mathbf{x}')$, donde $\mathbf{x}$ y $\mathbf{x}'$ son puntos pertenecientes al conjunto de datos de entrenamiento. De esta forma, podemos definir la siguiente función:

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T\Sigma_p\phi(\mathbf{x}'), \tag{4.12}$$

que recibe el nombre de función de covarianzas y no es más que un producto interno con respecto de Σ_p. Por tratarse de un producto escalar, dicha función es continua, definida positiva y simétrica. La ventaja que tiene definir una función de esta forma es que no es necesario definir explícitamente la transformación ϕ, sino que basta con elegir una función de covarianzas adecuada.

Resumiendo, asumiendo que se cumplen (2.34), (4.2) y (4.4), se obtienen las siguientes expresiones para la función de media a posteriori y la función de covarianza a posteriori de un proceso Gaussiano:

$$\bar{m} = k_{ZX}(k_{XX} + \sigma_n^2 I)^{-1}\mathbf{y}. \tag{4.13}$$

$$\bar{k} = k_{ZZ} - k_{ZX}(k_{XX} + \sigma_n^2 I)^{-1}k_{ZX}^T. \tag{4.14}$$

4.2. Espacio de funciones.

Una alternativa al punto de vista anterior es trabajar directamente en un espacio de funciones. La ventaja que ofrece este punto de vista es que no estamos asumiendo linealidad entre las variables originales, por lo que no debemos preocuparnos por si el modelo podrá o no ajustar los datos correctamente. Incluso si hay un gran número de datos, siempre existe cierta flexibilidad para escoger las funciones.

Consideramos el siguiente problema de regresión: sea $\mathcal{X} \subset \mathbb{R}^d$ un conjunto no vacío y sea $f : \mathcal{X} \to \mathbb{R}$ una función. Supongamos que tenemos un conjunto de datos $(\mathbf{x}_i, y_i)_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$ con $n \in \mathbb{N}$, que reciben el nombre de datos de entrenamiento, tales que

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, ..., n, \tag{4.15}$$

donde ϵ_i es una variable aleatoria de media cero que representa el ruido en las observaciones. Como se ha mencionado anteriormente, el objetivo de los problemas de regresión es encontrar la función f basándonos en los datos de entrenamiento. Dicha función se puede expresar como $f(x) = E[y|x]$, donde (x, y) es una variable aleatoria cuya distribución está dada por el modelo descrito en (4.15). Si no hay ruido en las observaciones, es decir, $\epsilon_i = 0$, entonces el problema recibe el nombre de interpolación y es posible encontrar una función que se ajuste exactamente a los datos de entrenamiento. Siguiendo con la notación empleada en la sección 4.1, denotaremos por X la matriz cuyas columnas son los vectores de entrada, es decir, $X = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathcal{X}^n$ y por $\mathbf{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$ al conjunto de datos de salida.

Los métodos de regresión basados en procesos Gaussianos son métodos Bayesianos no paramétricos, por lo que producen una distribución a posteriori para la función de regresión f que tratamos de determinar en función de los datos de entrenamiento $(X, \mathbf{y})$, una distribución a priori en f, $p(f)$, y una función de verosimilitud que denotamos por $l_{X,Y}(f)$. Dicha distribución a priori viene definida como un proceso Gaussiano con función de media $m : \mathcal{X} \to \mathbb{R}$ y función de covarianza $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,

$$f \sim \mathcal{GP}(m, k). \tag{4.16}$$

Puesto que esta distribución se usa como distribución a priori del problema, la función de media y la función de covarianza deben escogerse de tal forma que reflejen el conocimiento que se tiene de dicha función de regresión a priori. De nuevo, para que el problema sea manejable, asumimos que el ruido asociado a las observaciones son variables aleatorias independientes e idénticamente distribuidas con una distribución normal de media cero y varianza σ_n^2. Desde el punto de vista bayesiano los errores deberían ser aleatorios, por lo que carecen de distribución. Sin embargo, podremos justificar la distribución normal de éstos viendo que, al final, el valor de σ no va a ser tan relevante.

$$\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \quad i = 1, ..., n. \tag{4.17}$$

De esta forma, la función de verosimilitud, que no es más que la densidad de probabilidad de las observaciones conocidos los parámetros, se define de la siguiente forma:

$$p(y|f) = l_{X,Y}(f) = \prod_{i=1}^{n} \mathcal{N}(y_i|f(\mathbf{x}_i), \sigma^2), \tag{4.18}$$

donde $\mathcal{N}(\cdot|\mu, \sigma^2)$ denota la distribución normal de media μ y varianza σ^2. Usando ahora la regla de Bayes, podemos obtener la distribución a posteriori $p(f|y)$ dada por

$$p(f|y) = \frac{p(y|f)p(f)}{p(y)} \propto p(y|f)p(f) = \prod_{i=1}^{n} \mathcal{N}(y_i|f(\mathbf{x}_i), \sigma^2)p(f). \tag{4.19}$$

De acuerdo con el teorema que se muestra a continuación, dicha distribución a posteriori es también un proceso Gaussiano con media $\bar{m}$ y función de covarianza $\bar{k}$ [23, Teorema 3.1].

Teorema 4.2.1. *Supongamos que se cumplen (4.15), (4.16) y (4.17) y sean* $X = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathcal{X}^n$ *e* $\mathbf{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$. *Entonces, tenemos que*

$$f|\mathbf{y} \sim \mathcal{GP}(\bar{m}, \bar{k}),$$

donde $\bar{m} : \mathcal{X} \to \mathbb{R}$ *y* $\bar{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *vienen dadas por*

$$\bar{m}(\mathbf{x}) = m(\mathbf{x}) + k_{xX}(k_{XX} + \sigma^2 I_n)^{-1}(\mathbf{y} - m(\mathbf{x})), \ \mathbf{x} \in \mathcal{X}, \tag{4.20}$$

$$\bar{k}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - k_{xX}(k_{XX} + \sigma^2 I_n)^{-1} k_{Xx'}, \ \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \tag{4.21}$$

con $k_{Xx} = k_{xX}^T = (k(\mathbf{x}_1, \mathbf{x}), ..., k(\mathbf{x}_n, \mathbf{x}))^T$. *Por tanto, como* $\mathcal{GP}(\bar{m}, \bar{k})$ *es un proceso Gaussiano,* $\bar{m}$ *recibe el nombre de media a posteriori y* $\bar{k}$ *es la función de covarianzas a posteriori.*

Demostración. Sea $m \in \mathbb{N}$ y sea $Z = (\mathbf{z}_1, ..., \mathbf{z}_m) \in \mathcal{X}^m$ un conjunto finito de puntos. Entonces, las observaciones $\mathbf{y} \in \mathbb{R}^n$ y los valores de prueba del proceso Gaussiano $f_z = (f(\mathbf{z}_1), ..., f(\mathbf{z}_m))^T \in \mathbb{R}^m$ tienen una distribución conjunta normal dad por

$$\begin{bmatrix} \mathbf{y} \\ f_z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_x \\ m_z \end{bmatrix}, \begin{bmatrix} k_{XX} + \sigma^2 I_n & k_{XZ} \\ k_{ZX} & k_{ZZ} \end{bmatrix} \right).$$

Con la notación de la proposición 2.1.3, tenemos que $a = \mathbf{y}|X, b = f_z, \mu_a = m_X, \mu_b = m_Z, A = k_{XX} + \sigma^2 I_n, B = k_{ZZ}$ y $C = k_{XZ}$. Utilizando el resultado de dicha proposición tenemos que la distribución de f_z condicionada por $\mathbf{y}$ es

$$f_z|\mathbf{y} \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma}),$$

donde

$$\bar{\mu} := m_z + k_{ZX}(k_{XX} + \sigma^2 I_n)^{-1}(\mathbf{y} - m_X) \in \mathbb{R}^n. \tag{4.22}$$

$$\bar{\Sigma} := k_{ZZ} - k_{ZX}(k_{XX} + \sigma^2 I_n)^{-1} k_{XZ} \in \mathbb{R}^{m \times m}. \tag{4.23}$$

Denotando por $\bar{m} = \bar{\mu}$ y $\bar{k} = \bar{\Sigma}$, se tiene que

$$f_z|\mathbf{y} \sim \mathcal{N}(\bar{m}, \bar{k}).$$

Observar que esto es cierto para cualquier conjunto de puntos $Z = (\mathbf{z}_1, ..., \mathbf{z}_m) \in \mathcal{X}^m$ de cualquier tamaño $m \in \mathbb{N}$. Por tanto, en virtud del teorema de extensión Kolmogorov y la definición de proceso Gaussiano, se tiene que el proceso $f \sim \mathcal{GP}(m, k)$ condicionado sobre los datos de entrenamiento (X, $\mathbf{y}$) es una muestra del proceso Gaussiano $\mathcal{GP}(\bar{m}, \bar{k})$. □

Es interesante resaltar que las hipótesis de ruido gaussiano y distribución a priori gaussiana son las que permiten obtener expresiones cerradas para la media y la función de covarianza a posteriori, sin necesidad de recurrir al Teorema de Bayes. Esto es importante por las siguientes razones:

1. Puesto que ambas distribuciones, a priori y a posteriori, vienen definidas en un espacio de dimensión infinita, una aplicación directa del Teorema de Bayes no permite obtener las expresiones cerradas para la media y la covarianza dadas por (4.20) y (4.21).

2. En el caso de observaciones libres de ruido, $\sigma^2 = 0$, el Teorema de Bayes no se puede emplear ya que la función de similitud es degenerada.

Del teorema anterior se deduce que dado un dato de prueba $\mathbf{z}$, la mejor estimación de $f(\mathbf{z})$ se obtiene evaluando la función de media a posteriori, ya que $E[f(\mathbf{x})|X, y] = \bar{m}(\mathbf{x})$. Además, la función de covarianza a posteriori permite cuantificar la incertidumbre de los valores obtenidos. Esta es una de las ventajas de los modelos basados en procesos Gaussianos frente a otro tipo de métodos de regresión. Por definición, la covarianza a posteriori se define a partir de la función de media a posteriori dada por (4.20) y el valor $f(\mathbf{x})$, con $f \sim \mathcal{GP}(\bar{m}, \bar{k})$,

$$\bar{k}(\mathbf{x}, \mathbf{x}) = E_{f\sim\mathcal{GP}(\bar{m},\bar{k})}[(f(\mathbf{x}) - \bar{m}(\mathbf{x}))^2]. \tag{4.24}$$

Es decir, $\bar{k}(\mathbf{x}, \mathbf{x})$ se puede interpretar como *el error medio* en un punto x desde el punto de vista Bayesiano.

Por otro lado, vemos como tanto la función de media como la función de covarianza a posteriori dependen del núcleo escogido y de las funciones de media y covarianza a priori. Muchas veces, estas funciones de covarianza depende de hiperparámetros que se deben elegir y hay técnicas especificas para esa elección que se resumen en la sección 4.3.

Si consideramos ahora el caso libre de ruido, es decir, un problema de interpolación, en el que $y_i = f(\mathbf{x}_i), i = 1, ..., n$, nos encontramos con los siguientes inconvenientes. En primer lugar, la función de similitud es degenerada, y por consiguiente, no está bien definida, ya que la distribución de y_i condicionada por los valores de $f(\mathbf{x}_i)$ viene dada por la distribución de Dirac en $f(\mathbf{x}_i)$, que no tiene función de densidad asociada. Por consiguiente, no es valida la expresión que hemos empleado para obtener la distribución a posteriori en (4.19). No obstante, como hemos visto en el teorema 4.2.1, es posible derivar las expresiones de la función de media y de la función de covarianza a posteriori sin emplear el Teorema de Bayes, obteniendo de nuevo las expresiones dadas por (4.20) y (4.21). Este es el resultado que mostrado en el siguiente Teorema [23, Teorema 3.3].

Teorema 4.2.2. *Supongamos que se cumplen las condiciones dadas por (4.16) y sean $X = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathcal{X}^n$ y $f_X = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))^T \in \mathbb{R}^n$. Supongamos además que la matriz de covarianzas $k_{XX} = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$ es invertible. Entonces la distribución de f condicionada por (X, f_X) es un proceso Gaussiano dado por*

$$f|f_X \sim \mathcal{GP}(\bar{m}, \bar{k}),$$

donde $\bar{m} : \mathcal{X} \to \mathbb{R}$ y $\bar{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ vienen dadas por

$$\bar{m}(\mathbf{x}) = m(\mathbf{x}) + k_{xX}k_{XX}^{-1}(f(\mathbf{x}) - m(\mathbf{x})), \ \mathbf{x} \in \mathcal{X}. \tag{4.25}$$

$$\bar{k}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - k_{xX}k_{XX}^{-1}k_{Xx'}, \ \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{4.26}$$

Demostración. La demostración es inmediata a partir del teorema anterior, sin más que sustituir $k_{XX} + \sigma^2 I_n$ por k_{XX}. □

Para que (4.25) y (4.26) estén bien definidas, es necesario que k_{XX} sea invertible. Esto no será cierto si, por ejemplo, algunos elementos de $X = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathcal{X}^n$ son idénticos, o si el núcleo empleado es un núcleo polinómico de grado m con $n > m$. Para evitar posibles errores, es habitual introducir un pequeño error en las observaciones a modo de constante de regularización. Vemos que en el caso de que $m_z = 0$, las funciones de media y covarianza a posteriori obtenidas por ambos caminos, espacio de pesos, (4.13) y (4.14), y espacio de funciones, (4.22) y (4.23), coinciden. Por consiguiente, vemos que son dos

formas equivalentes de desarrollar el modelo de los procesos Gaussianos.

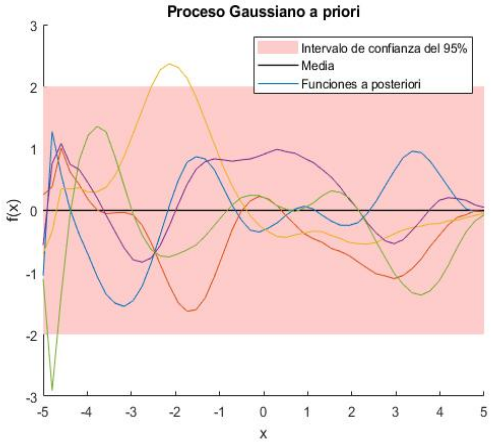
En las siguientes imágenes se observa un ejemplo de regresión basado en procesos Gaussianos tanto para el caso en el que las observaciones están libres de ruido (Figura 4.2.1) como para el caso en el que hay ruido presente (Figura 4.2.2). Para obtener estas imágenes hemos generado unos datos de prueba de forma equidistribuida en el intervalo $[-5, 5]$. Como función núcleo, hemos seleccionado la función de covarianza exponencial al cuadrado (SE), que es la que se usa con mas frecuencia debido a su flexibilidad y a que es infinitas veces diferenciable. Por último, para generar muestras con una distribución normal de media μ y función de covarianza k se procede de la siguiente forma:

1. Calculamos la descomposición de Cholesky de la función de covarianza k, $k = LL^T$, donde L es una matriz triangular inferior.

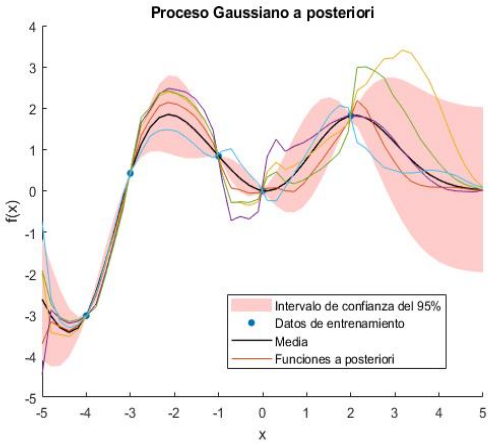
2. Generamos muestras normales de media cero y función de covarianza la identidad, $u \sim \mathcal{N}(0, I)$. Para ello empleamos una función que genere muestras Gaussianas.

3. Calculamos $f = \mu + Lu$, que tiene la distribución deseada, con media μ y matriz de covarianza $k = LL^T = LE[uu^T]L^T$.

Es posible que sea necesario añadir a la función de covarianza un pequeño múltiplo de la matriz identidad por motivos numéricos. Esto se debe a que los autovalores de k pueden ser muy próximos a cero, de manera que la descomposición de Cholesky es propensa a fallar. Una vez generadas las distribuciones Gaussianas a priori, usamos las ecuaciones (4.20) y (4.21) para obtener las funciones de media y de covarianza del proceso a posteriori. De esta forma, basta con repetir el procedimiento anterior para generar muestras normales con con esos parámetros.

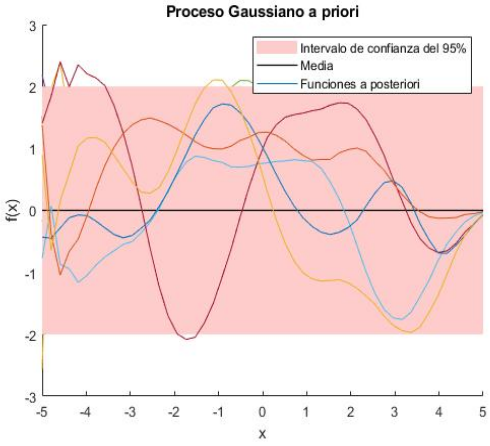


(a) Muestras de un proceso Gaussiano a priori.

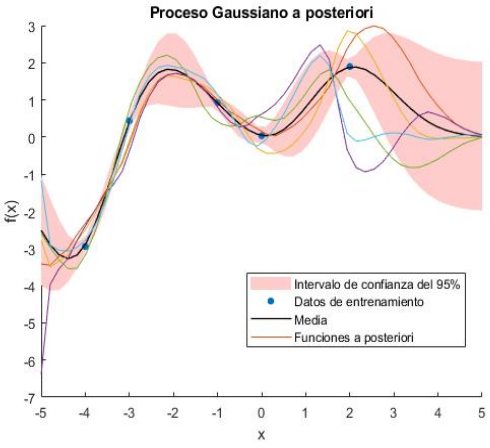


(b) Muestras de un proceso Gaussiano a posteriori.

Figura 4.2.1: A la izquierda, se muestran funciones correspondientes a un proceso Gaussiano a priori con media cero. A la derecha, funciones correspondientes a un proceso Gaussiano a posteriori, es decir, la distribución a priori condicionada por las observaciones libres de ruido. La zona sombreada se corresponde con la región de confianza del 95 %.



(a) Muestras de un proceso Gaussiano a priori.



(b) Muestras de un proceso Gaussiano a posteriori.

Figura 4.2.2: A la izquierda, se muestran funciones correspondientes a un proceso Gaussiano a priori con media cero. A la derecha, funciones correspondientes a un proceso Gaussiano a posteriori, es decir, la distribución a priori condicionada por las observaciones con ruido. La zona sombreada representa la región de confianza del 95 %.

4.3. Selección de los hiperparámetros.

Hemos visto que existen distintas familias de funciones de covarianza que pueden emplearse en los problemas de regresión basados en procesos Gaussianos. Cada una de estas familias consta de una serie de hiperparámetros cuyos valores deben ser determinados. Por consiguiente, escoger una determinada función de covarianza para un problema concreto supone fijar los hiperparámetros dentro de una familia, así como comprar funciones de covarianza de diferentes familias. Por ello es necesario introducir modelos de selección que nos permitan solventar ambos problemas. El problema de selección de modelos se conoce también como entrenamiento de un proceso Gaussiano. Es importante notar que el valor de los hiperparámetros determina el comportamiento de la función de covarianza. El objetivo es, por tanto, realizar inferencias sobre la forma y valor de los parámetros de una determinada función de covarianza a partir de unos datos de entrenamiento.

Aunque existen distintas técnicas de selección de modelos, todas ellas cumplen, por regla general, los siguientes principios:

1. Calcular la probabilidad del modelo dados los datos de entrenamiento.
2. Estimar el error de generalización.
3. Acotar el error de generalización.

En particular, el modelo de regresión basado en procesos Gaussianos analiza el conjunto de datos de entrenamiento y mediante técnicas de modelado bayesiano, infiere un valor para éstos de manera que proporcionen una buena capacidad de generalización. Para ello, se emplea el método de la maximización de la verosimilitud marginal (*marginal likelihood maximization o Maximum a Posteriori*).

Denotaremos por θ al conjunto de todos los hiperparámetros de una determinada función núcleo. Así pues, en el caso de la función de cuadrado exponencial, dada por $k(r) = \exp\left(-\frac{r^2}{2\sigma^2}\right)$, tenemos que $\theta = \sigma$. A partir del Teorema de Bayes, podemos representar la probabilidad a posteriori de los hiperparámetros como sigue:

$$p(\theta|X, \mathbf{y}) = \frac{p(\mathbf{y}|X, \theta)p(\theta)}{p(\mathbf{y}|X)}, \tag{4.27}$$

donde $p(\mathbf{y}|X, \theta)$ es la verosimilitud marginal que deseamos maximizar y $p(\theta)$ es la probabilidad a priori de los hiperparámetros. La verosimilitud puede calcularse de la siguiente forma [34]:

$$p(\mathbf{y}|X, \theta) = \int p(\mathbf{y}|f, X, \theta)p(f|X), \theta)df. \tag{4.28}$$

Generalmente, este tipo de integrales no pueden resolverse analíticamente, por lo que no es sencillo derivar buenas aproximaciones. No obstante, en el marco de los procesos Gaussianos con ruido Gaussiano, podemos asumir que todo sigue una distribución normal, de modo que

$$\mathbf{y} \sim \mathcal{N}(0, K_y),$$

donde $K_y = K_f + \sigma_n^2 I$ es la función de covarianza para las observaciones con ruido. De esta forma, podemos calcular la integral anterior de manera analítica. Calculando el logaritmo, obtenemos la siguiente expresión:

$$\log p(\mathbf{y}|X, \theta) = -\frac{1}{2}\mathbf{y}^T K_y^{-1}\mathbf{y} - \frac{1}{2}\log|K_y| - \frac{n}{2}\log 2\pi. \tag{4.29}$$

El primer término de esta expresión representa cómo de bien se ajusta el modelo a los datos, el segundo expresa la complejidad del modelo empleado y el tercero es una constante de normalización. Por consiguiente, encontrar los parámetros que maximizan esta función es equivalente a encontrar los parámetros que proporcionan un mejor acuerdo con los datos observados. Así pues, el problema de ajustar los hiperparámetros consiste en maximizar la función $L(\theta) = \log p(\mathbf{y}|X, \theta)$, o minimizar $-L(\theta)$. Esto se resuelve generalmente mediante un algoritmo de descenso por gradiente.

