Universidad de Valladolid

Facultad de Ciencias

Trabajo Fin de Máster

Máster de Investigación en Matemáticas

Garantías Estadísticas para Métodos de Aprendizaje con Penalización l_1

Autor: M. Tereso del Río Almajano

Tutor: Eustasio del Barrio Tellado

Garantías Estadísticas para Métodos de Aprendizaje con Penalización l_1

Miguel Tereso del Río Almajano

18 de julio de 2020

Índice general

Capítulo 1

Introducción

Se estima que más de una cuarta parte de las muertes en España son resultado de un cáncer [12], dado que las posibilidades de sobrevivir a esta enfermedad aumentan significativamente si se detecta de forma temprana sería de gran utilidad saber que personas son más propensas a desarrollar un tumor para realizarles análisis regularmente.

Es sensato pensar que se puede tratar de determinar la predisposición de un individuo al cáncer analizando su genoma, pues esta secuencia de 3.000 millones de letras repartidas en 25.000 genes alberga toda la información genética del individuo y en ocasiones ha sido suficiente para predecir con exactitud la predisposición a una determinada enfermedad, como por ejemplo la enfermedad neurológica conocida como la fenilcetonuria causada por una mutación en un único gen.

sin embargo, cuando esta predisposición no depende de un único gen el problema se vuelve más complejo, además en este tipo de estudios el número de pacientes suele ser reducido, lo cual complica la extracción de conclusiones mediante los métodos tradicionales. Este hecho motivó la introducción de métodos que tuviesen garantías estadísticas incluso en escenarios donde el tamaño muestral sea inferior al número de variables.

En particular en esta memoria se presenta el estimador lasso tal como fue propuesto por Robert Tibshirani en 1996 en [16] y se trata de dotar de garantías estadísticas a dicho estimador trabajando en el marco teórico desarrollado por Peter Bühlmann y Sara A. van de Geer entre otros.

1.1. Motivación histórica del lasso

1.1.1. Mínimos cuadrados

En problemas de predicción, una solución que minimice los cuadrados de los errores muestrales además de ser sencilla de obtener, pues consiste en minimizar una función convexa, es en muchas ocasiones la que más beneficios posee.

sin embargo, en el contexto de la estadística de alta dimensión, es decir, en problemas como el descrito anteriormente donde el número de variables p es mayor que el tamaño de la muestra n, el algoritmo de mínimos cuadrados no cuenta con las garantías teóricas deseables.

Intuitivamente esto se debe a que la función a minimizar posee una infinidad de mínimos en problemas de alta dimensión. En particular la solución en el problema de regresión con matriz de datos X en estadística clásica (donde $p \leq n$) no se puede replicar puesto que requiere del cálculo de la inversa de la matriz de Gram $X^T X$ que es singular en estadística de alta dimensión, o de forma más precisa la solución de un sistema lineal no determinado, lo que conlleva una explosión en la varianza de los estimadores.

Con el fin de comprender mejor las limitaciones del algoritmo de mínimos cuadrados se calcula la distribución de la solución devuelta por este algoritmo en los casos que se pueden estudiar, es decir, cuando el número de variables es menor que el tamaño de la muestra y la matriz X tiene rango máximo. Se supone además que los datos provienen del modelo lineal

$$Y = X\beta^0 + \epsilon,$$

donde β^0 es un vector de coeficientes, donde ϵ es un vector de errores independientes e igualmente distribuidos (i.i.d.) $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ y donde la matriz de X tiene rango máximo. En este escenario es conocido que la solución que minimiza el error cuadrático en la muestra es

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Para calcular la distribución de $\hat{\beta}$ se empieza por el cómputo de su esperanza

$$\begin{aligned} \mathbb{E}(\hat{\beta}) &= \mathbb{E}((X^T X)^{-1} X^T Y) \\ &= (X^T X)^{-1} X^T \mathbb{E}(X\beta^0 + \epsilon) \\ &= (X^T X)^{-1} (X^T X)\beta^0 = \beta^0. \end{aligned}$$

Y puesto que

$$\begin{aligned}\hat{\beta} - \beta^0 &= (X^T X)^{-1} X^T Y - (X^T X)^{-1}(X^T X)\beta^0 \\ &= (X^T X)^{-1} X^T (Y - X\beta^0) \\ &= (X^T X)^{-1} X^T \epsilon,\end{aligned}$$

la varianza de $\hat{\beta}$ toma el siguiente valor

$$\begin{aligned}\mathbb{E}\left[(\hat{\beta} - \beta^0)(\hat{\beta} - \beta^0)^T\right] &= \mathbb{E}\left[((X^T X)^{-1} X^T \epsilon\epsilon^T X((X^T X)^{-1})^T)\right] \\ &= (X^T X)^{-1} X^T \mathbb{E}[\epsilon\epsilon^T] X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} \\ &= (X^T X)^{-1}\sigma^2.\end{aligned}$$

Además, por ser $\hat{\beta}$ una combinación lineal de normales sigue una distribución normal, en concreto, $\hat{\beta}$ sigue la distribución $\mathcal{N}(\beta^0, (X^T X)^{-1}\sigma^2)$. Por tanto en el límite, cuando la matriz X tiende a dejar de tener rango máximo p, la varianza de $\hat{\beta}$ se dispara. Esto es suficiente argumento para descartar el uso del método de los mínimos cuadrados en problemas donde el número de variables sea mayor que el tamaño de la muestra, en los que claramente la matriz $X^T X$ es singular.

Mientras las columnas de X sean linealmente independientes la matriz $X^T X$ va a tener rango p, cuando $X^T X$ es aproximadamente singular, el determinante de $X^T X$ estará cerca de cero y por tanto el determinante de $(X^T X)^{-1}$ crece hacia infinito. La suma de las varianzas de los estimadores $\hat{\sigma}_j$ es la traza de $(X^T X)^{-1}\sigma$. Además si el determinante de una matriz es muy grande es porque alguno de los autovalores es muy grande y por lo tanto la traza también será grande, de donde se deduce que alguna de las varianzas debe ser muy grande. Así se comprende que cuando nos aproximamos a tener una matriz de Gram singular la varianza de los autovalores explota.

1.1.2. El criterio C_p de Mallows

En [13] Mallows introdujo un criterio para seleccionar el modelo de regresión lineal más apropiado cuando se dispone de una colección grande de regresores, en particular, cuando el número de regresores es mayor que el tamaño de la muestra. Este criterio es aplicable bajo la suposición de que los datos han sido extraídos de un modelo lineal normal

$$Y = X\beta^0 + \epsilon,$$

donde $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ y donde β^0 es un vector de coeficientes que únicamente tiene entradas no nulas en el conjunto activo $S_0 = \{j : \beta_j^0 \neq 0\}$. Para resaltar que β^0 es un vector disperso el modelo lineal normal se podría reescribir como

$$Y = X_{S_0}\beta^0_{S_0} + \epsilon, \tag{1.1}$$

donde X_{S_0} es la matriz de datos con ceros en las columnas asociadas al conjunto S_0^C.

A diferencia del criterio de los mínimos cuadrados que se centran únicamente en la predicción, este criterio permite realizar una selección de variables. Esto puede ser útil puesto que en ocasiones no solo interesa la predicción de la respuesta si no que también interesa conocer qué variables influyen en la respuesta. Por ejemplo, al tratar de diseñar el sistema de diagnóstico de un determinado tumor, es razonable tomar miles de características sobre la muestra, pero este análisis puede resultar costoso y por tanto sería deseable extraer unos pocos marcadores que permitan realizar una prueba barata y efectiva para la detección de dicho tumor.

En dicho artículo Mallows plantea considerar todas las regresiones posibles por subconjuntos de variables $S \subset \{1, 2, ..., p\}$ tales que $|S| \leq n$ y que X_S, la matriz X restringida a las columnas de S, tenga rango máximo.

Para cada conjunto $S \subset \{1, 2, ..., p\}$ que cumpla los requisitos previos el estimador de mínimos cuadrados de los coeficientes de regresión es

$$\hat{\beta}_S = (X_S^T X_S)^{-1} X_S^T Y$$

y los valores predichos con este modelo son las componentes del vector

$$\hat{Y}(S) = X_S \hat{\beta}_S.$$

Sería deseable seleccionar el mejor de estos modelos, pero para ello se debe concretar qué se entiende por una predicción de calidad. Para evaluar lo bueno que es cada uno de estos modelos prediciendo la etiqueta en una localización fija (con la matriz de datos X constante) se usa el riesgo de predicción

$$R(S) = \mathbb{E}\big[\|\hat{Y}(S) - Y^*\|_2^2\big],$$

donde $Y^* = \beta^0 X + \varepsilon^*$, siendo ε^* un vector de errores i.i.d. $N(0, \sigma^2)$ independientes del vector ε. Y^* representa el vector de observaciones generadas en las mismas condiciones que las de Y, de forma independiente.

Es inmediato que el conjunto S_0 minimizará el riesgo de predicción siempre que este conjunto cumpla los requisitos exigidos a los subconjuntos para realizar regresión sobre ellos. Sin embargo, S_0 es desconocido y el riesgo de predicción no es realmente útil para encontrarlo puesto que no es directamente observable.

Para estimar el riesgo de predicción se utiliza el error de entrenamiento

$$R_n = \|\hat{Y}(S) - Y\|^2,$$

como estimador, pero como se prueba a continuación este estimador no es insesgado. Para comprobarlo, se observa que

$$\hat{Y}(\mathcal{S}) - Y^* = -(I_n - X_{\mathcal{S}}(X_{\mathcal{S}}^T X_{\mathcal{S}})^{-1} X_{\mathcal{S}}^T) X_{\mathcal{S}_0}\beta_{\mathcal{S}_0} + X_{\mathcal{S}}(X_{\mathcal{S}}^T X_{\mathcal{S}})^{-1} X_{\mathcal{S}}^T \varepsilon - \varepsilon^*,$$

y de aquí se deduce que

$$R(\mathcal{S}) = \|(I_n - X_{\mathcal{S}}(X_{\mathcal{S}}^T X_{\mathcal{S}})^{-1} X_{\mathcal{S}}^T) X_{\mathcal{S}_0}\beta_{\mathcal{S}_0}\|_2^2 + \sigma^2(n + |\mathcal{S}|),$$

donde en el último cálculo se ha utilizado que si Z_1 y Z_2 son vectores aleatorios de la misma dimensión, independientes y al menos uno es centrado entonces

$$E\|Z_1 + Z_2\|_2^2 = E\|Z_1\|_2^2 + E\|Z_2\|_2^2$$

y que si Z es un vector aleatorio centrado con matriz de covarianza Σ entonces para toda constante a

$$E\|Z + a\|_2^2 = \|a\|_2^2 + \mathrm{Tr}(\Sigma),$$

además se ha utilizado que cuando ambas multiplicaciones entre matrices son posibles $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$.

Con un cálculo parecido se obtiene

$$\mathbb{E}\,[R_n] = \|(I_n - X_{\mathcal{S}}(X_{\mathcal{S}}^T X_{\mathcal{S}})^{-1} X_{\mathcal{S}}^T) X_{\mathcal{S}_0}\beta_{\mathcal{S}_0}\|_2^2 + \sigma^2(n - |\mathcal{S}|),$$

de donde se concluye que R_n es un estimador sesgado de $R(\mathcal{S})$, con sesgo $-2\sigma^2(|\mathcal{S}|)$.

Esta observación da lugar al estadístico de Mallows

$$C_p = R_n + 2\hat{\sigma}^2|\mathcal{S}|,$$

donde $\hat{\sigma}^2$ es un estimador de la varianza residual en el modelo completo, el modelo que considera todos los atributos. Es inmediato de los cálculos anteriores ver que C_p es un estimador de $R(S)$ aproximadamente insesgado.

Esto motiva la introducción del *El criterio C_p de Mallows* que escoge de los modelos de regresión considerados el que minimiza el estadístico C_p, es decir, el asociado al vector

$$\hat{\beta}_{C_p} = \underset{\beta}{\operatorname{argmin}} \Big[\|X\beta - Y\|_2^2 + 2\hat{\sigma}^2 \|\beta\|_0 \Big], \tag{1.2}$$

donde $\|\beta\|_0 = \sum_{j=1}^{p} (\beta_j)^0$ es el número de componentes no nulos de β.

En el criterio C_p de Mallows penaliza $2\sigma^2$ veces la 'norma' l_0 del vector asociado a la solución, por tanto se puede entender como un caso concreto del grupo de estimadores que penalizan por un múltiplo de la 'norma' l_0

$$\hat{\beta}_{MejorSubconjunto}(\lambda) = \underset{\beta}{\operatorname{argmin}} \Big[\|X\beta - Y\|_2^2 + \lambda \|\beta\|_0 \Big]. \tag{1.3}$$

Las penalizaciones de este tipo dan lugar a estimadores que se conocen como estimadores del mejor subconjunto, pues es evidente que estos estimadores buscan tener soporte en conjuntos lo más pequeños posibles, pues son penalizados en función del tamaño de los mismos.

Con la introducción de estos estimadores nace una idea muy potente, la idea de añadir una penalización en función de la complejidad del modelo. Esta penalización favorece a los modelos simples, evocando al concepto de la Navaja de Ockham, concepto que como se discute en el tercer capítulo de [4] tiene una interpretación interesante en el contexto de selección de modelos en aprendizaje automático

Por ejemplo, otra penalización posible y más común en la práctica es la que da lugar al estimador ridge, que penaliza por un múltiplo de la norma l_2 escogiendo el modelo lineal asociado al vector

$$\hat{\beta}_{ridge}(\lambda) = \underset{\beta}{\operatorname{argmin}} \|X\beta - Y\|_2^2 + \lambda \|\beta\|_2^2, \tag{1.4}$$

donde λ es fijo y se conoce como parámetro de penalización.

