Control del error y sobreajuste en los problemas de clasificación

Daniel Carpio Martín

Trabajo Fin de Grado
dirigido por Eustasio Del Barrio Tellado

Grado en Matemáticas
Universidad de Valladolid
Julio 2015

Universidad de Valladolid

Índice general

Capítulo 1

Introducción

Nos encontramos en la era del Big Data, donde cada vez más y más cantidad de datos son almacenados. No siempre se sabe sacar partido de esta abundancia de datos, y en muchos casos se pretende extraer conocimiento a partir de esta gran cantidad de información. El aprendizaje automático permite desarrollar métodos de predicción a partir de esta información almacenada.

Con este término genérico de aprendizaje automático se suele aludir a una serie de métodos estadísticos y computacionales que tratan de inferir información relativa a alguna variable respuesta a partir de atributos observables mediante el estudio de la relación entre ambos en un conjunto de datos (aprendizaje supervisado) o bien de encontrar patrones o estructuras más o menos ocultas en el conjunto de los datos(aprendizaje no supervisado).

Este trabajo se centra en problemas de aprendizaje supervisado, principalmente en el problema de clasificación. La disponibilidad computacional ha hecho aumentar de forma notable la cantidad y variedad de reglas de clasificación que se pueden ajustar. Es habitual encontrar situaciones en las que sobre un conjunto reducido de casos ($n \simeq 100,\ 1000$) se ha registrado una cantidad de información de variables respuesta ($p \simeq 10^5,\ 10^6$). En estas condiciones es muy fácil encontrar reglas que tengan un buen comportamiento sobre la muestra. Sin embargo, sin un buen entendimiento del comportamiento probabilístico de los métodos empleados para seleccionar tales reglas se puede llegar a aceptar reglas con capacidad predictiva muy limitada. Este trabajo se dedica precisamente a estudiar este comportamiento probabilístico y a analizar las propiedades de tres de los principios más habituales de selección de reglas de clasificación actualmente: minimización del riesgo estructural, máquinas de soporte vectorial y lasso.

En el desarrollo del trabajo se comprobará que para controlar la capacidad predictiva de una regla de clasificación es necesario estudiar el comportamiento probabilístico de la máxima desviación entre promedios empíricos y valores poblacionales, lo que de forma técnica se conoce como proceso empírico. El control probabilístico del supremo de un proceso empírico se efectúa mediante la combinación de dos técnicas. Por un lado, las desigualdades maximales acotan el valor medio de tales supremos. Por otra parte, las desigualdades de concentración acotan las probabilidades de que tales supremos se desvíen de los valores esperados. Aunque no son el objeto principal de este trabajo, parte de la memoria se dedica a describir alguna de estas herramientas.

En el capítulo 2, se proporcionará una pequeña descripción del aprendizaje automático, diferenciando entre sus dos versiones, el aprendizaje supervisado y el no supervisado y viendo algunos ejemplos de cada uno de los diferentes problemas.

En el capítulo 3, se analiza en más detalle el problema de clasificación binaria donde el objetivo será encontrar una función cuyo error de generalización se aproxime al denominado *clasificador de Bayes*, que es el clasificador que minimiza dicho riesgo, pero su interés será meramente teórico, pues dependerá de cantidades desconocidas. Se introducen ciertos resultados teóricos como las desigualdades de concentración o la teoría de Vapnik-Chervonenkis que serán utilizadas en la obtención de desigualdades maximales. Se introduce también el principio de minimización del riesgo estructural o penalizado, llegando a derivar una desigualdad tipo "oráculo".

A pesar del interés teórico de los resultados presentados en el capítulo 3, su utilidad práctica está seriamente limitada por el hecho de que las reglas propuestas requerirían, para su cálculo en la práctica, la solución de problemas de optimización combinatoria. Una alternativa computacionalmente factible, está dada por las máquinas de soporte vectorial(o SVM).

En el capítulo 4, se estudiará este método, comenzando por su planteamiento geométrico en diferentes situaciones y concluyendo con una visión probabilística del modelo. Parte de la popularidad de este método se debe a la manera computacionalmente eficiente que tiene de manejar de forma implícita transformaciones de los datos mediante el "truco del núcleo", propiedad que también se estudiará a lo largo del capítulo.

El buen comportamiento computacional del SVM no conlleva un buen comportamiento probabilístico. El capítulo 5, se dedica al método del lasso. Este es un método de penalización parecido al SVM pero con penalizaciones

l_1 en lugar de las penalizaciones l_2 del SVM. Veremos que esta elección permite recuperar un buen comportamiento probabilístico manteniendo a la vez buenas propiedades computacionales.

Finalmente, hay una breve sección de conclusiones y una lista con referencias bibliográficas empleadas.

Capítulo 2

El aprendizaje automático y sus aplicaciones

Cada vez es más frecuente encontrarse con tareas de supervisión, clasificación o toma de decisiones que no pueden ser realizadas por expertos humanos porque se tienen que hacer de forma masiva. Un ejemplo puede ser la detección de spam en emails o el uso de marcadores biológicos en chequeos de salud con el objeto de buscar indicadores de ciertas enfermedades.

En estas situaciones, lo normal es que no se pueda resolver de manera totalmente satisfactoria, pues no se puede establecer una relación directa entre las variables input y la respuesta. La forma de tratar estos problemas es utilizar ejemplos para aprender la relación existente entre las entradas y la salida que queremos hallar. En este contexto, aparece el *aprendizaje automático* (o *machine learning*), que se define como un conjunto de métodos o técnicas que permiten detectar comportamientos o patrones a partir de ejemplos ya estudiados y que se utilizarán para predecir nuevas situaciones o mejorar algunos tipos de decisión.

Dentro del *aprendizaje automático* se pueden distinguir dos tipos de problemas: el aprendizaje supervisado y el no supervisado. En el aprendizaje supervisado, los modelos parten de un conjunto de ejemplos cuyo comportamiento sí es conocido y que permite establecer una relación entre las características de las que se dispone a priori y las que que se pretenden predecir. Ejemplos de problemas de aprendizaje supervisado son la clasificación y la regresión.

En cambio en el aprendizaje no supervisado, sólo se tiene un conjunto de entradas de ejemplo y los modelos consisten en localizar alguna estructura

en dichos datos o encontrar algún patrón interesante (estructura en grupos, concentración en subespacios de dimensión reducida, etc). Ejemplos de aprendizaje no supervisado son el análisis clúster o el análisis de componentes principales.

Este tipo de problemas se utilizan en distintos campos como los negocios, la ciencia o la ingeniería. Algunas aplicaciones pueden ser en diagnósticos médicos, predicción en series temporales, reconocimiento de caracteres, clasificación de textos o para mejorar estrategias de mercado de las empresas, entre otras muchas.

Este trabajo se centrará en el caso del aprendizaje supervisado, que se detallará a continuación.

2.1. El aprendizaje supervisado

De manera más formal a la anteriormente vista, el aprendizaje supervisado consiste en encontrar una relación entre unas variables de entrada $\mathbf{x}$ y de salida y, a partir de un conjunto de ejemplos correctamente caracterizados $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ llamado conjunto de entrenamiento.

En el caso más sencillo, las observaciones $\mathbf{x}_i$ corresponderán a vectores numéricos d-dimensionales, pero también pueden representar imágenes, textos, etc.

En ocasiones, la relación entre los pares $(\mathbf{x}_i, y_i)$ se puede representar a través de una función llamada *función objetivo*, pero normalmente existe cierto ruido que no permite hallarla. Mediante los algoritmos de aprendizaje se estimará una función que será la solución del problema.

Se asume que es y es esencialmente (salvo alguna perturbación aleatoria) una función f de $\mathbf{x}$ desconocida y el objetivo del problema de aprendizaje consistirá en calcular una estimación $\hat{f}$ de dicha f a partir del conjunto de entrenamiento y después hacer predicciones a partir de esa estimación $\hat{y} = \hat{f}(\mathbf{x})$ en nuevos casos (nuevos valores de $\mathbf{x}$) no vistos anteriormente (a este proceso se le llama generalización). La elección de esta función $\hat{f}$ se realizará entre las funciones pertenecientes a algún subconjunto particular de todas las funciones posibles. A la elección del conjunto donde se buscará la estimación se denomina selección del modelo.

Dentro del aprendizaje supervisado, existen distintos tipos de problemas en función de los posibles valores que pueda tomar las variables respuesta

y_i, diferenciando entre los problemas de clasificación y los problemas de regresión. Esta separación se corresponde más a la orientación concreta de la aplicación que a los aspectos teóricos relacionados.

2.1.1. Problemas de clasificación

Los problemas de clasificación serán aquellos en los que la salida toma valores en un conjunto finito de posibilidades. Es decir, se buscará una aplicación entre las variables de entrada $\mathbf{x}$ y la de salida y, donde $y \in \{1, ..., C\}$, siendo C el número de clases posibles.

En la situación de que $C = 2$ estaremos en el caso de *clasificación binaria* y podemos tomar $y \in \{0, 1\}$ (alternativamente también se suele considerar el espacio $Y = \{-1, 1\}$). Si, en cambio, $C > 2$, estaremos ante un problema de *clasificación multiclase*.

Normalmente se asumirá que el par $(\mathbf{x}, y)$ tiene un comportamiento aleatorio (lo más habitual es que no haya posibilidad de predecir la etiqueta y con un 100 % de seguridad conociendo el atributo $\mathbf{x}$). Entonces, será interesante buscar reglas, es decir, funciones $f : \mathcal{X} \to \{1, ..., C\}$ que se equivoquen lo menos posible. Es decir, que minimicen $\mathbb{P}\{f(\mathbf{x}) \neq y\}$.

2.1.2. Problemas de regresión

El objetivo de los problemas de regresión es parecido al de la clasificación pero en este caso la variable a predecir y será continua, por ejemplo, que tome valores en $\mathbb{R}$. Algunas de las aplicaciones de los problemas de regresión pueden ser la predicción de temperaturas, de cantidades de algún antígeno en el cuerpo o de algunas características físicas cómo longitudes, edades, etc, a partir de otras variables de más fácil medición o control. De nuevo, lo normal es asumir que $(\mathbf{x}, y)$ tiene un comportamiento aleatorio y que el conocimiento de $\mathbf{x}$ no permite conocer el valor de y con total seguridad. Se buscará entonces funciones $f : \mathcal{X} \to \mathbb{R}$ que garanticen un error de aproximación pequeño. Este error es aleatorio, por lo que suele recurrirse a versiones promedio, por ejemplo, en la regresión de mínimos cuadrados, se trata de minimizar $\mathbb{E}\{(y - f(\mathbf{x}))^2\}$ entre las funciones f pertenecientes a cierta clase.

En los dos problemas, tanto regresión como clasificación, la función objetivo es desconocida, por lo que se debe estimar de alguna manera.

En este trabajo se tratará el caso de los problemas de clasificación y se profundizará en la situación del caso de clasificación binaria.

Capítulo 3

El problema de clasificación binaria

El problema de clasificación consiste en dada una observación, predecir una clase a la que pertenece. Para ello se buscarán posibles relaciones entre las características conocidas de dichas observaciones y las que se pretenden predecir.

Consideraremos cada observación un vector d-dimensional $\mathbf{x} \in \mathbb{R}^d$ de entradas y denotaremos la clase a la que pertenezca por y, cuyo valor será una de las n clases posibles C_i. Es decir, $y \in \mathcal{C} = \{C_i : 1 \leq i \leq n\}$.

En el proceso de clasificación, se tratará de buscar una función $f : \mathbb{R}^d \to \mathcal{C}$ de forma que a cada observación $\mathbf{x}$, le asigne un valor $f(\mathbf{x}) \in \mathcal{C}$ que puede coincidir con el valor real de y o no. A esta función se le denomina *clasificador* y diremos que comete un error cuando $f(\mathbf{x}) \neq y$, siendo y la verdadera clase a la que pertenece dicha observación.

Nos centraremos en el caso de la clasificación binaria, es decir, la situación en la que sólo hay dos posibles clases a las que pertenezca nuestra observación que denotaremos por 0 ó 1, de forma que $f(\mathbf{x}) \in \{0, 1\}$.

Comenzaremos el capítulo estudiando la versión poblacional del problema, es decir, el caso en que la distribución de $(\mathbf{x}, y)$ es conocida y buscamos, de entre todas las funciones $f : \mathbb{R}^d \to \{0, 1\}$, la que minimiza $\mathbb{P}\{f(\mathbf{x}) \neq y\}$. Después veremos el problema cuando estas distribuciones no se conocen y hay que recurrir a ejemplos ya clasificados para poder construir los modelos de clasificación, buscando soluciones a través de la minimización del riesgo empírico. También se expondrán ciertos resultados teóricos que permitirán

obtener cotas para el error de los clasificadores a partir de desigualdades maximales y de concentración. Por último veremos el método de la minimización del riesgo estructural a través del cuál se buscará el clasificador dentro de un modelo óptimo. En el desarrollo de este capítulo se han utilizado principalmente [1], [2] y [3].

3.1. El problema de Bayes

Introduciremos un par (X, Y) de variables aleatorias con valores en el espacio $\mathbb{R}^d \times \{0, 1\}$, con X una distribución dada e y_0 e y_1 las probabilidades condicionadas de que $Y = 0$ e $Y = 1$ cuando $X = \mathbf{x}$, también conocidas como *probabilidades a posteriori*.

$$\begin{aligned} y_0(\mathbf{x}) &= \mathbb{P}\{Y = 0 | X = \mathbf{x}\} \\ y_1(\mathbf{x}) &= \mathbb{P}\{Y = 1 | X = \mathbf{x}\} \end{aligned} \tag{3.1}$$

Puesto que $\mathbf{x}$ pertenecerá a una clase o la otra, se verificará $y_0 = 1 - y_1$.

Se denomina *clasificador de Bayes* a aquel que dado $\mathbf{x}$, clasifica en 1 cuando la probabilidad de que $Y = 1$ sea mayor que la de que $Y = 0$ y viceversa. Denotaremos a este clasificador f^*

$$f^*(\mathbf{x}) = \begin{cases} 1 & si \quad y_1(\mathbf{x}) \leq y_0(\mathbf{x}) \quad \Rightarrow \quad y_1(\mathbf{x}) \geq 1/2 \\ 0 & si \quad y_1(\mathbf{x}) \geq y_0(\mathbf{x}) \quad \Rightarrow \quad y_1(\mathbf{x}) < 1/2 \end{cases} \tag{3.2}$$

3.1 Definición. Sea f un clasificador cualquiera, la probabilidad de error de f vendrá dada por:

$$R(f) = \mathbb{P}\{f(X) \neq Y\} \tag{3.3}$$

3.2 Teorema. El clasificador de Bayes tiene la menor probabilidad de error, es decir, para cualquier $f : \mathbb{R}^d \to \{0, 1\}$, se cumple

$$R(f^*) \leq R(f) \tag{3.4}$$

Demostración. Dado $X = \mathbf{x}$, la probabilidad de error para cualquier f, puede expresarse como

$$\begin{aligned} &\mathbb{P}\{f(X) \neq Y | X = \mathbf{x}\} \\ &\quad = 1 - \mathbb{P}\{f(X) = Y | X = \mathbf{x}\} \\ &\quad = 1 - [\mathbb{P}\{f(X) = 1, Y = 1 | X = \mathbf{x}\} + \mathbb{P}\{f(X) = 0, Y = 0 | X = \mathbf{x}\}] \\ &\quad = 1 - [\mathbb{P}\{f(X) = 1 | X = \mathbf{x}\} \cdot \mathbb{P}\{Y = 1 | X = \mathbf{x}\} \\ &\qquad + \mathbb{P}\{f(X) = 0 | X = \mathbf{x}\} \cdot \mathbb{P}\{Y = 0 | X = \mathbf{x}\}] \\ &\quad = 1 - [\mathbb{P}\{f(X) = 1 | X = \mathbf{x}\} \cdot y_1(\mathbf{x}) + \mathbb{P}\{f(X) = 0 | X = \mathbf{x}\} \cdot (1 - y_1(\mathbf{x}))] \\ &\quad = 1 - \left[\mathbb{I}_{\{f(\mathbf{x})=1\}} \cdot y_1(\mathbf{x}) + \mathbb{I}_{\{f(\mathbf{x})=0\}} \cdot (1 - y_1(\mathbf{x}))\right] \end{aligned}$$

Dado esto, tenemos:

$$\begin{aligned}
&\mathbb{P}\{f(X) \neq Y | X = \mathbf{x}\} - \mathbb{P}\{f^*(X) \neq Y | X = \mathbf{x}\} \\
&\quad = y_1(\mathbf{x}) \left[\mathbb{I}_{\{f^*(\mathbf{x})=1\}} - \mathbb{I}_{\{f(\mathbf{x})=1\}}\right] + (1 - y_1(\mathbf{x})) \left[\mathbb{I}_{\{f^*(\mathbf{x})=0\}} - \mathbb{I}_{\{f(\mathbf{x})=0\}}\right] \\
&\quad = (2y_1(\mathbf{x}) - 1) \left[\mathbb{I}_{\{f^*(\mathbf{x})=1\}} - \mathbb{I}_{\{f^*(\mathbf{x})=0\}}\right] \geq 0
\end{aligned}$$

□

Integrando respecto de $\mu(dx)$,

$$\begin{aligned}
R(f) &= \int_{\mathbf{x} \in \mathbb{R}^d} \mathbb{P}\{f(X) \neq Y | X = \mathbf{x}\} \mu(dx) \\
&= 1 - \mathbb{E}\{\mathbb{I}_{\{f(X)=1\}} y_1(X) + \mathbb{I}_{\{f(X)=0\}} (1 - y_1(X))\}
\end{aligned} \tag{3.5}$$

y en el caso del clasificador de Bayes f^*,

$$\begin{aligned}
R^* = R(f^*) &= 1 - \mathbb{E}\{\mathbb{I}_{\{y_1(X) \geq 1/2\}} y_1(X) + \mathbb{I}_{\{y_1(X) < 1/2\}} (1 - y_1(X))\} \\
&= 1 - \mathbb{E}\{\text{máx}(y_1(X), 1 - y_1(X))\} \\
&= \mathbb{E}\{\text{mín}(y_1(X), 1 - y_1(X))\}
\end{aligned} \tag{3.6}$$

En todos estos cálculos, el clasificador de Bayes depende de la distribución del par (X, Y), por tanto, si esa distribución es desconocida (situación en la que nos encontraremos en la mayoría de los casos), f^* será desconocida.

Para construir nuestro modelo, recurriremos entonces a utilizar un conjunto de datos ya observados en el pasado que tengan la misma distribución que (X, Y).

3.2. La minimización del riesgo empírico

Como ya hemos mencionado en la sección 3.1, utilizaremos pares de datos (X_i, Y_i), $1 \leq i \leq n$, ya observados que serán variables i.i.d. con la misma distribución de (X, Y), denominado conjunto de entrenamiento.