4.4. Consideraciones sobre la regresión basada en GP.

Hasta ahora hemos supuesto que el ruido sigue una distribución Gaussiana. No obstante, este no es siempre el caso, sino que podemos asumir cualquier distribución de probabilidad. La ventaja de suponer ruido Gaussiano es que las integrales que aparecen en las distribuciones de probabilidad pueden resolverse analíticamente, sin necesidad de recurrir a técnicas más complejas para su aproximación, como por ejemplo, *Grid Integration*, *Monte Carlo Integration* o *Central Composite Design*. Por otro lado, hemos asumido que la media del proceso es cero. De nuevo, esto no tiene por qué ser así. Además, es posible introducir distribuciones de probabilidad a priori sobre los hiperparámetros. Estas modificaciones introducen complicaciones en el modelo que requieren de técnicas más complejas. Vemos que los procesos de regresión basados en procesos Gaussianos se basan en una teoría que tiene una gran flexibilidad de modelado mediante la función de covarianza, las distribuciones a priori, etc. Cuántas más modificaciones, más complejidad, por lo que más difícil será de entrenar y configurar adecuadamente el modelo.

Por otro lado, de acuerdo con las expresiones (4.20) y (4.21), la regresión basada en procesos Gaussianos tiene una complejidad computacional de $O(n^3)$. Por consiguiente, para conjuntos de datos con $n > 10000$ el coste computacional derivado de invertir matrices $n \times n$, así como la capacidad de memoria necesaria para almacenar la información, se vuelven prohibitivos. Una forma de resolver este problema es introduciendo una serie de variables, que reciben el nombre de *variables inducidas*. Se trata de valores del proceso Gaussiano correspondientes a un conjunto de datos de entrada X_u, llamados *puntos inducidos*. Existen diversas formas de escoger las variables inducidas. Una de ellas consiste en escoger un subconjunto del conjunto de entrenamiento (*SD, subset of data*).

Siguiendo con la notación introducida en secciones anteriores, con $Z = (\mathbf{z}_1, ..., \mathbf{z}_m) \in \mathcal{X}^m$ el conjunto de datos de prueba, siendo los valores de las observaciones asociadas $\mathbf{f}_* = (f(\mathbf{z}_1), ..., f(\mathbf{z}_m))^T \in \mathbb{R}^m$, podemos recuperar la distribución de probabilidad conjunta $p(\mathbf{f}_*, \mathbf{f})$ a partir de $p(\mathbf{f}_*, \mathbf{f}, u)$ de la siguiente forma:

$$p(\mathbf{f}_*, \mathbf{f}) = \int p(\mathbf{f}_*, \mathbf{f}, u)du = \int p(\mathbf{f}_*, \mathbf{f}|u)p(u)du \text{ donde } p(u) = \mathcal{N}(0, K_{u,u}). \tag{4.30}$$

Esta expresión es exacta. A continuación, se introducen las aproximaciones que dan lugar a la gran mayoría de *aproximaciones dispersas* (*sparse approximation* en inglés). Dicha aproximación consiste en asumir que $\mathbf{f}_*$ y $\mathbf{f}$ son condicionalmente independientes dado u, de forma que

$$p(\mathbf{f}_*, \mathbf{f}) \simeq q(\mathbf{f}_*, \mathbf{f}) = \int q(\mathbf{f}_*|u)q(\mathbf{f}|u)p(u)du. \tag{4.31}$$

Así pues, el nombre de variable inducida se debe a que $\mathbf{f}_*$ y $\mathbf{f}$ tan solo pueden comunicarse a través de u, por lo que u induce la dependencia entre los datos de entrenamiento y los datos de prueba.

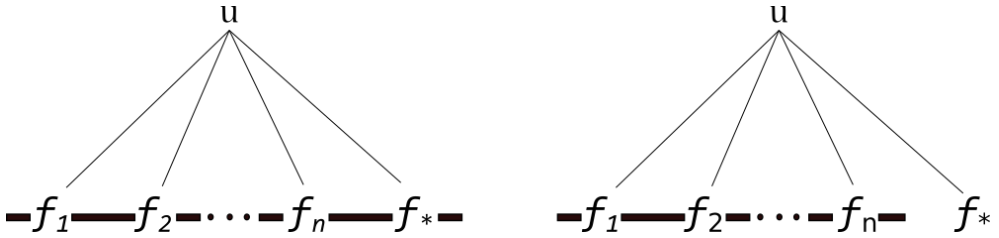


Figura 4.4.1: Representación gráfica de la relación entre las variables inducidas u, y las observaciones de entrenamiento y de prueba, $\mathbf{f}$ y $\mathbf{f}_*$.

Los diferentes métodos de aproximación consistirán en diferentes suposiciones añadidas sobre $q(\mathbf{f}_*|u)$ y $q(\mathbf{f}|u)$. Con estas condiciones, obtenemos expresiones muy similares a las ecuaciones (4.25) y (4.26):

$$p(\mathbf{f}|u) = \mathcal{N}(k_{f,u}k_{u,u}^{-1}u, k_{f,f} - Q_{f,f}) \tag{4.32}$$

$$p(\mathbf{f}_*|u) = \mathcal{N}(k_{f_*,u}k_{u,u}^{-1}u, k_{f_*,f_*} - Q_{f_*,f_*}), \tag{4.33}$$

donde $Q_{a,b} := k_{a,u}k_{u,u}^{-1}k_{u,b}$. La interpretación del término $k - Q$ es la siguiente: la covarianza a priori k menos una matriz Q que cuantifica cuánta información proporcionan las variables u sobre los valores de $\mathbf{f}_*$ y $\mathbf{f}$.

Presentamos en esta sección la aproximación basada en un subconjunto de los datos, ya que será la que empleemos en la simulación 7.3. Como su nombre indica, consiste en seleccionar

como variables inducidas un subconjunto de tamaño m de los datos de entrenamiento. De esta forma, la complejidad computacional pasa a ser $O(m^3)$, con $m < n$. Se trata de la aproximación dispersa más básica y sencilla, pero es la que peores resultados arroja, ya que estamos despreciando una gran cantidad de datos de entrenamiento. Esta aproximación da lugar a regiones de confianza de gran tamaño, pero tiene la ventaja de que es independiente de n.

4.5. Equivalencias entre regresión Ridge y regresión basada en GP.

De las expresiones (3.27) y (4.10), según las cuales

$$\begin{aligned} \hat{f} &= {k_{Xx}}^T (G + \lambda I_n)^{-1} \mathbf{y}, \\ \bar{m} &= k_{ZX} (k_{XX} + \sigma_n^2 I)^{-1} \mathbf{y}, \end{aligned}$$

parece evidente que existe una relación entre la regresión Ridge y la regresión basada en procesos Gaussianos. Esta equivalencia entre ambos métodos se recoge en la siguiente proposición [23, Proposición 3.6]:

Proposición 4.5.1. *Sea $\mathcal{X} \subset \mathbb{R}^d$ un espacio no vacío, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ una función núcleo y $\mathcal{D} = (\mathbf{x}_i, y_i) \subset \mathcal{X} \times \mathbb{R}$ un conjunto de datos de entrenamiento. Entonces, $\bar{m} = \hat{f}$ si $\sigma_n^2 = \lambda$ donde:*

- *$\bar{m}$ es la media a posteriori (4.10) de la regresión Gaussiana basada en el conjunto de datos de entrenamiento $\mathcal{D}$, la distribución a priori $f \sim \mathcal{GP}(m, k)$ y el modelo (4.15) con $\epsilon_i \sim \mathcal{N}(0, \sigma_n^2)$.*
- *$\hat{f}$ es la solución (3.27) de la regresión Ridge basada en los datos de entrenamiento $\mathcal{D}$ dada por (3.25) con parámetro de regularización $\lambda > 0$.*

De acuerdo con esta proposición, podemos decir que la hipótesis de ruido en el modelo (4.15) juega el papel de regularización el la regresión Ridge, ya que ambos estimadores coinciden cuando suponemos que el parámetro de regularización λ coincide con la varianza σ_n^2 de las variables ϵ_i. Por otro lado, cuánto mayor sea el parámetro de regularización, mayor será la flexibilidad de la función, es decir, la función óptima es más suave. Lo mismo ocurre con la varianza, ya que cuánto mayor sea el error en los datos más suave será la función que los representa.

Ambos métodos emplean el truco del núcleo para predecir la función que mejor se ajusta a unos datos. Por un lado, la regresión Ridge es capaz de predecir una función lineal en el espacio inducido por dicho núcleo, que se corresponde con una función no lineal en el espacio original. La función lineal en el espacio de características se escoge en base al error cuadrático medio con regularización Ridge. Por el contrario, los métodos basados en procesos Gaussianos utilizan la función núcleo para definir una covarianza asociada a una distribución a priori sobre los datos de entrenamiento. De esta forma, la regresión Ridge tan solo proporciona información sobre los valores de $\mathbf{y}$, mientras que los procesos Gaussianos son capaces de cuantificar la incertidumbre sobre $\mathbf{y}$ y generar muestras a posteriori. Por otro lado, en el caso de los procesos Gaussianos es posible escoger los hiperparámetros de las funciones núcleo de forma óptima mediante la optimización de la verosimilitud marginal por el método de descenso de gradiente.

CAPÍTULO 5

CLASIFICACIÓN BASADA EN PROCESOS GAUSSIANOS.

En este capítulo se introducen aspectos relacionados con la clasificación basada en procesos Gaussianos. Puesto que no es el objetivo principal del trabajo, se tratarán tan solo los conceptos más importantes. Se puede consultar más información en [34, Capítulo 3]. En la sección 5.1 se define el problema de clasificación. A continuación, en la sección 5.2 se introduce la Teoría de la Decisión para problemas de clasificación, a partir de la cual asignamos un dato a una determinada clase. Por último, en la sección 5.3 se muestran las ideas básicas en relación con la clasificación binaria, fácilmente extrapolables a casos más complejos.

5.1. El problema de clasificación.

En los capítulos anteriores hemos tratado tan solo problemas de regresión basados en procesos Gaussianos. Otro tipo de aplicaciones en las que este tipo de métodos resultan útiles son los problemas de clasificación, donde el objetivo es asignar cada dato de entrada $\mathbf{x}$ a una clase $\mathcal{C}_i$, $i = 1, ..., c$. En este capítulo nos centraremos en la clasificación probabilística, donde las predicciones de las muestras test tienen forma de clases probabilísticas. El punto de partida para estudiar problemas de clasificación es la probabilidad conjunta, $p(y, \mathbf{x})$, donde y denota la etiqueta de la clase. De acuerdo con la regla de multiplicación, es posible descomponer dicha probabilidad o bien como $p(y)p(\mathbf{x}|y)$ o como $p(\mathbf{x})p(y|\mathbf{x})$. Surgen por tanto dos aproximaciones diferentes al problema de clasificación. La primera recibe el nombre de aproximación generativa, y modela la distribución de probabilidad condicionada a cada clase, $p(\mathbf{x}|y)$, para $y = \mathcal{C}_1, ..., \mathcal{C}_c$, así como las probabilidades a priori de cada clase, para posteriormente calcular la probabilidad a posteriori:

$$p(y|\mathbf{x}) = \frac{p(y)p(\mathbf{x}|y)}{\sum_{i=1}^{c} p(\mathcal{C}_c)p(\mathbf{x}|\mathcal{C}_c)}. \tag{5.1}$$

En este caso, es frecuente modelar las densidades de probabilidad condicionadas a cada clase mediante distribuciones Gaussianas, $p(\mathbf{x}|\mathcal{C}_c) = \mathcal{N}(\mu_c, \Sigma_c)$. No obstante, esta aproximación introduce una suposición muy fuerte en cuanto a la forma de las densidades condicionadas a las clases, por lo que es posible que no funcione de manera óptima.

La otra vía, que se conoce como aproximación discriminativa, se centra en modelar directamente $p(y|\mathbf{x})$. Puesto que las probabilidades toman valores en el intervalo $[0, 1]$ y un

proceso estocástico $f \sim \mathcal{GP}(m, k)$ define una función estocástica que toma valores en el conjunto de los números reales, empleamos una función $\lambda : (-\infty, \infty) \to [0, 1]$, que sea suave y estrictamente creciente. Esta función recibe el nombre de función sigmoide y presenta la siguiente forma:

$$\lambda(z) = \frac{1}{1 + e^{-z}}. \tag{5.2}$$

De esta forma, tenemos que

$$p(\mathcal{C}_1|\mathbf{x}) = \lambda(\mathbf{x}^T\mathbf{w}) = \lambda(f(\mathbf{x})). \tag{5.3}$$

En este capítulo nos centraremos en los problemas de clasificación discriminativos.

5.2. Teoría de la Decisión para problemas de clasificación.

Los métodos presentados en la sección anterior nos proporcionan formas de obtener la probabilidad $p(y_*|\mathbf{x}_*)$. No obstante, es frecuente que tengamos que tomar una decisión en base a esta probabilidad para asignar dicho dato a una determinada clase. Al igual que sucedía con los problemas de regresión, empleamos la función de pérdida, $\mathcal{L}(c, c')$, que representa el error cometido al escoger la clase $\mathcal{C}_c$ cuando la correcta es $\mathcal{C}_{c'}$. Definimos entonces el riesgo esperado de escoger c' como $\mathcal{R}_\mathcal{L} = \sum_c \mathcal{L}(c, c')p(\mathcal{C}_c|\mathbf{x})$, de manera que la decisión óptima c^* será la que minimice el riesgo esperado. Es frecuente que el error de mala clasificación sea distinto, es decir, $\mathcal{L}(c, c') \neq \mathcal{L}(c', c)$, algo que hay que tener en cuenta a la hora de definir la función de error o función de pérdida.

El método de decisión planteado consta de dos pasos. Primero debemos calcular la distribución a posteriori y posteriormente combinar esta información con la función de pérdida para tomar una decisión. No obstante, es frecuente combinar ambos pasos de manera que si lo que nos interesa es tomar una decisión, basta con encontrar la función que minimiza el riesgo esperado, dado por:

$$\mathcal{R}_\mathcal{L}(c) = \int \mathcal{L}(y, c(\mathbf{x}))p(y, \mathbf{x})dyd\mathbf{x}, \tag{5.4}$$

donde $p(y, \mathbf{x})$ es la distribución de probabilidad conjunta y $c(\mathbf{x})$ es la función de clasificación que asigna cada vector de entrada $\mathbf{x}$ a una clase $\mathcal{C}$. Puesto que generalmente no concomeos la distribución de probabilidad conjunta, es frecuente minimizar una función que incluye el resigo empírico, $\sum_{i=1}^n \mathcal{L}(y_i, c(\mathbf{x}_i))$, así como un término de regularización.

5.3. Modelos lineales para clasificación.

En esta sección introducimos algunas ideas básicas sobre clasificación binaria, que constituye la base para los métodos de clasificación basados en procesos Gaussianos. En este caso, el objetivo es asignar a un vector $\mathbf{x}$ una clase entre dos posibles, lo cual se representa mediante una etiqueta $y(\mathbf{x})$ que puede tomar dos valores: 1 ó (-1). Sea $D = \{(\mathbf{x}_i, y_i), i = 1, ..., n\}$ un conjunto de datos de entrenamiento, con $y_i \in \{-1, 1\}$ y sea $\mathbf{z}$ un dato de prueba. En este caso, tenemos que $p(y = 1|\mathbf{z}, \mathbf{w}) = \sigma(f(\mathbf{z}))$. La función $\sigma : (-\infty, \infty) \to [0, 1]$ puede ser cualquier función sigmoide. Si tomamos $\sigma(z) = \lambda(z)$, siendo $\lambda(z)$ la función definida en (5.2), entonces el modelo recibe el nombre de regresión logística. Puesto que la suma de las probabilidades de ambas clases debe ser uno, tenemos que $p(y = -1|\mathbf{z}, \mathbf{w}) = 1 - p(y = 1|\mathbf{z}, \mathbf{w})$. Por consiguiente, para cualquier dato $(\mathbf{x}_i, y_i)$ la

función de verosimilitud viene dada por $\sigma(\mathbf{x}_i^T\mathbf{w})$ si $y_i = 1$ y $1 - \sigma(\mathbf{x}_i^T\mathbf{w})$ si $y_i = -1$.

Al igual que en los problemas de regresión, podemos obtener una expresión para la probabilidad a posteriori, que viene dada por:

$$\log p(\mathbf{w}|X, y) = -\frac{1}{2}\mathbf{w}^T\sigma_p^{-1}\mathbf{w} + \sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i, \mathbf{w}), \tag{5.5}$$

donde $D = \{(\mathbf{x}_i, y_i), i = 1, ..., n\}$ es un conjunto de datos de entrenamiento y $\mathbf{w} \sim \mathcal{N}(0, \Sigma_p)$. El problema que plantean los modelos de clasificación es que la probabilidad a posterior no tiene una expresión analítica sencilla, por lo que es frecuente usar métodos de aproximación.

Empleando la metodología de selección bayesiana usada en los capítulos anteriores, podemos dividir el proceso de inferencia en dos pasos: primero estimamos la distribución a posterior del proceso Gaussiano para un dato de prueba,

$$p(f_z|X, y, \mathbf{z}) = \int p(f_z|X, \mathbf{z}, f)p(f|X, y)df, \tag{5.6}$$

donde $p(f|X, y) = p(y|f)p(f|X)/p(y|X)$, y posteriormente calculamos la probabilidad en la que estamos interesados,

$$\pi(\mathbf{z}) = p(y = 1|X, y, \mathbf{z})) = \int \sigma(f_z)p(f_z|X, y, \mathbf{z})df_z. \tag{5.7}$$

A diferencia de lo que ocurría con los modelos de regresión en los que, asumiendo distribuciones Gaussianas, las integrales podían resolverse analíticamente, en este caso, el uso de distribuciones de probabilidad no Gaussianas impide resolver las integrales de forma exacta. En particular, la distribución a posteriori $p(f|X, y)$ no es normal. Por tanto, es necesario recurrir a métodos de aproximación, como aproximación numérica de integrales o modelos basados en simulaciones de Mote Carlo. Para el caso de clasificación binaria se suele emplear la aproximación de Laplace. El objetivo de este método es aproximar una distribución no Gaussiana $p(f|X, y)$ por una distribución Gaussiana $q(f|X, y)$. Para ello, realizamos una aproximación de Taylor de orden 2 del logaritmo de dicha distribución de probabilidad, $\log p(f|X, y)$, obteniendo:

$$q(f|X, y) \sim \mathcal{N}(f; \hat{f}, A^{-1}) \propto \exp\left(-\frac{1}{2}(f - \hat{f})^T A(f - \hat{f}))\right), \tag{5.8}$$

donde

$$\hat{f} = \arg\,\text{máx}_f q(f|X, y). \tag{5.9}$$

$$A = -D^2 \log p(f|X, y)|_{f=\hat{f}}. \tag{5.10}$$

De acuerdo con el Teorema de Bayes tenemos que $p(f|X, y) = p(y|f)p(f|X)/p(y|X)$. Puesto que el denominador es una constante y la función logaritmo es estrictamente creciente, maximizar la función $\log p(f|X, y)$ es equivalente a maximizar la siguiente función:

$$\Phi(f) = \log p(y|f) + \log(f|X). \tag{5.11}$$

Usando que la distribución a priori es Gaussiana, $f|X \sim \mathcal{N}(0, K)$, tenemos que:

$$\log p(f|X) = -\frac{1}{2}f^T K^{-1} f - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi. \tag{5.12}$$

De esta forma, la ecuación (5.11) se escribe de la siguiente forma:

$$\Phi(f) = \log p(y|f) - \frac{1}{2} f^T K^{-1} f - \frac{1}{2} \log |K| - \frac{n}{2} \log 2\pi. \tag{5.13}$$

Diferenciando la ecuación (5.13) con respecto de f, obtenemos:

$$D\Phi(f) = D \log p(y|f) - K^{-1} f, \tag{5.14}$$
$$D^2\Phi(f) = D^2 \log p(y|f) - K^{-1} = -W - K^{-1}, \tag{5.15}$$

donde $W := -D^2 \log p(y|f)$ es diagonal. El máximo de la función $\Phi(f)$ corresponde con

$$D\Phi(f) = 0 \rightarrow \hat{f} = K(D \log p(y|\hat{f})). \tag{5.16}$$

Esta ecuación no puede resolverse de forma directa, por lo que es necesario emplear métodos de aproximación. Un camino posible es usar la aproximación de Newton, descrita en [34, Sección 3.4]. Una vez hallado el máximo, es posible especificar la aproximación de Laplace de la distribución a posteriori, que viene dada por:

$$q(f|X, y) = \mathcal{N}(\hat{f}, (K^{-1} + W)^{-1})). \tag{5.17}$$

Conocida la aproximación de Laplace de la distribución a posteriori para los datos de entrenamiento, es posible calcular la distribución a posteriori $q(f_z|X, y, \mathbf{z})$ para un determinado dato de prueba $\mathbf{z}$. Teniendo en cuenta que para un proceso Gaussiano la media puede escribirse como $E[f_z|f, X, \mathbf{z}] = \bar{m}_z = k(\mathbf{z})^T (K+\sigma^2 I)^{-1} y$, de acuerdo con lo obtenido en (4.20) para $m_z = 0$, y usando la ecuación (5.16), obtenemos las siguientes expresiones para la media a posteriori:

$$E_q[f_z|X, y, \mathbf{z}] = k(\mathbf{z})^T K^{-1} \hat{f} = k(\mathbf{z})^T D \log p(y|\hat{f}), \tag{5.18}$$
$$E_p[f_z|X, y, \mathbf{z}] = \int E_[f_z|f, X, \mathbf{z}] p(f|X, y) df$$
$$= \int k(\mathbf{z})^T K^{-1} f p(f|X, y) df = k(\mathbf{z})^T K^{-1} E[f|X, y], \tag{5.19}$$

donde $E_q[f_z|X, y, \mathbf{z}]$ denota la media en aproximación de Laplace y $E_p[f_z|X, y, \mathbf{z}]$ denota la media exacta. Comparando la esperanza con su valor exacto, observamos que la aproximación supone sustituir $E[f|X, y]$ por $\hat{f}$, donde $E[f|X, y]$ denota la media a posteriori de f dados X e y.

A la vista de (5.18), vemos que valores positivos de los datos de entrenamiento dan lugar a valores positivos de las funciones núcleo y viceversa, ya que $D_i \log p(y_i|f_i) \geq 0$ en el primer caso y negativo en el segundo caso. Además, los datos de entrenamiento para los cuales $D_i \log p(y_i|f_i) \simeq 0$, y que por tanto están bien explicados por $\hat{f}$, no contribuyen de manera significativa en la predicción de datos de prueba.