Muchas otras penalizaciones se pueden considerar, pero una particularmente interesante es la que da lugar al estimador lasso. Este estimador introduce una penalización proporcional a la norma en l_1 del estimador

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left\{ \|X\beta - Y\|_2^2/n + \lambda \|\beta\|_1 \right\}, \tag{1.5}$$

donde λ es el parámetro de penalización.

El estimador lasso recibe su nombre del acrónimo en inglés de Least Absolute Shrinkage and Selection Operator, el operador que es mínimo global en disminución y selección (de variables) en castellano, los motivos de este nombre se destacan durante esta introducción.

En ambos estimadores el parámetro de penalización λ regula la importancia que se le otorga a la complejidad de un modelo con respecto a la calidad de las predicciones que realiza, a mayor λ más se favorece a los modelos simples. En la práctica este parámetro se suele escoger mediante validación cruzada (cross validation en la literatura inglesa), en cambio durante este trabajo en vez de buscar las garantías que se obtienen escogiendo λ de esta manera se tratan las garantías estadísticas que se pueden obtener para un parámetro λ fijo.

1.2. El estimador lasso

Esta memoria está centrada en el estudio del estimador lasso y en la obtención de garantías estadísticas sobre el mismo, en particular garantías sobre su capacidad de predicción y cuando sea posible sobre su capacidad de aproximación de la solución real.

Se distinguen en los capítulos posteriores dos clases de garantías, garantías generales que permiten asegurar la consistencia del estimador en escenarios genéricos y garantías bajo dispersión estadística válidas cuando el conjunto S_0 descrito previamente en (1.1) es pequeño, es decir, cuando únicamente unos pocos atributos sean relevantes a la hora de calcular la etiqueta.

Las garantías sobre dispersión estadística se recogen en *desigualdades oráculo*, estas desigualdades oráculo aseguran que el estimador lasso es capaz de adaptarse a la situación de dispersión y que su rendimiento estadístico es aproximadamente igual que el que obtendríamos aplicando el método de mínimos cuadrados al modelo reducido válido que es desconocido a priori. Para entender más sobre este tipo de resultados pueden leerse las discusiones sobre los Teoremas 2.3.2, 3.2.1 y 4.2.1.

El último capítulo de esta memoria se dedica a proporcionar las garantías estadísticas asociadas a reglas de aprendizaje automático obtenidas mediante la penalización l_1 con funciones de pérdida convexas generales. Esto incluye entre otros el problema de clasificación binaria con máquinas de soporte vectorial o con

regresión logística cuando se utiliza la penalización l_1 en las dos frases.

El primer beneficio que se observa al utilizar este estimador y una de las razones de su nombre es que una gran parte de los parámetros estimados valen 0. De hecho, como se prueba en [7] el número de parámetros no nulos es a lo sumo el tamaño de la muestra n, es decir, la clase de funciones que el lasso considera no es la clase de funciones lineales de los atributos completa, si no la clase de funciones lineales de a lo sumo n atributos, por lo que de manera natural el estimador lasso realiza una selección de atributos sobre la que se pueden alcanzar garantías.

Además en [3] se prueba que bajo un modelo lineal con parámetro real β^0 y siendo S_λ el conjunto de atributos seleccionados por el lasso con penalización λ

$$\mathbb{P}\left[\{j : \beta_j^0 \geq a\} \subset S_\lambda\right] \underset{n\to\infty}{\longrightarrow} 1,$$

para todo a fijo. Sin embargo, este tipo de resultados queda fuera de los márgenes del trabajo.

Para comprender intuitivamente por qué tantos parámetros se anulan vamos a comparar el lasso con el estimador ridge que se presentó en la sección anterior. La comparación es más sencilla si utilizamos la versión restringida en lugar de la versión penalizada. Es fácil probar, como se hace en el quinto capítulo de [2], que por la convexidad de las bolas en l_1 existe una equivalencia entre las soluciones de (1.5) y las de

$$\hat{\beta}_{equiv}(R) = \underset{\|\beta\|_1 \leq R}{\operatorname{argmin}} \left\{\|X\beta - Y\|_2^2/n\right\},$$

en particular, se prueba que dados X e Y existe una función biyectiva $f : \mathbb{R} \to \mathbb{R}$ que asigna un R a cada λ de tal forma que $\hat{\beta}(\lambda) = \hat{\beta}_{equiv}(f(\lambda))$. Y de forma análoga existe una equivalencia entre las soluciones de (1.4) y las de

$$\hat{\beta}_{ridge,equiv}(R) = \underset{\|\beta\|_2 \leq R}{\operatorname{argmin}} \left\{\|X\beta - Y\|_2^2/n\right\}.$$

En ambos casos se observa que la solución del problema equivalente es el vector que minimiza el error de entrenamiento dentro de una bola centrada en el origen.

En el caso de dos variables gracias a estas equivalencias se puede visualizar en el plano porqué en el estimador obtenido mediante lasso se anula una gran cantidad de coeficientes, pues como se observa en la figura 1.1, para que el primer coeficiente en el estimador obtenido mediante ridge se anule, la curva de nivel que pasa por el $(0, R)$ debe hacerlo con pendiente cero, es decir, se anula con probabilidad 0. Sin embargo, para que se anule el primer coeficiente en el estimador obtenido mediante lasso es suficiente que la curva de nivel que pasa por el $(0, R)$ lo haga

con una pendiente menor que -1, por lo que este se anula con una probabilidad bastante alta.

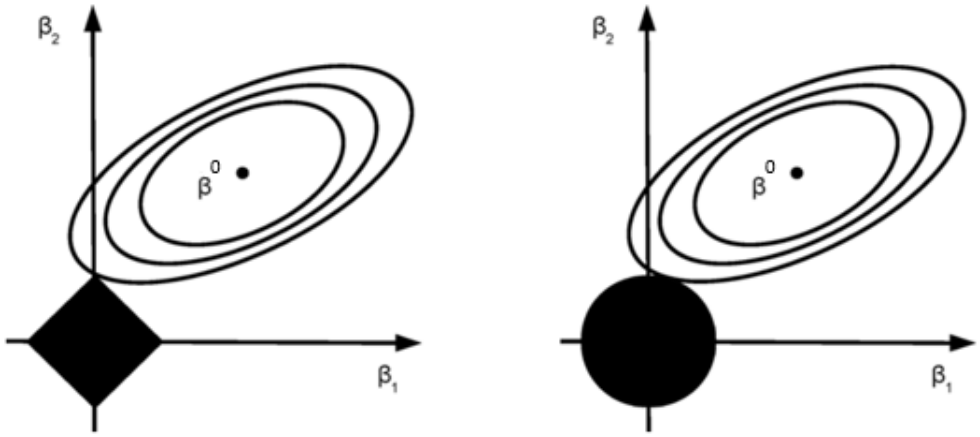

Figura 1.1: Comparación de las curvas de nivel del error de entrenamiento con las bolas de norma l_1 asociada al lasso (izquierda) y de norma l_2 asociada al ridge (derecha), figura extraída de [3]

Podría pensarse que se puede diseñar un método que haga una selección más fuerte que el lasso simplemente sustituyendo la penalización sobre la norma l_1 por una penalización sobre el criterio l_q con $q < 1$. Puesto que eso significaría que la condición sobre la pendiente de la curva de nivel en $(0, R)$ para que un coeficiente se anule sería más débil todavía, de hecho se podría debilitar tanto como se quiera tomando un q suficientemente pequeño. sin embargo, esto no es posible, puesto que las bolas en l_q con $q < 1$ no son convexas y esta era una condición necesaria para probar la equivalencia entre las dos formas de expresar los métodos vistos en esta sección.

Por el mismo motivo tampoco se puede encontrar una equivalencia directa entre las soluciones de estimadores de selección de variables, similares al obtenido del Criterio de Mallows, que penalicen por el criterio l_0 y las de

$$\hat{\beta}_{C_p,eqiv}(k) = \underset{\|\beta\|_0 \leq k}{\operatorname{argmin}} \left\{ \|X\beta - Y\|_2^2/n \right\}. \tag{1.6}$$

Además tanto (1.2) como (1.6) presentan problemas computacionales, puesto que no existen algoritmos capaces de encontrar los mínimos necesarios en tiempo

polinómico. Por ello es mucho más conveniente utilizar en la práctica el estimador ridge, para el que existe una solución explicita, o el estimador lasso, para el que como se ve a continuación existen algoritmos de coste computacional bajo capaces de aproximar dicho estimador. Esto es un alivio, pues aunque el estimador lasso no penaliza directamente sobre el tamaño del conjunto activo sí que posee propiedades de selección de variables (Teorema 2.1.1).

1.2.1. Aspectos computacionales

Volviendo al estudio del lasso, desde un punto de vista numérico es claro que aunque la función que tratamos de minimizar es convexa no se puede utilizar el método del descenso del gradiente en su forma clásica para aproximar el mínimo global puesto que la función carece de diferenciabilidad, por no ser diferenciable la norma l_1. Por ello se utiliza el método de descenso por coordenadas que trata de minimizar la función variando en cada paso una única coordenada.

El objetivo de esta sección es mostrar que el estimador lasso se puede calcular de manera sencilla mediante un método iterativo, no volveremos a tratar aspectos computacionales en el resto de esta memoria.

Veamos a continuación cómo funciona este método en el caso relevante para el uso del lasso con función de pérdida cuadrática, en donde se trata de minimizar en cada paso la función

$$f(\beta) = \|X\beta - Y\|_2^2/n + \lambda\|\beta\|_1$$

variando únicamente la coordenada j de β.

Para encontrar el mínimo se calcula el valor de la derivada en los puntos de continuidad ($\beta_j \neq 0$),

$$\begin{aligned}\frac{\partial f(\beta)}{\partial \beta_j} =& \frac{2}{n}\left(X^{(j)}\right)^T (X\beta - Y) + signo(\beta_j)\lambda \\ =& \frac{2}{n}\left(X^{(j)}\right)^T (X^{(j)}\beta_j + X\beta_{-j} - Y) + signo(\beta_j)\lambda \\ =& \frac{2}{n}\left|\left|X^{(j)}\right|\right|_2^2 \beta_j + \frac{2}{n}\left(X^{(j)}\right)^T (X\beta_{-j} - Y) + signo(\beta_j)\lambda,\end{aligned}$$

donde $X^{(j)}$ denota la columna j de la matriz X y β_{-j} denota el vector β con un cero en la coordenada j.

Por ser la derivada una función creciente respecto de β_j la función alcanzará el mínimo global en donde se anule y si no se anula lo alcanzará en $\beta_j = 0$ por ser el único punto de discontinuidad, es decir, el mínimo global se alcanza en:

$$\beta_j = \begin{cases} \frac{\lambda/2 - \frac{1}{n}\left(X^{(j)}\right)^T (X\beta_{-j} - Y)}{\frac{1}{n}\left|\left|X^{(j)}\right|\right|_2^2} & si \quad \frac{1}{n}\left(X^{(j)}\right)^T (X\beta_{-j} - Y) > \lambda/2 \\ 0 & si \quad \frac{1}{n}\left|\left(X^{(j)}\right)^T (X\beta_{-j} - Y)\right| < \lambda/2 \\ \frac{-\lambda/2 - \frac{1}{n}\left(X^{(j)}\right)^T (X\beta_{-j} - Y)}{\frac{1}{n}\left|\left|X^{(j)}\right|\right|_2^2} & si \quad \frac{1}{n}\left(X^{(j)}\right)^T (X\beta_{-j} - Y) < -\lambda/2 \end{cases}$$

Se puede ver de forma más clara definiendo $Z_j := \frac{1}{n}\left(X^{(j)}\right)^T (X\beta_{-j} - Y)$, utilizando la notación $(x)_+$ que denota la parte positiva de x y recordando que la matriz X está normalizada de acuerdo con (2.4),

$$\beta_j = signo(Z_j)(|Z_j| - \lambda/2)_+ \tag{1.7}$$

En la figura 1.2 se puede ver gráficamente para qué valor de β_j se alcanza el mínimo con respecto al valor de Z_j, esta gráfica sin duda refuerza la idea intuitiva presentada previamente de que el estimador lasso devuelve un vector con una gran cantidad de coeficientes nulos.

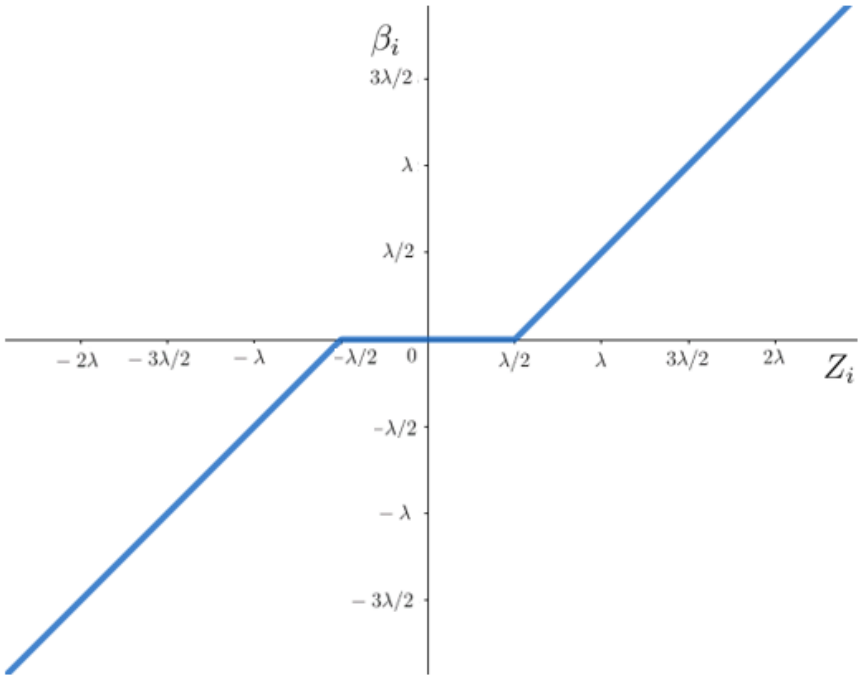

Figura 1.2: Actualización del coeficiente β_j

Esto es suficiente para describir el método de descenso por coordenadas que se utiliza para obtener una solución aproximada de (1.5) a partir de un β inicial, que consiste en repetir varias veces el siguiente bucle

for j **from** 1 **to** p:
    Se actualiza el valor de β_j de acuerdo con (1.7).