Al proceso de construcción de una función a partir de los datos ya observados se le denomina *aprendizaje supervisado*. Dicho clasificador será

$$f_n(X; X_1, Y_1, ..., X_n, Y_n) \tag{3.7}$$

y su probabilidad de error vendrá dada por

$$R_n = R(f_n) = \mathbb{P}\{f_n(X; X_1, Y_1, ..., X_n, Y_n) \neq Y | X_1, Y_1, ..., X_n, Y_n\} \tag{3.8}$$

que será una variable aleatoria por depender de los datos (X_i, Y_i).

Buscaremos estudiar el valor

$$\mathbb{E}R_n = \mathbb{P}\{f_n(X) \neq Y\} \tag{3.9}$$

que indicará la calidad de la secuencia de datos y que será especialmente útil cuando R_n esté concentrada alrededor de su media con una gran probabilidad.

Sea $\mathcal{C}$ un conjunto de clasificadores $f : \mathbb{R}^d \to \{0, 1\}$, el objetivo es hallar una función con una pequeña probabilidad de error. En el caso de que la distribución sea desconocida, será necesario recurrir a la estimación de las probabilidades de error de cada f a partir de los datos ya obtenidos. Una de las opciones más naturales es el denominado *error empírico de f*, dado por:

$$\hat{R}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{f(X_i) \neq Y_i\}}. \tag{3.10}$$

que se emplea la llamada función de pérdida $0 - 1$ por valer 0 en caso de clasificar bien y 1 en caso contrario. Veremos más adelante que en la práctica se utilizan otras funciones de pérdida que sean convexas y que permitirán resolver problemas de optimización.

Un buen método para buscar un buen clasificador sería hallar uno con una probabilidad de error próxima a la mínima probabilidad de error en la clase $\mathcal{C}$. Si el error de generalización está uniformemente bien estimado en la clase $\mathcal{C}$, el clasificador que minimice el error empírico tendrá un error de generalización próximo al óptimo en la clase.

Denominaremos f_n^* al clasificador que minimiza la probabilidad de error empírico en $\mathcal{C}$, es decir, se cumplirá

$$\hat{R}_n(f_n^*) \leq \hat{R}_n(f) \qquad \forall f \in \mathcal{C} \tag{3.11}$$

El siguiente lema establecerá cotas superiores para el error de dicho clasificador.

3.3 Lema.

$$R(f_n^*) - \inf_{f \in \mathcal{C}} R(f) \leq 2 \sup_{f \in \mathcal{C}} |\hat{R}_n(f) - R(f)| \tag{3.12a}$$

$$|\hat{R}_n(f_n^*) - R(f_n^*)| \leq \sup_{f \in \mathcal{C}} |\hat{R}_n(f) - R(f)| \tag{3.12b}$$

Demostración.

$$\begin{aligned} R(f_n^*) - \inf_{f\in\mathcal{C}} R(f) &= R(f_n^*) - \hat{R}_n(f_n^*) + \hat{R}_n(f_n^*) - \inf_{f\in\mathcal{C}} R(f) \\ &\leq R(f_n^*) - \hat{R}_n(f_n^*) + \sup_{f\in\mathcal{C}} |\hat{R}_n(f) - R(f)| \\ &\leq 2\sup_{f\in\mathcal{C}} |\hat{R}_n(f) - R(f)| \end{aligned}$$

La segunda ecuación es trivial.

□

De estas ecuaciones se obtiene una cota para la suboptimalidad de f_n^* en $\mathcal{C}$, es decir, una cota para la distancia entre la menor probabilidad de error empírico y la menor probabilidad de error real en $\mathcal{C}$ y otra para la distancia entre el error empírico y el error real del clasificador f_n^*. Por lo tanto, será necesario controlar el tamaño de $\mathcal{C}$, para que sea por un lado suficientemente grande cómo para que el clasificador que minimice el riesgo en $\mathcal{C}$ no esté muy alejado del óptimo(el clasificador de Bayes) y a la vez no excesivamente grande como para que el clasificador que minimice el riesgo empírico pueda estar muy lejos del clasificador que minimiza en la clase.

Salvo el factor multiplicativo $1/n$ en la variable $\hat{R}_n$, tendremos una suma de n ensayos independientes entre sí cuya probabilidad de que el valor de cada ensayo sea 1 viene dada por $R(f)$. Por este motivo, podemos concluir que la variable aleatoria $n\hat{R}_n(f)$ sigue una distribución binomial con parámetros $(n, R(f))$.

En la próxima sección estudiaremos las desviaciones uniformes de estas variables respecto a su media para comprobar que con gran probabilidad se distribuyen alrededor de su media y obtener unas cotas útiles.

3.3. Desigualdades de concentración

En esta sección y la siguiente estudiaremos desigualdades que permiten acotar la probabilidad de que los términos en el lado derecho de (3.12) tomen valores grandes. Los sumandos en $R_n(f)$ son acotados y se pueden tratar con la desigualdad de Hoeffding que se presenta en la subsección 3.3.1 y la desigualdad de diferencias acotadas, que se presenta en la subsección 3.3.3. La subsección 3.3.2 presenta variantes de la desigualdad de Hoeffding que tiene en cuenta la variabilidad de estos sumandos.

3.3.1. El método de Chernoff

Por la desigualdad de *Markov*, dada X una variable aleatoria y s un número positivo arbitrario, se verifica:

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{e^{sX} \geq e^{st}\} \leq \frac{\mathbb{E}e^{sX}}{e^{st}} \tag{3.13}$$

El método de *Chernoff* consiste en hallar el valor de s que minimiza la cota superior. En el caso de una suma de variables aleatorias independientes, por (3.13) se verifica

$$\begin{aligned}\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} &\leq e^{-st}\mathbb{E}\left\{e^{s\sum_{i=1}^n (X_i - \mathbb{E}X_i)}\right\} \\ &= e^{-st}\prod_{i=1}^n \mathbb{E}\left\{e^{s(X_i-\mathbb{E}X_i)}\right\} \qquad \text{(por independencia)} \end{aligned} \tag{3.14}$$

Y entonces, será necesario buscar una cota para la función generadora de momentos de $X_i - \mathbb{E}X_i$.

3.4 Lema. Sea X una variable aleatoria con $\mathbb{E}X = 0$, $a \leq X \leq b$. Entonces para todo $s > 0$,

$$\mathbb{E}\{e^{sX}\} \leq e^{s^2(b-a)^2/8} \tag{3.15}$$

Demostración. Por la convexidad de la función exponencial, se cumple

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa} \qquad \text{para } a \leq x \leq b$$

Utilizando que $\mathbb{E}X = 0$ y utilizando la notación $p = -a/(b-a)$,

$$\begin{aligned}\mathbb{E}e^{sX} &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= \left(1 - p + pe^{s(b-a)}\right)e^{-ps(b-a)} \\ &\stackrel{\text{def}}{=} e^{\phi(u)},\end{aligned}$$

donde $u = s(b-a)$ y $\phi(u) = -pu + \log(1-p+pe^u)$. Calcularemos el desarrollo de Taylor y para ello veremos las derivadas primera y segunda de ϕ

$$\begin{aligned}\phi'(u) &= -p + \frac{p}{p + (1-p)e^{-u}} \\ \phi''(u) &= \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2} \leq \frac{1}{4}\end{aligned}$$

Evaluando en $u = 0$ tendremos que $\phi(0) = \phi'(0) = 0$, y por tanto, para algún $\theta \in [0, u]$, el desarrollo de Taylor

$$\phi(u) = \phi(0) + u\phi'(u) + \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}$$

□

Continuando con el método de Chernoff, en (3.14), tendremos:

$$\begin{aligned}
\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} &\leq e^{-st} \prod_{i=1}^{n} \mathbb{E}\left\{e^{s(X_i - \mathbb{E}X_i)}\right\} \\
&\leq e^{-st} \prod_{i=1}^{n} e^{s^2(b_i-a_i)^2/8} \qquad \text{(por el Lema 3.4)} \\
&= e^{-st} e^{s^2 \sum_{i=1}^{n}(b_i-a_i)^2/8} \qquad (3.16)
\end{aligned}$$

que alcanzará su mínimo cuando $s = \frac{4t}{\sum_{i=1}^{n}(b_i-a_i)^2}$, y por tanto se cumplirá

$$\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} \leq e^{-2t^2/\sum_{i=1}^{n}(b_i-a_i)^2} \qquad (3.17a)$$

y también se verificará

$$\mathbb{P}\{S_n - \mathbb{E}S_n \leq -t\} \leq e^{-2t^2/\sum_{i=1}^{n}(b_i-a_i)^2} \qquad (3.17b)$$

Estas son las desigualdades de Hoeffding, que aplicadas a la distribución binomial, se obtiene:

$$\mathbb{P}\{S_n/n - p \geq t\} \leq e^{-2nt^2} \qquad (3.18a)$$

y por tanto, uniendo ambas desigualdades se concluye que

$$\mathbb{P}\{|S_n/n - p| \geq t\} \leq 2e^{-2nt^2} \qquad (3.18b)$$

La desigualdad (3.18b) es una desigualdad de concentración. Nos dice que la probabilidad de que una variable se separe mucho de su valor central es pequeña. Se puede emplear la desigualdad para obtener desigualdades maximales, es decir, desigualdades que dicen que el superior de una familia de variables aleatorias es grande con poca probabilidad o que es pequeño en valor esperado, como se ve en las desigualdades (3.19), (3.20a) y (3.20b) a continuación.

Aplicando la desigualdad (3.18b) a la minimización del riesgo empírico en el caso de que la clase $\mathcal{C}$ tenga un cardinal finito, se verificará el siguiente teorema:

3.5 Teorema. Supongamos que el cardinal de $\mathcal{C}$ está acotado por N. Entonces, para todo $t > 0$,

$$\mathbb{P}\left\{\sup_{f\in\mathcal{C}} |\hat{R}_n(f) - R(f)| > t\right\} \leq 2Ne^{-2nt^2} \tag{3.19}$$

3.6 Lema. Sea $\sigma > 0$, $n \geq 2$, y sean $Y_1, ..., Y_n$ variables aleatorias reales tales que para todo $s > 0$, se verifica que $\mathbb{E}\{e^{sY_i}\} \leq e^{s^2\sigma^2/2}$. Entonces, se cumple

$$\mathbb{E}\left\{\max_{i\leq n} Y_i\right\} \leq \sigma\sqrt{2\ln n} \tag{3.20a}$$

Si además, también se cumple $\mathbb{E}\{e^{s(-Y_i)}\} \leq e^{s^2\sigma^2/2}$, entonces

$$\mathbb{E}\left\{\max_{i\leq n} |Y_i|\right\} \leq \sigma\sqrt{2\ln(2n)} \tag{3.20b}$$

Demostración. Por la desigualdad de Jensen,

$$e^{s\mathbb{E}\{\max_{i\leq n} Y_i\}} \leq \mathbb{E}\{e^{s\max_{i\leq n} Y_i}\} = \mathbb{E}\left\{\max_{i\leq n} e^{sY_i}\right\} \leq \sum_{i=1}^{n} \mathbb{E}\{e^{sY_i}\} \leq ne^{s^2\sigma^2/2}$$

Por tanto, para todo $s > 0$,

$$\mathbb{E}\left\{\max_{i\leq n} Y_i\right\} \leq \frac{\ln n}{s} + \frac{s\sigma^2}{2}$$

y minimizando, para $s = \sqrt{2\ln n/\sigma^2}$ se obtiene la desigualdad de (3.20a). Para la cota (3.20b), sólo es necesario aplicar (3.20a) teniendo en cuenta que $\max_{i\leq n} |Y_i| = \max(Y_1, -Y_1, ..., Y_n, -Y_n)$.

□

Aplicando este resultado a nuestra variable $\sup_{f\in\mathcal{C}} |\hat{R}_n(f) - R(f)|$, obtenemos que

$$\mathbb{E}\sup_{f\in\mathcal{C}} |\hat{R}_n(f) - R(f)| \leq \sqrt{\frac{\ln(2N)}{2n}} \tag{3.21}$$

3.3.2. Otras desigualdades

Aunque no demostraremos los siguientes resultados, hay variantes a la desigualdad de Hoeffding que sí tienen en cuenta la variabilidad de los sumandos en S_n. Incluimos dos ejemplos a continuación, que se pueden demostrar usando también el método de Chernoff.

3.7 Teorema. La desigualdad de Bennet: Sean $X_1, ..., X_n$ variables aleatorias reales e independientes con media 0 y $|X_i| \leq c$ con probabilidad 1. Sea $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Var}\{X_i\}$. Entonces

$$\mathbb{P}\{S_n > t\} \leq \exp\left(-\frac{n\sigma^2}{c^2} h\left(\frac{ct}{n\sigma^2}\right)\right) \tag{3.22}$$

donde la función h está definida por $h(u) = (1+u)\log(1+u) - u$ para $u \geq 0$.

3.8 Teorema. La desigualdad de Bernstein: En las condiciones del teorema anterior y para todo $t > 0$,

$$\mathbb{P}\{S_n > t\} \leq \exp\left(-\frac{t^2}{2n\sigma^2 + 2ct/3}\right) \tag{3.23}$$

3.3.3. La desigualdad de las diferencias acotadas

Sea A un conjunto y tomamos $f : A^n \to \mathbb{R}$ una función medible de n variables. Desarrollaremos desigualdades para la diferencia entre $f(X_1, ..., X_n)$ y su valor esperado si $X_1, ..., X_n$ son variables aleatorias independientes con valores en A.

Estas desigualdades se obtendrán mediante un refinamiento del método de Chernoff y de su aplicación para la desigualdad de Hoeffding. La desigualdad que vamos a presentar no requiere que tratemos con funciones de tipo suma, $f(X_1, ..., X_n)$ puede ser de cualquier tipo con tal de que $X_1, ..., X_n$ sean variables aleatorias independientes y de que f satisfaga cierta condición de incrementos. Bajo tales condiciones, veremos que $f(X_1, ..., X_n)$ está exponencialmente concentrada en torno a su valor central.

La condición de incrementos se satisface por muchas funciones. En particular, es válida para el supremo del proceso empírico considerado en secciones anteriores.

Sea $f : A^n \to \mathbb{R}$ una función que verifique la hipótesis de las diferencias acotadas, es decir,

$$\sup_{\substack{x_1,\ldots,x_n,\\ x_i' \in A}} |f(x_1,\ldots,x_n) - f(x_1,\ldots,x_i',\ldots,x_n)| \leq c_i, \qquad 1 \leq i \leq n, \tag{3.24}$$

o lo que es lo mismo, que si cambiamos la variable i de f y mantenemos el resto fijas, el cambio que tiene el valor de la función no es mayor que c_i.

3.9 Teorema. La desigualdad de las diferencias acotadas o desigualdad de McDiarmid: Sean $X_1,\ldots,X_n$ variables aleatorias independientes. Supongamos que se verifica la hipótesis de las diferencias acotadas (3.24). Entonces, para todo $t > 0$,

$$\mathbb{P}\{f(X_1,\ldots X_n) - \mathbb{E}f(X_1,\ldots X_n) \geq t\} \leq e^{-2t^2/\sum_{i=1}^n c_i^2} \tag{3.25a}$$

y

$$\mathbb{P}\{\mathbb{E}f(X_1,\ldots X_n) - f(X_1,\ldots X_n) \geq t\} \leq e^{-2t^2/\sum_{i=1}^n c_i^2} \tag{3.25b}$$

La demostración del Teorema 3.9 utiliza el siguiente lema, que es una extensión del lema 3.4.

3.10 Lema. Sean V y Z dos variables aleatorias tales que $\mathbb{E}\{V|Z\} = 0$ con probabilidad uno, y para alguna función f y alguna constante $c \geq 0$, se cumple

$$f(Z) \leq V \leq f(Z) + c \tag{3.26}$$

Entonces, para todo $s > 0$

$$\mathbb{E}\{e^{sV}|Z\} \leq e^{s^2c^2/8} \tag{3.27}$$

Demostración (del Teorema 3.9). Sea $V = f - \mathbb{E}f$ y definimos para cada $i = 1,\ldots,n$, $V_i = \mathbb{E}(f|X_1,\ldots,X_i) - \mathbb{E}(f|X_1,\ldots X_{i-1})$. Se verifica

$$V = \sum_{i=1}^n V_i$$

Primero comprobaremos que cada V_i cumple (3.26), siendo $c = c_i$ y $Z = (X_1,\ldots,X_{i-1})$. Se verifica que $\mathbb{E}\{V_i|X_1,\ldots X_{i-1}\} = 0$. Para comprobar la otra hipótesis, definiremos primero

$$L_i = \inf_x \mathbb{E}\{f|X_1,\ldots,X_{i-1},x\} - \mathbb{E}\{f|X_1,\ldots,X_{i-1}\}$$

$$U_i = \sup_x \mathbb{E}\{f|X_1, ..., X_{i-1}, x\} - \mathbb{E}\{f|X_1, ..., X_{i-1}\}$$

Es fácil ver que $L_i \leq V_i \leq U_i$, y habrá que hallar una cota para la diferencia entre U_i y L_i.

$$\begin{aligned}
U_i - L_i &= \sup_x \mathbb{E}\{f|X_1, ..., X_{i-1}, x\} - \operatorname{ínf}_x \mathbb{E}\{f|X_1, ..., X_{i-1}, x\} \\
&= \sup_x \int f(X_1, ..., X_{i-1}, x, x_{i+1}, ..., x_n) dP(x_{i+1}, ..., x_n) \\
&\quad - \operatorname{ínf}_x \int f(X_1, ..., X_{i-1}, x, x_{i+1}, ..., x_n) dP(x_{i+1}, ..., x_n) \\
&= \sup_{x,y} \int f(X_1, ..., X_{i-1}, x, x_{i+1}, ..., x_n) \\
&\quad - f(X_1, ..., X_{i-1}, y, x_{i+1}, ..., x_n) dP(x_{i+1}, ..., x_n) \\
&\leq c_i
\end{aligned}$$

Finalmente, aplicando el método de Chernoff:

$$\begin{aligned}
\mathbb{P}\{f - \mathbb{E}f \geq t\} = \mathbb{P}\{\sum_{i=1}^{n} V_i \geq t\} &= \mathbb{P}\{e^{s\sum_{i=1}^n V_i} \geq e^{st}\} \\
&\leq e^{-st}\mathbb{E}\{e^{s\sum_{i=1}^n V_i}\} \\
&= e^{-st}\mathbb{E}\left\{e^{s\sum_{i=1}^{n-1} V_i}\mathbb{E}\{e^{sV_n}|X_1, ..., X_n - 1\}\right\} \\
&\leq e^{-st}e^{s^2c_n^2/8}\mathbb{E}\{e^{s\sum_{i=1}^{n-1} V_i}\} \\
&\leq e^{-st}e^{s^2\sum_{i=1}^n c_i^2/8} \qquad \text{(por iteración)}
\end{aligned}$$

La primera desigualdad se verifica tomando $s = 4t/\sum_{i=1}^n c_i^2$. La segunda se cumple por simetría.