Es posible determinar la expresión de la varianza de $f_z|X, y$, $V_q[f_z|X, y]$ empleando la aproximación Gaussiana [34, Ecuaciones 3.23 y 3.24]:

$$V_q[f_z|X, y] = k(\mathbf{z}, \mathbf{z}) - k(\mathbf{z})^T (K + W^{-1})^{-1} k(\mathbf{z}). \tag{5.20}$$

Conocidas la media y la varianza de f_z, podemos realizar predicciones a través de la siguiente integral:

$$\bar{\pi}_z \simeq E_q[\pi_z|X, y, \mathbf{z}] = \int \sigma(f_z) q(f_z|X, y, \mathbf{z}) df_z, \tag{5.21}$$

donde $q(f_z|X, y, \mathbf{z})$ es Gaussiana con media y varianza dadas por (5.18) y (5.20), respectivamente.

Siempre que estemos interesados en calcular no solo la clasificación más probable, sino también la probabilidad de acertar en la clasificación, será necesario calcular la integral (5.21). En caso contrario, bastará con calcular el valor de $E_q[f_z|X, y, \mathbf{z}]$, dado en (5.18).

CAPÍTULO 6

APRENDIZAJE PAC.

En teoría de aprendizaje estadístico, *aprendizaje probablemente aproximadamente correcto o aprendizaje PAC* (en inglés, *probably approximately correct learning*) es un marco para análisis matemático de aprendizaje de máquina. Éste fue propuesto en 1984 por Leslie Valiant [49]. En este marco, la técnica de aprendizaje recibe muestras y debe seleccionar una función de generalización, llamada hipótesis, de una clase de posibles funciones. El objetivo es que, con una alta probabilidad (la parte del "probablemente"), la función seleccionada tendrá un error de generalización bajo (la parte del "aproximadamente correcto"). La técnica de aprendizaje tiene que ser capaz de aprender el concepto con una proporción de aproximación arbitraria, probabilidad de éxito, o distribución de las muestras. En la sección 6.1 se describe con más detalle este modelo de aprendizaje PAC. Posteriormente, en la sección 6.2 se introduce el concepto de riesgo como herramienta que nos permite cuantificar la exactitud del modelo. En la sección 6.3 se habla sobre la consistencia de un modelo de aprendizaje, a través de la cual es posible determinar si realmente el riesgo empírico bajo implica riesgo real bajo. En la siguiente sección, 6.4, se definen los conceptos más importantes del aprendizaje PAC. Posteriormente, se describen la descomposición del riesgo, 6.5, y la dimensión VC, 6.6, una medida de la capacidad de clasificación estadística de los algoritmos. Por último, se introduce el aprendizaje PAC-Bayesiano en la sección 6.7, que combina el moldeo PAC con el modelo Bayesiano para compensar la generalidad del primero con la eficiencia del segundo. Este marco del aprendizaje PAC se planteó en un contexto no bayesiano. En las secciones 6.1 hasta 6.6 abandonamos temporalmente el marco bayesiano para describir los elementos principales de la teoría. Posteriormente, en la sección 6.7 trataremos de adaptar estas ideas del aprendizaje PAC al marco bayesiano.

6.1. Modelo de aprendizaje. Marco de trabajo de PAC.

En los capítulos anteriores hemos visto como podemos predecir la variable respuesta asociada a una cierta observación $\mathbf{x}_*$, tanto en un contexto de regresión como de clasificación. Aquí vamos a tratar de dar una descripción unificada de estos problemas y de otras opciones dentro del aprendizaje supervisado.

No parece razonable esperar que, a partir de una muestra finita de datos de entrenamiento, seamos capaces de determinar de manera exacta la distribución que los origina. Con el fin de modelar el proceso de aprendizaje, supongamos un modelo que consta de los

siguientes elementos:

1. Datos de entrada disponibles para el observador:
   - Dominio de datos de entrada: tenemos un conjunto arbitrario que denotamos por $\mathcal{X}$. Este dominio contiene el conjunto de objetos que queremos clasificar. Generalmente los puntos del dominio serán en realidad vectores que hacen referencia a las diferentes características medidas.
   - Etiquetas: son los posibles valores a los que podemos asignar las observaciones. Denotamos por $\mathcal{Y}$ al conjunto de las posibles etiquetas. En problemas de clasificación dicho conjunto será finito, como es por ejemplo $\{-1, 1\}$, mientras que en problemas de regresión puede ser el conjunto de los números reales.
   - Datos de entrenamiento: secuencia finita $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ de pares de valores en $\mathcal{X} \times \mathcal{Y}$.
2. Datos de salida proporcionados por el observador: el observador debe, a partir de los datos de entrenamiento, construir una regla $h : \mathcal{X} \longrightarrow \mathcal{Y}$ que permita predecir la etiqueta de nuevas observaciones. Esta regla recibe el nombre de función de clasificación, función de hipótesis o función de predicción.
3. Un modelo de generación de datos: denotaremos por $\mathcal{D}$ a la distribución conjunta sobre $\mathcal{X} \times \mathcal{Y}$. Podemos separar dicha distribución en dos contribuciones: la distribución $\mathcal{D}_x$ sobre los datos del dominio $\mathcal{X}$, y la distribución condicionada sobre las etiquetas de los puntos del dominio, $\mathcal{D}((\mathbf{x}, y)|\mathbf{x})$. Es importante notar que, en principio, el observador desconoce cualquier información acerca de esta distribución de probabilidad.
4. Medida del éxito: definimos el error del modelo como la probabilidad de que no prediga la etiqueta correcta en un conjunto de muestras aleatorias generadas por la distribución $\mathcal{D}$. En este sentido, introducimos la función de pérdida como una medida de este error, siendo ésta una aplicación $\mathcal{L} : \mathcal{H} \times Z \to \mathbb{R}_+$, donde Z es un cierto dominio (para problemas de predicción, $Z = \mathcal{X} \times \mathcal{Y}$) y $\mathcal{H}$ el conjunto que contiene a las funciones de hipótesis. Definimos entonces el riesgo como el valor esperado de la función de error o pérdida, que viene dado por:

$$\mathcal{R}_{\mathcal{D}}(h) = \int \mathcal{L}(h, (\mathbf{x}, y)) dD(\mathbf{x}, y), \tag{6.1}$$

   donde $\mathcal{L}(h, (\mathbf{x}, y))$ es la función de error y $D(\mathbf{x}, y)$ es la función de distribución conjunta. El riesgo es, por tanto, una forma de cuantificar el error cometido al aproximar el valor real, y_{real}, por el valor aproximado de acuerdo con la función de predicción o hipótesis h que hemos calculado, $y_{esperado}$.

Podemos decir entonces que el aprendizaje PAC es un modelo de aprendizaje basado en la capacidad de un observador para obtener una solución aproximada y con un alto grado de exactitud con una elevada probabilidad. Es importante resaltar que el aprendizaje PAC y el modelo de máxima verosimilitud son dos aproximaciones diferentes. En el primero estamos usando una aproximación discriminativa, de modo que tratamos de maximizar la verosimilitud condicional; por el contrario, en el segundo buscamos un modelo que nos permita describir los datos, es decir, se busca maximizar la verosimilitud del conjunto de datos dado el modelo aprendido.

6.2. Minimización del riesgo empírico.

Tal y como hemos comentado en las secciones anteriores, un modelo de aprendizaje se basa en una muestra de datos de entretenimiento S generados a partir de una cierta distribución $\mathcal{D}$ y tiene como objetivo determinar la función de clasificación o de predicción h_S. El objetivo del algoritmo es encontrar la función de predicción que minimice el riesgo con respecto a la distribución $\mathcal{D}$. Dicho problema se conoce como problema de minimización del riesgo empírico (en inglés, *Empirical Risk Minimization, ERM)*. Puesto que dicha distribución es desconocida para el observador, no es posible calcular el riesgo real. Trabajamos entonces con el riesgo empírico, que es el error de clasificación sobre los datos de entrenamiento:

$$\mathcal{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(h,(\mathbf{x},y)). \tag{6.2}$$

En el marco de la regresión lineal, la función de pérdida viene dada por $\mathcal{L}(h,(\mathbf{x},y)) = (y-h(\mathbf{x}))^2$. Por tanto, sustituyendo la función de error en la definición del riesgo empírico dada por (6.2), tenemos que

$$\mathcal{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n} (y-h(\mathbf{x}))^2. \tag{6.3}$$

El inconveniente que plantea el método tal y como lo hemos introducido es que es propenso a problemas de sobreajuste, ya que puede ser que minimice el error sobre los datos de la muestra, pero que, sin embargo, el error cometido sobre otras observaciones sea muy elevado. Esto ocurre cuando nuestra hipótesis h ajusta los datos de entrenamiento demasiado bien.

Un forma de evitar los problemas de sobreajuste asociados a la minimización del riesgo empírico es reducir el conjunto de hipótesis o funciones de predicción, $\mathcal{H}$. De acuerdo con esta restricción, se trata de encontrar la hipótesis h que reduzca el riesgo empírico, con $h \in \mathcal{H}$:

$$ERM_{\mathcal{H}}(S) \in \operatorname*{argmín}_{h\in\mathcal{H}} \mathcal{R}_S(h).$$

La elección del conjunto $\mathcal{H}$ se basa entonces en el conocimiento previo o a priori del problema que se desea estudiar. Esta restricción se conoce como restricción de sesgo inductivo. La restricción más sencilla que podemos imponer al conjunto de hipótesis posibles es mediante una cota superior en su tamaño. Es decir, limitamos el número de funciones de predicción en $\mathcal{H}$.

Observamos que $\mathcal{R}_S(h_S)$ depende de la muestra de entrenamiento, que se obtiene a partir de un proceso aleatorio. Por ello, el riesgo es una variable aleatoria. Por consiguiente, no es realista asumir que podemos, a partir de la muestra S, determinar de manera exacta la función de distribución $\mathcal{D}$, ya que siempre hay alguna probabilidad de que la muestra de entrenamiento no sea muy representativa. En este sentido, denotaremos por δ a la probabilidad de obtener una muestra no representativa, de modo que $1-\delta$ será el parámetro de confianza. Por otro lado, puesto que no es posible garantizar una predicción exacta, introducimos otro parámetro que da cuenta de la calidad de la predicción, ϵ, que recibe el nombre de parámetro de exactitud. De esta forma, consideraremos que el algoritmo es adecuado, aproximadamente correcto, si obtenemos $\mathcal{R}_S(h_S) \leq \epsilon$. Se trata entonces de encontrar una cota superior del riesgo empírico que nos permita determinar si el algoritmo

es adecuado, en el sentido de que da una buena aproximación del verdadero error. Esto será tratado en secciones posteriores.

6.3. Consistencia de un modelo de aprendizaje.

El objetivo de esta sección es determinar cuándo la minimización el riesgo empírico implica verdaderamente un riesgo bajo. Siguiendo con la notación presentada en la sección anterior introducimos la siguiente definición [50, Sección 2.1].

Definición 6.3.1 (Consistencia)**.** *Diremos que el método de minimización de riesgo empírico es consistente para el conjunto de funciones $h \in \mathcal{H}$, y para la función de distribución $D(\mathbf{x}, y)$ si las siguientes secuencias convergen en probabilidad hacia el mismo límite:*

$$R_{\mathcal{D}}(h_S) \xrightarrow[n\to\infty]{P} \inf_{h\in\mathcal{H}} R(h). \tag{6.4}$$

$$R_S(h_S) \xrightarrow[n\to\infty]{P} \inf_{h\in\mathcal{H}} R(h). \tag{6.5}$$

Es decir, diremos que el método de minimización del riesgo empírico es consistente si proporciona una serie de funciones $h \in \mathcal{H}$ para las cuales tanto el riesgo real como el riesgo empírico convergen hacia el menor valor de la función de riesgo. Por tanto, un método será consistente si al aumentar el número de datos de entrenamiento aseguramos una mejor predicción de la distribución que los genera.

Bajo ciertas condiciones, es posible asegurar que la regresión basada en procesos Gaussianos es consistente [14]. Estas condiciones se basan en la suavidad de la función de media y la función de covarianza, así como la suavidad de $E[y|\mathbf{x}]$. Adicionalmente, debemos suponer que el ruido de las observaciones tiene una distribución normal o Laplaciana. La consistencia es una propiedad interesante para métodos de aprendizaje supervisado. No obstante, no nos informa sobre cómo funcionará un cierto proceso para un determinado problema.

6.4. Aprendizaje PAC.

En esta sección introducimos formalmente el concepto de aprendizaje PAC y las condiciones bajo las cuales una clase es aprendible en este sentido [41, Definición 3.3].

Definición 6.4.1 (Clase aprendible)**.** *Diremos que una clase $\mathcal{H}$ es PAC aprendible con respecto a un dominio Z y una función de pérdida $\mathcal{L} : \mathcal{H} \times Z \longrightarrow \mathbb{R}_+$ si existe una función $m_{\mathcal{H}} : (0,1)^2 \longrightarrow \mathbb{N}$ y un algoritmo de aprendizaje verificando la siguiente propiedad: Para todos $\epsilon, \delta \in (0,1)$ y para toda distribución $\mathcal{D}$ sobre $\mathcal{X} \times \mathcal{Y}$, al efectuar el algoritmo de aprendizaje sobre m muestras de entrenamiento i.i.d. generadas por la distribución $\mathcal{D}$, tales que $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, es posible encontrar una hipótesis o función de predicción h tal que, con probabilidad al menos $1 - \delta$,*

$$\mathcal{R}_{\mathcal{D}}(h) \leq \min_{h'\in\mathcal{H}} \mathcal{R}_{\mathcal{D}}(h') + \epsilon,$$

donde $\mathcal{R}_{\mathcal{D}}(h) = E_{z\sim\mathcal{D}}[\mathcal{L}(h, z)]$.

La función $m_{\mathcal{H}} : (0,1)^2 \longrightarrow \mathbb{N}$ determina la complejidad de la muestra. Es decir, el tamaño de la muestra de entrenamiento S que garantiza que exista una solución probablemente

aproximadamente correcta. Es por tanto función de ϵ y de δ, así como de las propiedades del conjunto de hipótesis $\mathcal{H}$. Puesto que existen varias funciones que satisfacen estos criterios, generalmente nos referiremos a la mínima función, en el sentido de que $m_{\mathcal{H}}(\epsilon, \delta)$ es el mínimo valor para el cual se cumplen los requisitos para ser una clase aprendible.

Introducimos a continuación el concepto de convergencia uniforme [41, Definición 4.1], que nos permitirá demostrar que cualquier clase finita es aprendible, independientemente de la función de pérdida, siempre y cuando el rango de ésta esté acotado. La idea es que podamos garantizar que de manera uniforme sobre los miembros de $\mathcal{H}$ el riesgo empírico sea próximo al verdadero riesgo. Es decir, que una función de predicción h que minimice el riesgo empírico con respecto a la muestra S sea de hecho un minimizador del riesgo con respecto a la distribución de probabilidad real de los datos. Para ello, es necesario presentar previamente algunos conceptos y resultados.

Definición 6.4.2 (Muestra ϵ-representativa)**.** *Diremos que un conjunto de entrenamiento S es ϵ-representativo con respecto a un cierto dominio Z, un conjunto de hipótesis $\mathcal{H}$, una función de pérdida $\mathcal{L}$ y una función de distribución $\mathcal{D}$ si*

$$\forall h \in \mathcal{H}, |\mathcal{R}_S(h) - \mathcal{R}_{\mathcal{D}}(h)| \leq \epsilon.$$

El siguiente resultado muestra que siempre que la muestra de entrenamiento sea $(\epsilon/2)$-representativa, la minimización del riesgo empírico proporciona una buena función de predicción [41, Lema 4.2].

Lema 6.4.1. *Suponemos que la muestra de entrenamiento S es $(\epsilon/2)$-representativa, con respecto a un cierto dominio Z, una clase de hipótesis $\mathcal{H}$, función de pérdida $\mathcal{L}$, y distribución conjunta $\mathcal{D}$. Entonces, cualquier hipótesis h obtenida como resultado de la minimización del riesgo empírico $ERM_{\mathcal{H}}(S)$, $h_s \in \underset{h\in\mathcal{H}}{argmin}\mathcal{R}_{\mathcal{D}}(h) + \epsilon$, verifica*

$$\mathcal{R}_{\mathcal{R}}(h_S) \leq \min_{h\in\mathcal{H}} \mathcal{R}_{\mathcal{D}}(h) + \epsilon.$$

Demostración. Por ser S una muestra de entrenamiento $(\epsilon/2)$-representativa, tenemos que

$$\mathcal{R}_{\mathcal{D}}(h_S) \leq \mathcal{R}_S(h_S) + \epsilon/2 \leq \mathcal{R}_S(h) + \epsilon/2 \leq \mathcal{R}_{\mathcal{D}}(h) + \epsilon/2 + \epsilon/2 \leq \mathcal{R}_{\mathcal{D}}(h) + \epsilon,$$

donde la primera y la tercera desigualdad se deducen de que S es una muestra de entrenamiento $(\epsilon/2)$-representativa y la segunda de que h_S es un minimizador del riesgo empírico. □

La siguiente definición hace referencia al concepto de convergencia uniforme [41, Definición 4.3], necesario para caracterizar las clases PAC aprendibles.

Definición 6.4.3 (Convergencia uniforme)**.** *Diremos que un conjunto de hipótesis $\mathcal{H}$ tiene la propiedad de la convergencia uniforme si existe una función $m_{\mathcal{H}}^{UC} : (0,1)^2 \longrightarrow \mathbb{N}$ tal que para todos $\epsilon, \delta \in (0,1)$ y para toda función de distribución $\mathcal{D}$ sobre Z, si S es un conjunto de entrenamiento de $m \geq m_{\mathcal{H}}^{UC}$ muestras i.i.d. de acuerdo con $\mathcal{D}$ entonces, con probabilidad de al menos $1 - \delta$, S es ϵ-representativo.*

La función $m_{\mathcal{H}}^{UC}$ mide el tamaño mínimo de la muestra para el cual se tiene la propiedad de convergencia uniforme. Es decir, indica el tamaño mínimo necesario para que con probabilidad al menos $1 - \delta$ la muestra sea ϵ-representativa.

Veremos a continuación que las clases finitas son PAC aprendibles [41, Sección 4.2]. De acuerdo con la siguiente proposición [41, Corolario 4.4], será suficiente probar que toda clase finita tiene la propiedad de la convergencia uniforme.

Proposición 6.4.1. *Si la clase de hipótesis $\mathcal{H}$ tiene la propiedad de la convergencia uniforme para una cierta función $m_{\mathcal{H}}^{UC}$ entonces dicha clase es PAC aprendible para una muestra de tamaño $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$.*

Teorema 6.4.1. *Si la función de pérdida está acotada, toda clase de hipótesis finita tiene la propiedad de la convergencia uniforme.*

Demostración. Sin pérdida de generalidad, suponemos que la función de pérdida está acotada por uno. Fijamos $\epsilon y \delta$. Buscamos un número natural m tal que si la muestra de entrenamiento es de tamaño m, entonces para cualquier distribución $\mathcal{D}$, con probabilidad al menos $1-\delta$ se tiene que para toda hipótesis $h \in \mathcal{H}$, $|\mathcal{R}_S(h) - \mathcal{R}_{\mathcal{D}}(h)| \leq \epsilon$. Es decir,

$$\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |\mathcal{R}_S(h) - \mathcal{R}_{\mathcal{D}}(h)| \leq \epsilon\}) \geq \delta. \tag{6.6}$$

Equivalentemente, podemos probar que

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |\mathcal{R}_S(h) - \mathcal{R}_{\mathcal{D}}(h)| > \epsilon\}) < \delta. \tag{6.7}$$

Podemos escribir lo siguiente:

$$\{S : \exists h \in \mathcal{H}, |\mathcal{R}_S(h) - \mathcal{R}_{\mathcal{D}}(h)| > \epsilon\} = \cup_{h \in \mathcal{H}} \{S : |\mathcal{R}_S(h) - \mathcal{R}_{\mathcal{D}}(h)| > \epsilon\}.$$

Usando la desigualdad de Boole C.4,

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |\mathcal{R}_S(h) - \mathcal{R}_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |\mathcal{R}_S(h) - \mathcal{R}_{\mathcal{D}}(h)| > \epsilon\}). \tag{6.8}$$

Veremos a continuación que los sumandos de la derecha son suficientemente pequeños siempre y cuando el tamaño de la muestra m sea suficientemente grande. Es decir, veremos que el riesgo empírico difiere poco del riesgo real. Por una lado tenemos que $\mathcal{R}_{\mathcal{D}}(h) = E_{z \sim \mathcal{D}}[\mathcal{L}(h, z)]$ mientras que $\mathcal{R}_S(h) = \frac{1}{m}\sum_{i=1}^m \mathcal{L}(h, z_i)$. Puesto que las variables z_i son muestras aleatorias obtenidas a partir de la distribución $\mathcal{D}$, el valor esperado de $\mathcal{L}(h, z_i)$ es precisamente $\mathcal{L}_{\mathcal{D}}(h)$. Por linealidad de la esperanza, se tiene que $\mathcal{L}_{\mathcal{D}}(h)$ es también el valor esperado de $\mathcal{R}_S(h)$. Por consiguiente, $|\mathcal{R}_S(h) - \mathcal{R}_{\mathcal{D}}(h)|$ es una medida de la desviación de la variable aleatoria $\mathcal{R}_S(h)$ con respecto a su valor esperado. Si el tamaño de la muestra m tendiese a infinito, el resultado sería inmediato a partir de la ley de los grandes números. No obstante, el tamaño de la muestra es un número finito, por lo que dicho resultado no nos proporciona información acerca de la diferencia entre el riesgo real y el empírico. En su lugar, empleamos la desigualdad de Hoeffding C.2. Denotando por x_i a $\mathcal{L}(h, z_i)$, se tiene que $\mathcal{R}_S(h) = \frac{1}{m}\sum_{i=1}^m x_i$. Por otro lado, tenemos $\mathcal{R}_{\mathcal{D}}(h) = \mu$. Podemos entonces escribir lo siguiente:

$$\mathcal{D}^m(\{S : |\mathcal{R}_S(h) - \mathcal{R}_{\mathcal{D}}(h)| > \epsilon\}) = P\left[|\frac{1}{m}\sum_{i=1}^m x_i - \mu| > \epsilon\right] \leq 2\exp\left(-2m\epsilon^2\right) \tag{6.9}$$

Usando la ecuación (6.8), tenemos que

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |\mathcal{R}_S(h) - \mathcal{R}_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{i=1}^m 2\exp\left(-2m\epsilon^2\right) = 2|\mathcal{H}|\exp\left(-2m\epsilon^2\right). \tag{6.10}$$

Por tanto, si tomamos

$$m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2},$$

entonces,

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |\mathcal{R}_S(h) - \mathcal{R}_\mathcal{D}(h)| > \epsilon\}) \leq \delta. \tag{6.11}$$