Se ha demostrado que el algoritmo de descenso por coordenadas converge de forma rápida hacia el minimizador de la función, por lo que se tiene un algoritmo eficiente para el cálculo del estimador lasso. Para más detalles se puede consultar el quinto capítulo de [8].

1.3. Herramientas utilizadas

El elemento probabilístico fundamental para poder dar garantías estadísticas sobre el funcionamiento del lasso son las desigualdades maximales, es decir, buenas cotas para la probabilidad de que el máximo de un proceso estocástico esté por encima de un determinado valor. En el segundo y en el tercer capítulo (a excepción de la sección 2.1.2) estas cotas son relativamente simples, pues el proceso empírico que querremos controlar es el máximo de un número finito de variables normales y por tanto es suficiente con utilizar la Cota de Chernoff. En la sección 2.1.2 dejamos de suponer que los errores son Gaussianos por lo que será necesario combinar herramientas estadísticas más sofisticadas para controlar el tamaño del máximo de un número finito de variables no Gaussianas, en dicha sección utilizaremos la Desigualdad de momentos de Hoeffding junto con una versión simplificada de la Desigualdad de Simetrización (ambos están enunciados en el apéndice) para alcanzar un caso particular de la Desigualdad de momentos de Nemirovski.

Cuando nos salimos del marco más sencillo de regresión lineal y de errores Gaussianos tenemos que enfrentarnos con el proceso empírico general, cuando las funciones de perdida utilizadas son convexas comprobaremos que un truco simple nos permite reducir el problema a controlar el máximo de un proceso empírico en una bola para la norma l_1. En este caso debemos recurrir de nuevo a métodos de simetrización y aleatorización para producir una buena desigualdad maximal. Podríamos utilizar la desigualdad de Markov para encontrar cotas probabilísticas sencillas a partir de las desigualdades maximales para momentos, si se quieren obtener cotas más finas hay que recurrir a algún tipo de desigualdad de concentración, en este contexto una de las más utilizadas es la Desigualdad de Bousquet que se enuncia en el Teorema 4.1.2.

La obtención de buenas desigualdades maximales y de concentración es un campo muy amplio, aquí se ha tocado de una forma tangencial, nos hemos limitado a utilizar los resultados apropiados, por lo que en general no se han incluido demostraciones de estos resultados.

1.4. Estructura de la memoria

En las secciones anteriores hemos motivado el sentido que tiene modificar el criterio de mínimos cuadrados mediante la penalización lasso. En el resto del trabajo nos vamos a dedicar a justificar que este tipo de penalización l_1 tiene una serie de garantías estadísticas. En el capítulo dos vamos a estudiar las garantías que pueden darse en el caso más sencillo posible, es decir, cuando asumimos un modelo lineal. Después estudiamos garantías en una situación más general en la que dejamos de suponer que existe un modelo lineal correcto, en ese caso todavía podemos dar ciertas garantías sobre la calidad del lasso en predicción. Finalmente el último capítulo se dedica a estudiar las garantías que se pueden obtener sobre el lasso cuando se utiliza una función de pérdida convexa que no es necesariamente la función de pérdida cuadrática.

Durante toda la memoria se destaca en las *desigualdades oráculo* la sorprendente capacidad del lasso de adaptarse a situaciones de dispersión consiguiendo garantías casi tan buenas como si se dispusiese de información adicional.

Como conclusión en el capitulo cinco comparamos las garantías que se obtienen para este procedimiento frente a las que se podrían obtener bajo paradigmas de aprendizaje más restrictivos como por ejemplo el aprendizaje PAC descrito en la Definición 5.0.1. Hay que destacar que las cotas que se obtiene aquí están fuertemente asociadas a la métrica l_1 con la que se dota el espacio paramétrico, veremos en el capítulo cinco como este tipo de cotas proporcionan garantías mucho más finas de las que se podrían obtener mediante procedimientos más generales basados en la dimensión de Vapnik-Chervonenkis (ver [14]).

Capítulo 2

Garantías estadísticas para el lasso bajo un modelo lineal

Después de esta pequeña introducción al lasso pasamos al tema central del trabajo, la obtención de garantías teóricas sobre la solución obtenida mediante dicho estimador. En los últimos treinta años se han propuesto una gran cantidad de métodos para resolver distintas variantes en los problemas principales en *Machine Learning*, que, resumiendo mucho, podríamos agrupar en regresión y clasificación. Entre los métodos más usados hay que incluir a las máquinas de soporte vectorial (SVM), árboles y bosques aleatorios o redes neuronales de muy distintos tipos. Las redes neuronales son la base de una gran parte de los algoritmos que consiguen el mejor rendimiento en muchos problemas aplicados, como puede comprobarse con una consulta a la página `https://www.kaggle.com/competitions`. Sin embargo, hay situaciones en las que es crucial disponer de garantías estadísticas sobre el rendimiento de un método. No parece aceptable que, por ejemplo, un sistema de conducción autónomo se base en algoritmos aprendidos sobre un conjunto de entrenamiento sin disponer de una cuantificación de los riesgos asociados a las decisiones que se deban tomar.

La disponibilidad de garantías estadísticas es una fortaleza del lasso. Mientras que es reconocido que métodos como las redes neuronales funcionan como cajas negras, para las que no se dispone de garantías estadísticas (ver, por ejemplo, [9]), en el caso del lasso se dispone, tal como veremos en este trabajo, de garantías de distinto tipo sobre su funcionamiento. Por un lado, probaremos cotas sobre el error de predicción del estimador lasso y veremos que es capaz de funcionar incluso en situaciones de alta dimensión: si el logaritmo del número de variables

explicativas es pequeño frente al número de casos, entonces el lasso es todavía capaz de garantizar un error de predicción pequeño. Pero las ventajas del lasso son más aparentes bajo *dispersión*. Con este término nos referimos a la situación en la que el número de variables explicativas que realmente tiene influencia en la respuesta es pequeño. Si supiésemos a priori cuáles de esas variables influyen en la respuesta tendríamos una gran ventaja desde el punto de vista estadístico. sin embargo, hay cada vez más situaciones en las que se mide una cantidad enorme de variables sobre cada uno de los individuos en un estudio porque la tecnología lo permite, pero sin que se tenga una idea clara de los factores que pueden ayudar a explicar la respuesta. No es raro encontrar datos de un estudio en el que a unos pocos cientos de individuos se les ha efectuado un estudio genético por el que se dispone de 500000 mediciones sobre cada uno de ellos. La estimación precisa de los tantos coeficientes de regresión sería imposible con métodos más clásicos, pero veremos que el lasso es capaz de funcionar (casi) como si conociese de antemano las variables verdaderamente influyentes en la respuesta.

Comenzamos el estudio de este tipo de resultados con el caso en el que los datos provienen de un modelo lineal normal, es decir, en el que las etiquetas se obtienen, para cierto β^0, de

$$Y = X\beta^0 + \epsilon, \tag{2.1}$$

donde ϵ es un vector de errores independientes e igualmente distribuidos (i.i.d.), que durante este capítulo se supone que sigue una distribución $\mathcal{N}(0, \sigma^2 I)$.

Bajo este modelo es claro que la regla de Bayes es lineal, por tanto no solo se buscarán garantías sobre lo bien que aproxima la solución devuelta por el lasso,

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left\{ \|X\beta - Y\|_2^2/n + \lambda\|\beta\|_1 \right\}, \tag{2.2}$$

si no que puesto que en este caso es posible, también se estudia la proximidad entre $\hat{\beta}$ y el valor real del parámetro β^0.

No nos planteamos de momento la selección del valor λ, de forma que normalmente simplificaremos la notación y escribiremos β en lugar de $\beta(\lambda)$.

Escribiremos $Y = [Y_1, \ldots, Y_n]^T$, $X_i = [X_{i,1}, \ldots, X_{i,n}]^T$ para los vectores fila de X y $X^{(j)} = [X_{1,j}, \ldots, X_{n,j}]^T$ para los vectores columna. Si medimos el error en términos de pérdida cuadrática,

$$l(Y, X; \beta) = \left(Y - X^T\beta\right)$$

entonces la pérdida promedio

$$R_n(\beta) := \frac{1}{n}\sum_{j=1}^{n} l\left(Y_i, X_i; \beta\right) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i^T\beta\right) = \frac{1}{n}\|Y - X\beta\|^2$$

es la parte principal. de la función criterio que se minimiza para obtener β. $R_n(\beta)$ es el riesgo empírico asociado a β (y a la perdida cuadrática). El valor promedio de $R_n(\beta)$ es el riesgo asociado a β, es decir,

$$R(\beta) = \frac{1}{n}\mathbb{E}\left\|Y - X\beta\right\|^2 = \frac{1}{n}\mathbb{E}\left\|X\left(\beta - \beta_0\right) + \epsilon\right\|^2 = \frac{1}{n}\left\|X\left(\beta - \beta_0\right)\right\|^2 + \sigma^2, \quad (2.3)$$

donde σ^2 es independiente de β.

Si $\hat{\beta}$ es un estimador de β (el lasso o cualquier otro) entonces $R(\hat{\beta})$ es una variable aleatoria. Estaremos interesados en dar cotas probables para $R(\hat{\beta})$ (es decir, en probar que $R(\hat{\beta})$ es pequeño con alta probabilidad). A la vista de (2.3), esto es equivalente dar cotas para el *error de predicción*,

$$\frac{1}{n}\left\|X\left(\hat{\beta} - \beta_0\right)\right\|^2,$$

que serán el tipo de cotas que buscamos durante esta memoria.

Puesto que estamos asumiendo que las entradas de X son deterministas no hay pérdida de generalidad en asumir que las columnas de X están normalizadas, es decir, que

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i = 0 \\ \frac{1}{n}\sum_{i=1}^{n} X_{ij}^2 &= 1 \ \forall j \leq p. \end{aligned} \quad (2.4)$$

Notamos aquí que en el modelo que estamos utilizando no se suele incluir un término independiente, en la práctica se aplica el estimador lasso una vez que el vector de etiquetas ha sido centrado. Esto hace que haya un pequeño desfase entre la práctica habitual y la teoría que de desarrolla en esta memoria, pero este es el procedimiento habitual en la literatura estadística sobre el lasso (ver el sexto capítulo de [3]).

2.1. Garantías generales

En esta sección se estudian los resultados que pueden obtenerse suponiendo únicamente que la matriz de datos X está normalizada de acuerdo con (2.4). Asumiremos que X es determinista y escribiremos $\hat{\beta}$ en lugar de $\hat{\beta}(\lambda)$ para la solución de (2.2). La clave para obtener una cota para el error de predicción está en ser capaz de controlar la probabilidad de que cierto proceso (Gaussiano) tome valores grandes. Vamos a tratar de justificar esta afirmación. Por definición

$$\frac{1}{n}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\|Y - X\beta^0\|_2^2 + \lambda\|\beta^0\|_1.$$

Además, puesto que para $Y = X\beta^0 + \epsilon$ (asumimos que el modelo lineal es correcto)

$$\frac{1}{n}\|\epsilon + X(\beta^0 - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\|\epsilon\|_2^2 + \lambda\|\beta^0\|_1$$

y desarrollando la primera norma se obtiene

$$\frac{1}{n}\|\epsilon\|_2^2 + \frac{2}{n}\epsilon^T X(\beta^0 - \hat{\beta}) + \frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\|\epsilon\|_2^2 + \lambda\|\beta^0\|_1,$$

de donde se deduce de forma inmediata

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{2}{n}\epsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1. \tag{2.5}$$

Si denotamos por $X^{(j)}$ la columna j de X, de (2.5) se deduce

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \max_{1\leq j\leq p}\left|\frac{2}{n}\epsilon^T X^{(j)}\right|\|\hat{\beta} - \beta^0\|_1 + \lambda\|\beta^0\|_1. \tag{2.6}$$

El término $\max_{1\leq j\leq p}\left|\frac{2}{n}\epsilon^T X^{(j)}\right|$ es el máximo (valor absoluto) de un proceso Gaussiano centrado muy simple. Esta formado por p variables aleatorias normales estándar (por la condición de normalización de las columnas de X). Si X es una variable aleatoria normal estándar la cota de Chernoff (ver (A.1) en el Apéndice) nos dice que

$$P(|Z| \geq t) \leq 2e^{-\frac{t^2}{2}}, \quad t > 0.$$

Si $Z_j = \frac{1}{\sqrt{n}\sigma}\epsilon^T X^{(j)}$ entonces

$$P\Big(\max_{1\leq j\leq p}|Z_j| \geq t\Big) = P\Big(\bigcup_{j=1}^{p}(|Z_j| \geq t)\Big) \leq \sum_{j=1}^{p} P\Big(|Z_j| \geq t\Big) \leq 2pe^{-\frac{t^2}{2}}.$$

Esto nos dice que si

$$\lambda_0 := 2\sigma\sqrt{\frac{2\log(2p/\delta)}{n}}, \tag{2.7}$$

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

$$\frac{2\sigma}{\sqrt{n}} \max_{1\leq j\leq p} \left| \frac{1}{\sqrt{n}\sigma} \epsilon^T X^{(j)} \right| \leq \lambda_0 \tag{2.8}$$

y esto, junto con (2.6), implica que

$$\frac{1}{n}\|X(\hat{\beta}-\beta^0)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \lambda_0\|\hat{\beta}-\beta^0\|_1 + \lambda\|\beta^0\|_1. \tag{2.9}$$

Entonces, si tomamos $\lambda \geq 2\lambda_0$ tendremos

$$\begin{aligned}\frac{2}{n}\|X(\hat{\beta}-\beta^0)\|_2^2 &\leq \lambda(\|\hat{\beta}\|_1 + \|\beta^0\|_1) + 2\lambda\|\beta^0\|_1 - 2\lambda\|\hat{\beta}\|_1 \\ &\leq 3\lambda\|\beta^0\|_1,\end{aligned}$$

y de manera similar

$$\begin{aligned}\frac{2}{n}\|X(\hat{\beta}-\beta^0)\|_2^2 &\leq \lambda\|\hat{\beta}-\beta^0\|_1 + 2\lambda\|\hat{\beta}-\beta^0\|_1 \\ &\leq 3\lambda\|\hat{\beta}-\beta^0\|_1.\end{aligned}$$

De esta forma hemos demostrado un primer resultado de consistencia para el lasso, que resumimos a continuación.