□

En el caso de la variable $\sup_{f\in\mathcal{C}} |\hat{R}_n(f) - R(f)|$ como una función de n pares aleatorios independientes $(X_i, Y_i),\ i = 1, ..., n$, la hipótesis de las diferencias divididas se verificará para $c_i = 1/n$ y el teorema anterior verificará

$$\mathbb{P}\left\{\left|\sup_{f\in\mathcal{C}} |\hat{R}_n(f) - R(f)| - \mathbb{E}\sup_{f\in\mathcal{C}} |\hat{R}_n(f) - R(f)|\right| > t\right\} \leq 2e^{-2nt^2} \qquad (3.28)$$

De esta desigualdad podemos concluir que sea cual sea su valor esperado, la variable $\sup_{f\in\mathcal{C}} |\hat{R}_n(f) - R(f)|$ está concentrada alrededor de su media con una gran probabilidad. En la próxima sección estudiaremos el valor esperado a través de las desigualdades maximales.

3.4. Desigualdades maximales. La teoría de Vapnik-Chervonenkis

A lo largo del capítulo ya hemos visto

$$\mathbb{P}\left\{\sup_{f\in\mathcal{C}}|\hat{R}_n(f)-R(f)|>\epsilon\right\}\leq 2Ne^{-2n\epsilon^2} \tag{3.29}$$

y

$$\mathbb{E}\sup_{f\in\mathcal{C}}|\hat{R}_n(f)-R(f)|\leq\sqrt{\frac{\ln(2N)}{2n}} \tag{3.30}$$

Pero en el caso de que N tome valores muy grandes o que el cardinal de $\mathcal{C}$ sea infinito no son unas cotas útiles.

Introduciremos en esta sección la teoría de Vapnik-Chervonenkis que nos permitirá acotar estos valores en las situaciones antes planteadas.

Sean $X_1,...,X_n$ variables aleatorias independientes e igualmente distribuidas con valores en $\mathbb{R}^d$ con distribución

$$\mu(A)=\mathbb{P}\{X_1\in A\}\qquad(A\subset\mathbb{R}^d) \tag{3.31}$$

Se define la distribución empírica como

$$\mu_n(A)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{[X_i\in A]}\qquad(A\subset\mathbb{R}^d) \tag{3.32}$$

Sea $\mathcal{A}$ una clase de subconjuntos de $\mathbb{R}^d$. Por el teorema 3.9 y la desigualdad de las diferencias acotadas, hemos visto:

$$\mathbb{P}\left\{\left|\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|-\mathbb{E}\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|\right|>t\right\}\leq 2e^{-2nt^2} \tag{3.33}$$

Introduciremos un término nuevo, que nos permitirá dar una cota para $\mathbb{E}\{\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|\}$.

3.11 Definición. Se denomina coeficiente de saturación al máximo número de distintos subconjuntos de un conjunto de n puntos que pueden ser obtenidos por intersecciones suyas con elementos de $\mathcal{A}$:

$$\mathbb{S}_{\mathcal{A}}(n)=\max_{x_1,...,x_n\in\mathbb{R}^d}|\{\{x_1,...,x_n\}\cap A;A\in\mathcal{A}\}| \tag{3.34}$$

A partir de esta definición podemos introducir el siguiente teorema

3.12 Teorema. La desigualdad de Vapnik-Chervonenkis. En las condiciones anteriormente descritas,

$$\mathbb{E}\left\{\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|\right\}\leq 2\sqrt{\frac{\log 2\mathbb{S}_{\mathcal{A}}(n)}{n}} \tag{3.35}$$

Demostración. Introduzcamos una copia $X_1',...,X_n'$ que sea independiente de las variables $X_1,...,X_n$ y sean n variables de signo $\sigma_1,...,\sigma_n$ i.i.d. tales que se cumpla $\mathbb{P}\{\sigma_1=1\}=\mathbb{P}\{\sigma_1=-1\}=1/2$ y que sean independientes de $X_1,X_1',...,X_n,X_n'$. La distribución empírica respecto a $X_1',...,X_n'$ será $\mu_n'(A)=(1/n)\sum_{i=1}^n\mathbb{I}_{[X_i'\in A]}$. Entonces se cumplirá

$$\begin{aligned}
&\mathbb{E}\left\{\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|\right\}\\
&\quad=\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\mathbb{E}\left\{\mu_n(A)-\mu_n'(A)\middle|X_1,...,X_n\right\}\right|\right\}\\
&\quad\leq\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\mathbb{E}\left\{\left|\mu_n(A)-\mu_n'(A)\right|\middle|X_1,...,X_n\right\}\right\}\\
&\qquad\text{(Por la desigualdad de Jensen)}\\
&\quad\leq\mathbb{E}\left\{\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu_n'(A)|\right\}\\
&\qquad\text{(Por }\sup\mathbb{E}(\cdot)\leq\mathbb{E}\sup(\cdot))\\
&\quad=\frac{1}{n}\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n\sigma_i\left(\mathbb{I}_{[X_i\in A]}-\mathbb{I}_{[X_i'\in A]}\right)\right|\right\}\\
&\qquad\text{(Por ser } X_1,X_1',...,X_n,X_n'\text{i.i.d.)}\\
&\quad=\frac{1}{n}\mathbb{E}\left\{\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n\sigma_i\left(\mathbb{I}_{[X_i\in A]}-\mathbb{I}_{[X_i'\in A]}\right)\right|\middle|X_1,X_1',...,X_n,X_n'\right\}\right\}
\end{aligned}$$

Por la independencia de las variables σ_i's del resto de variables, fijaremos puntos $X_1=x_1,X_1'=x_1',...,X_n=x_n,X_n'=x_n'$, y estudiaremos el valor

$$\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^n\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i'\in A]}\right)\right|\right\}.$$

Sea $\hat{\mathcal{A}}\subset\mathcal{A}$ una colección de conjuntos tales que cualesquiera dos conjuntos de $\hat{\mathcal{A}}$ tienen distintas intersecciones con el conjunto $\{x_1,x_1',...,x_n,x_n'\}$ y además

cada posible intersección está representada una vez. Por tanto $|\hat{\mathcal{A}}| \leq \mathbb{S}_{\mathcal{A}}(2n)$, y

$$\mathbb{E}\left\{\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i'\in A]}\right)\right|\right\}=\mathbb{E}\left\{\max_{A\in\hat{\mathcal{A}}}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i'\in A]}\right)\right|\right\}.$$

Como cada variable aleatoria $\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i'\in A]}\right)$ tiene media 0 y además toma valores en $[-1,1]$, por el lema 3.4, se cumplirá que para cualquier $s>0$,

$$\mathbb{E}e^{s\sum_{i=1}^{n}\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i'\in A]}\right)}=\prod_{i=1}^{n}\mathbb{E}e^{s\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i'\in A]}\right)}\leq e^{ns^2/2}$$

y por tener $\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i'\in A]}\right)$ una distribución simétrica, podemos aplicar el lema 3.6 que implica

$$\mathbb{E}\left\{\max_{A\in\hat{\mathcal{A}}}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i'\in A]}\right)\right|\right\}\leq\sqrt{2n\ln 2\mathbb{S}_{\mathcal{A}}(2n)}$$

Por último, llegaremos a la desigualdad planteada, aplicando que $\mathbb{S}_{\mathcal{A}}(2n)\leq\mathbb{S}_{\mathcal{A}}(n)^2$.

□

Combinando el teorema con la desigualdad de concentración para el supremo, llegamos a

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu(A)|>t\right\}\leq 4\mathbb{S}_{\mathcal{A}}(2n)e^{-nt^2/8} \tag{3.36}$$

3.13 Definición. Llamamos dimensión de Vapnik-Chervonenkis o dimensión VC V de una clase $\mathcal{A}$ de conjuntos al mayor entero n tal que

$$\mathbb{S}_{\mathcal{A}}(n)=2^n \tag{3.37}$$

En el caso de que $\mathbb{S}_{\mathcal{A}}(n)=2^n\ \forall n$, entonces $V=\infty$.

Enunciaremos a continuación un resultado combinatorio no elemental del que omitimos la demostración.

3.14 Lema. Lema de Sauer: Sea $\mathcal{A}$ una clase de conjuntos con dimensión VC, V finita. Entonces para todo n,

$$\mathbb{S}_{\mathcal{A}}(n)\leq\sum_{i=0}^{V}\binom{n}{i} \tag{3.38}$$

De este lema se puede concluir que si $V < \infty$, el valor $\mathbb{S}_{\mathcal{A}}(n)$ crecerá de forma polinómica y se verificará que $\mathbb{S}_{\mathcal{A}}(n) \leq (n+1)^V$.

Utilizando este lema en la desigualdad de Vapnik-Chervonenkis, concluiremos que si $\mathcal{A}$ es una clase de conjuntos con dimensión VC, V finita, entonces

$$\mathbb{E}\left\{\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\right\} \leq 2\sqrt{\frac{V\log(n+1) + \log 2}{n}} \tag{3.39}$$

Es decir que si dado $\mathcal{A}$, $V < \infty$, la esperanza de la mayor de las desviaciones convergerá hacia cero con una tasa $O(\sqrt{\log n/n})$.

3.5. Selección modelos

Supongamos que a partir de un conjunto de entrenamiento se ha elegido una regla $\hat{f}_{\mathcal{A}}$ dentro de la clase $\mathcal{A}$. El exceso de riesgo $\mathscr{E}(\hat{f}_{\mathcal{A}})$, es decir, la diferencia entre el riesgo asociado a $\hat{f}_{\mathcal{A}}$ y el riesgo de Bayes, de una buena regla de clasificación debe de ser pequeño. Veremos que esta cantidad dependerá de dos factores influidos por la complejidad de la clase $\mathcal{A}$, siendo necesario hallar un equilibrio entre ambos términos controlando dicha complejidad. En esta sección estudiamos un método que permite conseguir un equilibrio adecuado. Si entre varios modelos escogiésemos aquel que minimizase el riesgo empírico, nos quedaríamos siempre con el modelo más complejo. Por tanto, añadir una penalización asociada a la complejidad del modelo puede permitir conseguir el equilibrio correcto.

3.5.1. Minimización del riesgo estructural

Sea $\mathcal{A}$ una clase de funciones clasificadoras. Podemos tomar una familia de subconjuntos de $\mathcal{A}$, a los que llamaremos modelos, $\{A_k\}_{k\in\mathcal{K}}$ tales que $\bigcup_{k\in\mathcal{K}} A_k = \mathcal{A}$. Sea V_k la dimensión VC de cada A_k, que mide la complejidad de cada modelo A_k.

Sea f_k^* el clasificador que minimiza el riesgo empírico para cada modelo $A_k, k \in \mathcal{K}$. A medida que la complejidad aumenta, es decir, V_k crece, es claro que el mínimo del riesgo empírico decrece, pero a su vez es más complicado hallar el mínimo.

Dada f_k^* podemos reescribir la diferencia entre su probabilidad de error y la de la regla de Bayes f^* como

$$R(f_k^*) - R(f^*) = \left(R(f_k^*) - \inf_{f \in A_k} R(f) \right) + \left(\inf_{f \in A_k} R(f) - R(f^*) \right) \tag{3.40}$$

de donde vemos que esta cantidad depende de dos factores, que dependerán de la estructura de A_k. Si es grande, el segundo término será pequeño pero el primero puede ser grande. En cambio, si es demasiado pequeño, no podremos asegurar que el error de aproximación $\inf_{f \in A_k} R(f) - R(f^*)$ sea pequeño, por lo que también será necesario controlar la complejidad de A_k.

Por este motivo, introduciremos un término $pen(k)$ que será una función de penalización en función de V_k y de la complejidad del modelo. Por tanto si $V_n < V_m$,

$$pen(n) < pen(m)$$

Sean f_k las reglas que minimizan el error en cada clase A_k, llamaremos regla oráculo a aquella f_k para la cual k minimice la cantidad

$$R(f_k) + pen(k) \tag{3.41}$$

Como no conocemos $R(f)$, usaremos el riesgo empírico para estimar dichas cantidades y llamaremos $\tilde{k}$ al valor de $k \in \mathcal{K}$ que minimice

$$\hat{R}_n(f_k^*) + pen(k) \tag{3.42}$$

y tomaremos el estimador penalizado

$$\tilde{f} = f_{\tilde{k}}^* \tag{3.43}$$

Vamos a probar que, con una adecuada elección del término $pen(k)$, el riesgo promedio de la regla penalizada $\tilde{f}$ no es, esencialmente, peor que la del oráculo.

Se verificarán

$$R_n(\tilde{f}) + pen(\tilde{k}) \leq \hat{R}_n(f_k) + pen(k) \qquad \forall k \in \mathcal{K} \tag{3.44a}$$

$$\hat{R}_n(\tilde{f}) + pen(\tilde{k}) \leq \hat{R}_n(f_{\tilde{k}}) + pen(\tilde{k}) \qquad \forall f_{\tilde{k}} \in A_k \tag{3.44b}$$

$$\hat{R}_n(\tilde{f}) + pen(\tilde{k}) \leq \hat{R}_n(\hat{f}_k) + pen(k) \qquad \forall k \in \mathcal{K} \tag{3.44c}$$

El objetivo será calcular el valor esperado de

$$R(\tilde{f}) + pen(\hat{k}) \tag{3.45}$$

Sumando y restando el término $\hat{R}_n(\tilde{f})$, se tiene

$$\begin{aligned} R(\tilde{f}) + pen(\tilde{k}) &= R(\tilde{f}) - \hat{R}_n(\tilde{f}) + \hat{R}_n(\tilde{f}) + pen(\tilde{k}) \\ &\leq R(\tilde{f}) - \hat{R}_n(\tilde{f}) + \hat{R}_n(f_k) + pen(k) \qquad \text{(por (3.44a))} \end{aligned}$$

y repitiendo el proceso con el término $R(f_k)$,

$$\begin{aligned} R(\tilde{f}) + pen(\tilde{k}) \leq & R(f_k) + pen(k) \\ & + \hat{R}_n(f_k) - R(f_k) \\ & + R(\tilde{f}) - \hat{R}_n(\tilde{f}) \end{aligned} \tag{3.46}$$

Por otro lado, dada una constante Σ, se consideran una serie de pesos no negativos $\{x_k\}_{k\in\mathcal{K}}$ tales que

$$\sum_{k\in\mathcal{K}} e^{-x_k} \leq \Sigma \tag{3.47}$$

y sea $z > 0$ dado. Por la desigualdad de McDiarmid o la desigualdad de las diferencias acotadas (3.28), para $t = \sqrt{\frac{z+x_k}{2n}}$, tendremos que $\forall k$:

$$\mathbb{P}\left\{ \sup_{f\in A_k} |R(f) - \hat{R}_n(f)| - \mathbb{E}\sup_{f\in A_k} |R(f) - \hat{R}_n(f)| \geq \sqrt{\frac{z+x_k}{2n}} \right\} \leq e^{-z}e^{-x_k} \tag{3.48}$$

Llamando $E_k = \mathbb{E}\sup_{f\in A_k} |R(f) - \hat{R}_n(f)|$,

$$\mathbb{P}\left\{ \sup_{f\in A_k} |R(f) - \hat{R}_n(f)| \geq \sqrt{\frac{z+x_k}{2n}} + E_k \right\} \leq e^{-z}e^{-x_k} \tag{3.49}$$

y, cambiando el sentido de la desigualdad:

$$\mathbb{P}\left\{ \sup_{f\in A_k} |R(f) - \hat{R}_n(f)| \leq \sqrt{\frac{z+x_k}{2n}} + E_k \right\} \geq 1 - e^{-z}e^{-x_k} \tag{3.50}$$

Por (3.47), $\Sigma = \sum_k e^{-x_k}$ y $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, por tanto:

$$\mathbb{P}\left\{ \sup_{f\in A_k} |R(f) - \hat{R}_n(f)| \leq \sqrt{\frac{z}{2n}} + \sqrt{\frac{x_{k'}}{2n}} + E_{k'} \right\} \geq 1 - \Sigma e^{-z} \qquad \forall k' \in \mathcal{K} \tag{3.51}$$

Volviendo a la desigualdad (3.46), y despejando $R(\tilde{f})$, tendremos que con probabilidad al menos $1-\Sigma e^{-z}$,

$$R(\tilde{f}) \leq R(f_k)+pen(k)+\hat{R}_n(f_k)-R(f_k)+\sqrt{\frac{z}{2n}}+\sqrt{\frac{x_{k'}}{2n}}+E_{\hat{k}}-pen(\tilde{k}) \quad (3.52)$$

Si consideramos $pen(k)$ de tal forma que sea mayor que $E_k+\sqrt{\frac{x_k}{2n}}$, se cumplirá con probabilidad $1-\Sigma e^{-z}$,

$$R(\tilde{f}) \leq R(f_k)+pen(k)+\hat{R}_n(f_k)-R(f_k)+\sqrt{\frac{z}{2n}} \quad (3.53)$$

Llamaremos $\bar{R}_n(f_k)=\hat{R}_n(f_k)-R(f_k)$, y despejando a un lado de la desigualdad $\sqrt{\frac{z}{2n}}$ y cambiando el sentido, tendremos:

$$\mathbb{P}\left\{R(\tilde{f})-R(f_k)-pen(k)+\bar{R}_n(f_k)>\sqrt{\frac{z}{2n}}\right\} \leq \Sigma e^{-z} \quad (3.54)$$

Podemos acotar el valor esperado de dicha variable por el de la esperanza de la parte positiva ($\mathbb{E}(X) \leq \mathbb{E}(X^+)$. Sea $\tilde{R}=R(\tilde{f})-R(f_k)-pen(k)+\bar{R}_n(f_k)$,

$$\begin{aligned}
\mathbb{E}(\tilde{R}) \leq & \int_0^\infty \mathbb{P}(\tilde{R}>y)dy \\
= & \frac{1}{\sqrt{2n}}\int_0^\infty \mathbb{P}\left\{\tilde{R}>\sqrt{\frac{dz}{2\sqrt{z}}}\right\} \\
\leq & \frac{\Sigma}{2\sqrt{2n}}\int_0^\infty z^{-1/2}e^{-z}dz \\
= & \frac{\Sigma}{2\sqrt{2n}}\Gamma\left(\frac{1}{2}\right)=\frac{\Sigma\sqrt{\pi}}{2\sqrt{2n}}
\end{aligned}$$

haciendo el cambio de variable $y=\sqrt{\frac{z}{2n}}$ y $\Gamma(\frac{1}{2})=\sqrt{\pi}$. Despejando $\mathbb{E}\{R(\tilde{f})\}$, tendremos:

$$\mathbb{E}\{R(\tilde{f})\} \leq \min_k\left(R(f_k)+pen(k)\right)+\frac{\Sigma\sqrt{\pi}}{2\sqrt{2n}} \quad (3.55)$$

Este desarrollo, ha demostrado el siguiente teorema:

3.15 Teorema. Sea una colección de modelos $\{A_k\}_{k\in\mathcal{K}}$. Se considera $\tilde{f} = f^*_{\tilde{k}}$, la función que minimiza el riesgo empírico en el modelo $A_{\tilde{k}}$, cuando $\tilde{k} = \arg\min_{k\in\mathcal{K}}(\hat{R}_n(f^*_k) + pen(k))$. Entonces se verifica:

$$\mathbb{E}\{R(\tilde{f})\} \leq \min_k \left(R(f_k) + pen(k)\right) + \frac{\Sigma\sqrt{\pi}}{2\sqrt{2n}} \tag{3.56}$$

donde Σ es una cota de $\sum_{k\in\mathcal{K}} e^{-x_k}$, $\{x_k\}_{k\in\mathcal{K}}$ es una serie de pesos asociados a cada modelo A_k y

$$pen(k) = E_k + \sqrt{\frac{x_k}{2n}}$$

3.5.2. Aplicación al caso de clasificadores lineales

Veremos en esta sección una aplicación de lo tratado anteriormente para ciertos modelos de clasificación.