□

6.5. Sobreajuste.

Una forma de evitar problemas de sobreajuste en un cierto modelo es restringir la clase de hipótesis $\mathcal{H}$, que proporciona información a priori sobre un cierto problema. Pero, ¿es esta información estrictamente necesaria para que un modelo funcione correctamente? Podríamos pensar que existe algún modelo capaz de resolver cualquier problema sin necesidad de incorporar ningún tipo de información a priori. Es posible probar que dicho algoritmo no existe, ya que para cualquier método, siempre existe una distribución para la cual éste falla. (Esta idea aparece en la literatura bajo el nombre de *No-Free-Lunch Theorem [41, Capítulo 5]*). De acuerdo con este resultado, no existe ningún algoritmo capaz de resolver con éxito cualquier problema. Por tanto, es necesario incluir cierta información a priori sobre la distribución $\mathcal{D}$ en nuestro modelo de aprendizaje. Una forma de incluir esta información es, por ejemplo, imponiendo que $\mathcal{D}$ pertenezca una familia de distribuciones paramétricas. Podemos también asumir que existe una cierta hipótesis h definida en una cierta clase de hipótesis $\mathcal{H}$ para la cual $\mathcal{R}_\mathcal{D}(h) = 0$. Una suposición algo más débil es asumir que $\min_{h\in\mathcal{H}} \mathcal{R}_\mathcal{D}(h)$ es pequeño. El objetivo de esta sección es estudiar los beneficios de incorporar dicha información a priori en el modelo. Definimos el exceso de riesgo de una regla de la siguiente forma:

$$\epsilon_\mathcal{D}(\mathcal{H}) = \mathcal{R}_\mathcal{D}(\mathcal{H}) - \mathcal{R}^*_\mathcal{D}(\mathcal{H}) \geq 0, \tag{6.12}$$

donde $\mathcal{R}^*_\mathcal{D}(\mathcal{H})$ denota el riesgo de Bayes sobre el conjunto total de la clase de hipótesis, y $\mathcal{R}_\mathcal{D}(\mathcal{H})$ denota el riesgo sobre una clase de hipótesis restringida. Buscamos entonces una descomposición del error $\mathcal{R}_\mathcal{D}(h_S) - \mathcal{R}^*_\mathcal{D}$ que nos permita seleccionar una clase de hipótesis $\mathcal{H}$ óptima para el modelo, encontrando un equilibrio entre modelos más o menos complejos que garanticen mejores resultados. Sumando y restando los términos $\mathcal{R}_\mathcal{D}(h)$ y $\mathcal{R}_S(h_S)$ tenemos que

$$\mathcal{R}_\mathcal{D}(h_S) - \mathcal{R}^*_\mathcal{D} = \mathcal{R}_\mathcal{D}(h_S) - \mathcal{R}_S(h_S) + \mathcal{R}_S(h_S) - \mathcal{R}_\mathcal{D}(h) + \mathcal{R}_\mathcal{D}(h) - \mathcal{R}^*_\mathcal{D}.$$

Puesto que h_S es un minimizador del riesgo empírico, se deduce lo siguiente:

$$\mathcal{R}_S(h_S) \leq \mathcal{R}_S(h).$$

Por tanto,

$$\mathcal{R}_\mathcal{D}(h_S) - \mathcal{R}^*_\mathcal{D} \leq (\mathcal{R}_\mathcal{D}(h_S) - \mathcal{R}_S(h_S)) + (\mathcal{R}_S(h) - \mathcal{R}_\mathcal{D}(h)) + (\mathcal{R}_\mathcal{D}(h) - \mathcal{R}^*_\mathcal{D}).$$

Acotando los dos primeros paréntesis por $\sup_{h\in\mathcal{H}} |\mathcal{R}_S(h) - \mathcal{R}_\mathcal{D}(h)|$ obtenemos la siguiente descomposición:

$$\mathcal{R}_\mathcal{D}(h_S) - \mathcal{R}^*_\mathcal{D} \leq \underbrace{2\sup_{h\in\mathcal{H}} |\mathcal{R}_S(h) - \mathcal{R}_\mathcal{D}(h)|}_{\text{Error de estimación}} + \underbrace{(\mathcal{R}_\mathcal{D}(h) - \mathcal{R}^*_\mathcal{D})}_{\text{Error de aproximación}}. \tag{6.13}$$

El error de estimación (*estimation error*), crece con el tamaño de $\mathcal{H}$, midiendo así el error debido a sobreajuste asociado con el tamaño o complejidad de la clase de hipótesis $\mathcal{H}$. Este error refleja la diferencia entre el error de aproximación y el error que obtenemos mediante minimización del riesgo empírico. Esta diferencia se debe a que el riesgo empírico es tan solo un estimador del riesgo real, por lo que el minimizador del riesgo empírico es tan solo un estimador del minimizador del riesgo real. Por el contrario, el error de aproximación (*model error*) decrece al aumentar $\mathcal{H}$ y refleja el conocimiento que tenemos a priori de la distribución $\mathcal{D}$, coincidiendo con el exceso de riesgo definido en (6.12). Este término refleja por tanto el mínimo error que podemos obtener para una hipótesis $h \in \mathcal{H}$, en función de las restricciones que impongamos sobre esta clase. Por consiguiente, el error de aproximación no depende del tamaño de la clase.

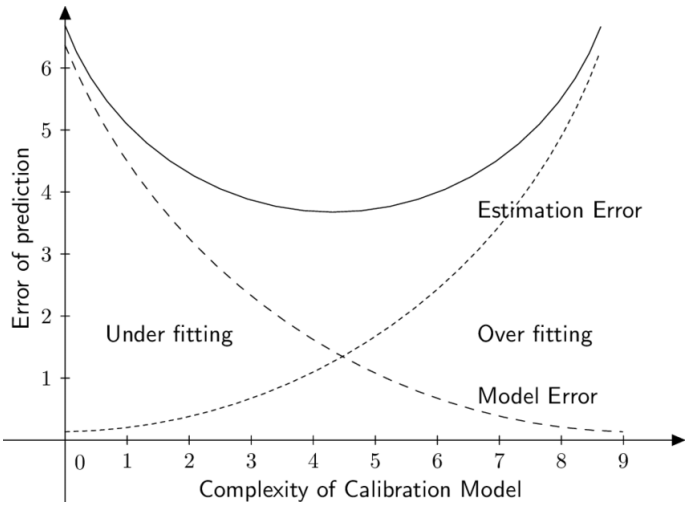


Figura 6.5.1: Error de estimación frente a error de aproximación. La curva negra representa el riesgo total [35].

Puesto que el objetivo es minimizar el riesgo total, debemos encontrar un equilibrio entre clases complejas $\mathcal{H}$, que aumentan el riesgo de sobreajuste pero disminuyen el sesgo, o clases más sencillas, que pueden aumentar el sesgo pero disminuir el riesgo de sobreajuste. Este problema se conoce también como equilibrio entre sesgo y varianza. La teoría de aprendizaje estudia por tanto como podemos escoger $\mathcal{H}$ de manera óptima manteniendo un error razonable.

La ventaja de esta descomposición es que el error de estimación es el único término aleatorio. El error de aproximación, como hemos señalado antes, depende de cómo de compleja sea la clase de hipótesis $\mathcal{H}$, pero no depende de la muestra de entrenamiento S.

6.6. Dimensión VC.

En las secciones anteriores hemos probado que una clase de hipótesis finita es aprendible, y por tanto ésta es una condición suficiente. No obstante, no es condición necesaria, pues es posible que $\mathcal{H}$ sea infinito y aun así sea aprendible. En esta sección introducimos el concepto de dimensión VC que nos permite caracterizar cuándo una clase es aprendible o no, así como cuantificar el poder de clasificación de un cierto algoritmo. La dimensión VC es una medida de la complejidad del conjunto de hipótesis $\mathcal{H}$ basada en el número

de muestras del conjunto $\mathcal{X}$ que pueden ser separadas por $\mathcal{H}$, en lugar de el número de hipótesis que contiene. No se pretende entrar en detalle, por lo que se explicarán los conceptos básicos a partir de los cuales es posible formular uno de los resultados fundamentales (*Teorema Fundamental del Aprendizaje Estadístico* (6.6.1)). Además, es importante resaltar que trataremos únicamente problemas de clasificación, ya que se trata de resultados más sencillos y manejables.

Supongamos que tenemos un cierto conjunto C y un conjunto de hipótesis $\mathcal{H}$. Cada hipótesis $h \in \mathcal{H}$ induce una partición en el conjunto C dada por dos subconjuntos: $\{x \in C : h(x) = 1\}$ y $\{x \in C : h(x) = 0\}$. Por tanto, para un cierto conjunto C existen $2^{|C|}$ posibles particiones de este. Diremos entonces que el conjunto $\mathcal{H}$ separa C. La siguiente definición recoge esta idea [41, Definición 6.3].

Definición 6.6.1 (Conjunto separado)**.** *Diremos que un conjunto de hipótesis $\mathcal{H}$ separa un conjunto finito $C \subset \mathcal{X}$ si la restricción de $\mathcal{H}$ a C es el conjunto de todas las funciones de C en $\{0, 1\}$. Es decir, $|\mathcal{H}_C| = 2^{|C|}$.*

Parece entonces razonable pensar que cuánto mayor sea el tamaño del conjunto C que $\mathcal{H}$ puede separar más compleja será ésta, y, por tanto, más información contendrá. La dimensión VC mide precisamente esta idea [29, Sección 7.4.2].

Definición 6.6.2 (Dimensión VC)**.** *La dimensión VC del espacio de hipótesis $\mathcal{H}$ definida sobre el dominio $\mathcal{X}$ es el tamaño del mayor subconjunto C de $\mathcal{X}$ que puede ser separado por $\mathcal{H}$. Si existe un subconjunto de $\mathcal{X}$ de dimensión arbitraria que puede ser separado por $\mathcal{H}$ diremos que VC($\mathcal{H}$)= ∞.*

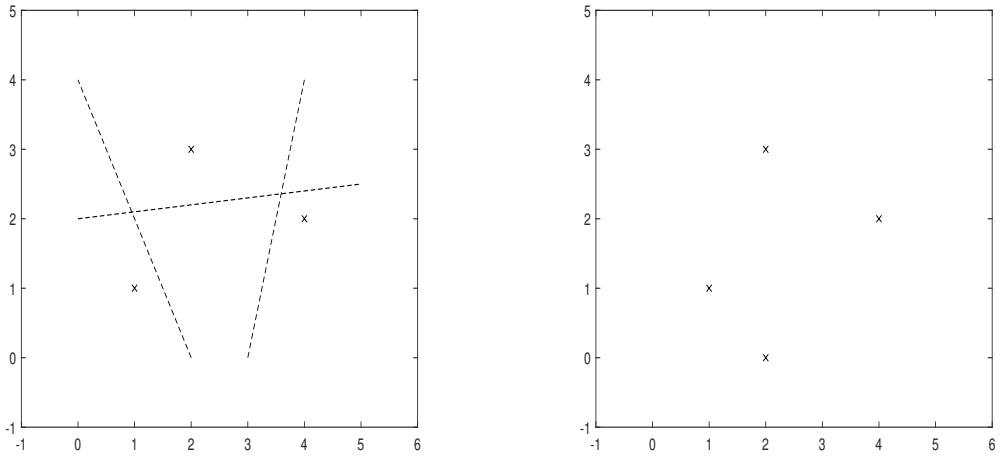


Figura 6.6.1: La dimensión VC de los semi-espacios en el plano es 3, ya que pueden separar tres puntos, pero no cuatro. Vemos que en la segunda figura los puntos que están en la misma vertical no pueden separase mediante una recta de los otros dos.

Podemos preguntarnos cuál ha de ser el tamaño mínimo de una muestra de entrenamiento para que podamos aprender de forma probablemente aproximadamente correcta un cierto concepto. En otras palabras, cuál ha de ser el tamaño de la muestra para que ésta sea ϵ-representativa con probabilidad al menos $1-\delta$. El siguiente teorema [21, Teorema 1.1], [29, Sección 7.4.3], que se conoce en la literatura como *Teorema Fundamental del Aprendizaje*

Estadístico, muestra que la dimensión VC de una clase de hipótesis determina el número de muestras necesarias para que una clase sea PAC aprendible. En particular, el teorema afirma que una clase es aprendible si su dimensión VC es finita.

Teorema 6.6.1. *Sea $\mathcal{H}$ una clase de hipótesis. Entonces,*

- *La clase $\mathcal{H}$ es PAC si y sólo si su dimensión VC de es finita.*
- *Si VC($\mathcal{H}$) = d, $d < \infty$, entonces*

$$m \geq \left(\frac{1}{\epsilon} 4 \log_2(2/\delta) + \frac{1}{\epsilon} 8\, VC(\mathcal{H}) \log_2(13/\epsilon) \right),$$

Es posible derivar también una cota superior para el número de muestras de entrenamiento necesarias para que una cierta clase sea PAC aprendible [29, Sección 7.4.3].

6.7. Aprendizaje PAC-Bayesiano.

Hasta el momento hemos tratado dos vías aparentemente separadas en la teoría del aprendizaje: la inferencia Bayesiana y el aprendizaje PAC. En ambos casos se pretende, a partir de unos datos iniciales, construir un modelo que pueda ser luego aplicado a datos de entrenamiento. En el caso del aprendizaje PAC, hemos visto que existen teoremas que garantizan que una cierta clase sea aprendible siempre y cuando las muestras sean independientes e igualmente distribuidas. En el caso de la inferencia Bayesiana, la exactitud del modelo depende de que los datos de prueba y de entrenamiento hayan sido generados de acuerdo con una cierta distribución a priori. De esta manera, el aprendizaje Bayesiano presupone que los datos han sido generados a partir de un cierto modelo que pertenece a la clase de hipótesis con la que estamos aprendiendo.

Cuando se cumplen las hipótesis sobre el conocimiento a priori de la distribución que origina los datos, los algoritmos Bayesianos son muy efectivos. No obstante, tienden a dar problemas de sobreajuste en caso contrario. Además, aplicados a problemas reales suelen ser demasiado optimistas, ya que no se tiene por qué tener información a priori de la clase que se está estudiando. Por consiguiente, los modelos Bayesianos requieren de un fuerte conocimiento a priori de la función que se desea estudiar. Por otro lado, el aprendizaje PAC basado en muestras i.i.d. no produce problemas de sobreajuste pero no puede usar de manera efectiva la información a priori que se tiene de un cierto modelo. Sin embargo, es capaz de proporcionar una serie de garantías sobre la exactitud del modelo cuando se aplica a nuevos datos. Además, el aprendizaje PAC tiende a ser un modelo genérico, ya que se garantiza cierta exactitud incluso en el peor de los escenarios. De esta forma, el aprendizaje PAC-Bayesiano es una manera de combinar ambos métodos para lograr una compensación entre la generalidad de uno (PAC) y la eficiencia del otro (Inferencia Bayesiana).

El inconveniente que presentan las cotas derivadas de la dimensión VC es que proporcionan un error entre el riesgo real y el empírico en función de la peor elección posible de $h \in \mathcal{H}$. Evidentemente, esta cota será válida para el algoritmo que usemos en un caso concreto, pero lo es además para cualquier otra forma de escoger $\mathcal{H}$, incluyendo el algoritmo más malicioso que, aún conociendo la distribución original, selecciona una hipótesis con un error máximo. Con el objetivo de evitar estos problemas, es frecuente emplear otras

medidas de la complejidad del algoritmo, a parte de la dimensión VC, que nos permitan introducir la información a priori que tenemos de la distribución. Además, la dimensión VC depende exclusivamente del tamaño de la muestra, pero no de la información que ésta proporciona. El aprendizaje PAC-Bayesiano puede solventar también este inconveniente.

En el aprendizaje PAC-Bayesiano el conocimiento a priori sobre la regla objetivo se expresa a través de una distribución a priori sobre el conjunto de hipótesis $\mathcal{H}$. Por consiguiente, asociamos una probabilidad (o densidad de probabilidad si $\mathcal{H}$ es continuo) $\mathcal{P}(h) \geq 0$ para cada $h \in \mathcal{H}$ y nos referiremos a $\mathcal{P}(h)$ como el peso a priori de h. De acuerdo con los métodos basados en inferencia Bayesiana, el resultado del algoritmo no es una sola hipótesis, sino una probabilidad a posteriori sobre $\mathcal{H}$ que denotaremos por $\mathcal{Q}$. Definimos entonces el error o función de pérdida de $\mathcal{Q}$ sobre una muestra z como el valor promedio de la función de error [41]:

$$\mathcal{L}(\mathcal{Q}) = \underset{h\sim\mathcal{Q}}{E}[\mathcal{L}(h, z)], \tag{6.14}$$

donde $h \sim \mathcal{Q}$ hace referencia a que h es una realización de la distribución $\mathcal{Q}$.

De forma similar, como consecuencia de la linealidad del valor esperado, definimos el riesgo real y el riesgo empírico [41]:

$$R_{\mathcal{D}} = \mathcal{R}_{\mathcal{D}}(\mathcal{Q}) = \underset{h\sim\mathcal{Q}}{E}[\mathcal{R}_{\mathcal{D}}(h)]. \tag{6.15}$$

$$R_S = \mathcal{R}_S(\mathcal{Q}) = \underset{h\sim\mathcal{Q}}{E}[\mathcal{R}_S(h)]. \tag{6.16}$$

El siguiente teorema [41, Teorema 31.1] nos muestra que la diferencia entre el riesgo real y el riesgo empírico de la distribución a posteriori $\mathcal{Q}$ está acotada por una expresión que depende de la divergencia de Kullback-Leibler entre $\mathcal{Q}$ y la distribución a priori $\mathcal{P}$. Se trata de la versión Bayesiana correspondiente al teorema (6.6.1). Este teorema sugiere que si queremos minimizar el riesgo real, debemos entonces minimizar el riesgo empírico y la distancia de Kullback-Leibler entre $\mathcal{Q}$ y la distribución a priori $\mathcal{P}$. Tal y como está enunciado, hace referencia a problemas de clasificación, aunque puede adaptarse a otros casos introduciendo las modificaciones adecuadas [34, Teorema 7.1].

Teorema 6.7.1 (Teorema PAC-Bayesiano de McAllester)**.** *Sea $\mathcal{D}$ una distribución arbitraria sobre un cierto dominio Z. Sea $\mathcal{H}$ un conjunto de hipótesis o clase de hipótesis y sea $\mathcal{L} : \mathcal{H} \times Z \longrightarrow [0, 1]$ una función de pérdida. Sea $\mathcal{P}$ una distribución a priori sobre $\mathcal{H}$ y sea $\delta \in (0, 1)$. Entonces, con probabilidad al menos $1 - \delta$ sobre un conjunto de entrenamiento $S = (z_1, ..., z_m)$ con z_i muestras i.i.d. distribuidas de acuerdo con $\mathcal{D}$, para toda distribución $\mathcal{Q}$ sobre $\mathcal{H}$, se tiene que*

$$\mathcal{R}_{\mathcal{D}}(\mathcal{Q}) \leq \mathcal{R}_S(\mathcal{Q}) + \sqrt{\frac{D(\mathcal{Q}||\mathcal{P}) + \ln(2m/\delta)}{2(m-1)}},$$

donde

$$D(\mathcal{Q}||\mathcal{P}) = \underset{h\sim\mathcal{Q}}{E}[\ln(\mathcal{Q}(h)/\mathcal{P}(h))]$$

es la divergencia de Kullback-Leibler.