Teorema 2.1.1 *Si se satisface el modelo lineal normal (2.1) y la normalización (2.4), con probabilidad al menos* $1-\delta$*,*

$$\frac{1}{n}\|X(\hat{\beta}-\beta^0)\|_2^2 \leq 6\sigma\|\beta^0\|_1\sqrt{\frac{2\log(2p/\delta)}{n}} \tag{2.10}$$

y

$$\frac{1}{n}\|X(\hat{\beta}-\beta^0)\|_2^2 \leq 6\sigma\|\hat{\beta}-\beta^0\|_1\sqrt{\frac{2\log(2p/\delta)}{n}}. \tag{2.11}$$

Destacamos que para probar este resultado únicamente hemos necesitado probar que con probabilidad al menos $1-\delta$

$$\frac{2\sigma}{\sqrt{n}} \max_{1\leq j\leq p} \left| \frac{1}{\sqrt{n}\sigma} \epsilon^T X^{(j)} \right| \leq \lambda_0.$$

Como consecuencia del Teorema 2.1.1, en particular como consecuencia de (2.10), que el error de predicción es pequeño, con alta probabilidad, siempre que $\log p$ sea pequeño frente a n. El error es de orden $O\left(\sqrt{\frac{\log p}{n}}\right)$. La cota depende del parámetro desconocido σ. Este parámetro se puede estimar y se puede emplear

esa estimación para dar una versión modificada de (2.10) que no dependa de δ, tal como se hace, por ejemplo, en la página 104 de [3] (Corollary 6.1). Sin embargo, consideramos inútil este esfuerzo porque seguirían interviniendo en la desigualdad parámetros desconocidos como la norma de β^0.

Al mismo tiempo, la desigualdad (2.11) no es tan interesante como (2.10), pues aunque comparte muchas similitudes con ella, el lado derecho de (2.11) no es constante porque interviene la variable aleatoria $\hat{\beta}$. Sin embargo, en la siguiente sección esta cota se probará útil, porque nos permitirá dar estimaciones mejoradas bajo la condición de dispersión estadística, es decir, cuando el vector de coeficientes β^0 tenga un número relativamente pequeño de entradas no nulas.

2.2. Errores no Gaussianos

Durante este capítulo, a excepción de esta sección, se asume que se satisface un modelo lineal normal, es decir, que la etiqueta Y se calcula como una función determinista de los atributos a la que se debe añadir un error Gaussiano ϵ. Esta hipótesis es estándar y se deriva de la idea de que es imposible adquirir toda la información necesaria para calcular de forma exacta la etiqueta. Puede haber errores pequeños con las mediciones de los atributos, pueden influir débilmente variables que no se han tenido en cuenta o incluso es posible que haya cierta aleatoriedad en la determinación de la etiqueta. Gracias al Teorema Central del Límite tiene sentido asumir que todas estas pequeñas diferencias agregadas siguen una distribución normal.

De todos modos, la teoría vista en estos capítulos es extensible a situaciones donde el modelo es únicamente lineal, es decir, cuando los errores son no Gaussianos. La clave para desarrollar esta teoría está en controlar el tamaño del siguiente máximo

$$\max_{1\leq j\leq p}\left|\epsilon^T X^{(j)}\right| = \max_{1\leq j\leq p}\left|\sum_{i=1}^{n}\epsilon_i X_{ij}\right|,$$

y por tanto se presenta con respecto a la sección previa el inconveniente de que el proceso que se estudia ya no es Gaussiano, a diferencia del proceso (2.8). Por ello la Cota de Chernoff ya no es suficiente y se requiere el uso de resultados estadísticos más potentes. En particular se utiliza la Desigualdad de Simetrización (Apéndice A.2) que requiere de la introducción del concepto de variable Rademacher, que son variables r tales que

$$\mathbb{P}\left[r=-1\right] = \mathbb{P}\left[r=1\right] = \frac{1}{2}. \tag{2.12}$$

Suponiendo que las variables ϵ_i son independientes y centradas para todo i, gracias a la Desigualdad de Simetrización se puede asegurar que siendo $r_1, r_2, ..., r_n$ variables i.i.d. Rademacher independientes del vector aleatorio ϵ

$$\mathbb{E}\left[\max_{1\leq j\leq p}\Big|\sum_{i=1}^{n}\epsilon_i X_{ij}\Big|\right] \leq 2\left(\mathbb{E}\left[\max_{1\leq j\leq p}\Big|\sum_{i=1}^{n} r_i\epsilon_i X_{ij}\Big|\right]\right). \tag{2.13}$$

También se usa la Desigualdad de momentos de Hoeffding (Apéndice A.4), que denotando por $\mathbb{E}_\epsilon$ a la esperanza condicionada a los valores de ϵ y utilizando nuevamente que las variables ϵ_i son centradas asegura que

$$\mathbb{E}_\epsilon\left[\max_{1\leq j\leq p}\Big|\sum_{i=1}^{n} r_i\epsilon_i X_{ij}\Big|\right] \leq \big(2\log(2p)\big)^{1/2}\max_{1\leq j\leq p}\Big(\sum_{i=1}^{n}\epsilon_i^2 X_{ij}^2\Big)^{1/2}.$$

Por último se supone que la varianza de los errores ϵ_i está acotada por 1 para todo i y que todas las entradas de la matriz X están acotadas por uno para alcanzar la siguiente cota sobre la esperanza en ϵ de lo anterior

$$\begin{aligned}\mathbb{E}\left[\max_{1\leq j\leq p}\Big|\sum_{i=1}^{n} r_i\epsilon_i X_{ij}\Big|\right] &\leq \big(2\log(2p)\big)^{1/2}\mathbb{E}\left[\max_{1\leq j\leq p}\sum_{i=1}^{n}\epsilon_i^2 X_{ij}^2\right]^{1/2}\\ &\leq \big(2n\log(2p)\big)^{1/2}\end{aligned}$$

Poniendo en conjunto este resultado con (2.13) se obtiene que si $|X_{ij}| \leq 1\ \forall i, j$ y además las variables ϵ_i son independientes, centradas y tienen varianza menor que uno para todo i entonces

$$\mathbb{E}\left[\max_{1\leq j\leq p}\Big|\sum_{i=1}^{n}\epsilon_i X_{ij}\Big|\right] \leq \big(8n\log(2p)\big)^{1/2}. \tag{2.14}$$

Esta cota es un caso particular de la Desigualdad de momentos de Nemirovski que está probada en su forma general en [3] (Lemma 14.24) y que se enuncia a continuación.

Teorema 2.2.1 *(Desigualdad de momentos de Nemirovski) Sean $Z_1, Z_2, ..., Z_n$ variables aleatorias y sean $\gamma_1, \gamma_2, ..., \gamma_p$ funciones del espacio donde viven dichas*

variables en los reales, entonces para todo $m > 1$ *y* $p > e^{m-1}$ *se satisface la desigualdad*

$$\mathbb{E}\left[\max_{1\leq j\leq p}\left|\sum_{i=1}^{n}\left(\gamma_j\left(Z_i\right)-\mathbb{E}\left[\gamma_j\left(Z_i\right)\right]\right)\right|^m\right]\leq\left[8\log(2p)\right]^{m/2}\mathbb{E}\left[\max_{1\leq j\leq p}\sum_{i=1}^{n}\gamma_j^2\left(Z_i\right)\right]^{m/2}.$$

Cabe destacar que aunque los errores no sean normales la cota (2.14) sobre la esperanza se puede mejorar hasta una cota probabilística, en particular, en el problema 6.2 de [3] basado en [6] se afirma que bajo las mismas hipótesis que hemos usado para probar (2.14) se tiene con probabilidad $1-\delta$

$$\frac{2}{n}\max_{1\leq j\leq p}\left|\sum_{i=1}^{n}\epsilon_i X_{ij}\right|\leq 4\sqrt{\frac{2\log(2p)}{\delta n}}.$$

Argumentos similares a los presentados en esta sección son válidos en cualquier problema donde los errores no se supongan Gaussianos, en concreto serían útiles en el tercer capítulo de esta memoria si dejásemos de suponer que los errores son normales.

2.3. Garantías bajo dispersión

Como se comenta al comienzo del capítulo es muy común en la práctica encontrarse en una situación de dispersión estadística, es decir, una situación en la que muchos de los p atributos considerados no guarden relación con la etiqueta.

Evidentemente el modelo subyacente será más simple cuanto menor sea el número de atributos relevantes, y si se conociesen dichos atributos la estimación sería más sencilla y tendría mejores garantías.

Sorprendentemente el lasso es capaz de ir más allá tal y como se ha mencionado posee garantías que mejoran cuanto más simple es el modelo subyacente casi al mismo ritmo que si tuviésemos información adicional. A las desigualdades que ofrecen garantías de este tipo se les conoce como desigualdades oráculo y para llegar a ellos se requiere de la siguiente notación.

Para distinguir los índices relevantes se define el conjunto activo $S_0 := \{j : \beta_j^0 \neq 0\}$ y $s_0 = |S_0$ el número de índices activos. Se introduce también la notación

β_S asociada al conjunto de índices $S \subset \{1, 2, ..., p\}$ que queda definida por

$$(\beta_S)_j := \beta_j \mathbb{I}(j \in S). \tag{2.15}$$

Teniendo esta notación en cuenta y definiendo $\overline{\beta^0}$ como la media de los valores absolutos de las entradas no nulas de β^0 se deduce de manera trivial de (2.10) que con probabilidad al menos $1 - \delta$,

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 \leq 6\hat{\sigma} s_0 \overline{\beta^0} \sqrt{\frac{2\log(2p/\delta)}{n}}. \tag{2.16}$$

Esta desigualdad ofrece garantías mejoran cuando disminuye el número de atributos relevantes. Sin embargo, las garantías no son lo suficientemente buenas para considerarla una desigualdad oráculo, pues como se verá después del Teorema 2.3.2 si conocemos los s_0 atributos relevantes se pueden encontrar estimadores que con probabilidad alta aseguren que el error es de orden $O(s_0/n)$ y esto queda realmente lejos de las cotas de orden $O\left(s_0\sqrt{\log(p)/n}\right)$ que ofrece este resultado.

Se entiende ahora mejor porqué se llama así a este tipo de resultados, que son capaces de ofrecer una cota en función de parámetros desconocidos y por tanto en la práctica solo se podría concretar la cota si un oráculo nos facilitase dichos parámetros, en este caso el número de atributos activos y la media de los valores absolutos de estos atributos.

2.3.1. La condición de compatibilidad

La notación β_S introducida en (2.15) es realmente útil a lo largo de toda la memoria para alcanzar resultados interesantes. Evidentemente esta notación siempre cumple

$$\|\beta\|_1 = \|\beta_S\|_1 + \|\beta_{S^C}\|_1$$

y en particular por la definición de S_0

$$\|\beta^0\|_1 = \|\beta^0_{S_0}\|_1.$$

Además, procediendo de igual manera que en la sección anterior, tomando λ_0 como en (2.7),

$$\lambda_0 := 2\sigma\sqrt{\frac{2\log(2p/\delta)}{n}},$$

con probabilidad al menos $1-\delta$ se cumple (2.9),

$$\frac{1}{n}\|X(\hat{\beta}-\beta^0)\|_2^2+\lambda\|\hat{\beta}\|_1 \leq \lambda_0\|\hat{\beta}-\beta^0\|_1+\lambda\|\beta^0\|_1.$$

De donde se deduce tomando $\lambda \geq 2\lambda_0$

$$\begin{aligned}\frac{2}{n}\|X(\hat{\beta}-\beta^0)\|_2^2 &\leq \lambda\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_1+\lambda\|\hat{\beta}_{S_0^C}\|_1+2\lambda\|\beta^0_{S_0}\|_1-2\lambda\|\hat{\beta}_{S_0}\|_1-2\lambda\|\hat{\beta}_{S_0^C}\|_1\\ &\leq 3\lambda\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_1-\lambda\|\hat{\beta}_{S_0^C}\|_1.\end{aligned} \tag{2.17}$$

Esta cota del error tiene un problema similar al que tenía la segunda parte del Teorema 2.1.1, no es aplicable directamente puesto que contiene la variable aleatoria $\hat{\beta}$ y se desconoce el tamaño de $\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_1$. Se busca por tanto una acotación de $\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_1$ que permita continuar la cadena de desigualdades.

Como se verá en (2.3.1) es particularmente deseable conseguir una cota de $\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_1$ por un múltiplo de $\|X(\hat{\beta}-\beta^0)\|_2^2$.

En esta acotación es útil la Desigualdad de Cauchy-Schwarz

$$\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_1 \leq \sqrt{s_0}\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_2, \tag{2.18}$$

donde $s_0=|S_0|$.