Sea el problema planteado al principio del capítulo 3. Sea $\mathbb{R}^d$ el espacio en el que estarán nuestras observaciones $\mathbf{x}$. Cada observación tendrá una clase $y \in \{0, 1\}$. Nuestro objetivo será hallar un clasificador $f : \mathbb{R}^d \to \{0, 1\}$ tal que la probabilidad de que $f(\mathbf{x}) \neq y$ sea pequeña.

Utilizaremos en nuestro caso clasificadores a partir de funciones lineales, es decir, los posibles clasificadores f, serán de la forma:

$$f(\mathbf{x}) = \mathbb{I}(\beta^T\mathbf{x} \leq b) \qquad \text{con } \beta \in \mathbb{R}^d \text{ y } b \in \mathbb{R}. \tag{3.57}$$

Cómo hemos visto en la sección 3.5.1, tomaremos distintas clases de subconjuntos del conjunto de clasificadores posibles. En nuestro caso, tomaremos como distintos modelos de clasificación los conjuntos A_k que sean los subconjuntos de funciones del tipo (3.57) tales que las coordenadas de β serán nulas a partir del término $k + 1$.

De esta forma, la familia de modelos A_k, será una serie de subconjuntos encajados tal que

$$A_1 \subset A_2 \subset ... \subset A_k \subset ... \subset A_d$$

cuyas dimensiones VC asociadas V_k, verifican

$$V_1 < V_2 < ... < V_k < ... < V_d.$$

Hemos visto en la sección de la teoría de Vapnik-Chervonenkis una cota para el valor esperado de la mayor desviación del riesgo empírico. Sea

V_k la dimensión de Vapnik-Chervonenkis para cada conjunto de funciones clasificadoras A_k, por la cota (3.39) anteriormente vista:

$$E_k = \mathbb{E}\left\{\sup_{f\in A_k} |R(f) - \hat{R}_n(f)|\right\} \leq 2\sqrt{\frac{V_k \log(n+1) + \log 2}{n}} \tag{3.58}$$

Trataremos de hallar el valor de V_k para cada conjunto A_k de funciones lineales sobre las k primeras componentes.

3.16 Teorema. Sea $\mathcal{G}$ un espacio vectorial de dimensión m de funciones definidas en $\mathbb{R}^d$ con valores en $\mathbb{R}$. Entonces, la clase de conjuntos

$$\mathcal{A} = \{\{\mathbf{x} : g(\mathbf{x}) \geq 0\}; g \in \mathcal{G}\}$$

tiene dimensión VC $V \leq m$.

Demostración. Basta ver que no hay ningún conjunto de tamaño $m+1$ que pueda ser saturado por conjuntos de la forma $\{\mathbf{x} : g(\mathbf{x}) \geq 0\}$. Si fijamos $m+1$ puntos arbitrarios $\mathbf{x}_1, ..., \mathbf{x}_{m+1}$ y definimos una aplicación lineal $L : \mathcal{G} \to \mathbb{R}^{m+1}$ de la forma

$$L(g) = (g(\mathbf{x}_1), ..., g(\mathbf{x}_{m+1}))$$

La imagen de $\mathcal{G}$ será un subespacio lineal de $\mathbb{R}^{m+1}$ de dimensión no superior a m. Por tanto, existirá un vector no nulo $\gamma = (\gamma_1, ..., \gamma_{m+1}) \in \mathbb{R}^{m+1}$ ortogonal a $L(\mathcal{G})$ que cumplirá para todo $g \in \mathcal{G}$

$$\gamma_1 g(\mathbf{x}_1) + ... + \gamma_{m+1} g(\mathbf{x}_{m+1}) = 0 \tag{3.59}$$

Podemos suponer que al menos una de las componentes del vector γ es negativa. Reajustando la igualdad de forma que todos los términos con γ_i no negativo queden a la izquierda, tendremos

$$\sum_{i:\gamma_i \geq 0} \gamma_i g(\mathbf{x}_i) = \sum_{i:\gamma_i < 0} -\gamma_i g(\mathbf{x}_i)$$

Ahora, supongamos que existe $g \in \mathcal{G}$ tal que el conjunto $\{\mathbf{x} : g(\mathbf{x}) \geq 0\}$ contiene exactamente los $\mathbf{x}_i$ de la izquierda de la igualdad. Entonces todos los términos de la izquierda serán no negativos mientras que los de la derecha deberán ser negativos, lo que es una contradicción, y por tanto queda demostrado que el conjunto de puntos $\mathbf{x}_1, ..., \mathbf{x}_{m+1}$ no puede ser saturado, demostrando el teorema.

□

3.17 Corolario. *Si $\mathcal{A}$ es la clase de todos los subespacios lineales, es decir, subconjuntos de $\mathbb{R}^d$ de la forma $\{\boldsymbol{x} : a^T x \geq b\}$, $\forall a \in \mathbb{R}^d$, $\forall b \in \mathbb{R}$ entonces $V \leq d+1$.*

Además en este caso se dará la igualdad $V = d+1$.

En nuestro caso, cada modelo A_k serán subconjuntos de $\mathbb{R}^d$ y por tanto $V_k = k+1,\ 1 \leq k \leq d$ y podremos acotar cada E_k por

$$E_k \leq 2\sqrt{\frac{(k+1)\log(n+1)+\log 2}{n}} \tag{3.60}$$

De acuerdo con (3.47), vamos a elegir x_k de forma que la serie $\sum_{k\in\mathcal{K}} e^{-x_k}$ sea convergente. Sin pérdida de generalidad, podemos elegir que la serie sume 1 y que todos los pesos sean iguales. En este caso, tomaremos $x_k = \log d$, $1 \leq k \leq d$ y

$$pen(k) = 2\sqrt{\frac{(k+1)\log(n+1)+\log 2}{n}} + \sqrt{\frac{\log d}{2n}} \tag{3.61}$$

Por el teorema 3.15, tendremos

$$\begin{aligned}\mathbb{E}\{R(\tilde{f})\} \leq \min_{1\leq k\leq d}\left(R(f_k) + 2\sqrt{\frac{(k+1)\log(n+1)+\log 2}{n}} + \sqrt{\frac{\log d}{2n}}\right) \\ + \frac{\sqrt{\pi}}{2\sqrt{2n}}\end{aligned} \tag{3.62}$$

En el caso de exista algún k_0 tal que $R(f_{k_0}) = 0$, entonces

$$\begin{aligned}\mathbb{E}\{R(\tilde{f})\} &\leq pen(k_0) + \frac{\sqrt{\pi}}{2\sqrt{2n}} \\ &= 2\sqrt{\frac{(k_0+1)\log(n+1)+\log 2}{n}} + \sqrt{\frac{\log d}{2n}} + \frac{\sqrt{\pi}}{2\sqrt{2n}}\end{aligned} \tag{3.63}$$

una cota que será pequeña aún cuando $d > n$.

Con estas familias anidadas de modelos existe el problema de que si hay buenos clasificadores que se basan sólo unas pocas variables pero hay alguna en una posición avanzada, el modelo tendrá que considerar todas las variables anteriores, aumentando esto la complejidad del modelo.

En cambio, podemos suponer como distintos modelos A_k todos los subconjuntos de $\mathcal{A}$ que representan las distintas posibles combinaciones de considerar o no cada variable. Tendremos entonces $2^d - 1$ modelos diferentes donde

buscar nuestro clasificador. En este caso, para que la serie de pesos x_k sea convergente, podemos tomar $x_k = \log(2^d - 1)$ y en vez de (3.63), tendremos

$$\begin{aligned}\mathbb{E}\{R(\tilde{f})\} &\leq pen(k_0) + \frac{\sqrt{\pi}}{2\sqrt{2n}} \\ &= 2\sqrt{\frac{V_k \log(n+1) + \log 2}{n}} + \sqrt{\frac{\log(2^d - 1)}{2n}} + \frac{\sqrt{\pi}}{2\sqrt{2n}} \end{aligned} \tag{3.64}$$

donde V_k será igual al número de variables consideradas más 1. Para que esta cota siga siendo suficientemente pequeña, se tiene que cumplir que $\log(2^n - 1) < n$.

Capítulo 4

Máquinas de soporte vectorial(SVM)

Las máquinas de soporte vectorial o SVMs (Support Vector Machines) son métodos de aprendizaje basados en la clase de funciones lineales

$$\mathcal{F} = \left\{ f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b, \quad \mathbf{w} \in \mathbb{R}^d, \quad b \in \mathbb{R} \right\} \tag{4.1}$$

El método SVM en el caso de clasificación binaria trata de buscar un hiperplano que permita separar los puntos $\mathbf{x}_i$ de forma que deje a un lado los puntos tales que $y_i = 0$ y al otro los que $y_i = 1$.

Este hiperplano no puede encontrarse siempre puesto que pueden existir conjuntos de puntos en el espacio $\mathbb{R}^d$ que no sean separables por ningún hiperplano. En este caso, la separabilidad puede mejorarse mediante una transformación del espacio de observaciones a otro de alta dimensión. Esta transformación (que puede ser difícil de calcular), puede manejarse de forma implícita a través del método de los núcleos.

Cuando se cumplen las condiciones para que exista un hiperplano que separe los puntos, es probable que existan más de uno, incluso infinitos. El objetivo de las máquinas de soporte vectorial es hallar de entre todos esos posibles hiperplanos, aquel que maximice el margen entre las clases y dicho hiperplano, es decir que la distancia de los puntos de cada clase más cercanos al hiperplano sea máxima. El proceso consistirá en resolver un problema de optimización, del que obtendremos su formulación dual asociada que nos permitirá ver que la solución sólo depende de unos pocos de puntos de todo el conjunto, denominados *vectores soporte*. Este problema se tratará en la sección 4.2.

Frecuentemente no es posible encontrar un hiperplano separante. En este caso se puede redefinir el problema de optimización, permitiendo la presencia de puntos mal clasificados y se buscará que el margen de mala clasificación sea lo menor posible. Este problema resultará equivalente a la minimización del riesgo con una función de pérdida particular y una penalización cuadrática. Esta situación se estudia en la sección 4.3.

Tal como se ha comentado, la separación lineal de un conjunto puede mejorarse mediante transformaciones a espacios de alta dimensión. Una propiedad computacionalmente conveniente es que la transformación se pueda manejar de forma implícita a través del uso de "núcleos". Este aspecto será tratado en la sección 4.4. Para el desarrollo de estas secciones han sido utilizados [4] y [5].

En la sección 4.5, se estudiarán cotas probabilísticas para el error de generalización de reglas de tipo SVM, a partir de ciertos resultados de [6].

Gran parte del desarrollo del capítulo se basa en resultados de optimización, incluyendo un principio de dualidad. Por esta razón, incluimos en la sección 4.1 una introducción teórica sobre estos conceptos.

4.1. Optimización y dualidad

Dado un conjunto Ω y funciones f, g y h, se plantea el problema de optimización

$$\begin{array}{lll} \min & f(w) & w \in \Omega \\ \text{sujeto a} & g_i(w) \leq 0 & i = 1, ..., k \\ & h_i(w) = 0 & i = 1, ..., m \end{array} \tag{4.2}$$

La teoría de Lagrange nos da un resultado en el caso de que las restricciones sean sólo de igualdad. Fueron más tarde Kuhn y Tucker los que desarrollaron una teoría que permitía estudiar las soluciones en el caso de restricciones de desigualdad. Veremos en esta sección estos resultados elementales para la optimización de estos tipos de problemas.

4.1 Definición. Dado un problema de optimización con función objetivo $f(w)$ y restricciones de igualdad $h_i(w) = 0$, $i = 1, ..., m$, se define la función *Lagrangiana* como

$$L(w, \beta) = f(w) + \sum_{i=1}^{m} \beta_i h_i(w) \tag{4.3}$$

y los β_i son llamados *multiplicadores de Lagrange*.

4.2 Teorema. (*Lagrange*) Dado el problema de minimizar $f(w)$ sujeto a que $h_i(w) = 0$, $i = 1, ..., m$, con $f, h_i \in C^1$, una condición necesaria para que w^* sea un mínimo es que se verifiquen

$$\frac{\partial L(w^*, \beta^*)}{\partial w} = 0 \tag{4.4a}$$

$$\frac{\partial L(w^*, \beta^*)}{\partial \beta} = 0 \tag{4.4b}$$

para algunos valores de β^*. Estas condiciones, serán además suficientes cuando $L(w, \beta^*)$ sea una función convexa de w.

Consideremos ahora el caso más general, en el que hay restricciones de igualdad y de desigualdad.

4.3 Definición. Dado el problema de optimización con restricciones de igualdad y desigualdad (4.2) definiremos el *Lagrangiano generalizado* como

$$\begin{aligned} L(w, \alpha, \beta) &= f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{m} \beta_i h_i(w) \\ &= f(w) + \alpha' g(w) + \beta' h(w) \end{aligned} \tag{4.5}$$

4.4 Definición. Dado el problema primal (4.2), definiremos el *problema dual* como

$$\begin{aligned} &\text{maximizar} \quad \theta(\alpha, \beta), \\ &\text{sujeto a} \quad \alpha \geq 0 \end{aligned} \tag{4.6}$$

donde $\theta(\alpha, \beta) = \text{ínf}_{w \in \Omega} L(w, \alpha, \beta)$. Al valor de la función objetivo en la solución óptima se le llama *valor del problema*.

4.5 Teorema. *(Teorema débil de dualidad)* Sea $w \in \Omega$ una solución factible del problema primal (4.2) y (α, β) una solución factible del problema dual (4.6). Entonces

$$f(w) \geq \theta(\alpha, \beta)$$

4.6 Corolario. *Si existen w^* y (α^*, β^*) tales que el valor de los dos problemas es igual, entonces w^* y (α^*, β^*) resuelven los problemas primal y dual respectivamente.*

A la diferencia entre el valor del problema primal y el problema dual se llama *salto de dualidad*. Si no existe esta diferencia, estaríamos en la situación del corolario anterior y tendríamos así soluciones para los dos problemas.

Una forma de detectar la ausencia de salto de dualidad es si existe un punto de silla del Lagrangiano del problema primal, es decir, que existe (w^*, α^*, β^*) que verifica

$$L(w^*, \alpha, \beta) \leq L(w^*, \alpha^*, \beta^*) \leq L(w, \alpha^*, \beta^*)$$

Si existe dicho punto de silla, el valor de los problemas primal y dual será

$$f(w^*) = \theta(\alpha^*, \beta^*)$$

4.7 Teorema. *(Teorema fuerte de dualidad).* Sea el problema de optimización (4.2) con dominio convexo $\Omega \subseteq \mathbb{R}^n$, f convexa, y con g_i y h_i funciones afines, es decir,

$$h(w) = Aw - b$$

para alguna matriz A y algún vector b. Entonces si las regiones primal y dual son factibles, el salto de dualidad es cero y se alcanza el mínimo en el problema primal y el máximo en el dual.

4.8 Teorema. *(Kuhn-Tucker).* Dado el problema de optimización (4.2) con un dominio convexo $\Omega \subseteq \mathbb{R}$, f convexa, y g_i, h_i afines, es condición necesaria y suficiente para que un punto w^* sea un óptimo, es la existencia de α^*, β^* que cumplan

$$\begin{aligned}
\frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial w} &= 0 \\
\frac{\partial L(w^*, \alpha^*, \beta^*)}{\partial \beta} &= 0 \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, ..., k \\
g_i(w^*) &\leq 0, \quad i = 1, ..., k \\
\alpha_i &\geq 0, \quad i = 1, ..., k
\end{aligned}$$

La tercera igualdad $\alpha_i g_i^*(w^*) = 0$, es la denominada condición complementaria de Karush-Kuhn-Tucker, que implica que para las restricciones activas ($g_i(w) = 0$), $\alpha_i^* \geq 0$, mientras que cuando la restricción es inactiva ($g_i(w) < 0$), entonces $\alpha_i^* = 0$.

A menudo el problema primal es difícil de resolver y una buena alternativa puede ser pasar al problema dual, que se obtiene introduciendo los multiplicadores de Lagrange, que serán las variables duales, que serán las principales incógnitas del problema. Para pasar al problema dual, tomaremos las derivadas del Lagrangiano respecto de las variables primales e igualaremos a 0

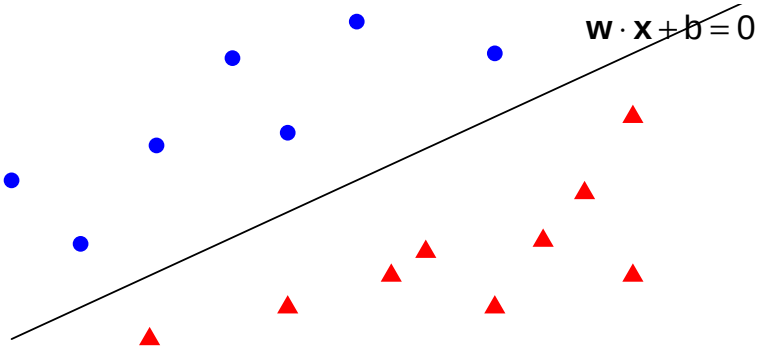

Figura 4.1: Puntos linealmente separables por un plano en $\mathbb{R}^2$.

obteniendo unas relaciones que, sustituyendo en el Lagrangiano nos permitirán eliminar la dependencia del problema respecto a las variables primales, y la función obtenida, será equivalente a calcular

$$\theta(\alpha, \beta) = \inf_{w \in \Omega} L(w, \alpha, \beta)$$

que contendrá solo las variables duales y deberá ser maximizada bajo restricciones más simples. Por la condición complementaria de Karush-Kuhn-Tucker (KKT), las únicas variables duales no nulas serán las relacionadas con la restricción $g(w) = 0$, por lo que tomando sólo los puntos para los que $\alpha_i > 0$, se puede reducir el número de vectores utilizados para la computación. Está técnica de resolver problemas de optimización mediante el paso al dual será utilizada en la teoría de las Máquinas de Soporte Vectorial (SVM).