Demostración. Para cualquier función f, de acuerdo con la desigualdad de Markov C.3 se tiene que

$$P_S[f(S) \geq \epsilon] = P_S[e^{f(S)} \geq e^{\epsilon}] \leq \frac{E_S[e^{f(S)}]}{e^{\epsilon}}. \tag{6.17}$$

Dentamos por $\Delta(h) = \mathcal{R}_{\mathcal{D}}(h) - \mathcal{R}_S(h)$ y aplicamos la ecuación anterior a la siguiente función:

$$f(S) = \sup_{\mathcal{Q}} \left(2(m-1) \underset{h\sim\mathcal{Q}}{E} (\Delta(h))^2 - D(\mathcal{Q}||\mathcal{P}) \right).$$

Buscaremos una cota para $E_S[e^{f(S)}]$. Para ello, buscaremos una expresión que no dependa de $\mathcal{Q}$ sino de la probabilidad a priori $\mathcal{P}$. Usando la definición de $D(\mathcal{Q}||\mathcal{P})$ tenemos la siguiente desigualdad:

$$\begin{aligned} 2(m-1) \underset{h\sim\mathcal{Q}}{E} (\Delta(h))^2 - D(\mathcal{Q}||\mathcal{P}) &= 2(m-1) \underset{h\sim\mathcal{Q}}{E} (\Delta(h))^2 - \underset{h\sim\mathcal{Q}}{E} [\ln(\mathcal{Q}(h)/\mathcal{P}(h))] = \\ \underset{h\sim\mathcal{Q}}{E} [\ln(e^{2(m-1)\Delta(h)^2}\mathcal{P}(h)/\mathcal{Q}(h))] &\leq \ln \underset{h\sim\mathcal{Q}}{E} [e^{2(m-1)\Delta(h)^2}\mathcal{P}(h)/\mathcal{Q}(h))] = \\ \ln \underset{h\sim\mathcal{P}}{E} [e^{2(m-1)\Delta(h)^2}], & \end{aligned} \tag{6.18}$$

donde la desigualdad se deduce de la desigualdad de Jensen C.3.1 y la convexidad de la función logaritmo. La última igualdad se obtiene a partir del teorema de Radon-Nikodym D.0.1. En concreto, se deduce a partir de la siguiente igualdad,

$$\int h dP = \int h \frac{dP}{dQ} dQ, \tag{6.19}$$

donde $f = \frac{dP}{dQ}$ es la derivada de Radon-Nikodym de P respecto de Q. Notar que si h es la función indicadora de un conjunto, se recupera la definición dada por el Teorema de Radon-Nikodym. Por consiguiente, se deduce que

$$E_S[e^{f(S)}] \leq E_S \underset{h\sim\mathcal{P}}{E} [e^{2(m-1)\Delta(h)^2}]. \tag{6.20}$$

Puesto que la probabilidad a priori no depende de la muestra S, podemos intercambiar los valores esperados, de modo que tenemos que

$$E_S[e^{f(S)}] \leq \underset{h\sim\mathcal{P}}{E} E_S[e^{2(m-1)\Delta(h)^2}]. \tag{6.21}$$

Necesitamos probar ahora que $E_S[e^{2(m-1)\Delta(h)^2}] \leq 2m$. Para ello, empleamos la desigualdad de Hoeffding C.2. Puesto que

$$E(z) = \int_0^\infty P(z > t) dt,$$

si escribimos $z = e^{2(m-1)x^2}$, tenemos que

$$P(z > t) = P\left(e^{2(m-1)x^2} > t \right).$$

Tomando logaritmos,

$$\begin{aligned} P(\ln(z) > \ln(t)) &= P(2(m-1)x^2 > \ln(t)) \to P(\ln(z) > \ln(t)) = P\left(x^2 > \frac{\ln(t)}{2(m-1)} \right) \\ &\Rightarrow P\left(|x| > \sqrt{\frac{\ln(t)}{2(m-1)}} \right). \end{aligned} \tag{6.22}$$

Empleando la desigualdad de Hoeffding, tenemos que $P_S[\Delta(h) \geq \epsilon] \leq e^{-2m\epsilon^2}$, por lo que, en este caso en particular,

$$P\left(|x| > \sqrt{\frac{\ln(t)}{2(m-1)}} \right) \leq \exp\left(\frac{-2m\ln(t)}{2(m-1)} \right). \tag{6.23}$$

Por la simetría de la función valor absoluto, podemos escribir

$$P\left(x > \sqrt{\frac{ln(t)}{2(m-1)}}\right) \leq 2\exp\left(\frac{-2m\ln(t)}{2(m-1)}\right) = 2\exp\left(\frac{-m\ln(t)}{(m-1)}\right)$$
$$= 2\exp\left(\ln(t)^{\frac{-m}{(m-1)}}\right) = 2\left(\frac{1}{t}\right)^{\frac{m}{m-1}}. \tag{6.24}$$

Por tanto,

$$E_S[z] = E_S[e^{2(m-1)\Delta(h)^2}] = \int_0^\infty 2\left(\frac{1}{t}\right)^{\frac{m}{m-1}} dt = 2\left[1 + \int_1^\infty \left(\frac{1}{t}\right)^{\frac{m}{m-1}} dt\right]$$
$$= 2\left[1 + m - 1\right] = 2m. \tag{6.25}$$

Combinado este resultado con la desigualdad dada por (6.21), tenemos que la ecuación (6.17) puede escribirse de la siguiente forma:

$$P_S[f(S) \geq \epsilon] \leq \frac{2m}{e^\epsilon}. \tag{6.26}$$

Denotando el lado derecho por δ, de manera que $\epsilon = \ln(2m/\delta)$, tenemos que

$$P_S[f(S) \leq \epsilon] \geq 1 - \delta.$$

Luego, con una probabilidad al menos $1-\delta$ para todo $\mathcal{Q}$ se tiene que

$$\sup_{\mathcal{Q}}\left(2(m-1)\underset{h\sim\mathcal{Q}}{E}(\Delta(h))^2 - D(\mathcal{Q}||\mathcal{P})\right) \leq \epsilon \Rightarrow \left(2(m-1)\underset{h\sim\mathcal{Q}}{E}(\Delta(h))^2 - D(\mathcal{Q}||\mathcal{P})\right) \leq \epsilon.$$

Reordenando y usando de nuevo la desigualdad de Jensen C.3.1, válida por ser la función x^2 convexa, tenemos que

$$\left(\underset{h\sim\mathcal{Q}}{E}\Delta(h)\right)^2 \leq \underset{h\sim\mathcal{Q}}{E}(\Delta(h))^2 \leq \frac{\ln(2m/\delta) + D(\mathcal{Q}||\mathcal{P})}{2(m-1)}. \tag{6.27}$$

Por consiguiente, recordando la definición de $\Delta(h) = \mathcal{R}_\mathcal{D}(h) - \mathcal{R}_S(h)$ y teniendo en cuenta la linealidad de la esperanza, tenemos que

$$\left(\underset{h\sim\mathcal{Q}}{E}[\mathcal{R}_\mathcal{D}(h) - \mathcal{R}_S(h)]\right)^2 = \left(\underset{h\sim\mathcal{Q}}{E}[\mathcal{R}_\mathcal{D}(h)] - \underset{h\sim\mathcal{Q}}{E}[\mathcal{R}_S(h)]\right)^2 \leq \frac{\ln\frac{2m}{\delta} + D(\mathcal{Q}||\mathcal{P})}{2(m-1)}. \tag{6.28}$$

Usando las definiciones del riesgo empírico promedio (6.15) y riesgo real promedio (6.16),

$$\mathcal{R}_\mathcal{D}(\mathcal{Q}) \leq \mathcal{R}_S(\mathcal{Q}) + \sqrt{\frac{D(\mathcal{Q}||\mathcal{P}) + \ln(2m/\delta)}{2(m-1)}}. \tag{6.29}$$

□

El resultado obtenido es válido para cualquier distribución a posteriori $\mathcal{Q}$. La idea más importante de este resultado es que, incluso si tenemos que ponderar respecto a una distribución $\mathcal{Q}$ mal elegida, el exceso de riesgo está controlado por $\sqrt{\ln m}/m$. Además, vemos que la cota de error depende del comportamiento de la divergencia de Kullback-Leibler,

cuya expresión se ha derivado en la sección 2.5 para el caso de que ambas distribuciones $\mathcal{P}$ y $\mathcal{Q}$ sean Gaussianas:

$$D(\mathcal{Q}||\mathcal{P}) = \frac{1}{2}\left[\ln\frac{|\Sigma_p|}{|\Sigma_q|} - m + (\mu_q - \mu_p)^T\Sigma_p^{-1}(\mu_q - \mu_p) + \mathrm{tr}\left(\Sigma_p^{-1}\Sigma_q\right)\right]. \tag{6.30}$$

Obtenemos entonces

$$\mathcal{R}_{\mathcal{D}}(\mathcal{Q}) \leq \mathcal{R}_S(\mathcal{Q}) + \sqrt{\frac{\frac{1}{2}\left[\ln\frac{|\Sigma_p|}{|\Sigma_q|} - m + (\mu_q - \mu_p)^T\Sigma_p^{-1}(\mu_q - \mu_p) + \mathrm{tr}\left(\Sigma_p^{-1}\Sigma_q\right)\right] + \ln\frac{2m}{\delta}}{2(m-1)}} \tag{6.31}$$

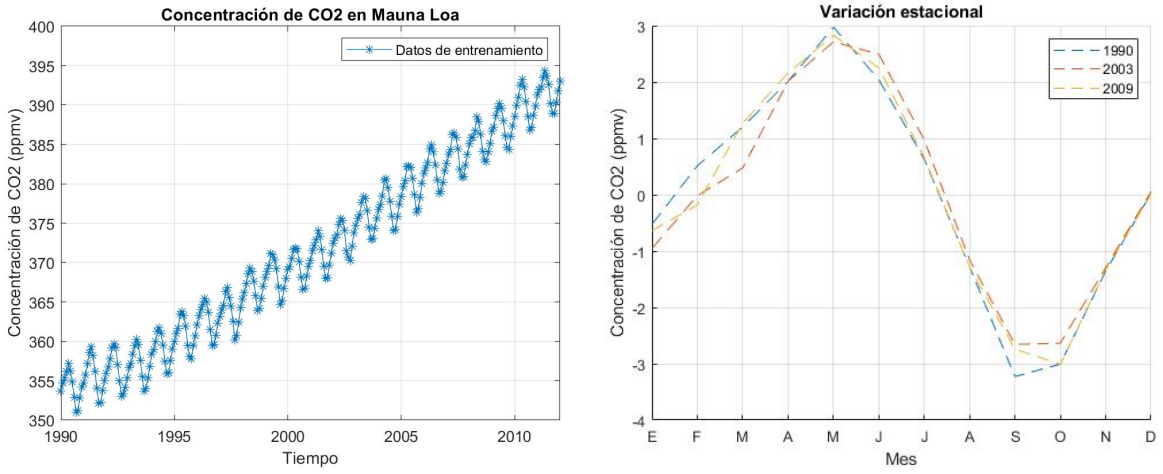
CAPÍTULO 7

SIMULACIONES.

En este capítulo se presentan las simulaciones realizadas para ilustrar los modelos de regresión basados en procesos Gaussianos. En las secciones 7.1 y 7.2 se muestran ejemplos en 2 y 3 dimensiones respectivamente, con un número pequeño de datos. Esto nos permite representar gráficamente la solución y visualizar el modelo de regresión. En la sección 7.3 se ha realizado una simulación para datos relacionados con el movimiento de un brazo robótico antropomórfico con siete grados de libertad. El objetivo es encontrar una aplicación que relacione el espacio de entrada de 21 dimensiones (7 posiciones articulares, 7 velocidades y 7 aceleraciones) con los correspondientes 7 movimientos de torsión. [1]

7.1. Ejemplo en 2D.

En esta sección realizamos una simulación sobre la concentración de dióxido de carbono en la atmósfera en Mauna Loa, un volcán de Hawai, para ilustrar el funcionamiento de los modelos de regresión basados en procesos Gaussianos.



(a) Concentración de CO2 en Mauna Loa. (b) Variación estacional de la concentración.

Figura 7.1.1: Representación de los datos sobre la concentración de CO_2 en Mauna Loa.

[1] Los datos han sido tomados de Sethu Vijayakumar y están disponibles en http://www.gaussianprocess.org/gpml/data/.

Se muestra cómo la verosimilitud marginal puede emplearse para determinar el valor de los hiperparámetros de la correspondiente función de covarianza. En concreto, se ha empleado un método de descenso de gradiente para la optimización de dichos hiperparámetros. Se ha seguido el esquema planteado en [34, Sección 5.4.3]. [2] En la Figura 7.1.1a se ha representado la poligonal que une los datos correspondientes a la concentración de dióxido de carbono desde 1990 hasta 2012. Por otro lado, en la Figura 7.1.1b se han centrado los datos correspondientes a las concentraciones en tres años diferentes para observar la variación estacional de esta concentración. Vemos que se observa un máximo en torno al mes de mayo y un mínimo en torno a los meses de septiembre y octubre.

El objetivo es modelar la concentración de CO_2 en el aire, y, como función del tiempo, x. De acuerdo con la figura 7.1.1a, es fácil ver que hay una tendencia de crecimiento con el paso de los años, una variación pronunciada con las estaciones y pequeñas irregularidades. Estas observaciones nos pueden facilitar la elección de una función de covarianza que refleje todas estas características. Dicha función de covarianza se define como suma de varios términos que dan cuenta de cada una de las características de los datos.

- Para modelar una suave tendencia de crecimiento con el paso de los años empleamos una función de covarianza Gaussiana (SE), que es de las más frecuentemente utilizadas. Ésta posee dos términos, θ_1 y θ_2, que controlan respectivamente la amplitud y la longitud característica.

$$k_1(x, x') = \theta_1^2 \exp\left(-\frac{(x - x')^2}{2\theta_2^2}\right). \tag{7.1}$$

  Un núcleo Gaussiano con un parámetro de escala grande hace que esta componente sea suave. Con los valores probados, se observa que la media a posteriori presenta una tendencia creciente.

- Usamos una función de covarianza periódica para simular la variación anual de los datos. Podemos usar una función de covarianza periódica menos rígida mediante el producto de una función de covarianza exponencial al cuadrado y una periódica. Obtenemos la siguiente función núcleo:

$$k_2(x, x') = \theta_3^2 \exp\left(-\frac{(x - x')^2}{2\theta_4^2} - \frac{2\sin^2(\pi(x - x'))}{\theta_5^2}\right), \tag{7.2}$$

  donde θ_3 representa la magnitud, θ_4 controla el decaimiento de la componente periódica, y θ_5 contiene la información sobre la periodicidad. El periodo se fija a un año, de acuerdo con lo observado en la figura 7.1.1a.

- Para modelar pequeñas irregularidades empleamos un núcleo racional cuadrático:

$$k_3(x, x`) = \theta_6^2 \left(1 + \frac{(x - x')^2}{2\theta_8\theta_7^2}\right)^{-\theta_8}, \tag{7.3}$$

  donde θ_6 es la magnitud, θ_7 es la longitud característica y θ_8 es el parámetro de forma que modifica la longitud característica.

[2] Los datos están disponibles en https://serc.carleton.edu/introgeo/interactive/examples/co2.html.

- Por último, podemos incluir un término de ruido mediante la suma de una covarianza exponencial al cuadrado con una componente independiente, de la siguiente forma:

$$k_4(x_p, x_q) = \theta_9^2 \exp\left(-\frac{(x_p - x_q)^2}{2\theta_{10}^2}\right) + \theta_{11}^2 \delta_{pq}, \tag{7.4}$$

  donde θ_9 es la magnitud de la componente de ruido correlada, θ_{10} es la escala de longitud y θ_{11} es la magnitud de la componente de ruido independiente.

La función de covarianza final se obtiene como suma de las anteriores:

$$k(x, x') = k_1(x, x') + k_2(x, x') + k_3(x, x') + k_4(x, x'). \tag{7.5}$$

Para obtener los hiperparámetros que optimizan la función de verosimilitud marginal debemos primero centrar los datos. Con este modelo no solo podemos determinar la función que mejor representa los datos, sino que además podemos predecir, con un cierto intervalo de confianza, el comportamiento futuro de la concentración de CO_2. A la vista de la Figura 7.1.2 vemos que el modelo proporciona predicciones con una alta credibilidad hasta 2020-2030. En la Tabla 7.1 se muestran los valores de los hiperparámetros obtenidos.

Hiperparámetro	Valor (log)
θ_1	-0.1884
θ_2	-1.1714
θ_3	4.5677
θ_4	0.2907
θ_5	0.5363
θ_6	0.000273
θ_7	1.2907
θ_8	4.4014
θ_9	4.46941
θ_{10}	3.0225
θ_{11}	-0.1884
θ_{12}	-1.1714
θ_{13}	-3.0019

Tabla 7.1: Valores obtenidos de los hiperparámetros.

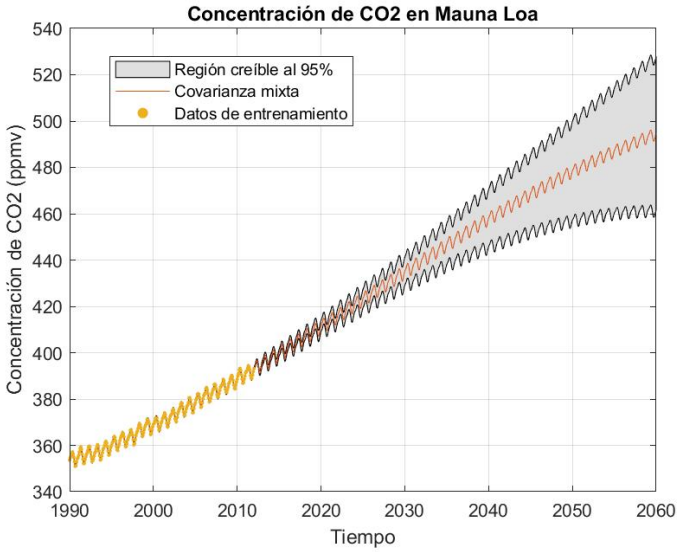


Figura 7.1.2: Predicción de la concentración de CO_2 en Mauna Loa.

La elección de la función de covarianza dada en (7.5) está basada en la observación de los datos de entrenamiento y un método de prueba y error. Es posible que existan muchas otras funciones de covarianza que proporcionen buenas predicciones. Esto se debe a que, como se ha comentado en capítulos anteriores, los métodos de regresión basados en procesos Gaussianos son muy flexibles y admiten muchas soluciones para un determinado problema.

Tal y como se ha comentado en capítulos anteriores, la elección de la función de covarianza es un punto muy importante para los modelos de ajuste y predicción basados en procesos Gaussianos. En la Figura 7.1.3 se observan otras posibles funciones de covarianza. Algunas de ellas dan lugar a predicciones poco creíbles. Vemos que la función de covarianza SE por sí sola modela la tendencia creciente pero no es capaz de simular las oscilaciones

asociadas con las estaciones. Por otro lado, la función de covarianza periódica no es capaz de detectar la tendencia creciente e interpreta las pequeñas oscilaciones como ruido en las observaciones.

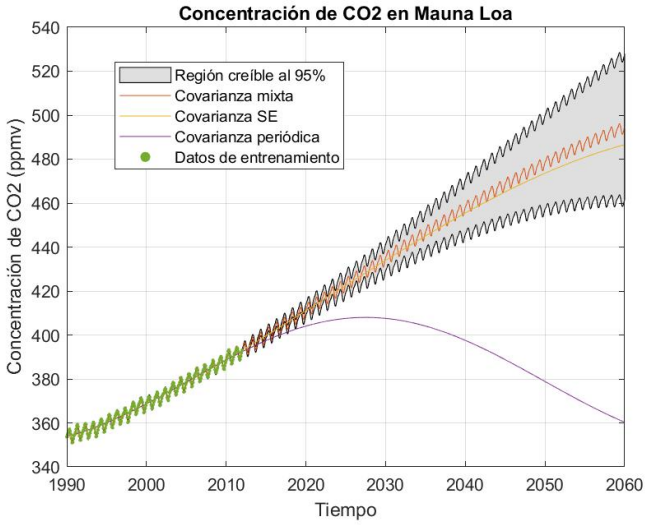


Figura 7.1.3: Predicción de la concentración de CO_2 usando diferentes funciones de covarianza.

Por otro lado, es importante escoger los hiperparámetros de forma adecuada, de manera que el problema de optimización no dé con mínimos/máximos locales. En la Figura 7.1.4 se muestra el resultado asociado a una mala elección de éstos.

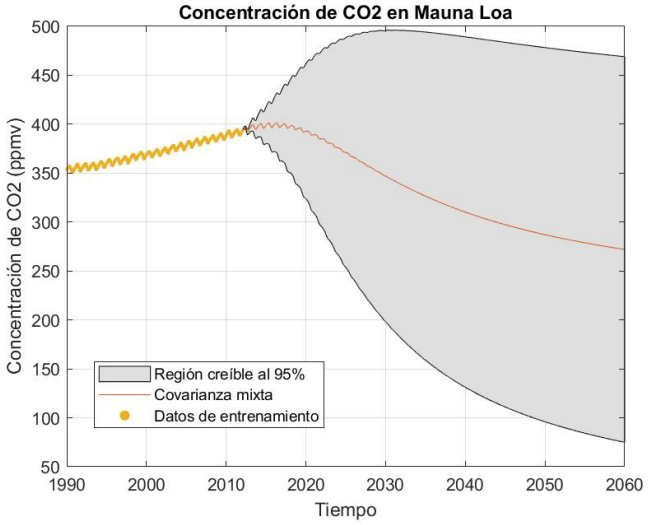


Figura 7.1.4: Predicción de la concentración de CO_2 en Mauna Loa partiendo de valores iniciales de los hiperparámetros poco adecuados .

A la vista de los resultados obtenidos, podemos comprobar la complejidad y flexibilidad de los métodos de regresión basados en procesos Gaussianos. Vemos además que es importante tener en cuenta el comportamiento a priori de los datos de entrenamiento para poder realizar buenas predicciones.

7.2. Ejemplo en 3D.

A continuación se muestra un ejemplo en tres dimensiones que permite visualizar el proceso de regresión basado en métodos Gaussianos. Se han empleado datos artificiales con los que poder modelar de forma sencilla y visual un problema de regresión. En concreto, se han empleado como datos de entrenamiento puntos correspondientes a la superficie dada por $y = \sqrt{x_1^2 + x_2^2}$, en la que se han introducido perturbaciones i.i.d normales. La representación de dichos datos de entrenamiento se ha representado en la Figura 7.2.1.

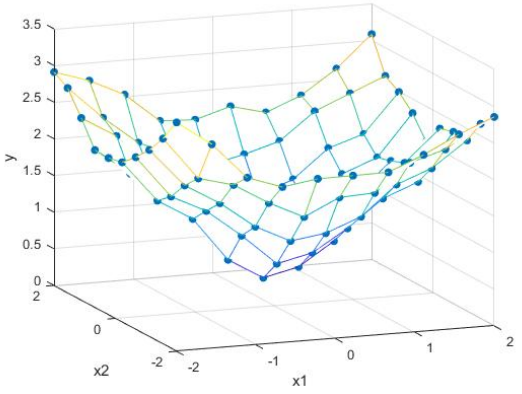


Figura 7.2.1: Representación 3D de los datos de entrenamiento.

Para el proceso de regresión se ha empleado un núcleo Gaussiano y una media cero. La matriz de datos de entrenamiento está formada por los pares cruzados correspondientes a las coordenadas de x_1 y x_2. Tras inicializar los hiperparámetros y optimizar el valor de éstos, se ha ejecutado el programa de regresión. En la Figura 7.2.2 se han representado tanto los datos iniciales como la predicción obtenida mediante la regresión basada en procesos Gaussianos.

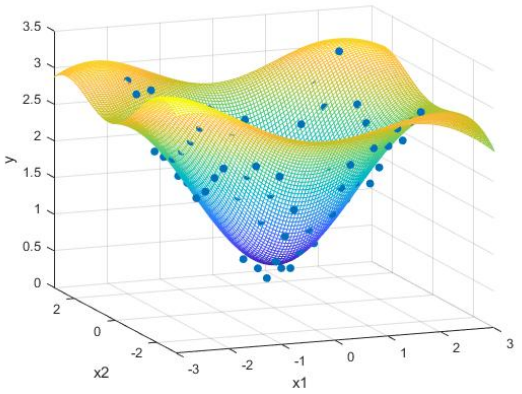


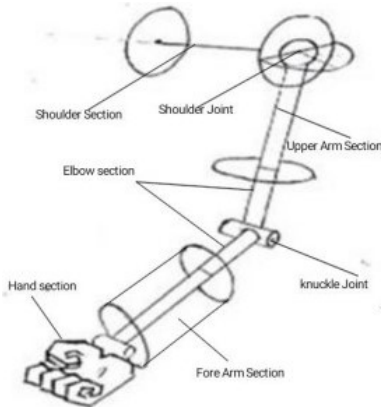
Figura 7.2.2: Representación 3D de la superficie predicha por el modelo de regresión empleado junto con los datos de entrenamiento.

En la zona central, donde disponemos de una mayor cantidad de datos de entrenamiento, es posible recuperar la función original. A medida que nos alejamos de esta zona, las predicciones son peores debido al problema de extrapolación.

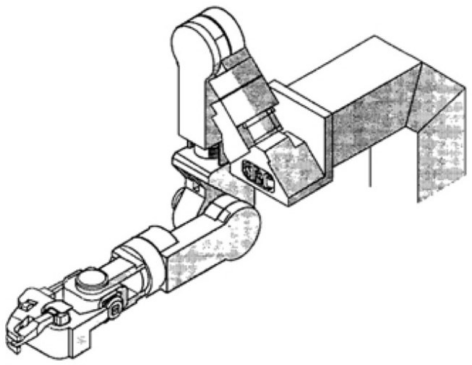
7.3. Modelo de un brazo robótico antropomórfico.