Además, definiendo $\hat{\Sigma}=X^TX/n$, se tiene

$$\frac{1}{n}\|X(\hat{\beta}-\beta^0)\|_2^2=(\hat{\beta}-\beta^0)\hat{\Sigma}(\hat{\beta}-\beta^0). \tag{2.19}$$

siempre que

$$(\hat{\beta}-\beta^0)\hat{\Sigma}(\hat{\beta}-\beta^0)>0 \text{ ó } \hat{\beta}=\beta^0, \tag{2.20}$$

Esto motiva la introducción de la condición de compatibilidad en el conjunto S_0 para $\hat{\beta}-\beta^0$ que se satisface si para algún $\phi_0>0$

$$\|\hat{\beta}_{S_0}-\beta^0_{S_0}\|_2^2 \leq (\hat{\beta}-\beta^0)\hat{\Sigma}(\hat{\beta}-\beta^0)/\phi_0^2. \tag{2.21}$$

Es evidente que esta condición es equivalente a que se satisfaga al menos una de las condiciones de (2.20), dando lugar a la siguiente observación.

Lema 2.3.1 *Si se cumple la condición de compatibilidad en el conjunto S_0 para $\hat{\beta} - \beta^0$, por la Desigualdad de Cauchy-Schwarz (2.18) y por (2.19) se obtiene la acotación deseada*

$$\begin{aligned}\|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1 &\leq \sqrt{s_0}\|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_2 \\ &\leq \sqrt{s_0(\hat{\beta} - \beta^0)\hat{\Sigma}(\hat{\beta} - \beta^0)/\phi_0^2} \\ &= \frac{\sqrt{s_0}}{\phi_0\sqrt{n}}\|X(\hat{\beta} - \beta^0)\|_2\end{aligned}$$

sin embargo, $\hat{\beta}$ es aleatorio y S_0 desconocido por lo que lo ideal sería garantizar que la condición de compatibilidad en todo conjunto $S \subset \{1, 2, ..., p\}$ se cumple para todo β.

En general se dice que un conjunto $S \subset \{1, 2, ..., p\}$ cumple la condición de compatibilidad si existe $\phi(S) > 0$ tal que para todo β

$$\|\beta_S\|_2^2 \leq \beta\hat{\Sigma}\beta|S|/\phi(S)^2. \tag{2.22}$$

Esta condición es una relajación de la condición de que la matriz X tenga rango máximo p, y la constante de compatibilidad $\phi(S)$ es una medida de lo bien acondicionada que está dicha matriz restringida a las columnas del conjunto S, a mayor ϕ_0 mejor acondicionada está.

Por ejemplo, idealizando el problema y suponiendo que $n = p$ y que las columnas de la matriz X son ortogonales se tiene que la matriz $\hat{\Sigma}$ es la identidad y por tanto para cualquier $\phi_0 \leq 1$ se satisface la condición de compatibilidad.

En general, se deberá tomar

$$\phi(S)^2 \leq \min_{v:\|v_S\|_2^2 \neq 0} \frac{\|Xv\|_2^2}{\|v_S\|_2^2}|S|.$$

Se resume por tanto que no se cumplirá la condición de compatibilidad para el conjunto S a no ser que todos los vectores v tales que $\|Xv\|_2^2 = 0$ cumplan $\|v_S\|_1 = 0$, es decir, a no ser que las relaciones que existen entre las columnas de X no involucren a los índices del conjunto S. En particular, el caso extremo en donde dos columnas son idénticas no significa necesariamente que no se cumpla la condición de compatibilidad a no ser que al menos una de las columnas esté asociada a un indice del conjunto activo. Esta condición es por tanto una relajación significativa en estadística de alta dimensión con respecto a la condición de que todos los autovalores de $\hat{\Sigma}$ sean estrictamente positivos, imposible cuando $p > n$.

Se puede relajar todavía más la condición de compatibilidad para el conjunto activo S_0, exigiendo la desigualdad (2.22) solo para los $\beta - \beta^0$ tales que $\|\beta_{S_0{}^C}\|_1 \leq 3\|\beta_{S_0} - \beta^0_{S_0}\|_1$, puesto que gracias a (2.17) se sabe que esta desigualdad se satisface para $\beta = \hat{\beta}$ en Ω_{λ_0}. Teniendo en cuenta esta última relajación se redefine la condición de compatibilidad.

Definición 2.3.1 *El conjunto S se satisface la condición de compatibilidad si existe $\phi(S)$ tal que para todo β que satisfaga*

$$\|\beta_{S^C}\|_1 \leq 3\|\beta_S - \beta^0_S\|_1 \tag{2.23}$$

se cumple

$$\|\beta_S - \beta^0_S\|_2^2 \leq (\beta - \beta^0)\hat{\Sigma}(\beta - \beta^0)|S|/\phi(S)^2. \tag{2.24}$$

Suponiendo que se satisface la condición de compatibilidad para el conjunto S_0 y tomando $\lambda \geq 2\lambda_0$, donde

$$\lambda_0 := 2\sigma\sqrt{\frac{2\log(2p/\delta)}{n}},$$

del Lema 2.3.1 y de (2.17) se tiene

$$\begin{aligned}\frac{2}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\hat{\beta} - \beta^0\|_1 &= \frac{2}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\hat{\beta}_{S_0^C}\|_1 + \lambda\|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1 \\ &\leq 4\lambda\|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1 \\ &\leq 4\lambda\frac{\sqrt{s_0}}{\phi_0\sqrt{n}}\|X(\hat{\beta} - \beta^0)\|_2.\end{aligned}$$

Además utilizando la desigualdad $4uv \leq 2u^2 + 2v^2$

$$\frac{2}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\hat{\beta} - \beta^0\|_1 \leq \frac{2}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 + 2\lambda^2\frac{s_0}{\phi_0^2},$$

de donde se deduce el siguiente resultado, en el que se usa la segunda parte del Teorema 2.1.1, (2.11), para obtener (2.26).

Teorema 2.3.2 *(Resultado oráculo) Si se satisface el modelo lineal normal (2.1), la normalización (2.4) y además se satisface la condición de compatibilidad para el conjunto activo S_0 con constante de compatibilidad $\phi_0 := \phi(S_0)$, entonces con probabilidad al menos $1 - \delta$ se tiene*

$$\|\hat{\beta} - \beta^0\|_1 \leq 8\sigma\frac{s_0}{\phi_0^2}\sqrt{\frac{2\log(2p/\delta)}{n}} \tag{2.25}$$

y

$$\frac{1}{n}\|X(\hat{\beta}-\beta^0)\|_2^2 \leq 96\sigma^2\frac{\log(2p/\delta)}{\phi_0^2}\frac{s_0}{n}. \tag{2.26}$$

Al igual que en el Teorema 2.1.1 la cota depende del parámetro desconocido σ y de nuevo podemos utilizar una estimación del parámetro para alcanzar una cota que no dependa de σ, sin embargo, consideramos nuevamente inútil este esfuerzo porque en la cota intervienen otros parámetros desconocidos.

Se observa que se ha mejorado bastante con la introducción de la condición de compatibilidad, pues gracias a este resultado se puede asegurar que el error de estimación es a lo sumo una $O\Big(s_0\log(p)/n\Big)$ frente a la $O\left(s_0\sqrt{\log(p)/n}\right)$ que se garantizaba en (2.16) que se cumple con la misma probabilidad aunque sin necesidad de asumir la condición de compatibilidad.

Otro aspecto destacable de este resultado es que ofrece una cota consistente del error de estimación(desigualdad (2.25)) . Y aunque esta desigualdad ha quedado eclipsada por su acompañante es un resultado muy fuerte y puede ser realmente útil en la práctica en un escenario en el que se busca entender de que forma cada uno de los atributos influye sobre la variable respuesta.

Se estudia ahora lo lejos que queda esta cota de las que se tendrían si se conociesen de antemano los s_0 atributos relevantes y se pudiese utilizar el algoritmo de mínimos cuadrados sobre ellos con garantías, es decir, si s_0 es menor que el tamaño de la muestra n.

Como ya se menciona en la introducción es conocido que la solución a los mínimos cuadrados vendría dada en este caso por

$$\ddot{\beta} = (X_0^T X_0)^{-1} X_0^T Y,$$

donde X_0 es la matriz X restringida a los s_0 atributos relevantes.

Además puesto que se está suponiendo que el modelo subyacente es lineal satisfaciendo

$$Y = X\beta^0 + \epsilon$$

se tiene

$$\ddot{\beta} = \beta^0 + (X_0^T X_0)^{-1} X_0^T \epsilon.$$

Esto permite calcular la esperanza del error de estimación

$$\mathbb{E}\left[\frac{1}{n}\|X_0\ddot{\beta} - X_0\beta^0\|_2^2\right] = \frac{1}{n}\mathbb{E}\left[\|X_0(X_0^TX_0)^{-1}X_0^T\epsilon\|_2^2\right].$$

Llamando H a la matriz de proyección $X_0(X_0^TX_0)^{-1}X_0$ y utilizando que las matrices de proyección son autoadjuntas e idempotentes

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\|X_0\ddot{\beta} - X_0\beta^0\|_2^2\right] &= \frac{1}{n}\mathbb{E}\left[\epsilon^TH^TH\epsilon\right] \\ &= \frac{1}{n}\mathbb{E}\left[\epsilon^TH\epsilon\right].\end{aligned}$$

Ahora usando que $\epsilon^TH\epsilon$ es escalar, la linealidad de la traza y del producto de matrices y que siendo A y B matrices $Tr(AB) = Tr(BA)$ si ambas multiplicaciones son posibles

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\|X_0\ddot{\beta} - X_0\beta^0\|_2^2\right] &= \frac{1}{n}\mathbb{E}\left[Tr(\epsilon^TH\epsilon)\right] \\ &= \frac{1}{n}\mathbb{E}\left[Tr(H\epsilon\epsilon^T)\right] \\ &= \frac{1}{n}Tr\left(\mathbb{E}\left[H\epsilon\epsilon^T\right]\right) \\ &= \frac{1}{n}Tr\left(H\mathbb{E}\left[\epsilon\epsilon^T\right]\right)\end{aligned}$$

y por la hipótesis sobre la varianza del error

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\|X_0\ddot{\beta} - X_0\beta^0\|_2^2\right] &= \frac{1}{n}Tr\left(H\sigma^2\right) \\ &= \frac{1}{n}Tr\left(H\right)\sigma^2 \\ &= \frac{1}{n}Tr\left(X_0(X_0^TX_0)^{-1}X_0^T\right)\sigma^2 \\ &= \frac{1}{n}Tr\left((X_0^TX_0)^{-1}X_0^TX_0\right)\sigma^2 \\ &= \frac{1}{n}Tr\left(I_{s_0}\right)\sigma^2 \\ &= s_0\sigma^2/n.\end{aligned}$$

Finalmente la Desigualdad de Markov asegura que

$$\mathbb{P}\left[\frac{1}{n}\|X_0\ddot{\beta} - X_0\beta^0\|_2^2 \leq \frac{s_0\sigma}{\delta n}\right] \geq 1 - \delta.$$

Por tanto con δ fijo con probabilidad $1-\delta$ la regla obtenida mediante mínimos cuadrados en un problema con s_0 atributos y tamaño muestral n tiene un error de estimación que es a lo sumo una $O(s_0/n)$, mientras que como se ha visto en el Teorema 2.3.2, con alta probabilidad en un problema con p atributos y tamaño muestral n donde el modelo subyacente tiene únicamente s_0 atributos activos la regla obtenida mediante el lasso tiene un error de estimación que es a lo sumo $O(s_0 \log(p)/n)$.

Es decir, la desigualdad oráculo (Teorema 2.3.2) permite alcanzar garantías similares para el lasso a las que se obtendrían mediante mínimos cuadrados conociendo los s_0 atributos activos de antemano, pagando el precio del logaritmo del número de atributos totales.

Capítulo 3

Garantías estadísticas para el lasso con modelo mal especificado

Se generalizará en esta sección la teoría vista en la sección previa para cuando dejamos de suponer que la etiqueta ha sido obtenida a partir de una combinación lineal de los atributos a la que se ha añadido cierto ruido.

Ahora la etiqueta puede haber sido generada de cualquier forma, en este escenario la regla de Bayes es

$$f^0(X) := \mathbb{E}[Y|X].$$

Se supone a lo largo de este capítulo que para un vector de errores ϵ i.i.d. normales estándar

$$Y = f^0(X) + \epsilon \tag{3.1}$$

En este capítulo trataremos de encontrar una desigualdad oráculo similar a la que ofrecía el Teorema 2.3.2, sin embargo, ahora nos encontramos que el modelo está mal especificado y que por tanto debemos definir un modelo reducido que haga una buena predicción. Una vez que se posee el modelo reducido llamaremos resultado oráculo a cualquier resultado que pruebe que el estimador lasso ofrece garantías similares a las de dicho modelo. En particular, al final de este capítulo escogeremos como modelo reducido el asociado a

$$\beta^* := \underset{\beta}{\operatorname{argmin}}\{\varepsilon(f_\beta) + \lambda\|\beta\|_1\}$$

y trataremos de probar que con probabilidad alta las garantías para el estimador lasso son casi tan buenas como las que podríamos obtener conociendo el 'verdadero

valor' de β^*.

3.1. Garantías incondicionales

En esta sección se busca un resultado que garantice la consistencia para el estimador lasso bajo el modelo (3.1). Para ello se supondrá únicamente que los datos están normalizados de acuerdo con (2.4) y por tanto se tomarán las X deterministas.

Procedemos a continuación igual que se hizo en el capítulo previo para el modelo (2.1), con la diferencia de que ahora se debe arrastrar el término $\frac{1}{n}\|X\beta^* - f^0(X)\|_2^2$ que se anulaba en $\beta^* = \beta^0$ bajo la hipótesis previa, como se observa a continuación.