4.2. SVMs para conjuntos separables

4.9 Definición. Un conjunto de puntos $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ con $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$, se denomina separable si existe, al menos, un hiperplano en $\mathbb{R}^d$ que separa al conjunto de puntos $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ en una región que contiene a los puntos con etiqueta $y_i = 0$ y en otra con los puntos $y_1 = 1$.

En función de la definición anterior, dado un conjunto separable, tendremos al menos un hiperplano de la forma

$$\pi : \mathbf{w} \cdot \mathbf{x} + b = 0 \tag{4.7}$$

que separará los vectores $\mathbf{x}_i$ según su clase como en la figura 4.1.

Los métodos SVM tratan de buscar de entre todos estos hiperplanos existentes aquel que haga máxima la distancia de separación con los dos subconjuntos.

Dado el hiperplano, se puede realizar un reescalado de $\mathbf{w}$ y b de tal forma que la mínima separación entre los vectores $\mathbf{x}_i$ de cada clase y el hiperplano sea 1

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1 \qquad \text{si } y_i = 1 \tag{4.8a}$$
$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 \qquad \text{si } y_i = 0 \tag{4.8b}$$

que se pueden reagrupar en una simple inecuación

$$(2y_i - 1)(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \qquad i = 1, ..., n \tag{4.8c}$$

La distancia de los puntos $\mathbf{x}_i$ al plano $\mathbf{w} \cdot \mathbf{x}_i + b = 0$ vendrá dada por

$$d(\mathbf{x}_i, \pi) = \frac{|\mathbf{w} \cdot \mathbf{x}_i + b|}{\|\mathbf{w}\|}$$

El margen estará definido por la distancia de los puntos de cada clase más cercanos al hiperplano, que serán aquellos que verifiquen las igualdades (4.8) y que estarán a una distancia $1/\|\mathbf{w}\|$, siendo $2/\|\mathbf{w}\|$ la distancia entre las dos clases. Como el objetivo es localizar el hiperplano tal que este margen sea máximo, la SVM tratará de resolver el problema de optimización consistente en

$$\begin{aligned} &\underset{\mathbf{w} \in \mathbb{R}^d}{\text{maximizar}} && \frac{1}{\|\mathbf{w}\|} \\ &\text{sujeto a} && (2y_i - 1)(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \end{aligned} \tag{4.9}$$

que será equivalente a

$$\begin{aligned} &\underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimizar}} && \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{sujeto a} && (2y_i - 1)(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \end{aligned} \tag{4.10}$$

El problema (4.10) se podrá resolver por el método de los multiplicadores de Lagrange descrito en la sección 4.1. Trataremos de hallar el problema dual. Para ello, observamos que el Lagrangiano del problema es,

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[(2y_i - 1)(w \cdot \mathbf{x}_i + b) - 1\right] \tag{4.11}$$

con $\alpha_i \geq 0$, los multiplicadores de Lagrange. Derivaremos el Lagrangiano respecto a las variables primales $\mathbf{w}, b$ e igualaremos a 0

$$0 = \frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i (2y_i - 1)\mathbf{x}_i \tag{4.12a}$$

$$0 = \frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i (2y_i - 1) \tag{4.12b}$$

y reemplazando las relaciones que hemos obtenido de las variables primales en el Lagrangiano, obtenemos:

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[(2y_i - 1)(\mathbf{w} \cdot \mathbf{x}_i + b) - 1\right] \\ &= \frac{1}{2} \sum_{i,j=1}^{n} z_i z_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j - \sum_{i,j=1}^{n} z_i z_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^{n} \alpha_i \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} z_i z_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \end{aligned} \tag{4.13}$$

siendo $z_i = (2y_i - 1)$ para simplificar la notación y transformando las dos posibles clases a $\{-1, 1\}$.

El problema dual será

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximizar}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} z_i z_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \\ \text{sujeto a} \quad & \sum_{i=1}^{n} \alpha_i z_i = 0 \\ & \alpha_i \geq 0, \quad i = 1, ..., n \end{aligned} \tag{4.14}$$

La versión dual del problema será un problema de optimización cuadrática que dependerá sólo de las etiquetas y_1 y de los productos escalares $\mathbf{x}_i \cdot \mathbf{x}_j$ y su solución será el vector $\boldsymbol{\alpha}^*$. Por (4.12a), la solución al problema primal $\mathbf{w}$, podremos escribirla como $\mathbf{w} = \sum_{i=1}^{n} z_i \alpha_i^* \mathbf{x}_i$ y podemos obtener

$$b^* = -\frac{\text{máx}_{y_i=0}(\mathbf{w}^* \cdot \mathbf{x}_i) + \text{mín}_{y_i=1}(\mathbf{w}^* \cdot \mathbf{x}_i)}{2} \tag{4.15}$$

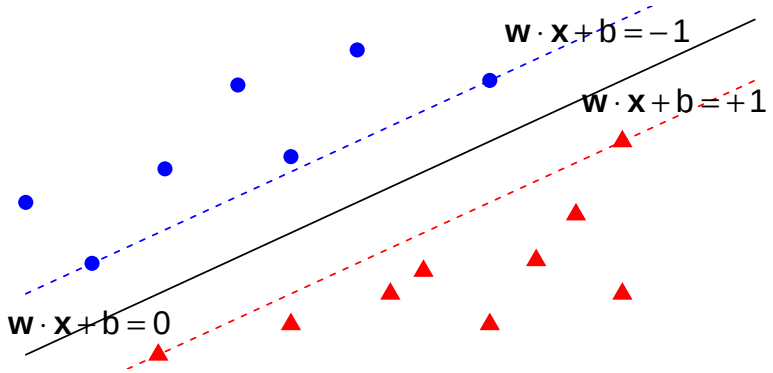

Figura 4.2: Planos que indican el clasificador óptimo y los márgenes en un conjunto de puntos separables en $\mathbb{R}^2$.

La condición complementaria KKT será

$$\alpha_i^* \left[z_i(\mathbf{w}^* \cdot \mathbf{x}_i + b) - 1 \right] = 0, \qquad i = 1, ..., n \tag{4.16}$$

que podemos interpretar como que los valores α_i^* no nulos, se corresponderán con los vectores $\mathbf{x}_i$ cuya distancia al hiperplano separador es 1, y serán los únicos puntos influyentes en el cálculo del plano separador.

En el caso de la clasificación , se tomará como función clasificadora $h(\mathbf{x}) = \mathbb{I}(f(\mathbf{x}) \geq 0)$ en el caso de que las etiquetas tomen valores 0 o 1 (o $h(\mathbf{x}) = \text{sgn}(f(\mathbf{x}))$ si $y \in \{-1, 1\}$), donde

$$f(\mathbf{x}) = \sum_{i \in \mathsf{vs}} z_i \alpha_i^* \mathbf{x}_i \cdot \mathbf{x} + b^*$$

donde vs representa el conjunto de índices de los vectores soporte.

En la figura 4.2 podemos ver el plano óptimo que separa las dos clases de puntos y los hiperplanos que indican el margen entre las clases y el separador óptimo que vendrán dados por los vectores soporte.

4.3. SVMs para conjuntos no separables

En la mayoría de los casos el conjunto de datos de entrenamiento no será separable, es decir, existirán puntos $\mathbf{x}_i$ que no verificarán las condiciones de separabilidad, puede ser porque el punto esté en la región correspondiente

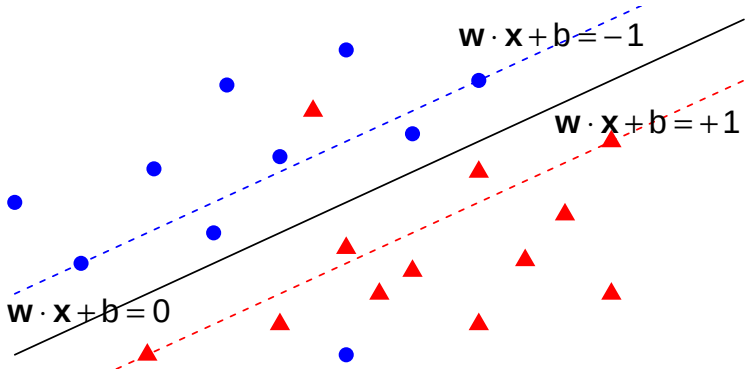

Figura 4.3: Ejemplo de hiperplano clasificador donde algunos puntos no cumplen la condición del margen o están mal clasificados.

a su clase pero que no cumpla la distancia a la que se encuentra del hiperplano, o que esté en la región de la clase contraria (véase la figura 4.3 como ejemplo). En estos casos no existe una solución factible para el problema (4.9). En esta situación, se introducirán unas *variables de holgura* ξ_i y nuevas restricciones

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1 - \xi_i \qquad \text{si } y_i = 1 \tag{4.17a}$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 + \xi_i \qquad \text{si } y_i = 0 \tag{4.17b}$$

que permiten la existencia de algunos datos que no verifiquen que la distancia sea 1 o incluso que estén en la región delimitada por el hiperplano incorrecta. Como en el caso anterior, podemos reunir las restricciones (4.17) en una sola tal que

$$(2y_i - 1)(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \qquad \xi_i \geq 0, \quad i = 1, ..., n \tag{4.17c}$$

siendo $\sum_{i=1}^{n} \xi_i$ un valor que permite medir el coste en función del número de observaciones mal clasificadas y por tanto será necesario añadir al problema de optimización un término que penalice la cantidad de errores, planteando el problema

$$\begin{aligned} \underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimizar}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i \\ \text{sujeto a} \quad & (2y_i - 1)(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{4.18}$$

donde C será un parámetro que controlará la importancia relativa de minimizar la norma de $\mathbf{w}$ y de incumplir las restricciones del margen para cada

punto. Si C es pequeño, los puntos que no satisfagan la restricción del margen influirán poco. En cambio, si C crece, el peso de los errores será mayor y por tanto el margen será menor pero habrá menos puntos que no cumplan las restricciones. En el caso de que $C = \infty$, estaremos ante el problema de clasificación sin errores.

El problema (4.18) admite una formulación equivalente que nos permitirá interpretar el método SVM como un método de minimización del riesgo empírico penalizado.

Si despejamos ξ_i, y volvemos a realizar el cambio $z_1 = 2y_i - 1$, tendrá que cumplirse

$$\begin{aligned} \xi_i &\geq 1 - z_i(\mathbf{w} \cdot \mathbf{x}_i + b) && \text{(4.19a)} \\ \xi_i &\geq 0 && \text{(4.19b)} \end{aligned}$$

así que en el óptimo podremos tomar

$$\xi_i = (1 - z_i(\mathbf{w} \cdot \mathbf{x}_i + b))_+ = h\left(z_i(\mathbf{w} \cdot \mathbf{x}_i + b)\right) \tag{4.19c}$$

siendo $h(\mathbf{x}) = (1-\mathbf{x})_+ = \max(0, 1-\mathbf{x})$ la función conocida como *pérdida hinge*, una función que contabilizará los errores en la clasificación en función de la distancia a su clase. Podremos reformular el problema como

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimizar}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(1 - z_i(\mathbf{w} \cdot \mathbf{x}_i + b))_+ \tag{4.20}$$

que será equivalente a

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimizar}} \quad \frac{1}{n}\sum_{i=1}^{n} h\left(z_i(\mathbf{w} \cdot \mathbf{x}_i + b)\right) + \lambda\|\mathbf{w}\|^2 \tag{4.21}$$

tomando $\lambda = 1/(2Cn)$.

Con esta forma, el problema consiste en la minimización de una función de pérdida más un término de penalización que representa la complejidad del modelo, cambiando la función de pérdida $0-1$ anteriormente vista por la función de pérdida *hinge*, que será convexa. teniendo así un problema de optimización en el que habrá que minimizar una función que contabiliza los errores más una función que representa la complejidad, entrando en conexión con lo visto anteriormente donde nuestro objetivo era minimizar el riesgo empírico más una penalización, habiendo pasado de utilizar la función de

pérdida $0-1$ a la función *hinge*, que es una función convexa y derivable salvo en $x = 0$.

Veamos la obtención del problema dual respecto a este problema dado de la forma (4.18). Primero plantearemos el Lagrangiano generalizado. En este caso será necesario añadir otra variable dual pues tenemos la restricción $\xi_i \geq 0$.

$$\begin{aligned} L(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[z_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i\right] \\ &\quad - \sum_{i=1}^{n}\beta_i\xi_i \end{aligned} \tag{4.22}$$

Derivando e igualando a 0 obtenemos

$$0 = \frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n}\alpha_i z_i \mathbf{x}_i \tag{4.23a}$$

$$0 = \frac{\partial L}{\partial b} = -\sum_{i=1}^{n}\alpha_i z_i \tag{4.23b}$$

$$0 = C - \alpha_i - \beta_i \tag{4.23c}$$

de donde obtenemos las relaciones $\mathbf{w} = \sum_{i=1}^{n}\alpha_i z_i \mathbf{x}_i$, $\sum_{i=1}^{n}\alpha_i z_i = 0$ y $\alpha_i = C - \beta_i$ y cómo $\beta_i \geq 0$, entonces $\alpha_i \leq C$. Como ya hemos mencionado, pocos $\alpha_i \neq 0$, y por tanto tendremos una representación de $\mathbf{w}$ dispersa en el sentido de que dependerá de pocos puntos $\mathbf{x}_i$. Sustituyendo estas relaciones en el Lagrangiano, tendremos

$$\begin{aligned} L(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}\sum_{i,j=1}^{n} z_i z_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j + C\sum_{i=1}^{n}\xi_i - \sum_{i,j=1}^{n} z_i z_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \\ &\quad - b\sum_{i=1}^{n}\alpha_i z_i + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\beta_i\xi_i \\ &= -\frac{1}{2}\sum_{i,j=1}^{n} z_i z_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\xi_i(C - \alpha_i - \beta_i) \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} z_i z_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \end{aligned}$$

Por lo tanto, la forma dual del problema (4.17) será

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximizar}} \quad & L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} z_i z_j \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \\ \text{sujeto a} \quad & \sum_{i=1}^{n} \alpha_i z_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, ..., n \end{aligned} \tag{4.24}$$

Las condiciones KKT complementarias en este caso serán

$$\alpha_i \left[z_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i \right] = 0, \qquad i = 1, ..., n \tag{4.25a}$$

$$\xi_i (\alpha_i - C) = 0, \qquad i = 1, ..., n \tag{4.25b}$$

De estas dos condiciones y de que $0 \leq \alpha_i \leq C$, podemos diferenciar tres casos distintos en función del valor que tome α_i. Si $\alpha_i = 0$, entonces $\xi_i = 0$ y el punto $\mathbf{x}_i$ puede estar en el margen (si está a distancia 1 del hiperplano) o fuera de él si $z_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 1$. Si $0 < \alpha_i < C$, el punto $\mathbf{x}_i$ será un vector soporte que estará en uno de los dos márgenes y por tanto estará a distancia 1. Y si por último, $\alpha_i = C$, el punto $\mathbf{x}_i$ puede ser un vector soporte o estar a una distancia menor que 1 del hiperplano separante en función del valor de ξ_i.

Prestamos ahora atención a aspectos estadísticos relacionados con el problema (4.20). ¿Cuál será el efecto de cambiar la función de pérdida $0-1$ tratada en el capítulo 3 por la pérdida hinge? A continuación veremos que el minimizador del riesgo con pérdida hinge es el mismo que con pérdida $0-1$, y por tanto, equivalente al riesgo de Bayes.

Si nuestro espacio de etiquetas es $Y = \{0, 1\}$, pasaremos a uno $Z\{-1, 1\}$ a través del cambio de variable $z_i = 2y_i - 1$. Hemos visto que la regla de Bayes f^* viene dada por

$$f^*(\mathbf{x}) = \begin{cases} 1 & \text{si } \eta(\mathbf{x}) \geq 1/2 \\ 0 & \text{si } \eta(\mathbf{x}) < 1/2 \end{cases} \tag{4.26}$$

con $\eta(\mathbf{x}) = \mathbb{P}\{Z = 1 | X = x\}$.

Sea la función de pérdida hinge

$$l(t) = (1 - t)_+ = \text{máx}\{0, 1 - t\} \tag{4.27}$$

El riesgo empírico vendrá dado por

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(1 - z_i f(x_i))_+ \tag{4.28}$$

y el riesgo esperado

$$R(f) = \mathbb{E}R_n(f) = \mathbb{E}\{(1 - Zf(X))_+\} \tag{4.29}$$

que se puede escribir como

$$\mathbb{E}\left\{\mathbb{E}\{(1 - Zf(X))_+|X\}\right\} \tag{4.30}$$

y

$$\mathbb{E}\{(1 - zf(x))_+|X = x\} = \eta(x)(1 - f(x))_+ + (1 - \eta(x))(1 + f(x))_+ \tag{4.31}$$

Por tanto, tendremos que

$$R(f) = \int \left[\eta(x)(1 - f(x))_+ + (1 - \eta(x))(1 + f(x))_+\right] dQ(x) \tag{4.32}$$

Si llamamos $s = f(x)$ y $p = \eta(x)$, buscaremos el valor de s que minimiza esa función que podremos reescribir como

$$p(1 - s)_+ + (1 - p)(1 + s)_+ = \begin{cases} (1 - p)(1 + s)_+ & si \quad s \geq 1 \\ 1 + (1 - 2p)s & si \quad -1 < s < 1 \\ p(1 - s)_+ & si \quad s \leq -1 \end{cases} \tag{4.33}$$

que será una función convexa y alcanzará el mínimo en $s = 1$ si $p \geq 1/2$ y en $s = -1$ en caso contrario. Por tanto vemos que la función de perdida hinge tendrá el mismo minimizador que la regla $0 - 1$, que es la regla de Bayes.

Por tener el mismo minimizador, no hay problema en cambiar la función de pérdida $0 - 1$ por la pérdida hinge que al ser convexa permitirá resolver el problema de optimización.

A continuación veremos como resolver el problema cuando el conjunto no es separable a través de una transformación del espacio de los $\mathbf{x}_i$ en otro espacio de alta dimensión.

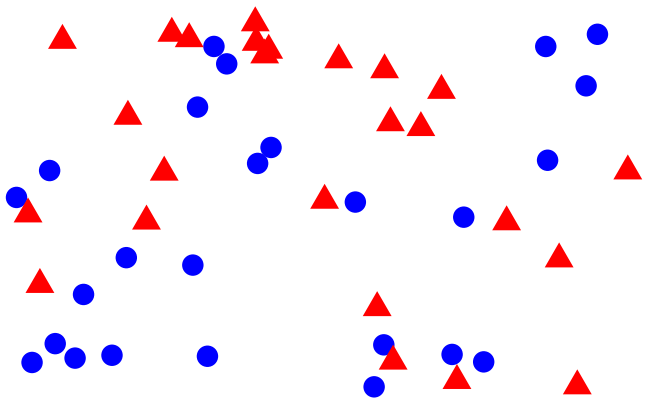

Figura 4.4: Conjunto de puntos linealmente no separables en $\mathbb{R}^2$.