Una vez introducidos ejemplos sencillos, en esta sección se pretende mostrar un problema de regresión real. En particular, se va a estudiar un brazo robótico. Los robots humanoides son sistemas de elevada dimensión de movimientos para los cuales las técnicas habituales de control son insuficientes debido a irregularidades en dichos sistemas. Con el objetivo de solventar estos problemas, se emplean modelos de aprendizaje con vistas a desarrollar robots completamente autónomos. Entre las principales características de los problemas de aprendizaje de robots destacamos la elevada dimensión del conjunto de entrenamiento, con datos posiblemente redundantes, distribuciones no estacionarias y la necesidad de un aprendizaje continuo.

Se han empleado datos relacionados con el movimiento de un brazo robótico antropomórfico con siete grados de libertad, como el que se muestra en la Figura 7.3.1:



(a) Sección de un brazo robótico con 7 grados de libertad (DOF) [43].



(b) Brazo robótico SARCOS [51].

Figura 7.3.1: Brazo robótico con 7 grados de libertad.

Los grados de libertad del brazo se corresponden con tres articulaciones:

- Tres grados de libertad de la articulación del hombro, que permiten la abducción-aducción, flexión-extensión y la rotación interna-externa.
- Un grado de libertad del codo, que posibilita la extensión del brazo.
- Tres grados de libertad de la muñeca, que posibilitan la extensión-flexión, supinación-pronación y la desviación del cubital y del radial.

Para cada uno de los grados de libertad del brazo se dispone de valores de la velocidad y aceleración correspondientes. De esta forma, se tienen 21 variables de entrada. Por otro lado, se han calculado los momentos de torsión asociados a cada posición, obteniendo así 7 etiquetas o variables respuesta. Disponemos de 48.933 datos de entrada con sus respectivas etiquetas, de los cuales 44.484 han sido usados como conjunto de entrenamiento y los 4.449 restantes como conjunto de prueba En este caso, nos centraremos en los datos relacionados con la articulación del hombro, correspondientes a la primera columna de etiquetas. Puesto que el conjunto de datos es de elevada dimensión, será necesario emplear aproximaciones para el modelo de regresión, como la planteada en la sección 4.4. Aunque a primera vista se podría pensar que no es necesario un modelo de aprendizaje para este problema, ya que

existen modelos físicos que permiten calcular estos momentos de torsión de forma teórica, hay problemas técnicos con su implementación, por lo que emplear modelos de aprendizaje resulta conveniente.

Para la realización de las predicciones, se han seleccionado subconjuntos de datos de dimensión m de forma aleatoria y sin reemplazamiento. Posteriormente, se han ajustado los hiperparámetros optimizando la verosimilitud marginal mediante un método de descenso de gradiente. Con el objetivo de medir la calidad de las predicciones se ha calculado el error cuadrático medio estandarizado. Para ello, primero se han centrado y estandarizado los datos, de modo que estos tengan media cero y varianza igual a la unidad. En la Figura 7.3.2 se muestra el error cuadrático medio estandarizado obtenido para los diferentes tamaños de las muestras. Dicho error se ha calculado de la siguiente forma:

$$\text{SMSE} = \frac{1}{m}\sum_{i=1}^{m}(y_{s_i} - \bar{y_{s_i}})^2, \tag{7.6}$$

donde m es el tamaño de la muestra, y_{s_i} son las etiquetas asociadas a los datos de prueba e $\bar{y_{s_i}}$ son las etiquetas predichas. Los valores de las etiquetas han sido previamente centrados y estandarizados para poder obtener el error cuadrático medio estandarizado.

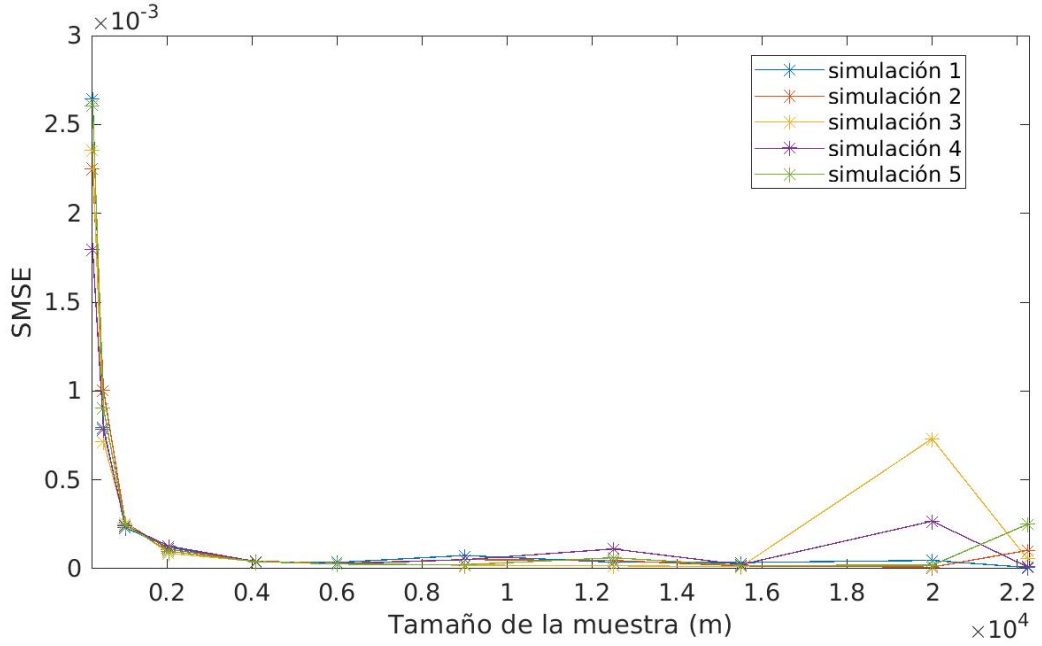


Figura 7.3.2: Representación del error cuadrático medio estandarizado en función del tamaño de la muestra para varias simulaciones.

A la vista de la Figura 7.3.2, observamos que no es necesario considerar muestras de tamaño elevado, ya que no hay mejoras significativas en las predicciones y el coste computacional y temporal es mayor (el programa se ha ejecutado durante unas 6 horas). Podemos ver además que el codo del brazo se alcanza para muestras de unos 1000 datos, mucho antes de considerar si quiera la mitad de los datos de entrenamiento. Esto puede indicar que hay una gran cantidad de datos redundantes entre el conjunto de entrenamiento. Por otro lado, los picos que aparecen pueden deberse a la aleatoriedad del procedimiento. En la

Figura 7.3.3 se ha representado el valor medio de SMSE para cada valor de m junto con la desviación asociada. La Tabla 7.2 recoge los valores obtenidos.

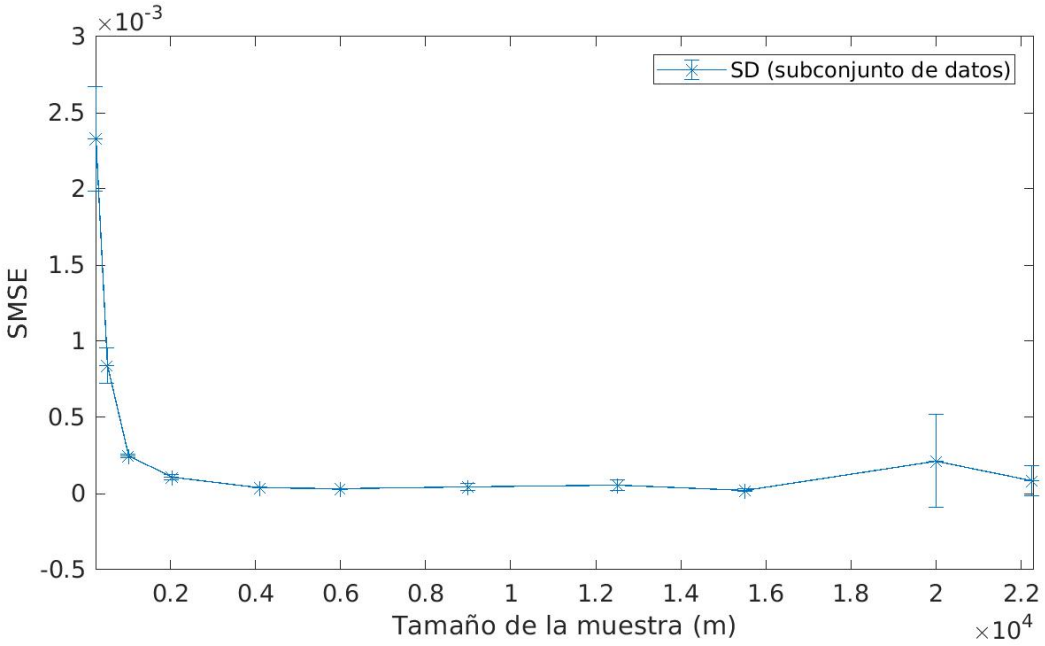


Figura 7.3.3: Representación del error cuadrático medio estandarizado en función del tamaño de la muestra.

m	SMSE
256	0.00230 ± 0.00034
512	0.00084 ± 0.00011
1024	0.00025 ± 0.00001
2048	0.00011 ± 0.00002
4096	0.00004 ± 0.00000
6000	0.00003 ± 0.00000
9000	0.00004 ± 0.00002
12500	0.00005 ± 0.00004
15500	0.00002 ± 0.00000
20000	0.00021 ± 0.00031
22242	0.00008 ± 0.00010

Tabla 7.2: Valores del SMSE obtenidos junto con la desviación asociada para cada valor de m.

A la vista de los resultados obtenidos, se ha repetido el procedimiento diez veces para cada valor de m, $m = \{256, 512, 1024, 2048, 4096\}$. Se ha tomado como valor del error cuadrático medio estandarizado el valor medio asociado a cada subconjunto y se ha calculado la desviación asociada. Los datos aparecen recogidos en la Tabla 7.3, y su representación se muetsra en la Figura 7.3.4.

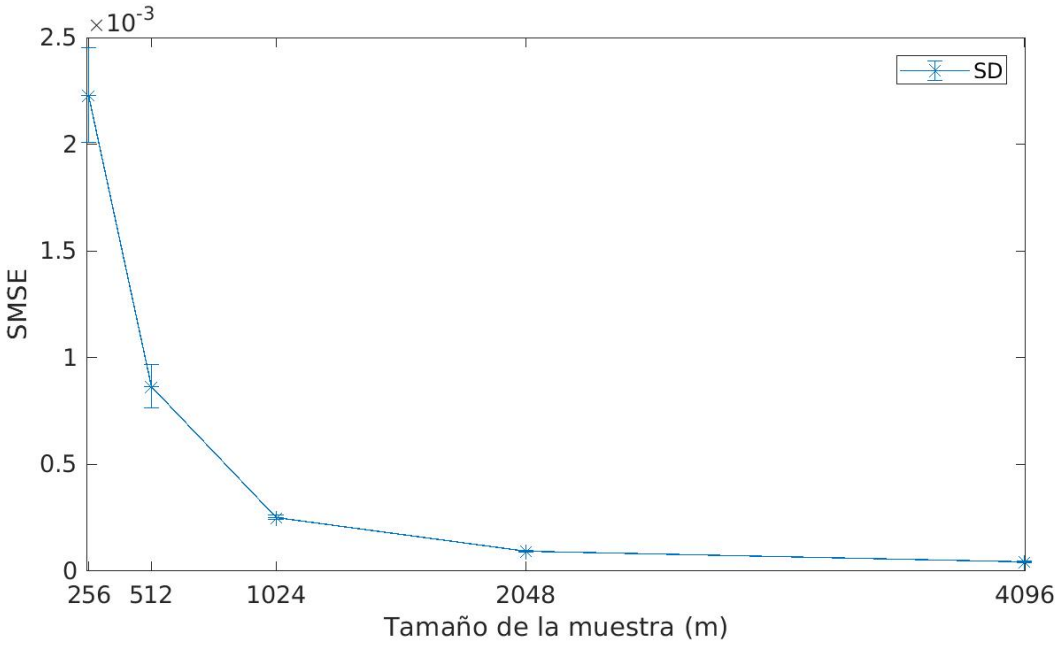


Figura 7.3.4: Representación del error cuadrático medio estandarizado en función del tamaño de la muestra para muestras de menor tamaño.

m	SMSE	Coste de almacenamiento	Coste computacional	Media	Varianza
256	0.00240 ± 0.00039				
512	0.00079 ± 0.00005				
1024	0.00024 ± 0.00001	$\mathcal{O}(m^2)$	$\mathcal{O}(m^3)$	$\mathcal{O}(m)$	$\mathcal{O}(m^2)$
2048	0.00009 ± 0.00001				
4096	0.00005 ± 0.00002				

Tabla 7.3: Valores de SMSE obtenidos junto con la desviación asociada para cada valor de m. Se muestra también el orden del coste computacional, de almacenamiento y de cálculo de la media y de la varianza.

A la vista de los resultados obtenidos, parece razonable pensar que hay redundancia en los datos de entrenamiento, lo cual justifica que aun tomando muestras de pequeño tamaño, el error cometido sea pequeño.

CAPÍTULO 8

CONCLUSIONES.

En este trabajo hemos estudiado los métodos de regresión basados en procesos Gaussianos. Hemos visto cómo podemos entender éstos a partir del problema de regresión Ridge, sin más que considerar los coeficientes de dicho modelo como variables aleatorias. Además, a lo largo del trabajo hemos visto que los procesos Gaussianos constituyen un modelo que permite ajustar de forma no paramétrica modelos de regresión, así como manejar problemas de clasificación. Constituyen por tanto métodos muy flexibles que permiten resolver diferentes problemas. Se han introducido diferentes perspectivas de los métodos de regresión basados en procesos Gaussianos, siendo éstas el espacio de pesos y el espacio de funciones. Se han obtenido resultados equivalentes en ambos casos. Además, se han mostrado vías diferentes en relación con el problema de aprendizaje, siendo éstas la inferencia Bayesiana y el aprendizaje PAC. En ambos casos se pretende construir un modelo a partir de unos datos de entrenamiento. Mediante una combinación de las dos vías, es posible obtener modelos más flexibles que son capaces de incorporar la información a priori (inferencia Bayesiana) junto con ciertas garantías de exactitud del modelo (aprendizaje PAC) al aplicarse a nuevos datos. Por otro lado, se han proporcionado resultados teóricos sobre el funcionamiento de los procedimientos de regresión basados en procesos Gaussianos basándonos en el concepto de aprendizaje PAC.

Con vistas a la implementación de estos modelos, se han analizado tres problemas diferentes. Dos de ellos constituyen ejemplos sencillos que facilitan la visualización de los procedimientos. A través de éstos es posible ilustrar la gran flexibilidad de los métodos de regresión basados en procesos Gaussianos. Hemos visto que la elección de la función de covarianza, así como de los hiperparámetros de ésta, determina en gran medida la credibilidad de la predicción obtenida. Por último, se ha presentado un problema real sobre los movimientos de un brazo robótico con el objetivo de mostrar uno de los principales inconvenientes de estos modelos: el coste computacional asociado con muestras de entrenamiento de gran tamaño. Hemos visto que es necesario recurrir a técnicas de aproximación que permitan la implementación de los procesos, que de otra forma tendrían costes prohibitivos.

ANEXOS A

EL TEOREMA DE EXTENSIÓN DE KOLMOGOROV.

Teorema A.0.1 (Teorema de extensión de Kolmogorov)**.** *[10, Teorema 36.1] Sea T un conjunto no vacío y supongamos que, para cada subconjunto finito no vacío V de T, P_V es una probabilidad en $\mathbb{R}^n$ si V tiene n elementos. Supongamos que estas probabilidades satisfacen la siguiente condición de consistencia:*

Para cada subconjunto $U = \{t_{i_1}, ..., t_{i_r}\}$ no vacío de V, la distribución de probabilidad de $pr_{(V,U)}$ respecto a P_V es P_U, donde, $pr_{(V,U)}$ es la aplicación $(x_{t_1}, ..., x_{t_n}) \in \mathbb{R}^n \to (x_{t_{i_1}}, ..., x_{t_{i_r}}) \in \mathbb{R}^r$ y pr_V es la aplicación $x \in \mathbb{R}^T \to (x_{t_1}, ..., x_{t_n}) \in \mathbb{R}^n$.

Entonces existe una única probabilidad P en $\mathcal{R}^T$ tal que, para cada subconjunto finito V de T, la distribución de pr_V respecto a P coincide con P_V, es decir, tal que para cada $n \in \mathbb{N}$, cada sucesión finita creciente $t_1 < ... < t_n$ en T y cada $H \in \mathcal{R}^n$ se verifica que $P(x \in \mathbb{R}^T : (x_{t_1}, ..., x_{t_n}) \in H) = P_{(t_1,...,t_n)}(H)$.

Consideremos ahora las aplicaciones de proyección $\pi_t : x \in \mathbb{R}^T \to x_t \in \mathbb{R}$. Si $(P_V)_{V_{\text{finito}} \subset T}$ es una familia de probabilidades que satisface las hipótesis del teorema anterior y si P es la probabilidad en $\mathcal{R}^T$ que proporciona dicho teorema, entonces para cada $n \in \mathbb{N}$, cada sucesión finita creciente $t_1 < ... < t_n$ en T y cada $H \in \mathcal{R}^n$ se verifica que

$$P[(\pi_{t_1}, ..., \pi_{t_n}) \in H] = P_{(t_1,...,t_n)}(H).$$

Por tanto, $(\mathbb{R}^T, \mathcal{R}^T, P, (\pi_t)_{t \in T})$ es un proceso estocástico cuyas distribuciones finito dimensionales son precisamente las P_V. Podemos entonces enunciar el siguiente teorema, que asegura la existencia de un proceso estocástico con unas distribuciones finito dimensionales dadas de antemano, suponiendo que verifican una condición de consistencia.

Teorema A.0.2 (Teorema de extensión de Kolmogorov, versión 2)**.** *[10, Teorema 36.2] Si $(P_V)_{V\,finito \subset T}$ es una familia de probabilidades que satisfacen la condición de consistencia del teorema anterior, entonces existe un proceso estocástico $(\Omega, F, P, (X_t)_{t \in T})$ cuyas distribuciones finito dimensionales son precisamente las P_V.*

ANEXOS B

IDENTIDADES MATRICIALES.

B.1. Inversión de matrices.

Lema B.1.1 (Fórmula de Woodbury)**.** *[31] Sean A,U,C y V matrices, todas ellas invertibles, con* $A \in \mathcal{M}^{n\times n}$, $U \in \mathcal{M}^{n\times k}$, $C \in \mathcal{M}^{k\times k}$ *y* $V \in \mathcal{M}^{k\times n}$. *Entonces, se tiene la siguiente identidad*

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1},$$

asumiendo que las inversas implicadas existen

Lema B.1.2. *Sean A y C dos matrices invertibles de tamaño* $n \times n$. *Entonces se verifica la siguiente igualdad:*

$$(A+C)^{-1} = A^{-1} - A^{-1}(C^{-1}+A^{-1})^{-1}A^{-1}. \tag{B.1}$$

Existe una expresión similar para los determinantes [34, Anexo A.3].

Lema B.1.3. *Sean A,U,C y V matrices, todas ellas invertibles, con* $A \in \mathcal{M}^{n\times n}$, $U \in \mathcal{M}^{n\times k}$, $C \in \mathcal{M}^{k\times k}$ *y* $V \in \mathcal{M}^{k\times n}$. *Entonces, se tiene la siguiente identidad*

$$|A+UCV| = |A||C||C^{-1}+VA^{-1}U|. \tag{B.2}$$

B.2. Descomposición de Cholesky.

La descomposición de Cholesky [34, Anexo A.4] de una matriz simétrica, semidefinida positiva A, permite escribir ésta como producto de dos matrices, L y L^T, siendo L una matriz triangular inferior que recibe el nombre de factor de Cholesky.

$$LL^T = A.$$

La descomposición de Cholesky es útil para resolver sistemas lineales con matriz de coeficientes A simétrica y semidefinida positiva. Para resolver el sistema $A\mathbf{x} = \mathbf{b}$ primero resolvemos $L\mathbf{y} = \mathbf{b}$ y posteriormente resolvemos el sistema triangular $L^T\mathbf{x} = \mathbf{y}$. Podemos entonces escribir la solución como $\mathbf{x} = L^T/(L/\mathbf{b})$, donde la notación $A/\mathbf{b}$ representa el vector $\mathbf{x}$ solución de $A\mathbf{x} = \mathbf{b}$. La resolución de los dos sistemas descritos requiere un total de $n^2/2$ operaciones, cuando A es de tamaño $n \times n$.

ANEXOS C

DESIGUALDADES BÁSICAS EN PROBABILIDAD.

Sean $X_1, ..., X_n$ una secuencia de variables aleatorias i.i.d. y sea μ su media. De acuerdo con la ley de los grandes números, si n tiende a infinito la media empírica, $\frac{1}{n}\sum_{i=1}^{n} X_i$tiende hacia su valor esperado, μ, con probabilidad 1. Las desigualdades de concentración de la medida cuantifican la desviación de la media empírica con respecto del valor esperado cuando n tiende a infinito.

C.1. Desigualdad de Markov.

Sea X una variable aleatoria no negativa. El valor medio de X puede escribirse como [41, Apéndice B.1]:

$$E[X] = \int_{x=0}^{\infty} P[X \geq x]dx. \tag{C.1}$$

Puesto que $P[X \geq x]$ es monótona no creciente, es posible escribir lo siguiente:

$$\forall a \geq 0, E[X] \geq \int_{x=0}^{a} P[X \geq x]dx \geq \int_{x=0}^{a} P[X \geq a]dx = P[X \geq a]. \tag{C.2}$$

Reordenando la desigualdad, obtenemos la desigualdad de Markov:

$$\forall a \geq 0, P[X \geq a] \leq \frac{E[X]}{a}. \tag{C.3}$$

C.2. Desigualdad de Hoeffding.

Teorema C.2.1. *[41, Lema B.6] Sea $X_1, X_2, ..., X_n$ una sucesión de variables aleatorias i.i.d. tales y sea $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Supongamos que que $E[\overline{X_i}] = \mu$ y $P[a \leq X_i \leq b] = 1$ para todo i. Entonces, para todo $\epsilon > 0$ se tiene que*

$$P\left[|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| > \epsilon\right] \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

C.3. Desigualdad de Jensen.