De la definición de $\hat{\beta}$ se tiene que para todo β^*

$$\frac{1}{n}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\|Y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1$$

sustituyendo Y por $f^0(X) + \epsilon$ y desarrollando los módulos

$$\begin{aligned}&\frac{1}{n}\left(\|\epsilon\|_2^2 + 2\epsilon(f^0(X) - X\hat{\beta}) + \|f^0(X) - X\hat{\beta}\|_2^2\right) + \lambda\|\hat{\beta}\|_1 \\ \leq &\frac{1}{n}\left(\|\epsilon\|_2^2 + 2\epsilon(f^0(X) - X\beta^*) + \|f^0(X) - X\beta^*\|_2^2\right) + \lambda\|\beta^*\|_1\end{aligned}$$

De donde se deduce de forma trivial

$$\frac{1}{n}\|X\hat{\beta} - f^0(X)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{2}{n}\epsilon^T X(\hat{\beta} - \beta^*) + \lambda\|\beta^*\|_1 + \frac{1}{n}\|X\beta^* - f^0(X)\|_2^2, \quad (3.2)$$

Al igual que en el capítulo pasado se desea eliminar los términos aleatorios de la desigualdad, por lo que se busca controlar el tamaño del proceso

$$\max_{1\leq j\leq p} \frac{2}{n}|\epsilon^T X^{(j)}|.$$

Por suerte, a diferencia del proceso que nos interesaba bajo el modelo únicamente lineal (no normal) este proceso vuelve a ser Gaussiano. De hecho es el mismo proceso que ya se estudió después de (2.6), por lo que ya se conoce que si se toma

$$\lambda_0 = 2\sigma\sqrt{\frac{2\log(2p/\delta)}{n}},$$

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

$$\frac{2\sigma}{\sqrt{n}} \max_{1\leq j\leq p} \left| \frac{1}{\sigma\sqrt{n}} \epsilon^T X^{(j)} \right| \leq \lambda_0. \tag{3.3}$$

Combinando esto con (3.2) se obtiene que para todo β^* con probabilidad al menos $1-\delta$

$$\frac{1}{n}\|X\hat{\beta} - f^0(X)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \lambda_0\|\hat{\beta} - \beta^*\|_1 + \lambda\|\beta^*\|_1 + \frac{1}{n}\|X\beta^* - f^0(X)\|_2^2,$$

además tomando $\lambda \geq 4\lambda_0$ y utilizando las propiedades de la notación β_S definida en (2.15) y definiendo $S^* = \{j : \beta_j^* \neq 0\}$

$$\frac{4}{n}\|X\hat{\beta} - f^0\|_2^2 + 4\lambda\|\hat{\beta}\|_1 \leq \lambda\|\hat{\beta}_{S^*} - \beta^*_{S^*}\| + \lambda\|\hat{\beta}_{S^{*c}}\| + 4\lambda\|\beta^*_{S^*}\|_1 + \frac{4}{n}\|X\beta^* - f^0\|_2^2,$$

y por tanto

$$\frac{4}{n}\|X\hat{\beta} - f^0\|_2^2 + 3\lambda\|\hat{\beta}_{S^{*c}}\|_1 \leq 5\lambda\|\hat{\beta}_{S^*} - \beta^*_{S^*}\|_1 + \frac{4}{n}\|X\beta^* - f^0\|_2^2. \tag{3.4}$$

Del hecho de que la desigualdad anterior se cumple para todo β^* con probabilidad $1-\delta$ se deduce el siguiente resultado

Teorema 3.1.1 *Bajo el modelo* (3.1)*, siendo* $\hat{\beta}$ *el estimador obtenido mediante el lasso con* $\lambda = 8\sigma\sqrt{\frac{2\log(2p/\delta)}{n}}$ *se tiene que con probabilidad al menos* $1-\delta$

$$\frac{1}{n}\|X\hat{\beta} - f^0\|_2^2 \leq \min_{\beta} \left[\frac{1}{n}\|X\beta - f^0\|_2^2 + 10\sigma\sqrt{\frac{2\log(2p/\delta)}{n}}\|\hat{\beta}_{S_\beta} - \beta_{S_\beta}\|_1 \right],$$

donde $S_\beta := \{j : \beta_j \neq 0\}$.

Este último resultado prueba la consistencia del lasso con modelo mal especificado. Sin embargo, esta está muy lejos de ser la desigualdad ideal, porque el lado derecho contiene el estimador $\hat{\beta}$, al igual que ocurría en la segunda parte del Teorema 2.1.1. Buscamos por tanto una cota para el tamaño de $\|\hat{\beta}_{S^*} - \beta^*_{S^*}\|_1$, esta cota se conseguirá al igual que en el capítulo anterior mediante la introducción de la condición de compatibilidad.

3.2. Garantías bajo dispersión

Como ya se ha comentado es común que muchos de los atributos conocidos no tengan influencia en el valor de la etiqueta. Sin embargo, a diferencia de lo que ocurría en el segundo capítulo los atributos ya no tienen porqué modificar la etiqueta de forma lineal. Por tanto para trabajar con más facilidad estableceremos un β^* (en Definición 3.2.1) que describirá un buena aproximación lineal de la función f^0. Seguirá siendo común en la práctica que muchos de los atributos no intervengan en dicha aproximación, de nuevo los atributos que intervienen forman el conjunto activo $S_* = \{j : \beta_j^* \neq 0\}$.

Es evidente que si se conociese el conjunto activo sería más fácil lograr una aproximación buena de β^* y por ende de f^0, pero este conjunto es desconocido. Sin embargo, al igual que en el capítulo anterior se presentan desigualdades oráculo capaces de ofrecer garantías para el lasso similares a las que se obtendrían conociendo el conjunto activo S^*.

Para alcanzar esta desigualdad oráculo se distinguen dos casos, el primero, que será el más favorable, cuando β^* cumple

$$\lambda\|\hat{\beta}_{S^*} - \beta^*_{S^*}\|_1 \leq \frac{1}{n}\|X\beta^* - f^0\|_2^2.$$

En este caso de (3.4) se deduce que con probabilidad al menos $1-\delta$ se satisface la desigualdad ideal

$$\frac{4}{n}\|X\hat{\beta} - f^0\|_2^2 + 3\lambda\|\hat{\beta}_{S^*_C}\|_1 \leq \frac{9}{n}\|X\beta^* - f^0\|_2^2, \tag{3.5}$$

de donde se deduce de forma inmediata

$$\frac{4}{n}\|X\hat{\beta} - f^0\|_2^2 + 3\lambda\|\hat{\beta} - \beta^*\|_1 \leq \frac{9}{n}\|X\beta^* - f^0\|_2^2 + 3\lambda\|\hat{\beta}_{S^*} - \beta^*_{S^*}\|_1. \tag{3.6}$$

Como veremos posteriormente esta desigualdad es suficiente para probar el resultado deseado en este primer caso.

El segundo caso, cuando β^* cumple

$$\frac{1}{n}\|X\beta^* - f^0\|_2^2 \leq \lambda\|\hat{\beta}_{S^*} - \beta^*_{S^*}\|_1$$

da lugar a la desigualdad

$$\frac{4}{n}\|X\hat{\beta} - f^0\|_2^2 + 3\lambda\|\hat{\beta}_{S^*_C}\|_1 \leq 9\lambda\|\hat{\beta}_{S^*} - \beta^*_{S^*}\|_1, \tag{3.7}$$

y por tanto a

$$\frac{4}{n}\|X\hat{\beta} - f^0\|_2^2 + 3\lambda\|\hat{\beta} - \beta^*\|_1 \leq 12\lambda\|\hat{\beta}_{S^*} - \beta^*_{S^*}\|_1. \tag{3.8}$$

Conseguir extraer una desigualdad interesante en este caso será más laborioso y requerirá de hipótesis adicionales porque ahora no ha desaparecido el estimador $\hat{\beta}$ del lado derecho. Para tratar con dicho estimador introducimos de nuevo la condición de compatibilidad, pero al igual que en el capítulo previo queremos encontrar una desigualdad que $\hat{\beta}$ satisfaga siempre que se cumpla (3.3), para que esta juegue el papel de la desigualdad (2.23) en la Definición 2.21 y que no sea necesario imponer la condición de compatibilidad para todo β. La desigualdad que buscamos se puede extraer directamente de (3.7), pues

$$\lambda\|\hat{\beta}_{S^*_C}\|_1 \leq 3\lambda\|\hat{\beta}_{S^*} - \beta^*_{S^*}\|_1$$

cumple todos los requisitos.

Por simplicidad a esta desigualdad se le ha hecho coincidir con la del capítulo anterior mediante la elección de un λ que cuadriplique a λ_0. Esta elección se puede modificar para obtener resultados ligeramente distintos pero con el mismo espíritu.

Se recuerda por tanto que la condición de compatibilidad se cumple el conjunto de índices $S \subset \{1, 2, .., p\}$ si existe una constante $\phi(S)$ tal que para todo β satisfaciendo

$$\lambda\|\hat{\beta}_{S^*_C}\|_1 \leq 3\lambda\|\hat{\beta}_{S^*} - \beta^*_{S^*}\|_1$$

se tiene

$$\|\beta_S\|_1^2 \leq \frac{|S|}{\phi^2(S)}\left(\beta^T\hat{\Sigma}\beta\right).$$

Es inmediato ver que la Nota 2.3.1 sigue siendo aplicable pues no requiere de ninguna hipótesis sobre el modelo subyacente.

Y ahora que disponemos de la condición de compatibilidad se puede deducir de (3.8) que si el conjunto S^* la cumple entonces en el segundo caso considerado con probabilidad $1 - \delta$ se tiene

$$\frac{4}{n}\|X\hat{\beta} - f^0\|_2^2 + 3\lambda\|\hat{\beta} - \beta^*\|_1 \leq \frac{12\lambda\sqrt{s_*}\|X(\hat{\beta} - \beta^*)\|_2}{\sqrt{n}\phi_*}$$

y utilizando las desigualdades $12uv \leq 18u^2 + 2v^2$ y $12uv \leq 6u^2 + 6v^2$ se tiene

$$\leq 24\frac{\lambda^2 s_*}{\phi_*^2} + \frac{2}{n}\|X\hat{\beta} - f^0\|_2^2 + \frac{6}{n}\|X\beta^* - f^0\|_2^2,$$

y por tanto

$$\frac{2}{n}\|X\hat{\beta} - f^0\|_2^2 + 3\lambda\|\hat{\beta} - \beta^*\|_1 \leq \frac{6}{n}\|X\beta^* - f^0\|_2^2 + 24\frac{\lambda^2 s_*}{\phi_*^2}. \tag{3.9}$$

Los resultados que hemos visto hasta ahora en este capítulo son válidos para todo β^*, tomamos ahora un β^* que optimice las cotas dadas en un resultado concreto, con la única restricción de que dicho β^* cumpla la condición de compatibilidad, por ello se debe escoger como se desee Ψ, una colección de conjuntos de índices S que cumplen la condición de compatibilidad. Esta colección nos permitirá escoger el β^* que optimiza el resultado que se presenta posteriormente

Definición 3.2.1 *Sea Ψ una colección de conjuntos de índices S que cumplen la condición de compatibilidad*

$$\beta^* = \underset{\beta: S_\beta \in \Psi}{argmin} \left\{ \frac{1}{n}\|X\beta - f^0(X)\|_2^2 + \frac{4\lambda^2 s_\beta}{\phi^2(S_\beta)} \right\},$$

donde $S_\beta := \{j : \beta_j \neq 0\}$ y $s_\beta = |S_\beta|$.

Es claro que β^* es un estimador similar al estimador de selección del mejor subconjunto (1.3) puesto que penaliza por un múltiplo del tamaño del conjunto de índices activos entre la constante de compatibilidad de dicho conjunto $s_\beta/\phi^2(S_\beta)$. Se espera por tanto que β^* tenga un conjunto soporte pequeño con una constante de compatibilidad no demasiado pequeña.

Por todo esto β^* será un buen candidato para jugar el papel que β^0 tenía en la desigualdad oráculo (Teorema 2.3.2) visto en el capítulo pasado. Y al igual que en aquel resultado se busca una cota que, sin conocer la función f^0, mejore cuanto mejor se comporte dicha función, en este caso cuanto mejor se pueda aproximar por una función lineal de un conjunto (pequeño) de atributos que satisfaga la condición de compatibilidad.

Gracias al estudio realizado anteriormente ahora bastará con conseguir una desigualdad que englobe tanto a (3.6) como a (3.9) para encontrar un resultado que se cumpla en los dos casos que se han distinguido. Además, como cada uno de los dos términos que aparecen en estas desigualdades tienen valor únicamente por separado, para poder alcanzar desigualdades más finas sobre cada uno de estos términos se presentan dos desigualdades,

$$\frac{1}{n}\|X\hat{\beta} - f^0\|_2^2 \leq 12\frac{\lambda^2 s_*}{\phi_*^2} + \frac{3}{n}\|X\beta^* - f^0\|_2^2$$

y

$$\|\hat{\beta} - \beta^*\|_1 \leq \frac{2}{n}\frac{\|X\beta^* - f^0\|_2^2}{\delta} + 8\frac{\lambda_*^s}{\phi_*^2},$$

que claramente se satisface en ambos casos y de las que se deduce el siguiente resultado.

Teorema 3.2.1 *(Resultado oráculo) Sea Ψ una colección de conjuntos de índices que cumplen la condición de compatibilidad, si se toma $\lambda \geq 4\lambda_0$ donde*

$$\lambda_0 = 2\sigma\sqrt{\frac{2\log(2p/\delta)}{n}},$$

entonces con probabilidad al menos $1-\delta$

$$\|\hat{\beta} - \beta^*\|_1 \leq \min_{\beta: S_\beta \in \Psi}\left(\frac{1}{4\sigma\sqrt{2n\log(2p/\delta)}}\|X\beta - f^0\|_2^2 + 64\sigma\sqrt{\frac{2\log(2p/\delta)}{n}}\frac{s_\beta}{\phi_{S_\beta}^2}\right),$$

y

$$\frac{1}{n}\|X\hat{\beta} - f^0\|_2^2 \leq \min_{\beta: S_\beta \in \Psi}\left(\frac{3}{n}\|X\beta - f^0\|_2^2 + 2^9 3\sigma^2\frac{\log(2p/\delta)}{n}\frac{s_\beta}{\phi_{S_\beta}^2}\right), \tag{3.10}$$

donde β^* está definida como en la Definición 3.2.1 y donde $s_\beta = |S_\beta|$.