4.4. SVMs no lineales

En muchos casos el conjunto de entrenamiento para nuestro SVM no será separable ni siquiera a excepción de pocos puntos del espacio X, como por ejemplo en la figura 4.4 En estos casos, el método utilizado consiste en transformar nuestros datos en el espacio X a un espacio de Hilbert $\mathcal{H}$ (generalmente de dimensión superior), denominado espacio de características, donde la transformación de dichos puntos si que pueda ser separada por un hiperplano. Dicha transformación, vendrá dada por una función no lineal

$$\begin{aligned} \Phi : & X \to \mathcal{H} \\ & \mathbf{x} \to \Phi(\mathbf{x}) = [\phi_1(\mathbf{x}), ..., \phi_m(\mathbf{x})] \end{aligned} \tag{4.34}$$

y se podrá construir un hiperplano en este espacio de características $\mathcal{H}$ que será de la forma:

$$\mathbf{w} \cdot \Phi(\mathbf{x}) + b = w_1\phi_1(\mathbf{x}) + ... + w_m\phi_m(\mathbf{x}) + b \tag{4.35}$$

Tendremos de esta forma un hiperplano que separará los puntos $\Phi(\mathbf{x}_i)$ en el espacio de características, pero que definirá fronteras entre las regiones de decisión no lineales en el espacio inicial de los puntos $\mathbf{x}_i$. Como ya he hemos visto anteriormente, la regla de decisión se puede calcular a partir de los productos internos entre los datos de entrenamiento y el punto a clasificar, que tras realizar la transformación Φ, será

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i z_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b \tag{4.36}$$

El cálculo de los puntos transformados $\Phi(\mathbf{x})$ puede ser complicado, pero si existe una función $K(\mathbf{x}, \mathbf{z})$, tal que permita calcular directamente dicho producto interno sin tener que hacer la transformación Φ, la complejidad del cálculo no se vería afectada por la dimensión de $\mathcal{H}$.

4.10 Definición. Una función *núcleo (o kernel)*, es una función $K(\mathbf{x}, \mathbf{z})$ que cumple

$$K(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{z}), \qquad \forall\ \mathbf{x}, \mathbf{z} \in X$$

siendo Φ una transformación (posiblemente no lineal) de X en un espacio de características $\mathcal{H}$ (que será un espacio de Hilbert).

Si podemos definir una función $K(\mathbf{x}, \mathbf{z})$ asociada a la transformación Φ, la regla de decisión se podría calcular de la forma

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i z_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{4.37}$$

Por tanto, cuando nos encontramos con un conjunto no separable linealmente y se realiza una transformación Φ a otro espacio de características $\mathcal{H}$, donde $K(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{z})$, el problema a resolver en su versión dual será

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximizar}} \quad & L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} z_i z_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{sujeto a} \quad & \sum_{i=1}^{n} \alpha_i z_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, ..., n \end{aligned} \tag{4.38}$$

y las decisiones se tomarán a partir de 4.37. Esta posibilidad de poder plantear el problema y hallar la solución solo en función de la función núcleo sin tener que computar la transformación Φ se denomina *truco del núcleo*.

4.11 Ejemplo. Veamos un ejemplo para la construcción de un SMV en un problema de clasificación binaria sin error cuya separación lineal sólo es posible a través de una transformación. Supongamos que se trata de buscar el hiperplano óptimo que clasifica correctamente el siguiente conjunto de datos:

$$\begin{aligned} \mathbf{x}_1 &= [1, 1] & y_1 &= 1 \\ \mathbf{x}_2 &= [1, -1] & y_2 &= 0 \\ \mathbf{x}_3 &= [-1, -1] & y_3 &= 1 \\ \mathbf{x}_4 &= [-1, 1] & y_4 &= 0 \end{aligned}$$

Si tomamos como núcleo la función

$$K(\mathbf{x}, \mathbf{z}) = [\mathbf{x} \cdot \mathbf{z} + 1]^2$$

asociada a la transformación $\Phi(a_1, a_2) = \left[1, \sqrt{2}a_1, \sqrt{2}a_2, \sqrt{2}a_1a_2, a_1^2, a_2^2\right]$, con a_i las coordenadas de cada punto $\mathbf{x}$, $\mathbf{x} = [a_1, a_2]$. Para hallar la frontera de decisión, será necesario resolver el problema de optimización 4.38, que en este caso será

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{maximizar}} \quad & L(\boldsymbol{\alpha}) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + -\frac{1}{2}\sum_{i,j=1}^{4} z_i z_j \alpha_i \alpha_j K_{ij} \\ \text{sujeto a} \quad & \alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 = 0 \\ & \alpha_i \geq 0, \quad i = 1, ..., 4 \end{aligned} \tag{4.39}$$

siendo $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, los elementos de la matriz

$$K = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}$$

La solución será $\alpha_1^* = \alpha_2^* = \alpha_3^* = \alpha_4^* = 0{,}125$ y en este caso, dado que la dimensión del espacio de características es pequeña, podemos escribir la función de decisión como

$$f(\mathbf{x}) = w_1^* + w_2^*\sqrt{2}a_1 + w_1^*\sqrt{2}a_2 + w_1^*\sqrt{2}a_1a_2 + w_1^*a_1^2 + w_1^*a_2^2$$

y podremos obtener $\mathbf{w}^*$ como

$$\mathbf{w}_i^* = \sum_{i=1}^{4} \alpha_i^* y_i \Phi(\mathbf{x}_i) = \left[0, 0, 0, \frac{\sqrt{2}}{2}, 0, 0\right]$$

y por tanto la función de decisión en este caso será

$$D(\mathbf{x}) = a_1a_2$$

□

A continuación se discutirá sobre que funciones son núcleos y si pueden manejarse sin hacer referencia explícita a la transformación ϕ. Por ser un producto escalar, $K(\mathbf{x}, \mathbf{z})$ deberá cumplir algunas propiedades como

1. $K(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{z}) = \Phi(\mathbf{z}) \cdot \Phi(\mathbf{x}) = K(\mathbf{z}, \mathbf{x})$ (Por simetría)

2. $K(\mathbf{x},\mathbf{z})^2 \leq K(\mathbf{x},\mathbf{x})K(\mathbf{z},\mathbf{z})$ (Por la desigualdad de Cauchy-Schwarz)

4.12 Definición. Una función $K : X \times X \to \mathbb{R}$ es simétrica y semidefinida positiva si para todo conjunto de puntos $\{x_1, ..., x_n\}$ la matriz

$$\mathbf{K} = (K(x_i, x_j))_{i,j=1}^{n} \tag{4.40}$$

es simétrica y semidefinida positiva.

Veamos a continuación que es esta propiedad la que caracteriza a las funciones núcleo.

4.13 Teorema. La función $K : X \times X \to \mathbb{R}$ será un núcleo si y sólo si es simétrica y semidefinida positiva.

Demostración. La implicación de que si K es núcleo, entonces es simétrica y semidefinida positiva es trivial. Veamos entonces la otra implicación. Dada una función simétrica y semidefinida positiva, construiremos una transformación $\boldsymbol{\phi}$ en un espacio de Hilbert en el que K sea el núcleo.

Definimos el espacio

$$\tilde{\mathcal{H}} = \left\{ \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \cdot) : \ \alpha_i \in \mathbb{R}, \ n \geq 1, \ \mathbf{x}_i \in X, \ i = 1, ..., n \right\} \tag{4.41}$$

que es un espacio vectorial. Definiremos ahora el producto interno

$$\left\langle \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \cdot), \sum_{j=1}^{m} \beta_j K(\mathbf{y}_j, \cdot) \right\rangle = \sum_{i,j} \alpha_i \beta_j K(\mathbf{x}_i, \mathbf{y}_j) \tag{4.42}$$

que satisface las propiedades de producto interno. Si tomamos el espacio $\mathcal{H}$ como aquel que es el completado de $\tilde{\mathcal{H}}$, es decir, el conjunto en el que las sucesiones de Cauchy de elementos de $\tilde{\mathcal{H}}$ son convergentes, tendremos el espacio de características, que será un espacio de Hilbert y ahora sólo será necesario definir la transformación $\boldsymbol{\phi}$

$$\boldsymbol{\phi} \ : \ X \to \mathcal{H} \tag{4.43}$$

$$\mathbf{x} \to \phi(\mathbf{x}) = K(\mathbf{x}, \cdot) \tag{4.44}$$

Sea

$$\langle f, \phi(\mathbf{x}) \rangle = \langle f, K(\mathbf{x}, \cdot) \rangle = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}) = f(\mathbf{x}) \tag{4.45}$$

que se denomina *propiedad reproductiva* y al espacio $\mathcal{H}$, el *RKHS*(Reproducing Kernel Hilbert Space) de la función K.

□

Hay otros resultados que permiten concluir que una función es un núcleo. Un ejemplo es el teorema de Mercer.

4.14 Teorema. *(*Mercer) . Sea X un subconjunto compacto de $\mathbb{R}^n$. Sea K una función simétrica y continua tal que el operador integral $T_K : L_2(X) \to L_2(X)$,

$$(T_K f)(\cdot) = \int_X K(\cdot, \mathbf{x}) f(\mathbf{x}) d\mathbf{x}, \tag{4.46}$$

es decir, que cumpla

$$\int_{X \times X} K(\mathbf{x}, \mathbf{z}) f(\mathbf{x}) f(\mathbf{z}) d\mathbf{x} d\mathbf{z} \geq 0 \tag{4.47}$$

para toda función $f \in L_2(X)$. Entonces, se puede expandir $K(\mathbf{x}, \mathbf{z})$ en una serie con convergencia uniforme sobre $X \times X$ en términos de T_K funciones propias $\phi_j \in L_2(X)$, de tal forma que $\|\phi_j\|_{L_2} = 1$ y con valores propios asociados positivos $\lambda_j \geq 0$,

$$K(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\mathbf{x}) \phi_j(\mathbf{z}) \tag{4.48}$$

Como consecuencia del Teorema de Mercer vemos que cualquier función $K(\mathbf{x}, \mathbf{z})$ que satisfaga las condiciones del teorema, es núcleo.

Para verlo, tomamos el espacio de sucesiones de números reales tales que $\sum \lambda_j x_j^2 < \infty$ con el producto interno

$$\langle \{\mathbf{x}\}, \{\mathbf{z}\} \rangle = \sum \lambda_j x_j z_j \tag{4.49}$$

y la transformación Φ dada por $\Phi(\mathbf{x}) = \{\phi_j(\mathbf{x})\}_{j=1}^{\infty}$, entonces

$$\langle \Phi(\mathbf{x}), \Phi(\mathbf{z}) \rangle = \sum_{j=1}^{\infty} \lambda_j \phi_j(\mathbf{x}) \phi_j(\mathbf{z}) = K(\mathbf{x}, \mathbf{z}), \tag{4.50}$$

luego K es un núcleo.

Por ejemplo vemos así que de núcleo puede ser $K(\mathbf{x}, \mathbf{z}) = \mathbf{x} \wedge \mathbf{z}$ es un núcleo.

Veremos a continuación ciertas propiedades que permitirán construir nuevos núcleos a partir de otros.

4.15 Proposición. *Sean K_1 y K_2 núcleos en $X \times X$, $X \subseteq \mathbb{R}^n$, $a \in \mathbb{R}^+$, $f(\cdot)$ una función real en X, la transformación ϕ con K_3 un núcleo en $\mathbb{R}^m \times \mathbb{R}^m$ y $\mathbf{B}$ una matriz simétrica positiva semidefinida de tamaño $n \times n$. Entonces las siguientes funciones son núcleos:*

1. $K(\boldsymbol{x}, \boldsymbol{z}) = K_1(\boldsymbol{x}, \boldsymbol{z}) + K_2(\boldsymbol{x}, \boldsymbol{z})$
2. $K(\boldsymbol{x}, \boldsymbol{z}) = aK_1(\boldsymbol{x}, \boldsymbol{z})$
3. $K(\boldsymbol{x}, \boldsymbol{z}) = K_1(\boldsymbol{x}, \boldsymbol{z})K_2(\boldsymbol{x}, \boldsymbol{z})$
4. $K(\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{x})f(\boldsymbol{z})$
5. $K(\boldsymbol{x}, \boldsymbol{z}) = K_3(\phi(\boldsymbol{x}), \phi(\boldsymbol{z}))$
6. $K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}'\mathbf{B}\boldsymbol{x}$

Demostración. 1 y 2 son triviales. 4, 5 y 6 se deducen de forma elemental. Demostraremos aquí la propiedad 3.

Fijado un conjunto de puntos $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ y sean $\mathbf{K}_1$ y $\mathbf{K}_2$ las matrices resultantes de restringir K_1 y K_2 a dichos puntos, veamos que la matriz del producto $K_1(\mathbf{x}, \mathbf{z})K_2(\mathbf{x}, \mathbf{z})$ es semidefinida positiva.

Sea

$$\mathbf{K} = \mathbf{K}_1 \bigotimes \mathbf{K}_2$$

el producto tensorial de las matrices $\mathbf{K}_1$ y $\mathbf{K}_2$. Esta matriz, al ser producto de dos matrices semidefinidas positivas, será semidefinida positiva, pues sus autovalores serán todos los pares de productos de los autovalores de las dos componentes. Por otro lado, la matriz $\mathbf{H}$ de la función K_1K_2 será la matriz cuyas entradas son los productos de las entradas de las dos componentes con mismos índices. Por tanto, esta matriz será una submatriz principal de $\mathbf{K}$. Se cumplirá entonces que para cualquier $\alpha \in \mathbb{R}^n$, existirá un $\alpha_1 \in \mathbb{R}^{n^2}$, tal que cumplirá

$$\alpha'\mathbf{H}\alpha = \alpha_1'\mathbf{K}\alpha_1 \geq 0$$

y entonces $\mathbf{H}$ será semidefinida positiva.

□

4.16 Corolario. *Sea $K_1(\boldsymbol{x}, \boldsymbol{z})$ un núcleo sobre $X \times X$, $\boldsymbol{x}$, $\boldsymbol{z} \in X$, y $p(x)$ un polinomio con coeficientes positivos. Las funciones siguientes serán núcleos también:*

1. $K(\boldsymbol{x}, \boldsymbol{z}) = p(K_1(\boldsymbol{x}, \boldsymbol{z}))$
2. $K(\boldsymbol{x}, \boldsymbol{z}) = \exp(K_1(\boldsymbol{x}, \boldsymbol{z}))$
3. $K(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\|\boldsymbol{x} - \boldsymbol{z}\|^2/\sigma^2)$

Demostración. La primera propiedad es trivial combinando las propiedades 1,2,3 y 4 de (4.15). La segunda se obtiene por ser la función exponencial límite de suma de polinomios. Veamos la tercera. Podemos descomponer la función en

$$\exp(-\|\mathbf{x} - \mathbf{z}\|^2/\sigma^2) = \exp(-\|\mathbf{x}\|^2/\sigma^2)\exp(-\|\mathbf{z}\|^2/\sigma^2)\exp(2\ \mathbf{x}\cdot\mathbf{z}/\sigma^2) \quad (4.51)$$

De aquí, el producto de los dos primeros términos será un núcleo por el apartado 4 de (4.15) y el último lo será por el apartado 2 de este corolario.

□

El tercer núcleo aquí definido es conocido como *núcleo gaussiano*, también denominado *núcleo RBF* al utilizar funciones de base radial. Otra forma de escribir este núcleo es

$$K(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{\sigma^2}\right) = e^{-\gamma\|\mathbf{x}-\mathbf{z}\|^2} \quad (4.52)$$

que proyecta los puntos en un espacio de características de dimensión infinita. Cuando γ es pequeño, un punto $\mathbf{x}$ tendrá núcleo positivo en relación con cualquiera de los vectores soporte, y por tanto todos estos vectores afectarán al valor de la función clasificadora en $\mathbf{x}$, lo que originará una frontera suave entre las regiones de clasificación. A medida que γ crezca, aumentará la curvatura de la región de la frontera de decisión y para valores grandes de γ las regiones estarán cada vez más concentradas en torno a los vectores soporte. En este caso se dará una clara situación de sobreajuste.

4.5. Cotas probabilísticas

Dado nuestro problema en la versión

$$\underset{\mathbf{w}}{\text{minimizar}}\ \frac{1}{n}\sum_{i=1}^{n}(1 - z_i f_{\mathbf{w}}(\mathbf{x}_i))_+ + \lambda\|\mathbf{w}\|^2 \quad (4.53)$$

donde $f_{\mathbf{w}} = \mathbf{w}\cdot\phi(\mathbf{x}) + b$.

Podemos reescribir este problema como

$$\min_{R>0}\left[\min_{\mathbf{w}:\|\mathbf{w}\|\leq R}\ \frac{1}{n}\sum_{i=1}^{n}(1-z_i f_{\mathbf{w}}(\mathbf{x}_i))_+ + \lambda R^2\right] \tag{4.54}$$

Podemos definir para cada R el conjunto de funciones

$$\mathcal{F}_R = \left\{ f(\mathbf{x}) = \sum_{i=1}^{n}\lambda_i K(\mathbf{x}_i,\mathbf{x})\ ;\ \sum_{i,j=1}^{n}\lambda_i\lambda_j K(\mathbf{x}_i,\mathbf{x}_j)\leq R^2 \right\} \tag{4.55}$$

De esta forma, podemos resolver el problema de optimización buscando los mínimos dentro de bolas de radio R en el espacio $\mathcal{H}$, lo que permite tener un problema de la forma de la minimización del riesgo estructural.

De la desigualdad de las diferencias acotadas (3.28), tendremos que si $\mathcal{F}$ es una clase de funciones con valores en $[-1,1]$, entonces con probabilidad al menos $1-\delta$,

$$\sup_{f\in\mathcal{F}}|R(f)-R_n(f)| \leq \mathbb{E}\sup_{f\in\mathcal{F}}|R(f)-R_n(f)| + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}. \tag{4.56}$$

Por otro lado, si introducimos una copia $X_1',...,X_n'$ independiente de las X_i e igualmente distribuidas, por la desigualdad de Jensen,

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}|R(f)-R_n(f)| &= \mathbb{E}\sup_{f\in\mathcal{F}}\left(\mathbb{E}\left[|R_n'(f)-R_n(f)|\,\middle|\,X_1,...,X_n\right]\right)\\ &\leq \mathbb{E}\sup_{f\in\mathcal{F}}|R_n'(f)-R_n(f)|.\end{aligned}$$

Si introducimos las variables aleatorias $\sigma_1,...,\sigma_n$ independientes tales que $\mathbb{P}\{\sigma_i=1\}=\mathbb{P}\{\sigma_i=-1\}=1/2$,

$$\mathbb{E}\sup_{f\in\mathcal{F}}|R(f)-R_n(f)| = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}(f(X_i')-f(X_i))\right|\right] \tag{4.57}$$

$$= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i(f(X_i')-f(X_i))\right|\right] \tag{4.58}$$

$$\leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i f(X_i)\right|\right] \tag{4.59}$$

Introduciremos a continuación unos conceptos nuevos.