Definición C.3.1 (Función convexa)**.** *[11, Definición 3.1.1] Una función $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ es convexa si para todo $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ y $t \in [0, 1]$, se tiene que*

$$f((1-t)\mathbf{x} + t\mathbf{y}) \leq (1-t)f(\mathbf{x}) + tf(\mathbf{y}). \tag{C.4}$$

Teorema C.3.1 (Desigualdad de Jensen)**.** *[11, Sección 3.1.8.] Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad, f una función real integrable con valores en un intervalo $I \subset \mathbb{R}$ y sea $\phi : I \longrightarrow \mathbb{R}$ una función convexa. Entonces*

$$\phi\left(\int_\Omega f dP\right) \leq \int_\Omega \phi \circ f dP. \tag{C.5}$$

Lema C.3.1. *[10, Ecuación 5.31] Sea $I \subset \mathbb{R}$ un intervalo, $f : I \longrightarrow \mathbb{R}$ una función convexa y sea X una variable aleatoria acotada. Entonces*

$$f(E[X]) \leq E[f(X)]. \tag{C.6}$$

C.4. Desigualdad de Boole.

En teoría de la probabilidad, la desigualdad de Boole [10, Ecuación 2.10] estipula que para toda familia finita o numerable de sucesos, la probabilidad de que al menos uno de esos sucesos ocurra es menor o igual a la suma de las probabilidades de los sucesos individuales. De manera más formal:

Teorema C.4.1. *Para una familia finita o numerable de sucesos $A_1, A_2, A_3, ...$ se cumple*

$$P\left(\underset{n}{\cup} A_n\right) \leq \sum_n P(A_n). \tag{C.7}$$

ANEXOS D

TEOREMA DE RADON-NIKODYM.

Definición D.0.1. *[10, Página 422.] Sea $(X, \mathcal{A}, \mu)$ un espacio de medida y sea ν una medida compleja definida en el mismo espacio medible. Diremos que ν es absolutamente continua respecto de μ y escribimos $\nu << \mu$ si se cumple*

$$\nu(M) = 0 \text{ para todo } M \in \mathcal{A} \text{ tal que } \mu(M) = 0. \tag{D.1}$$

Teorema D.0.1 (Teorema de Radon-Nikodym)**.** *[15, C.7.] Sea $(X, \mathcal{A}, \mu)$ un espacio de medida σ-finito, ν una medida compleja definida en el espacio medible $(X, \mathcal{A})$ y absolutamente continua respecto de μ. Existe una función integrable $f \in \mathcal{L}^1(\mu, \mathbb{C})$ que cumple:*

$$\nu(A) = \int_A f d\mu, \text{ para todo } A \text{ medible.} \tag{D.2}$$

Se dice que f es la derivada de Radon-Nikodym de ν respecto de μ y se denota por $\frac{d\nu}{d\mu}$.

ANEXOS E

PROGRAMAS DE MATLAB Y R.

E.1. Regresión Ridge.

El siguiente código se ha empleado para la construcción de las figuras 3.5.1 y 3.5.2. Se ha utilizado como referencia el código presentado en [9].

```
% Seleccionamos los datos
x1=(50-10).*rand(100,1) + 10;
x2=(50-10).*rand(100,1) + 10;
y=x1+x2 + 0.3.*rand(100,1);
x=[x1,x2];
n=length(x1);

% Construimos el núcleo Gaussiano y el lineal
K=zeros(n,n);
for j=1:n
 for i=1:n
    K(i,j)=exp(-norm(x(j,:)-x(i,:)));
 end
end

Klineal=(x*x');

% Algoritmo
% f_gaussian : Función de regresión para un núcleo Gaussiano
% f_lineal :  Función de regresión para un núcleo lineal
% mse : Mean Square Error
% A\b en lugar de inv(A)*b (más rápido)

f_gaussian=zeros(n,1);
f_lineal=zeros(n,1);
mse=[];
intvl=0.1;
for lambda=intvl:intvl:1

    for i=1:n
        f_lineal(i,1)= y'*((Klineal+ lambda*eye(n))\Klineal(i,:)');
        f_gaussian(i,1)= y'*((K+ lambda*eye(n))\K(i,:)');
    end

    mse=[mse; norm(f_gaussian-y)^2/n,norm(f_lineal-y)^2/n];
end

% Buscamos minimizar el error cuadrático medio
mse_=min(mse(:,2));
[mse_,linear_indx]=min(mse(:,2));
```

```
fprintf('Error cuadrático medio (Núcleo lineal) : %f',mse_)

[mse_,gaussain_indx]=min(mse(:,1));
fprintf('Mínimo error cuadrático medio (núcleo Gaussiano) : %f',mse_)

% Valor óptimo del parámetro de regularización
lambda_optimal_gaussian=intvl*gaussain_indx
lambda_optimal_linear=intvl*linear_indx

% Valores reales y predichos
Predicted=[y,f_gaussian,f_lineal]

% Valor del parámetro alpha y de los pesos w :
alpha_gaussian=inv((K+lambda_optimal_gaussian*eye(n)))*y
w_gaussian=alpha_gaussian.*x

alpha_linear=inv((K+lambda_optimal_linear*eye(n)))*y
w_linear=alpha_linear.*x

% Representación
figure(1)
hold on
grid on
scatter3(x1,x2,y,'g')
view([-21.1 15.4 ])
scatter3(x1,x2,f_lineal,'.r')
title('Kernel Ridge Regression')
xlabel('x1')
ylabel('x2')
zlabel('y')
legend('Datos de entrenamiento','Datos predichos')
hold off

figure(2)
hold on
grid on
scatter3(x1,x2,y,'g')
view([-21.1 15.4 ])
scatter3(x1,x2,f_gaussian,'.r')
title('Kernel Ridge Regression')
xlabel('x1')
ylabel('x2')
zlabel('y')
legend('Datos de entrenamiento','Datos predichos')
hold off
```

E.2. Regresión lineal basada en procesos Gaussianos.

El siguiente código se corresponde con las figuras 4.2.2 y 4.2.1.

```
% Generamos los puntos de prueba. En este caso, los generamos
% de forma equidistribuida entre -5 y 5.
xx=linspace(-5,5,50);
n=length(xx);
muxx=zeros(n,1);

% La función kernel kxx=k(Xx,Xx) es el núcleo de cuadrado
% exponencial (SE). En este ejemplo, fijamos
% los hiperparámetros, de modo que sigma_f=1, l=1, sigma_n=0.
for i=1:n
    for j=1:n
```

```
12          kxx(i,j)=exp(-1/2*(xx(i)-xx(j))^2);
13      end
14  end
15
16  %Introducimos una corrección para poder emplear la
17  %descomposición de Cholesky. (Los autovalores estimados
18  %deben ser mayores que cero).
19  epsilon=0.0001;
20  kxx=kxx+epsilon*eye(n);
21
22  %Generamos muestras independientes u~N(0,I) de modo que,
23  %mediante una transformación del tipo y=mu+L*u, tenemos
24  %una distribución Gaussiana multivariante de parámetros
25  %mu y sigma. En este caso, generamos 1000 muestras.
26  uxx=normrnd(0,1,[n,1000]);
27
28  %Descomposición de Cholesky
29  Lxx=chol(kxx);
30  %Valores de f, Gaussiana multivariante.
31  f_prior=muxx+Lxx*uxx;
32
33  %Representamos gráficamente algunas de las Gaussianas
34  %obtenidas anteriormente, así como el intervalo de
35  %confianza del 95% y la media puntual.
36  figure(1)
37  hold on
38
39  %Intervalo de confianza del 95%
40  xxconf = [xx xx(end:-1:1)] ;
41  yyconf = [muxx+2 ; muxx(end:-1:1)-2]';
42
43  p = fill(xxconf,yyconf,'red');
44  p.FaceColor = [1  0.8  0.8];
45  p.EdgeColor = 'none';
46
47  %Rerpesentamos la media de la muestra
48  plot(xx,muxx,'k', 'LineWidth',1)
49
50  for i=1:5
51      hold on
52      plot(xx,f_prior(:,i))
53  end
54
55  xlabel('x')
56  ylabel('f(x)')
57  legend('Intervalo de confianza del 95%','Datos de entrenamiento↩
        ','Media','Funciones a posteriori')
58  title('Proceso Gaussiano a priori')
59
```

```
% Generamos ahora la distribución a posteriori a partir de los
% datos de entrenamiento. Asumimos inicialmente que los datos
% están libres de ruido. De esta forma, las Gaussianas que
% consideremos pasan por los datos de entrenamiento.
% Una posible opción sería generar infinitas Gaussianas y
% rechazar aquellas que no pasen por estos  puntos.
% No obstante, eso tendría un coste computacional muy alto.
% Por tanto, lo que hacemos es usar la probabilidad
% condicionada para generar una nueva distribución Gaussiana
% multivariante, en la que tanto la media como la matriz
% de covarianzas se obtienen considerando los datos de
% entrenamiento, (fx|X,Xx,f). En caso de tener ruido, tendremos
% que calcular los valores y = f(x) + e, donde e es el ruido,
% que asumimos que tiene una distribución Gaussiana de  media
% cero y varianza sigma^2.

% Conjunto de datos: Generamos puntos aleatorios.
x=[-4,-3,-1,0,2]';
f=x.*sin(x);
m=length(x);

% Datos con ruido
noise=0.1;
y=f+noise*rand(m,1);

% Calculamos kx=k(Xx,X) y k=k(X,X)
for i=1:n
    for j=1:m
        kx(i,j)=exp(-1/2*(xx(i)-x(j))^2);
    end
end

for i=1:m
    for j=1:m
        k(i,j)=exp(-1/2*(x(i)-x(j))^2);
    end
end
k_noise=k+noise^2*eye(m);

% Calculamos la media y la matriz de covarianza de la
% probabilidad condicionada fx|X,Xx,f
mu_post=kx*inv(k)*f;
sigma_post=kxx-kx*inv(k)*kx';

mu_post_noise=kx*inv(k_noise)*y;
sigma_post_noise=kxx-kx*inv(k_noise)*kx';

% Repetimos el procedimiento anterior para obtener los valores
% de fx.
```

```
109 Lx=chol(sigma_post);
110 ux=normrnd(0,1,[n,1000]);
111 fx=mu_post+Lx*ux;
112
113 Lx_noise=chol(sigma_post_noise);
114 ux=normrnd(0,1,[n,1000]);
115 fx_noise=mu_post_noise+Lx_noise*ux;
116
117 figure(2)
118 hold on
119
120 %En este caso, para calcular el intervalo de confianza tenemos
121 %que sumar y restar 2 veces la desviación estándar de cada
122 %punto a la media puntual. Para ello, construimos un vector
123 %cuyas componentes son la raíz cuadrada de los elementos
124 %diagonales de la matriz de covarianzas, que se corresponden
125 %con la desviación estándar de cada punto xx.
126 B=sqrt(diag(sigma_post'));
127 xconf = [xx xx(end:-1:1)] ;
128 yconf = [mu_post+2*B ; mu_post(end:-1:1)-2*B(end:-1:1)]';
129
130 B_noise=sqrt(diag(sigma_post_noise'));
131 xconf_noise = [xx xx(end:-1:1)] ;
132 yconf_noise = [mu_post_noise+2*B_noise ; mu_post_noise(end↩
       :-1:1)-2*B_noise(end:-1:1)]';
133
134 %Sin ruido
135 p = fill(xconf,yconf,'red');
136 p.FaceColor = [1 0.8 0.8];
137 p.EdgeColor = 'none';
138
139 %Representamos los datos de entrenamiento.
140 plot(x,f,'.', 'MarkerSize',15)
141 %Representamos la media.
142 plot(xx,mu_post,'k', 'LineWidth',1)
143
144 for i=1:5
145     plot(xx,fx(:,i))
146 end
147
148 xlabel('x')
149 ylabel('f(x)')
150 legend('Intervalo de confianza del 95%','Datos de entrenamiento↩
       ','Media','Funciones a posteriori')
151 title('Proceso Gaussiano a posteriori')
152
153 %Con ruido
154 figure(3)
155 hold on
```

```
156
157 p_noise = fill(xconf_noise,yconf_noise,'red');
158 p_noise.FaceColor = [1  0.8  0.8];
159 p_noise.EdgeColor = 'none';
160
161 %Representamos los datos de entrenamiento.
162 plot(x,y,'.', 'MarkerSize',15)
163 %Representamos la media.
164 plot(xx,mu_post_noise,'k', 'LineWidth',1)
165
166 for i=1:5
167     plot(xx,fx_noise(:,i))
168 end
169
170 xlabel('x')
171 ylabel('f(x)')
172 legend('Intervalo de confianza del 95%','Datos de entrenamiento↩
    ','Media','Funciones a posteriori')
173 title('Proceso Gaussiano a posteriori')
```

E.3. Funciones de covarianza.

En esta sección se muestran los códigos empleados para generar las trayectorias correspondientes a las diferentes funciones de covarianza presentadas en la sección 3.4.3.

E.3.1. Función de covarianza de Matern.

```
1  library(MASS)
2  library(fields)
3
4  d<- seq( 0,3,,200)
5  y<- Matern( d, range=1.5, smoothness=1.0)
6
7  plot( d,y, type=l)
8  # Representamos diferentes funciones de covarianza de
9  # Matern empleando diferentes  parámetros.
10
11 r1<- Matern.cor.to.range( 10, nu=.5, cor.target=.1)
12 r2<- Matern.cor.to.range( 10, nu=1.0, cor.target=.1)
13 r3<- Matern.cor.to.range( 10, nu=2.0, cor.target=.1)
14
15 d<- seq( 0, 15,,200)
16 y<- cbind(  Matern( d, range=r1, nu=.5),
17             Matern( d, range=r2, nu=1.0),
18             Matern( d, range=r3, nu=2.0))
19
```

```
matplot( d, y, type=l, lty=1, lwd=2, xlab=Distancia entre datos,
    r, ylab = Función de covarianza,
)
legend( x = topright, legend=c(nu=0.5, nu=1, nu=2), title = Valores de
    nu,lty=1, lwd=2, col = 1:3)

# Generamos matrices de covarianza para puntos 'x' usando una
# determinada función núcleo.
cov_matrix <- function(x, kernel_fn, ...) {
  outer(x, x, function(a, b) kernel_fn(a, b, ...))
}

# Dadas las coordenadas de x, dibujamos N funciones de
# covarianza en esos puntos.
draw_samples <- function(x, N, seed = 1, kernel_fn, ...) {
  Y <- matrix(NA, nrow = length(x), ncol = N)
  set.seed(seed)
  for (n in 1:N) {
    K <- cov_matrix(x, kernel_fn, ...)
    Y[, n] <- mvrnorm(1, mu = rep(0, times = length(x)), Sigma ↩
        = K)
  }
  Y
}
x <- seq(0, 2, length.out = 201)  # coordenadas de x
N <- 3  # número de muestras
col_list <- c(red, blue, black)  # colores de las muestras

# Función de covarianza de Matern.
matern_kernel <- function(x, y, nu = 1.5, sigma = 1, l = 1) {
  if (!(nu %in% c(0.5, 1.5, 2.5))) {
    stop(p must be equal to 0.5, 1.5 or 2.5)
  }
  p <- nu - 0.5
  d <- abs(x - y)
  if (p == 0) {
    sigma^2 * exp(- d / l)
  } else if (p == 1) {
    sigma^2 * (1 + sqrt(3)*d/l) * exp(- sqrt(3)*d/l)
  } else {
    sigma^2 * (1 + sqrt(5)*d/l + 5*d^2 / (3*l^2)) * exp(-sqrt↩
        (5)*d/l)
  }
}
par(mfrow = c(1, 3))
for (nu in c(0.5, 1.5, 2.5)) {
  Y <- draw_samples(x, N, kernel_fn = matern_kernel, nu = nu)

  plot(range(x), range(Y), xlab = x, ylab = y, type = n,
```

```
65          main = paste(Matern kernel, nu =, nu * 2, / 2))
66    for (n in 1:N) {
67      lines(x, Y[, n], col = col_list[n], lwd = 1.5)
68    }
69  }
70
71  # Función de covarianza SE
72  se_kernel <- function(x, y, sigma = 1, length = 1) {
73    sigma^2 * exp(- (x - y)^2 / (2 * length^2))
74  }
75
76  # Dibujamos ejemplos de funciones de un proceso Gaussiano
77  Y <- draw_samples(x, 1, kernel_fn = matern_kernel, nu = 0.5)
78  plot(x,Y, xlab = datos de entrada, x, ylab = datos de salida,
       y, col=1,type=l)
79
80  Y <- draw_samples(x, 1, kernel_fn = matern_kernel, nu = 1.5)
81  lines(x, Y, col = 2, type=l)
82
83  Y <- draw_samples(x, 1, kernel_fn = matern_kernel, nu = 2.5)
84  lines(x, Y, col = 3, type=l)
85
86  Y <- draw_samples(x, 1, kernel_fn = se_kernel, length = 1)
87  lines(x, Y, col = 4, type=l)
88
89  legend( x = topleft, legend=c(nu=0.5, nu=1.5, nu=2.5, SE), title = ↩
       Valores de
       nu,lty=1, lwd=2, col = 1:4)
```

E.3.2. Función de covarianza Gamma-exponencial.

```
1  # Gamma-exponencial
2  d<- seq( 0,3,,200)
3  l=1;
4
5  y1=exp(-(d/l)^1)
6  y2=exp(-(d/l)^1.5)
7  y3=exp(-(d/l)^2)
8
9  plot(d, y1, type=l, lty=1, lwd=2, xlab=Distancia entre datos,
       r, ylab = Función de covarianza)
10 lines(d, y2, col=2, type=l, lty=1, lwd=2)
11 lines(d, y3, col=3, type=l, lty=1, lwd=2)
12 legend( x = topright, legend=c(gamma=1, gamma=1.5, gamma=2), ↩
       title = Valores de
       gamma,lty=1, lwd=2, col = 1:3)
13
14 # Muestras
```

```
15 library(MASS)
16
17 # Generamos matrices de covarianza para puntos 'x' usando una
18 # determinada función núcleo.
19 cov_matrix <- function(x, kernel_fn, ...) {
20   outer(x, x, function(a, b) kernel_fn(a, b, ...))
21 }
22 # Dadas las coordenadas de x, dibujamos N funciones de
23 # covarianza en esos puntos.
24 draw_samples <- function(x, N, seed = 1, kernel_fn, ...) {
25   Y <- matrix(NA, nrow = length(x), ncol = N)
26   set.seed(seed)
27   for (n in 1:N) {
28     K <- cov_matrix(x, kernel_fn, ...)
29     Y[, n] <- mvrnorm(1, mu = rep(0, times = length(x)), Sigma ↩
         = K)
30   }
31   Y
32 }
33 x <- seq(-5, 5, length.out = 201)
34 gamma_exp <- function(x, y, gamma, length = 1) {
35    exp(- (x - y)^gamma / (length^gamma))
36 }
37 Y <- draw_samples(x, 1, kernel_fn = gamma_exp, gamma=1)
38 plot(x,Y, xlab = datos de entrada, x, ylab = datos de salida,
    y, col=1,type=l)
39 Y <- draw_samples(x, 1, kernel_fn = gamma_exp, gamma=3/2)
40 lines(x, Y, col = 2, type=l)
41 Y <- draw_samples(x, 1, kernel_fn = gamma_exp, gamma=2)
42 lines(x, Y, col = 3, type=l)
43
44 legend( x = topleft, legend=c(gamma=1, gamma=2), title = Valores de
    gamma,lty=1, lwd=2, col = c(1,3))
```

E.3.3. Función de covarianza periódica.

```
1 # Función de covarianza periódica
2
3 period_kernel <- function(x, y, length = 1, p=1) {
4   exp(-2 * sin(pi * abs(x - y) / p)^2 / length^2)
5 }
6 N=3
7 par(mfrow = c(1, 3))
8 for (p in c(0.5, 1, 2)) {
9   Y <- draw_samples(x, N, kernel_fn = period_kernel, p = p)
10
11   plot(range(x), range(Y), xlab = x, ylab = y, type = n,
12        main = paste(Periodic kernel, p =, p))
```

```
13    for (n in 1:N) {
14      lines(x, Y[, n], col = col_list[n], lwd = 1.5)
15    }
16  }
17
18  d<- seq( 0,3,,200)
19  l=1;
20
21  y1=exp(-2 * sin(pi * abs(d) / 0.5)^2 / l^2)
22  y2=exp(-2 * sin(pi * abs(d) / 1)^2 / l^2)
23  y3=exp(-2 * sin(pi * abs(d) / 2)^2 / l^2)
24
25  plot(d, y1, type=l, lty=1, lwd=2, xlab=Distancia entre datos,
        r, ylab = Función de covarianza)
26  lines(d, y2, col=2, type=l, lty=1, lwd=2)
27  lines(d, y3, col=3, type=l, lty=1, lwd=2)
28  legend( x = topright, legend=c(p=0.5, p=1, p=2), title = Valores de
        p,lty=1, lwd=2, col = 1:3)
29
30  # Muestras
31  library(MASS)
32
33  # Generamos matrices de covarianza para puntos 'x' usando una
34  # determinada función núcleo.
35  cov_matrix <- function(x, kernel_fn, ...) {
36    outer(x, x, function(a, b) kernel_fn(a, b, ...))
37  }
38  # Dadas las coordenadas de x, dibujamos N funciones de
39  # covarianza en esos puntos.
40  draw_samples <- function(x, N, seed = 1, kernel_fn, ...) {
41    Y <- matrix(NA, nrow = length(x), ncol = N)
42    set.seed(seed)
43    for (n in 1:N) {
44      K <- cov_matrix(x, kernel_fn, ...)
45      Y[, n] <- mvrnorm(1, mu = rep(0, times = length(x)), Sigma ↩
          = K)
46    }
47    Y
48  }
49  x <- seq(-1, 1, length.out = 201)
50  Y <- draw_samples(x, 1, kernel_fn = period_kernel, p = 0.5)
51  plot(x,Y, xlab = datos de entrada, x, ylab = datos de salida,
        y, col=1,type=l, ylim=c(-0.5,2))
52  Y <- draw_samples(x, 1, kernel_fn = period_kernel, p = 1)
53  lines(x, Y, col = 2, type=l)
54  Y <- draw_samples(x, 1, kernel_fn = period_kernel, p = 2)
55  lines(x, Y, col = 3, type=l)
56
57  legend( x = topright, legend=c(p=0.5, p=1, p=2), title = Valores de
```

```
    p,lty=1, lwd=2, col = 1:3)
```

E.3.4. Función de covarianza racional cuadrática.

```
# Función de covarianza racional cuadrática (RQ)
d<- seq( 0,3,,2000)
rq_kernel <- function(d, alpha, length = 1) {
    (1 + (d)^2 / (2 * alpha * length^2))^(-alpha)
}
y1=rq_kernel(d,0.5)
y2=rq_kernel(d,2)
y3=rq_kernel(d,10000)

plot(d, y1, type=l, lty=1, lwd=2, xlab=Distancia entre datos,
    r, ylab = Función de covarianza, ylim = c(0,1))
lines(d, y2, col=2, type=l, lty=1, lwd=2)
lines(d, y3, col=3, type=l, lty=1, lwd=2)

legend( x = topright, legend=c(alpha=0.5, alpha=2, alpha=infinity), ↩
    title = Valores de
    alpha,lty=1, lwd=2, col = 1:3)

#Muestras
library(MASS)

# Generamos matrices de covarianza para puntos 'x' usando una
# determinada función núcleo.
cov_matrix <- function(x, kernel_fn, ...) {
  outer(x, x, function(a, b) kernel_fn(a, b, ...))
}
# Dadas las coordenadas de x, dibujamos N funciones de
# covarianza en esos puntos.
draw_samples <- function(x, N, seed = 1, kernel_fn, ...) {
  Y <- matrix(NA, nrow = length(x), ncol = N)
  set.seed(seed)
  for (n in 1:N) {
    K <- cov_matrix(x, kernel_fn, ...)
    Y[, n] <- mvrnorm(1, mu = rep(0, times = length(x)), Sigma ↩
        = K)
  }
  Y
}
x <- seq(-5, 5, length.out = 2001)
rq_kernel <- function(x,y, alpha, length = 1) {
  (1 + (x-y)^2 / (2 * alpha * length^2))^(-alpha)
}

Y <- draw_samples(x, 1, kernel_fn = rq_kernel, alpha=0.5)
```

```
plot(x,Y, xlab = datos de entrada, x, ylab = datos de salida,
    y, col=1,type=l,ylim = c(-3,3))
Y <- draw_samples(x, 1, kernel_fn = rq_kernel, alpha=2)
lines(x, Y, col = 2, type=l)
Y <- draw_samples(x, 1, kernel_fn = rq_kernel, alpha=1000)
lines(x, Y, col = 3, type=l)
legend( x = topright, legend=c(alpha=0.5, alpha=2, alpha=infinity), ↩
    title = Valores de
    alpha,lty=1, lwd=2, col = 1:3)
```

E.3.5. Función de covarianza exponencial cuadrada (SE).

```
# Función de covarianza exponencial cuadrada (SE)
x<- seq( -5, 5,,200)

y1=exp(-x^2/(2*1^2))
y2=exp(-x^2/(2*2^2))
y3=exp(-x^2/(2*0.5^2))

y<- cbind( y1,y2,y3)

matplot( x, y, type=l, lty=1, lwd=2, xlab=Distancia entre datos,
    r, ylab = Función de covarianza)
legend( x = topright, legend=c(l=1, l=2, l=0.5), title = Valores de
    nu,lty=1, lwd=2, col = 1:3)

x <- seq(-5, 5, length.out = 201)
N <- 3
col_list <- c(red, blue, black)
se_kernel <- function(x, y, sigma = 1, length = 1) {
  sigma^2 * exp(- (x - y)^2 / (2 * length^2))
}

#Dibujamos ejemplos de funciones de un proceso Gaussiano.
Y <- draw_samples(x, 1, kernel_fn = se_kernel, length = 1)
plot(x,Y, xlab = datos de entrada, x, ylab = datos de salida,
    y, col=1,type=l)

Y <- draw_samples(x, 1, kernel_fn = se_kernel, length = 2)
lines(x, Y, col = 2, type=l)

Y <- draw_samples(x, 1, kernel_fn = se_kernel, length = 0.5)
lines(x, Y, col = 3, type=l)
```

E.4. Simulaciones.