Puede sorprender el tamaño de las constantes en algunos de los resultados de la memoria, como por ejemplo en la desigualdad (3.10) del Teorema previo. Cabe destacar por tanto que durante este trabajo se otorga mayor relevancia a los ratios de convergencia, que nos permiten comprobar que el estimador lasso continua aprendiendo de forma adecuada incluso en situaciones adversas, como por ejemplo cuando el modelo está mal especificado como en el resultado previo. De hecho en (3.10) se puede considerar más preocupante el término con la constante más pequeña puesto que por muy grande que sea el tamaño de la muestra seguirá teniendo un tamaño aproximadamente constante.

Destacamos aquí que si el modelo lineal fuese correcto, es decir $f^0(X) = X\beta^0$ y además $S_{\beta^0} \in \Psi$ entonces la desigualdad (3.10) tomaría la forma

$$\frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 \leq \min_{\beta: S_\beta \in \Psi}\left(+2^9 3\sigma^2\frac{\log(2p/\delta)}{n}\frac{s_\beta}{\phi_{S_\beta}^2}\right),$$

nos damos cuenta de que aunque se ha trabajado en un marco mucho más general para alcanzar el Teorema 3.2.1, cuando hemos supuesto que el modelo lineal es correcto, se ha recuperado la tasa de convergencia correcta $(\log(p)/n)$, aunque se ha pagado cierto precio en forma del tamaño de la constante.

Capítulo 4

Funciones de pérdida convexas

Hasta ahora siempre se ha trabajado con la función de pérdida cuadrática, pero como se menciona en la introducción según el problema se pueden escoger distintas funciones de pérdida. Durante esta sección se buscan las garantías que tiene el lasso cuando se utiliza una función de pérdida convexa genérica l, en particular nos interesan las funciones de pérdida hinge y logística que permitirán estudiar las garantías que tiene el lasso en el problema de clasificación binaria donde la etiqueta $Y \in \{-1, 1\}$.

Las funciones de pérdida suelen coger como argumentos a f, X e Y, pero durante este capítulo las funciones de pérdida tendrán como argumentos a $f(X)$ y a Y para facilitar la descripción de propiedades sobre la función de pérdida. De esta forma la función de pérdida hinge, representada en la figura 4.1 en rojo, está dada por

$$l_{hinge}(f(X), Y) = \text{máx}\{0, 1 - f(X) * Y\}$$

es particularmente útil para la clasificación binaria mediante máquinas de soporte vectorial duras (hard SVM en la literatura inglesa) puesto que no solo es una función de pérdida convexa que acota la función de pérdida $0-1$ si no que además casa muy bien con la filosofía de estos algoritmos, pues vale 0 para todas las predicciones correctas que superan el margen y para el resto de predicciones la función de pérdida toma el valor de la distancia del punto al margen correcto, por lo que los errores se pueden calcular con productos escalares, permitiendo así transformaciones a dimensiones más grandes en las que se trabajará mediante el núcleo de la transformación.

La otra función relevante es la función de pérdida logística, esta función viene

dada por

$$l_{log}(f(X), Y) = -Yf(X) + \log(1 + exp(f(X)))$$

si la etiqueta $Y \in \{0, 1\}$ y equivalentemente viene dada por

$$l_{log}(f(X), Y) = log_2(1 + exp(-f(X) * Y))$$

si la etiqueta $Y \in \{-1, 1\}$. Esta última función se representa en la figura 4.1 en verde.

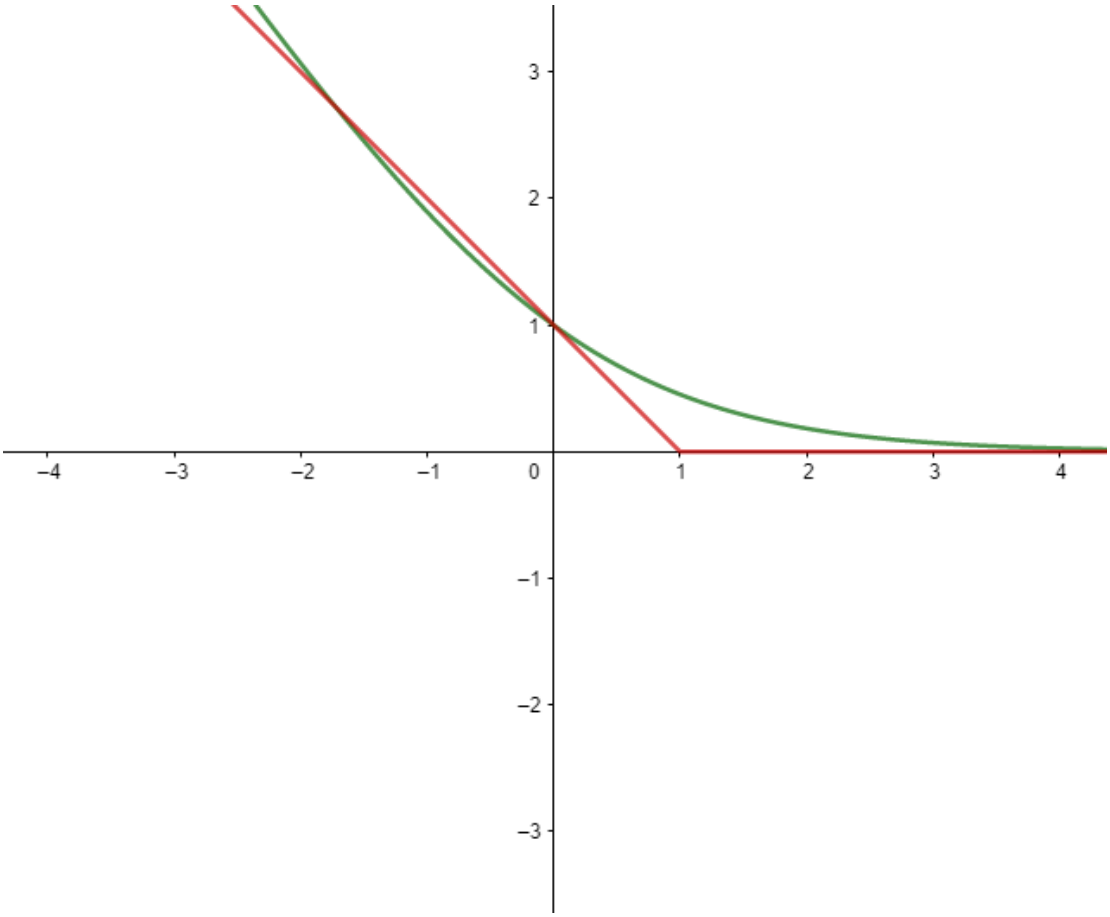

Figura 4.1: Funciones de pérdida hinge (rojo) y logística (verde).

Ambas funciones son útiles para el problema de clasificación binaria porque son acotaciones convexas de la función de pérdida 0-1. Aunque la función de pérdida hinge es una cota más fina la función de pérdida logística es más suave.

Ahora se describe la teoría que nos permitirá dar garantías sobre el estimador

lasso tanto para la función de pérdida hinge, la logística o cualquier otra función de pérdida convexa.

Empezamos por describir el riesgo para una función de pérdida l de una regla f, que viene dado por

$$R(f) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} l(f(X), Y)\right]$$

La regla óptima es

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

donde $\mathbf{F}$ es el espacio de parámetros. El exceso de riesgo de una regla f viene dado por

$$\varepsilon(f) = R(f) - R\left(f^0\right)$$

En la práctica es imposible calcular el riesgo de una regla f, porque depende de la distribución (desconocida) de las observaciones. En su lugar se trabaja con el riesgo empírico

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} l\left(f(X_i)Y_i\right)$$

Sobre la función de pérdida asumiremos convexidad en f, es decir, que para cada Y fijo la función

$$z \to l(z, Y)$$

es convexa. Esto tiene sentido si asumimos que $\mathbf{F}$ es un espacio normado de funciones (con norma que denotaremos $\|\cdot\|$). Este espacio puede ser muy grande, pero la regla f se busca en un subespacio $\mathcal{F} = \{f_\beta : \beta \in \mathbb{R}^p\} \subset \mathbf{F}$ y supondremos, además, que la aplicación $\beta \mapsto f_\beta$ es lineal (típicamente $f_\beta(X^{(i)}) = X^{(i)}\beta$ con $X^{(i)} \in \mathbb{R}^p$. La mejor aproximación f^0 en $\mathcal{F}$ es

$$f^0_{\mathrm{GLM}} = f_{\beta^0_{\mathrm{GLM}}}, \quad \text{con} \quad \beta^0_{\mathrm{GLM}} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} R\left(f_\beta\right)$$

En este marco el estimador lasso es

$$\hat{\beta} = \operatorname{argmin}\left(R_n\left(f_\beta\right) + \lambda\|\beta\|_1\right)$$

En particular, definiendo

$$\pi(x) = \mathbb{P}\left[Y = 1 | X = x\right] \tag{4.1}$$

y denotando $a = f(x)$, para la función de pérdida logística con etiquetas $Y \in \{0, 1\}$ el riesgo de la regla f para un x fijo, viene dado por

$$\mathbb{E}\left[\frac{1}{n}(-Yf(X) + \log(1 + exp(f(X))))|X = x\right] = \frac{1}{n}(-\pi(x)a + \log(1 + exp(a))). \tag{4.2}$$

Y derivando respecto de a e igualando a cero se puede calcular fácilmente que valor de a que minimiza (4.2) para cada x, de donde se deduce la regla de Bayes que en este caso viene dada por

$$f^0_{log}(x) = \frac{\pi(x)}{1 - \pi(x)}.$$

Por supuesto esta regla no se conoce de forma explícita por ser desconocida la función $\pi(x)$, pero posteriormente se verá que con ciertas suposiciones sobre $\pi(x)$ y sobre el modelo se pueden alcanzar resultados interesantes para el estimador lasso con función de pérdida logística en el problema de clasificación binaria, pero por el momento se vuelve al estudio de funciones convexas genéricas.

4.1. Garantías generales

Para alcanzar cualquier garantía en este escenario se debe estudiar, al igual que en capítulos anteriores, un proceso empírico. Sabemos que la aleatoriedad en este problema está introducida la discordancia entre el riesgo y el riesgo empírico que se denota para un β concreto por

$$v_n(\beta) := R_n(\beta) - R(\beta).$$

En particular, como veremos en (4.6) nos interesa el superior en β y β^* del proceso empírico

$$v_n(\beta) - v_n(\beta^*) = v_n(\beta - \beta^*),$$

pero este proceso es imposible de controlar de forma global (para todo β y β^*), principalmente porque tanto β como β^* varían en un espacio finito dimensional. Sin embargo, fijando β^* (4.4) y restringiendo β a una bola l_1 de radio M centrada en β^* denotaremos el superior del proceso empírico por

$$Z_M := \sup_{\|\beta - \beta^*\|_1 \leq M} |v_n(\beta) - v_n(\beta^*)|, \tag{4.3}$$

y podremos aprovechar la convexidad de las funciones de pérdida junto con alguna hipótesis adicional para trasladar el problema al problema de controlar el tamaño

del superior en un conjunto finito del proceso empírico, que como veremos seremos capaces de resolver bajo ciertas hipótesis. Tomaremos

$$\beta^* := \underset{\beta}{\operatorname{argmin}}\{\varepsilon(f_\beta) + \lambda\|\beta\|_1\} \tag{4.4}$$

con esperanzas de que este vector sea próximo a $\hat{\beta}$ por ser hacia donde tiende el estimador cuando el tamaño de la muestra tiende hacia infinito, además este vector tiene evidentemente muy buenas propiedades estadísticas por lo que es deseable comparar a nuestro estimador con él.

Procedemos ahora a plasmar el procedimiento descrito en el párrafo anterior con más detalle.