4.17 Definición. Dado $A \subset \mathbb{R}^n$ un conjunto acotado de vectores $a = (a_1, ..., a_n)$, se define como *promedio de Rademacher* a la cantidad

$$\mathcal{R}_n(A) = \mathbb{E} \sup_{a \in A} \frac{2}{n} \left| \sum_{i=1}^{n} \sigma_i a_i \right| \tag{4.60}$$

4.18 Definición. Dada una clase $\mathcal{F}$, llamaremos *complejidad empírica de Rademacher* de la clase $\mathcal{F}$, $\hat{\mathcal{R}}_n(\mathcal{F})$ al promedio de Rademacher asociado al conjunto $\mathcal{F}(X_1, ..., X_n) = \{(f(X_1), ..., f(X_n)) : f \in \mathcal{F}\}$ y la *complejidad de Rademacher* de $\mathcal{F}$ será $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}(\hat{\mathcal{R}}_n(\mathcal{F}))$.

De esta definición y de (4.59),

$$\mathbb{E} \sup_{f \in \mathcal{F}} |R(f) - R_n(f)| \leq \mathcal{R}_n(\mathcal{F}) \tag{4.61}$$

y entonces, con probabilidad al menos $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |R(f) - R_n(f)| \leq \mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}. \tag{4.62}$$

Por otro lado, las complejidades de Rademacher también verifican la condición de las diferencias acotadas para $c_i = 4/n$, por lo que por la desigualdad de las diferencias acotadas, se cumple que con probabilidad al menos $1 - \delta/2$, se cumplirá que

$$\mathcal{R}_n(\mathcal{F}) \leq \hat{\mathcal{R}}_n(\mathcal{F}) + \sqrt{\frac{8 \log \frac{2}{\delta}}{n}} \tag{4.63}$$

y combinando con (4.56), podemos concluir que con probabilidad $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |R(f) - R_n(f)| \leq \hat{\mathcal{R}}_n(\mathcal{F}) + 3\sqrt{\frac{2 \log \frac{2}{\delta}}{n}}. \tag{4.64}$$

En la situación de que las funciones $f \in \mathcal{F}$ no tomasen valores en $[-1, 1]$ se podría normalizar dichas funciones con un factor $1/\|f\|_\infty$ y en ese caso la desigualdad sería

$$\sup_{f \in \mathcal{F}} |R(f) - R_n(f)| \leq \hat{\mathcal{R}}_n(\mathcal{F}) + 3\|f\|_\infty \sqrt{\frac{2 \log \frac{2}{\delta}}{n}}. \tag{4.65}$$

Si tomamos ahora las clases $\mathcal{F}_R$, su complejidad empírica de Rademacher será

$$\begin{aligned}\hat{\mathcal{R}}_n(\mathcal{F}) &= \frac{2}{n}\mathbb{E}\sup_{\|f\|\leq R}\sum_{i=1}^n \sigma_i f(\mathbf{x}_i) && (4.66)\\ &= \frac{2}{n}\mathbb{E}\sup_{\|f\|\leq R}\sum_{i=1}^n \sigma_i \langle f, K(\mathbf{x}_i,\cdot)\rangle && (4.67)\\ &= \frac{2R}{n}\mathbb{E}\left\|\sum_{i=1}^n \sigma_i K(\mathbf{x}_i,\cdot)\right\| && (4.68)\end{aligned}$$

Podremos aplicar la desigualdad de Kahane-Khinchine que dice que para cualquier conjunto de vectores $a_1, ..., a_n$ en un espacio de Hilbert, se cumple

$$\frac{1}{\sqrt{2}}\mathbb{E}\left\|\sum_{i=1}^n \sigma_i a_i\right\|^2 \leq \left(\mathbb{E}\left\|\sum_{i=1}^n \sigma_i a_i\right\|\right)^2 \leq \mathbb{E}\left\|\sum_{i=1}^n \sigma_i a_i\right\|^2 \tag{4.69}$$

Además se cumplirá

$$\mathbb{E}\left\|\sum_{i=1}^n \sigma_i a_i\right\|^2 = \mathbb{E}\sum_{i,j=1}^n \sigma_i\sigma_j\langle a_i, a_j\rangle = \sum \|a_i\|^2 \tag{4.70}$$

Por tanto, podremos acotar el término $\hat{\mathcal{R}}_n(\mathcal{F}_R)$ por

$$\frac{2R}{n\sqrt{2}}\sqrt{\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)} \leq \hat{\mathcal{R}}_n(\mathcal{F}_R) \leq \frac{2R}{n}\sqrt{\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)} \tag{4.71}$$

y podremos dar la siguiente cota:

4.19 Teorema. Con probabilidad al menos $1-\delta$ se tiene que para toda $f \in \mathcal{F}_R$,

$$R(f) \leq R_n(f) + 2\frac{R}{n}\sqrt{\sum_{i=1}^n K(\mathbf{x}_i,\mathbf{x}_i)} + 3\|f\|_\infty\sqrt{\frac{2\log\frac{2}{\delta}}{n}} \tag{4.72}$$

La desigualdad del Teorema 4.19 permite asegurar que el riesgo, con pérdida hinge, de la regla elegida por el SVM (dentro de una bola de radio R) no es muy grande si $R_n(f)$ es pequeño y R no es grande, pero no nos da una indicación sobre el tamaño de R que debemos elegir ni nos indica nada sobre el exceso de riesgo. Estas deficiencias se corregirán con el método del lasso en el siguiente capítulo.

Capítulo 5

El método del Lasso

El método del *lasso* (acrónimo de *least absolute shrinkage and selection operator*) fue propuesto por Tibshirani (1996). Es un método de penalización utilizado en las técnicas de aprendizaje supervisado y que puede ser implementado para muchos modelos. En este trabajo nos centraremos en el caso de modelos lineales, basándonos principalmente en resultados de [7].

En el caso del *lasso*, la penalización es de tipo l_1. En la situación de clasificación binaria con pérdida hinge el objetivo será resolver el problema de optimización:

$$\underset{\mathbf{w}}{\text{minimizar}} \quad \frac{1}{n}\sum_{i=1}^{n}\left(1 - z_i(\mathbf{w}\cdot\mathbf{x}_i + b)\right)_+ + \lambda\|\mathbf{w}\|_1 \tag{5.1}$$

siendo $\|\mathbf{w}\|_1 = \sum_{j=1}^{k}|w_j|$.

El método puede aplicarse en muchas situaciones. Primero veremos la descripción del método en el caso de regresión y cuyo objetivo es minimizar el error cuadrático y después lo plantearemos en la situación de la clasificación binaria.

5.1. El lasso en regresión lineal

Supongamos que nuestro modelo es lineal y viene dado por

$$Y_i = \sum_{j=1}^{p} X_i^{(j)}\beta + \epsilon_i, \qquad i = 1, ..., n \tag{5.2}$$

siendo ϵ_i términos de error. Por simplicidad, usaremos la notación matricial

$$Y = \mathbf{X}\beta + \epsilon \tag{5.3}$$

donde $\mathbf{X}_{n\times p}$ es la matriz de todas nuestras observaciones, también llamada matriz de diseño, $Y_{n\times 1}$, el vector de respuestas y $\epsilon_{n\times 1}$ un vector de errores que podemos suponer con distribución $N(0, \sigma^2 I_n)$.

Si asumimos que el modelo es correcto, entonces existirá β^0 que será el verdadero valor del parámetro y se cumplirá $Y = \mathbf{X}\beta^0 + \epsilon$. El estimador del método del lasso incluirá un término de regularización que será la norma l_1 del vector β y vendrá dado por

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{\|Y - \mathbf{X}\beta\|^2}{n} + \lambda\|\beta\|_1 \right\} \tag{5.4}$$

Una ventaja del método del lasso en comparación con otros métodos es que debido al término de penalización l_1 en función del λ existirán algunos $\beta_j = 0$, para algún j. Por esta característica, se suele decir que el lasso conduce a estimadores "dispersos" que propicia de forma natural una selección de variables (aquellas asociadas con los coeficientes $\hat{\beta}_j$ no nulos).

El resto de esta sección se dedicará al estudio probabilístico del lasso. El análisis parte de la siguiente desigualdad:

5.1 Lema. Desigualdad básica: Dado el modelo $Y = \mathbf{X}\beta + \epsilon$, β^0 el verdadero valor de β y $\hat{\beta}$ el estimador lasso, se verificará la siguiente desigualdad:

$$\frac{1}{n}\|\mathbf{X}(\hat{\beta} - \beta^0)\|^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{2}{n}\epsilon^T\mathbf{X}(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 \tag{5.5}$$

Al término $\|\mathbf{X}(\hat{\beta} - \beta^0)\|^2/n$ se le llama error de predicción.

Demostración. Por (5.4) se verificará

$$\frac{1}{n}\|Y - \mathbf{X}\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\|Y - \mathbf{X}\beta^0\|^2 + \lambda\|\beta^0\|_1$$

Desarrollando las normas cuadráticas de la igualdad tenemos

$$\begin{aligned} \|Y - \mathbf{X}\hat{\beta}\|^2 &= \|\mathbf{X}(\beta^0 - \hat{\beta}) + \epsilon\|^2 \\ &= \|\mathbf{X}(\beta^0 - \hat{\beta})\|^2 + \|\epsilon\|^2 + 2\epsilon^T\mathbf{X}(\beta^0 - \hat{\beta}) \\ \|Y - \mathbf{X}\beta^0\|^2 &= \|\epsilon\|^2 \end{aligned}$$

Sustituyendo, llegamos a la desigualdad que queríamos tener.

□

A la vista de esta desigualdad, para poder controlar el error de predicción del lasso, habrá que controlar, por tanto, el término $2\epsilon^T\mathbf{X}(\hat{\beta}-\beta^0)/n$, que podremos acotar por

$$\frac{2}{n}\epsilon^T\mathbf{X}(\hat{\beta}-\beta^0)=\frac{2}{n}\sum_{j=1}^{p}\left(\epsilon^T\mathbf{X}^{(j)}\right)\left(\hat{\beta}-\beta^0\right)_j\leq\frac{2}{n}\max_{1\leq j\leq p}|\epsilon^T\mathbf{X}^{(j)}|\|\hat{\beta}-\beta^0\|_1 \tag{5.6}$$

donde $\mathbf{X}^{(1)},...,\mathbf{X}^{(p)}$ representan las columnas de la matriz $\mathbf{X}$.

Sea la matriz de Gram $\hat{\Sigma}=\mathbf{X}^T\mathbf{X}/n$. Normalizando, si es necesario, las columnas de $\mathbf{X}$, podemos suponer que $\hat{\Sigma}_{j,j}=\hat{\sigma}_j=1$. Si tomamos la variable

$$V_j=\frac{\epsilon^T\mathbf{X}^{(j)}}{\sigma\sqrt{n}} \tag{5.7}$$

tendrá una distribución $\mathcal{N}\left(0,\mathbf{X}^{(j)}\mathbf{X}^{(j)}/n\right)=\mathcal{N}(0,1)$ y

$$\mathbb{P}\left\{\max_{1\leq j\leq p}|V_j|>\sqrt{t^2+2\log p}\right\}\leq 2pe^{-\frac{t^2+2\log p}{2}}=2e^{-\frac{t^2}{2}} \tag{5.8}$$

Consideremos el suceso

$$\mathcal{T}:=\left\{\max_{1\leq j\leq p}\frac{2}{n}|\epsilon^T\mathbf{X}^{(j)}|\leq\lambda_0\right\} \tag{5.9}$$

por las ecuaciones (5.7) y (5.8), tendremos que para

$$\lambda_0=2\sigma\sqrt{\frac{t^2+2\log p}{n}}$$

se cumplirá

$$\mathbb{P}(\mathcal{T})\geq 1-2e^{-\frac{t^2}{2}}$$

y se verificará que en el conjunto $\mathcal{T}$

$$\frac{1}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2+\lambda\|\hat{\beta}\|_1\leq\lambda_0\|\hat{\beta}-\beta^0\|_1+\lambda\|\beta^0\|_1 \tag{5.10}$$

5.2 Proposición. *Sea* $\lambda=4\hat{\sigma}\sqrt{\frac{t^2+2\log p}{n}}$ *siendo* $\hat{\sigma}$ *un estimador de* σ. *Entonces con probabilidad al menos* $1-\alpha$, *donde*

$$\alpha=2e^{-t^2/2}+\mathbb{P}(\hat{\sigma}\leq\sigma),$$

se tiene que

$$\frac{2}{n}\|\boldsymbol{X}(\hat{\beta}-\beta^0)\|^2\leq 3\lambda\|\beta^0\|_1 \tag{5.11}$$

Demostración. Si $\hat{\sigma} > \sigma$ y $\lambda_0 = 2\sigma\sqrt{\frac{t^2+2\log p}{n}}$ entonces se cumple que $\lambda > 2\lambda_0$ y en el conjunto $\mathcal{T} \cap (\hat{\sigma} > \sigma)$, se cumplirá

$$\begin{aligned}\frac{1}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2 &\leq \lambda_0\|\hat{\beta}-\beta^0\|_1 - \lambda\|\hat{\beta}\|_1 + \lambda\|\beta^0\|_1 \\ &\leq \frac{\lambda}{2}(\|\hat{\beta}-\beta^0\|_1 - \|\hat{\beta}\|_1) + \lambda\|\beta^0\|_1 \\ &\leq \frac{3\lambda}{2}\|\beta^0\|_1\end{aligned}$$

□

En esta situación será necesario que el estimador $\hat{\sigma}$ sea escogido de forma que $\mathbb{P}\{\hat{\sigma} \leq \sigma\}$ sea pequeña pero que a la vez $\hat{\sigma}$ no sea demasiado grande para que el valor de λ no sea grande también.

Con el mismo argumento vemos que si $\lambda \geq 2\lambda_0$, entonces con probabilidad al menos $1 - 2e^{-t^2/2}$, se cumplirá

$$\frac{2}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2 \leq 3\lambda\|\beta^0\|^2$$

Esto significa que se puede elegir $\lambda \simeq \sqrt{\log p/n}$ de forma que con alta probabilidad

$$\frac{2}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2 \leq C\sqrt{\frac{\log p}{n}}.$$

Es decir, el lasso puede garantizar un error de predicción pequeño incluso cuando $p > n$.

En la práctica, el término de regularización λ, se obtiene mediante el método de validación cruzada, que consiste en hacer particiones del conjunto de entrenamiento en K subconjuntos, construyendo un modelo para cada subconjunto menos para el último que se utilizará para evaluar dichos modelos y a partir de aquí calcular el valor del parámetro que mejor ajusta cada modelo y calcular el promedio de ellos que será el utilizado en el verdadero modelo.

La situación más habitual es que no todas las variables influyan en nuestro modelo, por lo tanto, si suponemos que β^0 es disperso y que depende de una pequeña parte de las variables, se puede establecer una cota más refinada.

Sea $S_0 = \{j : \beta_j^0 \neq 0\}$ el conjunto de componentes de β^0 no nulas y sea $s_0 = card(S_0)$. Dado $S \subset \{1, ..., p\}$ denotaremos por

$$\beta_{j,S} = \beta_j \mathbb{I}(j \in S)$$

y el vector

$$\beta_S = (\beta_{j,S})_{1\leq j\leq p}$$

5.3 Proposición. *En las condiciones anteriores, si* $\lambda \geq 2\lambda_0$*, entonces*

$$\frac{2}{n}\|\boldsymbol{X}(\hat{\beta}-\beta^0)\|^2 + \lambda\|\hat{\beta}_{S_0^C}\|_1 \leq 3\lambda\|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1 \tag{5.12}$$

Demostración. Hemos visto que en el conjunto $\mathcal{T}$ se verifica la desigualdad

$$\frac{2}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2 + 2\lambda\|\hat{\beta}\|_1 \leq \lambda\|\hat{\beta}-\beta^0\|_1 + 2\lambda\|\beta^0\|_1$$

Se cumple que

$$\begin{aligned}
\|\hat{\beta}\|_1 &= \|\hat{\beta}_{S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1 \\
&\geq \|\beta^0_{S_0}\|_1 - \|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1 \quad \text{(desigualdad triangular)} \\
\|\hat{\beta}-\beta^0\|_1 &= \|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1
\end{aligned}$$

Por tanto, tendremos

$$\begin{aligned}
\frac{2}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2 &+ 2\lambda\|\beta^0_{S_0}\|_1 + 2\lambda\|\hat{\beta}_{S_0^c}\|_1 - 2\lambda\|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1 && (5.13)\\
&\leq \frac{2}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2 + 2\lambda\|\hat{\beta}\|_1 && (5.14)\\
&\leq \lambda\|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1 + \lambda\|\hat{\beta}_{S_0^c}\|_1 + 2\lambda\|\beta^0\|_1 && (5.15)
\end{aligned}$$

Con lo que se concluye la desigualdad anterior.

□

Una vez dada esta cota, se pueden suponer ciertas condiciones que la mejoren.

5.4 Definición. Sea $S_0 \subset \{1,...,p\}$, diremos que satisface la *condición de compatibilidad*(c.c.) si para algún ϕ_0 y todo β que verifique $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$ se cumple

$$\|\beta_{S_0}\|_1^2 \leq \frac{s_0}{\phi_0^2}\beta^T\hat{\Sigma}\beta \tag{5.16}$$

5.5 Teorema. Sea S_0 un conjunto de índices que satisface la condición de compatibilidad. Entonces si $\lambda \geq 2\lambda_0$, en $\mathcal{T}$ se verifica

$$\frac{1}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2 + \lambda\|\hat{\beta}-\beta^0\|_1 \leq 4\frac{\lambda^2 s_0}{\phi_0^2} \tag{5.17}$$

Demostración. Usando la desigualdad de 5.3, tendremos

$$\begin{aligned}
&\frac{2}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2+\lambda\|\hat{\beta}-\beta^0\|_1\\
&\quad=\frac{2}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2+\lambda\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_1+\lambda\|\hat{\beta}_{S_0^c}\|_1\\
&\quad\leq 4\lambda\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_1\\
&\quad\leq 4\lambda\frac{\sqrt{s_0}}{\phi_0}\frac{1}{\sqrt{n}}\|\mathbf{X}(\hat{\beta}-\beta^0)\| \qquad \text{(por la c.c.)}
\end{aligned}$$

Si consideramos $v=\lambda\sqrt{s_0}/\phi_0$ y $u=\|\mathbf{X}(\hat{\beta}-\beta^0)\|/\sqrt{n}$ y la desigualdad $4uv\leq u^2+4v^2$, podemos concluir que

$$\frac{2}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2+\lambda\|\hat{\beta}-\beta^0\|_1\leq\frac{1}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2+4\frac{\lambda^2 s_0}{\phi_0^2} \tag{5.18}$$

□

De este teorema se obtienen dos cotas interesantes,

$$\frac{1}{n}\|\mathbf{X}(\hat{\beta}-\beta^0)\|^2\leq\frac{4\lambda^2 s_0}{\phi_0^2} \tag{5.19a}$$

y

$$\|\hat{\beta}-\beta^0\|_1\leq 4\frac{\lambda s_0}{\phi_0^2} \tag{5.19b}$$

De (5.19a), podemos ver que salvo un factor constante, el error de generalización será del orden de λ^2, es decir como $\log p/n$ y por tanto esta cota será menor que en el caso general visto en la proposición 5.2. Aparte, de(5.19b), tenemos una cota para $\|\hat{\beta}-\beta^0\|_1$, que es del orden de $\sqrt{\log p/n}$ que permite controlar $\hat{\beta}$ y, por tanto, el modelo.