E.4.1. Simulación de la concentración de CO2 (2 dimensiones).

Para realizar la simulación se han empleado los códigos de software de [34] (http://www.gaussianprocess.org/gpml/). Con la ayuda de éstos, se ha simulado un problema de regresión para la concentración de CO_2 en Mauna Loa (https://serc.carleton.edu/introgeo/interactive/examples/co2.html).

```
clear all;clc;
% Primero definimos los valores de x e y
x=linspace(1990,2012,265)';
y=data(373:637);

% Representamos los datos de entrenamiento
figure(1)
plot(x,y,'-*')
ylabel('Concentración de CO2 (ppmv)')
xlabel('Tiempo')
axis([1990 2012 350 400])
legend('Datos de entrenamiento')
title('Concentración de CO2 en Mauna Loa')
grid on

% Centramos los datos para optimizar los parámetros
y1=y-repmat( mean(y), [265,1] );

% Definimos los puntos de test -> queremos ver tendencias en
% el futuro
xs=linspace(1990,2025,10000)';

% Definimos la función de media y de covarianzas
meanfunc = [];

prod={'covProd',{'covSEiso','covPeriodic'}};
sum={'covSum',{'covSEiso','covNoise'}};
covfunc = {'covSum',{'covSEiso',prod,'covRQiso',sum}};
likfunc = @likGauss;

% Inicializamos los hiperparámetros
hyp = struct('mean',[],'cov',[0 0 3 0 0 0 1 3 5 3 0 0 -3],'lik',-1)
hyp2 = minimize(hyp,@gp,-100,@infGaussLik, meanfunc,covfunc,likfunc,x,y1);

[mu s2] = gp(hyp2,@infGaussLik,meanfunc, covfunc, likfunc, x, y, xs);

figure(2)
```

```
f = [mu+2*sqrt(s2); flipdim(mu-2*sqrt(s2),1)];
fill([xs; flipdim(xs,1)], f, [7 7 7]/8)
hold on; plot(xs, mu); plot(x, y, '+')

ylabel('Concentración de CO2 (ppmv)')
xlabel('Tiempo')
legend('Intervalo de confianza del 95%','Datos de prueba','↩
   Datos de entrenamiento')
title('Concentración de CO2 en Mauna Loa')
grid on
```

E.4.2. Simulación en 3 dimensiones.

Se muestra el código para el ejemplo en tres dimensiones. Se trata de un caso sencillo que permite visualizar el procedimiento. En este caso, se han seleccionado como datos los correspondientes a una función bidimensional, los cuales han sido perturbados, con el objetivo de comprobar si puede recuperarse la función original.

```
clear all;clc;
%Primero definimos los valores de x e y
x1=linspace(-2,2,10)';
x2=linspace(-2,2,10)';
n=length(x1);
xz=[x1 x2];

[x1,x2]=meshgrid(x1,x2);
y=sqrt(x1.^2+x2.^2)+ 0.3.*rand(10,10);

x1c = reshape(x1,[],1);
x2c = reshape(x2,[],1);
x=[x1c x2c];
yc=reshape(y,[],1);

figure(1)
grid on
hold on
mesh(x1,x2,y)
scatter3(x1c,x2c,yc,'filled');
view([-21.1 15.4 ])
xlabel('x1')
ylabel('x2')
zlabel('y')

%Centramos los datos para optimizar los parámetros
ym=yc-repmat( mean(yc), [100,1] );

%Definimos los puntos de test --> queremos ver tendencias en
%el futuro
xs1=linspace(-3,3,100)';
```

```
xs2=linspace(-3,3,100)';
[xs1,xs2]=meshgrid(xs1,xs2);

xs1c = reshape(xs1,[],1);
xs2c = reshape(xs2,[],1);
xs=[xs1c xs2c];

% Definimos la función de media y de covarianzas
meanfunc = [];
covfunc = @covSEiso;
likfunc = @likGauss;

% Inicializamos los hiperparámetros
 hyp = struct('mean', [], 'cov', [0 0], 'lik', -1)
 hyp2 = minimize(hyp, @gp, -100, @infGaussLik, meanfunc, ↩
    covfunc, likfunc, x, ym);

 [ymu ys2] = gp(hyp2, @infGaussLik, meanfunc, covfunc, likfunc,↩
     x, yc, xs);

figure(2)
ymu1=reshape(ymu,[],100); mesh(xs1,xs2,ymu1)
view([-21.1 15.4 ])
xlabel('x1')
ylabel('x2')
zlabel('y')

figure(3)
grid on
hold on
view([-21.1 15.4 ])
mesh(xs1,xs2,ymu1)
scatter3(x1c,x2c,yc,'filled')
xlabel('x1')
ylabel('x2')
zlabel('y')

figure(4)
grid on
hold on
view([-21.1 15.4 ])
mesh(xs1,xs2,ymu1)
surf(x1,x2,y)
xlabel('x1')
ylabel('x2')
zlabel('y')
```

E.4.3. Simulación del brazo robótico 7 DOF-SARCOS.

Se muestra el código empleado para la simulación del problema de regresión basado en datos reales. En concreto, se han seleccionado los datos correspondientes a los diferentes movimientos de un brazo robótico. Las funciones empleadas pueden encontrarse en [34].

```
clear all;clc;
%Importamos los datos de entrenamiento y de test
sarcos_inv=importdata('sarcos_inv.mat');
sarcos_inv_test=importdata('sarcos_inv_test.mat');

%Asignamos los correspondientes valores a las variables.
x=zeros(size(sarcos_inv,1),size(sarcos_inv,2));
for i=1:size(sarcos_inv,2)
    x(:,i)=sarcos_inv(:,i);
end
y=sarcos_inv(:,22);

xs=zeros(size(sarcos_inv_test,1),size(sarcos_inv_test,2));
for i=1:size(sarcos_inv_test,2)
    xs(:,i)=sarcos_inv_test(:,i);
end
ys=sarcos_inv_test(:,22);

%Centramos y estandarizamos los datos
x=(x-repmat(mean(x),size(sarcos_inv,1),1))./(repmat(std(x),size(sarcos_inv,1),1));
y=(y-repmat(mean(y),size(sarcos_inv,1),1))./(repmat(std(y),size(sarcos_inv,1),1));

xs=(xs-repmat(mean(xs),size(sarcos_inv_test,1),1))./(repmat(std(xs),size(sarcos_inv_test,1),1));
ys=(ys-repmat(mean(ys),size(sarcos_inv_test,1),1))./(repmat(std(ys),size(sarcos_inv_test,1),1));

%Repetimos el procedimiento para distintos valores de m
n=[256,512,1024,2048,4096];
hiperparametros=zeros(2,size(n,2));
media=zeros(size(sarcos_inv_test,1),size(n,2));
desviacion=zeros(size(sarcos_inv_test,1),size(n,2));

for i=1:size(n,2)
    for j=1:10
        m=n(i);
        meanfunc = [];
        covfunc = @covSEiso;
        likfunc = @likGauss;
        inf = @infGaussLik

```

```
40          % Inicializamos los hiperparámetros
41          hyp = struct('mean', [], 'cov', [5 3], 'lik', -1)
42
43          %u contiene el subconjunto de datos que vamos
44          %a emplear. Se escogen aleatoriamente. Tenemos
45          %que calcular los y asociados
46          %a dicho conjunto.
47          nu = fix(m/2); iu = randperm(m); iu=iu(1:nu); u=x(iu,:)↩
                ; yu=y(iu);
48          covfuncF = {@apxSparse, {covfunc}, u};
49          inf = @(varargin) inf(varargin{:},struct('s',0.0));
50
51          hyp2 = minimize(hyp, @gp, -100, inf, meanfunc, covfuncF↩
                , likfunc, u, yu);
52          [mF s2F] = gp(hyp2,inf,meanfunc, covfuncF,likfunc, x, y↩
                ,xs);
53
54          hiperparametros(:,i)=hyp2.cov;
55          media(:,i)=mF;
56          desviacion(:,i)=s2F;
57
58          %Para ver cómo de buena es la aproximación podemos
59          %usar el error cuadrático medio.
60          MSE_matrix(j,i)=(1/size(sarcos_inv_test,2))*(mF-ys).'*(↩
                mF-ys);
61          SMSE(j,i)=(MSE_matrix(i))/sqrt(var(ys));
62      end
63      MSE=mean(MSE_matrix);
64      error=std(MSE_matrix);
65  end
66
67  for i=1:size(n,2)
68      fprintf('El error cuadrático medio para m=%d es: %e  \n', n↩
            (i),MSE(i));
69  end
70
71  %Representamos el MSE frente al tama o de la muestra escogido↩
        .
72  errorbar(n,MSE,error,'-*');
73  hold on
74  ylabel('SMSE')
75  xlabel('Tama o de la muestra (m)')
76  legend('SD')
77  xticks([256, 512 1024 2048 4096])
78  xlim([250,4100])
79  hold off
```

BIBLIOGRAFÍA

[1] Abramowitz, Milton: *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables,*. Dover Publications, Inc., USA, 1974, ISBN 0486612724.

[2] Agarwal, Shivani: *Excess Error, Approximation Error, and Estimation Error*. En *Statistical Learning Theory Lecture Notes*, páginas 1–5. 2011.

[3] Ando, Tomohiro, Sadanori Konishi y Seiya Imoto: *Nonlinear regression modeling via regularized radial basis function networks*. Journal of Statistical Planning and Inference, 138(11):3616–3633, Noviembre 2008, ISSN 03783758.

[4] Ash, Robert B.: *Topics in Stochastic Processes*. Probability and Mathematical Statistics ; 27. Academic Press, New York [etc, 1975, ISBN 0120652706.

[5] Banerjee, S. y A. E. Gelfand: *On smoothness properties of spatial processes*, Enero 2003. ISSN 0047259X.

[6] BARTLETT PeterBartlett, Peter L: *Model Selection and Error Estimation **. Informe técnico, 2002.

[7] Bell, William, Peter J. Brockwell y Richard A. Davis: *Time Series: Theory and Methods.*, volumen 84. 1989, ISBN 9781441903198.

[8] Berlinet, Alain y Christine Thomas-Agnan: *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. 2004, ISBN 9781461347927.

[9] Bhartendu: *Kernel Ridge Regression - File Exchange - MATLAB Central*. https://es.mathworks.com/matlabcentral/fileexchange/63122-kernel-ridge-regression.

[10] Billingsley, Patrick: *Probability and measure*. Wiley series in probability and mathematical statistics. John Wiley & Sons, New York [etc, 3rd ed. edición, 1995, ISBN 0471007102.

[11] Boyd, Stephen y Lieven Vandenberghe: *Convex Optimization*. ISBN 9780521833783. http://www.cambridge.org.

[12] Brandt Petersen Michael Syskind Pedersen, Kaare, Bill Baxter, Brian Templeton, Christian Rishøj, Christian Schröppel, Dan Boley, Douglas L Theobald, Esben Hoegh-Rasmussen, Evripidis Karseras, Georg Martius, Glynne Casteel, Jan Larsen, Jun

Bin Gao, Jürgen Struckmeier, Kamil Dedecius, Karim T Abou-Moustafa, Korbinian Strimmer, Lars Christiansen, Lars Kai Hansen, Leland Wilkinson, Liguo He, Loic Thibaut, Markus Froeb, Michael Hubatka, Miguel Barão, Ole Winther, Pavel Sakov, Stephan Hattinger, Troels Pedersen, Vasile Sima y Vincent Rabaud: *The Matrix Cookbook*. Informe técnico.

[13] Buhmann, Martin D.: *Radial Basis Functions*. Cambridge University Press, Julio 2003, ISBN 9780521633383. https://www.cambridge.org/core/product/identifier/9780511543241/type/book.

[14] Choi, Taeryon y Mark J. Schervish: *On posterior consistency in nonparametric regression problems*. Journal of Multivariate Analysis, 98(10):1969–1987, Noviembre 2007, ISSN 0047259X.

[15] Conway, John B.: *A Course in Functional Analysis*. Graduate Texts in Mathematics, 96. Springer New York, New York, NY, 2nd ed. 20 edición, 2007, ISBN 1-4757-4383-1.

[16] Data, Towards: *Understanding Gaussian Process, the Socratic Way - Towards Data Science*. (X):1–3, 2020. https://towardsdatascience.com/understanding-gaussian-process-the-socratic-way-ba02369d804.

[17] Devroye, Luc.: *A Probabilistic Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability, 31. Springer New York, New York, NY, 1st ed. 19 edición, 1996, ISBN 1-4612-0711-8.

[18] Dudley, R. M. (Richard M.): *Real analysis and probability*. Cambridge studies in advanced mathematics ; 74. Cambridge University Press, Cambridge, second edi edición, 2002, ISBN 9780511755347.

[19] Fasshauer, Gregory E: *Meshfree Approximation Methods with Matlab*, volumen 6 de *Interdisciplinary Mathematical Sciences*. WORLD SCIENTIFIC, Abril 2007, ISBN 978-981-270-633-1. https://www.worldscientific.com/worldscibooks/10.1142/6437.

[20] Gikhman, Iosif I.: *The Theory of Stochastic Processes I*. Classics in Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st ed. 20 edición, 2004, ISBN 3-642-61943-6.

[21] Goldberg, Paul W. y Mark R. Jerrum: *Bounding the Vapnik-Chervonenkis Dimension of Concept Classes Parameterized by Real Numbers*. Machine Learning, 18(2):131–148, 1995, ISSN 15730565.

[22] Haussler, David: *Probably Approximately Correct learning and decision-theoretic generalizations*. Mathematical perspectives on neural networks. Hillsdale, NJ: Lawrence Erlbaum Publishers, 1996.

[23] Kanagawa, Motonobu, Philipp Hennig, Dino Sejdinovic y Bharath K Sriperumbudur: *Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences*. arXiv, Julio 2018. http://arxiv.org/abs/1807.02582.

[24] Karimi, N., S. Kazem, D. Ahmadian, H. Adibi y L. V. Ballestra: *On a generalized Gaussian radial basis function: Analysis and applications*. Engineering Analysis with Boundary Elements, 112:46–57, Marzo 2020, ISSN 09557997.

[25] Kullback, S.: *Information theory and statistics - Solomon Kullback.pdf*, 1968, ISBN 0844656259.

[26] Mackay, D.: *Introduction to Gaussian processes*. 1998.

[27] McAllester, David A.: *PAC-Bayesian stochastic model selection*. Machine Learning, 51(1):5–21, 2003, ISSN 08856125.

[28] McAllester, David A., Jonathan Baxter y Nicoì O Cesa-Bianchi: *Some PAC-Bayesian theorems*. Machine Learning, 37(3):355–363, 1999, ISSN 08856125.

[29] Mitchell, Tom M.: *Machine learning*. MacGraw-Hill, New York [etc, 1997, ISBN 0070428077.

[30] Pavliotis, G A: *STOCHASTIC PROCESSES AND APPLICATIONS*. Informe técnico, 2015.

[31] Press, William H, Saul A Teukolsky, William T Vetterling y Brian P Flannery: *The Art of Scientific Computing Second Edition Cambridge New York Port Chester Melbourne Sydney*. Informe técnico, 1988, ISBN 0521431085. http://www.nr.com.

[32] Quiñonero, Joaquin, Qui˜ Quiñonero-Candela, Carl Edward Rasmussen y Carl@tuebingen Mpg De: *A Unifying View of Sparse Approximate Gaussian Process Regression*. Informe técnico, 2005.

[33] Rasmussen, Carl Edward y Hn@tue Mpg De: *Gaussian Processes for Machine Learning (GPML) Toolbox Hannes Nickisch*. Informe técnico, 2010. http://www.kyb.tuebingen.mpg.de/bs/people/carl/code/minimize/.

[34] Rasmussen, Carl Edward y Christopher K. I. Williams: *Gaussian Processes for Machine Learning*. The MIT Press, Diciembre 2018.

[35] Rimal, Raju: *Evaluation of Models for predicting the average monthly Euro versus Norwegian krone exchange rate from financial and commodity information*. Tesis de Doctorado, Diciembre 2014.

[36] Rohilla, Cosma, Shalizi With y Aryeh Kontorovich: *Almost None of the Theory of Stochastic Processes A Course on Random Processes, for Students of Measure-Theoretic Probability, with a View to Applications in Dynamics and Statistics*. Informe técnico.

[37] Sandberg, Ellen y Trygve Almøy: *Evaluation of Models for predicting the*. Informe técnico, Mayo 2014. https://nmbu.brage.unit.no/nmbu-xmlui/handle/11250/283547.

[38] Seeger, Matthias: *PAC-Bayesian generalisation error bounds for Gaussian process classification*. Journal of Machine Learning Research, 3(2):233–269, 2003, ISSN 15324435.

[39] Seldin, Yevgeny: *A PAC-Bayesian Approach to Structure Learning*. Informe técnico.

[40] Shalev-Schwartz, Shai y Shai Ben-David: *Understanding machine learning*, volumen 128. 2017, ISBN 9781107298019.

[41] Shalev-Shwartz, Shai: *Understanding machine learning : from theory to algorithms*. Cambridge University Press, Cambridge, 2014, ISBN 9781107298019.

[42] Shawe-Taylor, John: *Kernel methods for pattern analysis*. Cambridge University Press, Cambridge [etc, 2004, ISBN 0-521-81397-2.

[43] Singh, Surya Dev y Sachin Gupta: *HYDRAULIC AND LINEAR ACTUATOR MOTOR OPERATED 7 DOF HUMANOID ROBOTIC ARM WITH DEXTEROUS HAND*. Informe técnico. http://www.ripublication.com.

[44] Stein, Michael L.: *Interpolation of Spatial Data Some Theory for Kriging*. Springer Series in Statistics. Springer New York, New York, NY, 1st ed. 19 edición, 1999, ISBN 1-4612-1494-7.

[45] Stein, Michael L.: *Interpolation of Spatial Data Some Theory for Kriging*. Springer Series in Statistics. Springer New York, New York, NY, 1st ed. 19 edición, 1999, ISBN 1-4612-1494-7.

[46] Steinwart, Ingo.: *Support Vector Machines*. Information Science and Statistics. Springer New York, New York, NY, 1st ed. 20 edición, 2008, ISBN 1-281-92704-X.

[47] Theodoridis, Sergios: *Machine learning : a Bayesian and optimization perspective*. .NET Developers Series. Academic Press, Amsterdam, [Netherlands, first edit edición, 2015, ISBN 0-12-801722-8.

[48] Uddin, Shahadat, Arif Khan, Md Ekramul Hossain y Mohammad Ali Moni: *Comparing different supervised machine learning algorithms for disease prediction*. BMC Medical Informatics and Decision Making, 19(1), Diciembre 2019, ISSN 14726947. https://pubmed.ncbi.nlm.nih.gov/31864346/.

[49] Valiant, G: *A Theory of the Learnable*. Informe técnico.

[50] Vapnik, Vladimir Naumovich: *The nature of statistical learning theory*. Springer, New York, 1995, ISBN 1-4757-2440-3.

[51] Vijayakumar, Sethu, Aaron D'souza, Tomohiro Shibata, Jörg Conradt y Stefan Schaal: *Statistical learning for humanoid robots*. Autonomous Robots, 12(1):55–69, Enero 2002, ISSN 09295593. https://link.springer.com/article/10.1023/A:1013258808932.

[52] Warmuth, Manfred: *Sample compression, learnability, and the Vapnik-Chervonenkis dimension*. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1208:1–2, 1997, ISSN 16113349.