5.2. El lasso en clasificación

En esta sección, trataremos la adaptación de las ideas vistas anteriormente al marco de la clasificación binaria.

Sean (X_i,Y_i) nuestros datos observados en un espacio $X\times Y$ y sea $(F,\|\cdot\|)$ un espacio normado en el que F es un espacio de funciones definidas en X.

Dada $f\in F$, llamaremos función de pérdida a una función $\rho_f:\ X\times Y\to\mathbb{R}$ y asumiremos que $f\mapsto\rho_f(z)$ es una función convexa para cada z.

Según el tipo de problema a tratar habrá distintas funciones de pérdida

1. En el caso de que $y \in \mathbb{R}$, podemos tomar la función de pérdida cuadrática en regresión lineal
$$\rho_f(\cdot, y) = (y - f(\cdot))^2$$
2. En el caso de que $y \in \{-1, 1\}$, podemos tomar la función de pérdida hinge en el caso de clasificación binaria
$$\rho_f(\cdot, y) = (1 - yf(\cdot))_+$$
3. En el caso de que $y \in \{0, 1\}$, podemos tomar la función de pérdida logística para el caso de regresión logística
$$\rho_f(\cdot, y) = -yf(\cdot) + \log(1 + \exp(f(\cdot)))$$

Veamos la relación de esta función de pérdida logística con la clasificación.

Sean $(X_1, Y_1), ..., (X_n, Y_n)$ variables con distribución de Bernoulli y sea $\eta(x_i) = \mathbb{E}\{Y_i | X_i = x_i\} = \mathbb{P}\{Y_i = 1 | X_i = x_i\}$. Si tratamos de estimar el parámetro η por la función de verosimilitud, tendremos

$$L(\eta) = \prod_{i=1}^{n} \eta(x_i)^{y_i} (1 - \eta(x_i))^{(1-y_i)} \tag{5.20}$$

y la función de log-verosimilitud será

$$l(\eta) = \log(L(\eta)) = \sum_{i=1}^{n} [y_i \log(\eta(x_i)) + (1 - y_i) \log(1 - \eta(x_i))] \tag{5.21}$$

Sea $\text{logit}(\eta(x)) = (f_\beta(x))$, por tanto

$$f_\beta(x) = \log\left(\frac{\eta(x)}{1 - \eta(x)}\right) \tag{5.22}$$

de donde se deduce que

$$\eta(x) = \frac{e^{f_\beta(x)}}{1 + e^{f_\beta(x)}} \tag{5.23}$$

Desarrollando (5.21)

$$\frac{1}{n} l(\eta) = \frac{1}{n} \sum_{i=1}^{n} [y_i(\log(\eta(x_i)) - \log(1 - \eta(x_i))) + \log(1 - \eta(x_i))] \tag{5.24}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\log\left(\frac{\eta(x_i)}{1 - \eta(x_i)}\right) + \log(1 - \eta(x_i))\right] \tag{5.25}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[y_i f_\beta(x_i) + \log(1 + e^{f_\beta(x_i)})\right] \tag{5.26}$$

Veamos ahora que es una función convexa. Sea $h(z) = yz + \log(1 + e^z)$, veamos el signo de su segunda derivada:

$$h'(z) = y + \frac{e^z}{1 + e^z} \tag{5.27}$$

$$h''(z) = \frac{e^z}{(1 + e^z)^2} = \frac{e^z}{1 + e^z}\left(1 - \frac{e^z}{1 + e^z}\right) \geq 0 \tag{5.28}$$

que es positivo, y por tanto, la función $h(z)$ es convexa. Se concluye que la función de pérdida logística encaja en el marco general. Observamos finalmente que la regresión logística se puede usar como método de clasificación y la regla de decisión a utilizar será entonces

$$g(x) = \begin{cases} 1 & \text{si} \quad \frac{e^{f_\beta}}{1+e^{f_\beta}} \geq \frac{1}{2} \\ \\ 0 & \text{si} \quad \frac{e^{f_\beta}}{1+e^{f_\beta}} < \frac{1}{2} \end{cases} \tag{5.29}$$

Veamos ahora una adaptación del método del lasso en el caso de clasificación binaria con funciones de pérdida convexas.

Si $z_i = (x_i, y_i)$, definimos el riesgo empírico

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \rho_f(z_i) \tag{5.30}$$

y el riesgo esperado

$$R(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\rho_f(z_i) \tag{5.31}$$

El objetivo será localizar una f^0 tal que minimice el riesgo, es decir

$$f^0 = \underset{f \in F}{\operatorname{argmin}}\ R(f) \tag{5.32}$$

Para cualquier $f \in F$, definimos el exceso de riesgo como

$$\mathscr{E}(f) = R(f) - R(f^0) \tag{5.33}$$

Consideramos ahora el subespacio lineal

$$\mathcal{F} = \{f_\beta : \beta \in \mathbb{R}^d\} \tag{5.34}$$

donde f_β depende linealmente de β.

Sea la mejor aproximación al objetivo f_{β^0} con $\beta^0 = \underset{\beta}{\operatorname{argmin}}\ R(f_\beta)$. Sea el estimador del lasso $\hat{\beta}$ dado por

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ R_n(f_\beta) + \lambda\|\beta\|_1 \tag{5.35}$$

y $\hat{f} = f_{\hat{\beta}}$.

Para cada f_β, denotamos a la diferencia entre el riesgo empírico y el riesgo promedio por

$$\nu_n(\beta) = R_n(f_\beta) - R(f_\beta) \tag{5.36}$$

A partir de esta definición se cumplirá la siguiente desigualdad básica

5.6 Lema.

$$\mathscr{E}(f_{\hat{\beta}}) + \lambda\|\hat{\beta}\|_1 \leq -\Big[\nu_n(\hat{\beta}) - \nu_n(\beta)\Big] + \lambda\|\beta\|_1 + \mathscr{E}(f_\beta) \tag{5.37}$$

Demostración. Por la definición de $\hat{\beta}$ se cumplirá

$$R_n(f_{\hat{\beta}}) + \lambda\|\hat{\beta}\|_1 \leq R_n(f_\beta) + \lambda\|\beta\|_1$$

Si restamos a ambos lados de la desigualdad el término $R(f_{\beta^0})$, la desigualdad no cambiará. Y lo mismo sucederá si en el lado izquierdo sumamos y restamos el término $R(f_{\hat{\beta}})$ y repetimos en el lado derecho pero con el término $R(f_\beta)$. De esta forma, reagrupando términos tendremos

$$\mathscr{E}(f_{\hat{\beta}}) + \nu_n(\hat{\beta}) + \lambda\|\hat{\beta}\|_1 \leq \mathscr{E}(f_\beta) + \nu_n(\beta) + \lambda\|\beta\|_1$$

obteniendo la desigualdad a probar

$$\mathscr{E}(f_{\hat{\beta}}) + \lambda\|\hat{\beta}\|_1 \leq -\Big[\nu_n(\hat{\beta}) - \nu_n(\beta)\Big] + \lambda\|\beta\|_1 + \mathscr{E}(f_\beta)$$

□

Veamos la consistencia del método en esta situación.

Notación. Usaremos la notación abreviada

$$Z_M(\beta_0) = \sup_{\|\beta-\beta_0\|_1 \leq M} |\nu_n(\beta) - \nu_n(\beta_0)| \tag{5.38}$$

Sea β^* el minimizador del riesgo penalizado

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \ (R(\beta) + \lambda\|\beta\|_1) = \underset{\beta}{\operatorname{argmin}} \ (\mathscr{E}(f_\beta) + \lambda\|\beta\|_1) \tag{5.39}$$

y tomamos $f^* = f_{\beta^*}$. Podemos plantear el siguiente teorema:

5.7 Teorema. Sean

$$M^* = \frac{1}{\lambda_0}\left(\mathscr{E}(f^*) + 2\lambda\|\hat{\beta}\|_1\right) \tag{5.40}$$

y el conjunto

$$\mathcal{T} = \left\{Z_{M^*}(\hat{\beta}) \leq \lambda_0 M^*\right\} \tag{5.41}$$

Entonces si $\lambda \geq 4\lambda_0$, en el conjunto $\mathcal{T}$ se verifica

$$\mathscr{E}(\hat{f}) + \lambda\|\hat{\beta}\|_1 \leq 2\left(\mathscr{E}(f^*) + 2\lambda\|\beta^*\|_1\right) \tag{5.42}$$

Demostración. Sean

$$t = \frac{M^*}{M^* + \|\hat{\beta} - \beta^*\|_1}, \qquad \text{y} \qquad \tilde{\beta} = t\hat{\beta} + (1-t)\beta^*$$

Por la convexidad de R_n, se verificará

$$R_n(f_{\tilde{\beta}}) \leq tR_n(f_{\hat{\beta}}) + (1-t)R_n(f_{\beta^*}) \tag{5.43}$$

y además

$$\lambda\|\tilde{\beta}\|_1 \leq \lambda t\|\hat{\beta}\|_1 + \lambda(1-t)\|\beta^*\|_1 \tag{5.44}$$

Entonces se cumplirá

$$R_n(f_{\tilde{\beta}}) + \lambda\|\tilde{\beta}\|_1 \leq tR_n(f_{\tilde{\beta}}) + \lambda t\|\hat{\beta}\|_1 + (1-t)R_n(f_{\beta^*}) + \lambda(1-t)\|\beta^*\|_1 \tag{5.45}$$

y por ser $\hat{\beta}$ el minimizador de $R_n(f_\beta) + \lambda\|\beta\|_1$,

$$R_n(f_{\tilde{\beta}}) + \lambda\|\tilde{\beta}\|_1 \leq R_n(f_{\beta^*}) + \lambda\|\beta^*\|_1 \tag{5.46}$$

Añadiendo los términos correspondientes a cada lado de la desigualdad como en la demostración del lema 5.6, tendremos

$$\mathscr{E}(f_{\tilde{\beta}}) + \lambda\|\tilde{\beta}\|_1 \leq \mathscr{E}(f_{\beta^*}) + \lambda\|\beta^*\|_1 + \nu_n(\beta^*) - \nu_n(\hat{\beta}) \tag{5.47}$$

Por la definición de $\tilde{\beta}$, $\|\tilde{\beta} - \beta^*\|_1 = t\|\hat{\beta} - \beta^*\|_1 \leq 1$. Por tanto, podemos acotar la diferencia $\nu_n(\beta^*) - \nu_n(\hat{\beta})$ por $Z_{M^*}(\beta^*)$ y por estar en el conjunto $\mathcal{T}$, tendremos

$$\mathscr{E}(f_{\tilde{\beta}}) + \lambda\|\tilde{\beta}\|_1 \leq \lambda_0 M^* + \mathscr{E}(f_{\beta^*}) + \lambda\|\beta^*\|_1 \tag{5.48}$$

También se verificará

$$\mathscr{E}(f_{\tilde{\beta}}) + \lambda\|\tilde{\beta} - \beta^*\|_1 - \lambda\|\beta^*\|_1 \leq \mathscr{E}(f_{\tilde{\beta}}) + \lambda\|\tilde{\beta}\|_1 \tag{5.49}$$

y uniendo las dos desigualdades, obtenemos

$$\begin{aligned} \mathscr{E}(f_{\tilde{\beta}}) + \lambda\|\tilde{\beta} - \beta^*\|_1 &\leq \lambda_0 M^* + \mathscr{E}(f_{\beta^*}) + 2\lambda\|\beta^*\|_1 = 2\lambda_0 M^* \quad (5.50) \\ &\leq \frac{\lambda M^*}{2} \quad (5.51) \end{aligned}$$

y como $\mathscr{E}(f_{\tilde{\beta}}) \geq 0$, se verificará que $\|\tilde{\beta} - \beta^*\|_1 \leq M^*/2$. Por la definición de $\tilde{\beta}$ tendremos

$$\|\tilde{\beta} - \beta^*\|_1 = t\|\hat{\beta} - \beta^*\|_1 = \frac{M^*\|\hat{\beta} - \beta^*\|_1}{M^* + \|\hat{\beta} - \beta^*\|_1} \leq \frac{M^*}{2} \tag{5.52}$$

desigualdad que sólo se verificará en el caso de que $\|\hat{\beta} - \beta^*\|_1 \leq M^*$. Por este motivo podemos replicar el razonamiento y concluir que

$$\mathscr{E}(\hat{f}) + \lambda\|\hat{\beta}\|_1 + \nu_n(\hat{\beta}) \leq \mathscr{E}(f_{\beta^*}) + \lambda\|\beta^*\|_1 + \nu_n(\beta^*) \tag{5.53}$$

y por tanto

$$\mathscr{E}(f_{\hat{\beta}}) + \lambda\|\hat{\beta}\|_1 \leq \lambda_0 M^* + \mathscr{E}(f_{\beta^*}) + \lambda\|\beta^*\|_1 \tag{5.54}$$

que se seguirá verificando si añadimos un término $\lambda\|\beta^*\|_1$ a la derecha y sustituyendo M^* por el valor establecido en (5.40),

$$\mathscr{E}(f_{\hat{\beta}}) + \lambda\|\hat{\beta}\|_1 \leq 2\left(\mathscr{E}(f_{\beta^*}) + 2\lambda\|\beta^*\|_1\right) \tag{5.55}$$

□

A partir de este teorema, si suponemos que el modelo lineal es correcto, existirá β^0 tal que $f^0 = f_{\beta^0}$ y por tanto $\mathscr{E}(f^0) = 0$ y se verificará

$$\mathscr{E}(f_{\hat{\beta}}) + \lambda\|\hat{\beta}\|_1 \leq 4\lambda\|\beta^0\|_1 \tag{5.56}$$

de donde vemos que el riesgo de generalización para el parámetro estimado $\hat{\beta}$ se alejará del riesgo del modelo correcto una cantidad del orden de λ y además su norma también estará acotada por cuatro veces la del verdadero parámetro.

Bajo ciertas condiciones, se puede acotar la probabilidad del conjunto $\mathcal{T}$. En el caso de pérdida hinge, las desigualdades de simetrización-aleatorización

junto con desigualdades de concentración apropiadas, permiten probar que si las f_β son de la forma

$$f_\beta(\mathbf{x}) = \sum_{j=1}^{p} \beta_j \psi_j(\mathbf{x}) \quad \text{y} \quad \|\psi_j\|_\infty \leq 1$$

entonces

$$\mathbb{P}\left\{ Z_M(\beta^*) \leq M \left[\lambda(n,p) + \frac{t}{3n} + \sqrt{\frac{2t}{n}} \sqrt{1 + 8\lambda(n,p)} \right] \right\} \geq 1 - e^{-t} \quad (5.57)$$

donde

$$\lambda(n,p) = \sqrt{\frac{2\log(2p)}{n}} + \frac{\log(2p)}{3n}$$

Este es un resultado desarrollado por Bühlmann y van de Geer (2011) cuando las funciones de pérdida son de Lipschitz (en este caso $L = 1$), (ver [7, Teorema 14.5]).

De esta desigualdad se deduce que podemos elegir $\lambda \simeq \sqrt{\log p/n}$ de forma que

$$\mathscr{E}(f_{\hat{\beta}}) + \lambda \|\hat{\beta}\|_1 \leq 2\left(\mathscr{E}(f^*) + 2\lambda \|\beta^*\|_1\right)$$

Esto significa que, igual que en el caso de regresión, el exceso de error de la regla lasso es, salvo términos de orden $\sqrt{\log p/n}$ menor que el doble de la regla oráculo. Si además, una regla lineal clasifica sin error ($\mathscr{E}(f^*) = 0$) entonces la regla lasso satisface

$$\mathscr{E}(\hat{f}) \leq C\sqrt{\frac{\log p}{n}}$$

que muestra que la regla lasso también proporciona buena clasificación en casos en los que $p >> n$.

Capítulo 6

Conclusiones

A lo largo de este trabajo, se han explorado los fundamentos matemáticos del problema de clasificación y analizado el comportamiento probabilístico de los métodos tratados.

Se ha comprobado que una buena regla de clasificación requiere de un equilibrio a la hora de elegir la complejidad de la clase de reglas a considerar, que se puede conseguir mediante métodos de penalización.

La elección adecuada de la penalización, requiere del uso de herramientas probabilísticas como las desigualdades maximales y de concentración. El desarrollo de nuevas desigualdades permitirá entender correctamente el funcionamiento de otros métodos de aprendizaje supervisado.

Se ha podido comprobar que los buenos resultados teóricos y el buen comportamiento computacional no siempre van juntos, pues la minimización del riesgo estructural se comporta muy bien desde el punto de vista teórico pero es de escasa utilidad práctica.

Por otro lado, los métodos SVM tienen buenas propiedades computacionales pero su comportamiento probabilístico no parece tan satisfactorio.

Finalmente se ha visto que el lasso parece ser un método que combina bien ambos aspectos, por lo que está justificada la popularización de su uso en los últimos años.

Bibliografía

[1] L. Devroye, L. Györfi, G. Lugosi; *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.

[2] G. Lugosi; Pattern Classification and Learning Theory, *Principles of Nonparametric Learning*, International Centre for Mechanical Sciences Vol. 434 págs 1-56, Springer, 2002.

[3] P. Massart; *Concentration Inequalities and Model Selection*, Lecture Notes in Mathematics Vol. 1896. Ecole d'Eté de Probabilités de Saint-Flour XXXIII- 2003, Springer.

[4] N. Cristianini, J. Shawe-Taylor; *An introduction to Support Vector Machines and other kernel-based learning algorithms*, Cambridge University Press, 2000.

[5] J. Shawe-Taylor, N. Cristianini; *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[6] S. Boucheron, O. Bousquet, G. Lugosi; *Theory of Classification: A Survey of Some Recent Advances*, ESAIM: Probability and Statistics, Vol. 9 págs 323-375, 2005.

[7] P. Bühlmann, S. van de Geer; *Statistics for High-Dimensional Data*, Springer, 2011.