Es claro que

$$\mathbb{E}\left[Z_M\right] = \mathbb{E}\left[\sup_{\|\beta-\beta^*\|_1\leq M} |v_n(\beta) - v_n(\beta^*)|\right]$$

y que denotando la iésima fila de X mediante $X_{(i)}$, v_n se puede reescribir como

$$v_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(l(f_\beta(X_{(i)}), Y_i) - \mathbb{E}\left[l(f_\beta(X_{(i)}), Y_i)\right]\right).$$

Se tiene por tanto que la esperanza de Z_M es la esperanza del superior cuando $\|\beta - \beta^*\|_1 \leq M$ de

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\left(l(f_\beta(X_{(i)}), Y_i) - l(f_{\beta^*}(X_{(i)}), Y_i)\right) - \mathbb{E}\left[l(f_\beta(X_{(i)}), Y_i) - l(f_{\beta^*}(X_{(i)}), Y_i)\right]\right)\right|.$$

En este contexto la Cota de Chernoff que se ha utilizado en otros capítulos no es suficiente y tampoco lo es la versión que se utilizó de la Desigualdad de Simetrización para tratar errores no Gaussianos en la sección 2.2, si no que se debe utilizar la versión más general de este Teorema en la que hay uniformidad en las funciones $l(f_*(X_i), Y_i)$. Utilizando la Desigualdad de Simetrización (Apéndice A.2) donde $r_1, r_2, ..., r_n$ son variables Rademacher (2.12) independientes entre ellas y respecto de la muestra se deduce de la ecuación anterior

$$\mathbb{E}\left[Z_M\right] \leq 2\mathbb{E}\left[\sup_{\|\beta-\beta^*\|_1\leq M}\left|\frac{1}{n}\sum_{i=1}^{n} r_i\left(l(f_\beta(X_{(i)}), Y_i) - l(f_{\beta^*}(X_{(i)}), Y_i)\right)\right|\right].$$

Suponiendo ahora que las funciones de pérdida son Lipschitz con respecto a la primera componente con constante L, la Desigualdad de Contracción (Apéndice A.3) asegura que

$$\mathbb{E}\left[Z_M\right] \leq 4L\mathbb{E}\left[\sup_{\|\beta-\beta^*\|_1\leq M}\left|\frac{1}{n}\sum_{i=1}^{n} r_i\left(f_\beta(X_{(i)}) - f_{\beta^*}(X_{(i)})\right)\right|\right].$$

Y puesto que $\|\beta - \beta^*\|_1 \leq M$ para todo i

$$|f_\beta(X_{(i)}) - f_{\beta^*}(X_{(i)})| \leq M \max_{1\leq j\leq p} X_{ij},$$

se tiene

$$\mathbb{E}\left[Z_M\right] \leq 4ML\mathbb{E}\left[\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n} r_i X_{ij}\right|\right].$$

Además por la Desigualdad de momentos de Hoeffding (Apéndice A.4)

$$\mathbb{E}\left[Z_M\right] \leq 4ML\sqrt{\frac{2\log(2p)}{n}}\mathbb{E}\left[\max_{1\leq j\leq p}\|X^{(i)}\|_2^2\right],$$

y finalmente utilizando la Desigualdad de Markov se obtiene el siguiente resultado:

Lema 4.1.1 *Sea Z_M como en (4.3), si la función de pérdida $l(f(X), Y)$ es convexa con respecto de la primera componente y Lipschitz con respecto de la primera componente con constante L, entonces con probabilidad al menos $1-\delta$*

$$Z_M \leq \frac{4ML}{\delta}\sqrt{\frac{2\log(2p)}{n}}\mathbb{E}\left[\max_{1\leq j\leq p}\|X^{(i)}\|_2^2\right]. \tag{4.5}$$

Se podrían obtener mejores cotas probabilísticas utilizando la siguiente desigualdad de concentración debida a Bousquet, en vez de la Desigualdad de Markov que es claramente subóptima.

Teorema 4.1.2 *(Desigualdad de Bousquet) Si*

$$\mathbb{E}\gamma\left(Z_i\right) = 0 \ \forall\gamma \in \Gamma, \ \forall i,$$

y

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{\gamma\in\Gamma}\mathbb{E}\gamma^2\left(Z_i\right) \leq 1$$

y además para alguna constante positiva K,

$$\|\gamma\|_\infty \leq K, \ \forall\gamma \in \Gamma.$$

Entonces para todo $t > 0$,

$$\mathbf{P}\left(Z \geq \mathbb{E}Z + \frac{tK}{3} + \sqrt{2}t\sqrt{1 + 2K\mathbb{E}Z}\right) \leq \exp[-nt].$$

De todos modos la acotación de Z_M que se dá en (4.5) es bastante potente y permite obtener de una forma elemental (pero muy fina) un resultado de consistencia del estimador lasso utilizando funciones de pérdida convexa (Teorema 4.1.3). Para alcanzar este resultado comenzamos destacando que por definición para todo β y en particular para β^*

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

y por tanto que

$$\varepsilon(f_{\hat{\beta}}) + \lambda\|\hat{\beta}\|_1 \leq \varepsilon(f_{\beta^*}) + \lambda\|\beta^*\|_1 - \left[v_n(\hat{\beta}) - v_n(\beta^*)\right], \tag{4.6}$$

por lo que si se consiguiese probar que efectivamente $\|\hat{\beta} - \beta^*\|_1 \leq M$ se obtendría el resultado deseado.

Buscando probar esta proximidad entre $\hat{\beta}$ y β^* se toma para cierto valor $0 \leq t \leq 1$ un nuevo vector de coeficientes $\tilde{\beta} := t\hat{\beta} + (1-t)\beta^*$. Gracias a la convexidad (en algunos casos heredada de la función de pérdida) de todas las funciones que intervienen se deduce de (4.6) y de la evidencia

$$\varepsilon(f_{\beta^*}) + \lambda\|\beta^*\|_1 \leq \varepsilon(f_{\beta^*}) + \lambda\|\beta^*\|_1 - [v_n(\beta^*) - v_n(\beta^*)]$$

que para este nuevo vector de coeficientes se satisface

$$\varepsilon(f_{\tilde{\beta}}) + \lambda\|\tilde{\beta}\|_1 \leq \varepsilon(f_{\beta^*}) + \lambda\|\beta^*\|_1 - \left[v_n(\tilde{\beta}) - v_n(\beta^*)\right].$$

Destacamos que esta cota es mucho mejor que la que se hubiese obtenido de (4.6) sustituyendo β^* por $\tilde{\beta}$ para resaltar la importancia de que la función de pérdida utilizada sea convexa.

Ahora tomando astutamente

$$t := \frac{M^*}{M^* + \|\hat{\beta} - \beta^*\|_1}$$

es claro que

$$\|\tilde{\beta} - \beta^*\|_1 = t\|\hat{\beta} - \beta^*\|_1 < M^*,$$

y que por tanto

$$\varepsilon(f_{\tilde{\beta}}) + \lambda\|\tilde{\beta}\|_1 \leq \varepsilon(f_{\beta^*}) + \lambda\|\beta^*\|_1 + Z_{M^*},$$

de donde se obtiene

$$\lambda\|\tilde{\beta} - \beta^*\|_1 \leq \varepsilon(f_{\beta^*}) + 2\lambda\|\beta^*\|_1 + Z_{M^*},$$

Además gracias al Lema 4.1.1 que nos permitía controlar el tamaño de este proceso empírico sabemos que tomando

$$\lambda_0 = \frac{4L}{\delta}\sqrt{\frac{2\log(2p)}{n}}\mathbb{E}\left[\max_{1\leq j\leq p}\|X^{(i)}\|_2^2\right]$$

con probabilidad al menos $1-\delta$

$$\lambda\|\tilde{\beta}-\beta^*\|_1 \leq \varepsilon(f_{\beta^*}) + 2\lambda\|\beta^*\|_1 + \lambda_0 M^* \leq 2\lambda_0 M^*,$$

de donde si tomamos $\lambda \geq 4\lambda$ se deduce que con la misma probabilidad

$$\|\hat{\beta}-\beta^*\|_1 = \frac{1}{t}\|\tilde{\beta}-\beta^*\|_1 \leq \frac{M^*+\|\hat{\beta}-\beta^*\|_1}{M^*}\frac{M^*}{2} = \frac{M^*+\|\hat{\beta}-\beta^*\|_1}{2} \leq M^*,$$

y esto es justo lo que necesitábamos para terminar de probar el siguiente resultado

Teorema 4.1.3 *Tomando* $\lambda \geq 4\lambda_0$, *donde*

$$\lambda_0 = \frac{4L}{\delta}\sqrt{\frac{2\log(2p)}{n}}\mathbb{E}\left[\max_{1\leq j\leq p}\|X^{(i)}\|_2^2\right].$$

Con probabilidad al menos $1-\delta$

$$\varepsilon(f_{\hat{\beta}}) + \lambda\|\hat{\beta}\|_1 \leq \underset{\beta}{argmin}\big(2\left(\varepsilon(f_\beta) + 2\lambda\|\beta\|_1\right)\big) \tag{4.7}$$

Si se tuviese por ejemplo que las columnas de X tienen norma más pequeña que 1, entonces

$$\lambda_0 = \frac{4L}{\delta}\sqrt{\frac{2\log(2p)}{n}},$$

y (4.7) se convertiría en

$$\varepsilon(f_{\hat{\beta}}) + \lambda\|\hat{\beta}\|_1 \leq \underset{\beta}{\operatorname{argmin}}\big(2\left(\varepsilon(f_\beta) + \frac{32L}{\delta}\sqrt{\frac{2\log(2p)}{n}}\|\beta\|_1\right)\big).$$

Esto es suficiente para ver la consistencia del lasso en este escenario, sin embargo, para mejorar este resultado se deben añadir hipótesis tal y como se discute en la sección venidera.

4.2. La condición sobre el margen

En la sección previa para obtener garantías sobre distintas funciones de pérdida se ha requerido convexidad a dichas funciones, imitando las propiedades de la función de pérdida cuadrática. Una de las mayores ventajas de la función de pérdida cuadrática es que tomando una clase de reglas suficientemente pequeña el exceso de riesgo crece estrictamente de forma cuadrática alrededor del óptimo.

Situaciones similares a esta son muy deseables pues cuanto más 'marcada' sea la función del riesgo entorno al óptimo, menos relevantes serán las perturbaciones que muestra el exceso de riesgo empírico con respecto al real. Para formalizar las situaciones más deseables se presenta la siguiente definición.

Definición 4.2.1 *Dada una clase de reglas $\mathcal{H}$, la condición sobre el margen se satisface para la función estrictamente convexa G si*

$$\varepsilon(f) \geq G(\|f - f^0\|) \tag{4.8}$$

para todo $f \in F_\mu := \{f \in \mathcal{H} : \|f - f^0\| \leq \mu\}$.

Además, diremos que el margen es cuadrático si

$$\varepsilon(f) \geq c\|f - f^0\|_2^2$$

con $c > 0$ para todo $f \in F_\mu := \{f \in \mathcal{H} : \|f - f^0\| \leq \mu\}$.

Para la función de pérdida logística podemos probar que se satisface la condición sobre el margen cuadrático asegurando que para todo x en el mínimo de (4.2), $a = \frac{\pi(x)}{1-\pi(x)}$, la segunda derivada de (4.2) con respecto de a es mayor que una constante $c > 0$, es decir, asegurando que

$$\pi(x)(1 - \pi(x)) > c$$

para todo x.
Por lo que si para alguna constante t se tiene

$$t \leq \pi(x) \leq 1 - t$$

entonces para la función de pérdida logística se satisface la condición sobre el margen cuadrático con constante

$$c = t(1 - t).$$

Cabe destacar que de la condición sobre el margen no se deduce necesariamente la convexidad, sin embargo, es un argumento en general más fuerte que la convexidad a la hora de probar resultados. Destacamos también que gracias a que se pide que (4.8) se satisfaga para una función G convexa se pueden aprovechar algunas propiedades de la función G y de su conjugada convexa. Esta conjugada coincide con la Transformada de Legendre, y viene dada por

$$G^*(v) := \sup_u \{uv - G(u)\}, v \geq 0.$$

Una propiedad particularmente interesante que se deduce directamente de la definición es que

$$G^*(v) + G(u) := uv \ \forall u, v > 0,$$

además se menciona como curiosidad (véase [1]) que

$$(G^*)^* = G.$$

A lo largo de esta sección se estudian las garantías que se pueden obtener para el estimador lasso si se satisface tanto la condición sobre el margen como la condición de compatibilidad. Esta última condición se redefine a continuación de tal forma que relacione la norma L_1 con una norma $\|\cdot\|$ en el espacio de funciones lineales.

Definición 4.2.2 *La condición de compatibilidad se satisface para un conjunto S si para todo $\|\beta_{S^C}\|_1 \leq 3\|\beta_S\|_1$ existe una constante $\phi(S) > 0$ tal que*

$$\|\beta_S\|_1^2 \leq \|f_\beta\|^2 |S|/\phi^2(S).$$

Para mejorar los resultados obtenidos en la sección anterior partiremos de nuevo de (4.6)), que asegura que

$$\varepsilon(f_{\hat{\beta}}) + \lambda\|\hat{\beta}\|_1 \leq \varepsilon(f_{\beta^*}) + \lambda\|\beta^*\|_1 - \left[v_n(\hat{\beta}) - v_n(\beta^*)\right],$$

para todo β^*. Y paralelamente a la sección previa definimos

$$Z_M := \sup_{\|\beta-\beta^*\|_1 \leq M} |v_n(\beta) - v_n(\beta^*)|,$$

donde β^* viene dado por

$$\beta^* = \underset{\beta: S_\beta \in \Psi}{\operatorname{argmin}} \left(3\varepsilon(f_{\beta^*}) + 2G^*\left(\frac{4\lambda\sqrt{s_*}}{\phi_*}\right)\right).$$

Además por simplicidad en las cuentas tomamos la notación

$$err^* := \frac{3\delta\left(f_{\beta^*}\right)}{2} + G^*\left(\frac{4\lambda\sqrt{s_*}}{\phi_*}\right)$$

y

$$M^* := err^*/\lambda_0,$$

que nos permitirá probar el siguiente resultado con mayor facilidad.

Teorema 4.2.1 *(Resultado oráculo)*

Sea Ψ una colección de conjuntos que cumplen la condición de compatibilidad y suponiendo que las columnas de X tienen norma menor que 1.

Si se satisface la condición sobre el margen en F_μ con función estrictamente convexa G y que $f_{\beta^*} \in F_\mu$ y $f_\beta \in F_\mu$ para todo β con $\|\beta - \beta^*\|_1 \leq M^*$. Y si además la función de pérdida $z \to l(z, y)$ es Lipschitz con constante L entonces con probabilidad al menos $1 - \delta$ se tiene

$$\|\hat{\beta} - \beta^*\|_1 \leq \min_{\beta: S_\beta \in \Psi}\left(\frac{3\sigma\sqrt{n}}{16L\sqrt{2\log(2p)}}\varepsilon\left(f_\beta\right) + 4G^*\left(\frac{4L\sqrt{s_\beta}}{\phi_{S_\beta}}\right)\right)$$

y

$$\varepsilon(f_{\hat{\beta}}) \leq \min_{\beta: S_\beta \in \Psi}\left(6\varepsilon\left(f_\beta\right) + 4G^*\left(\frac{2^7 L\sqrt{2\log(2p)s_\beta}}{\sigma\phi_{S_\beta}\sqrt{n}}\right)\right).$$

Demostración. Para empezar destacamos que del Lema 4.1.1 es claro que si

$$\lambda_0 = \frac{4L}{\delta}\sqrt{\frac{2\log(2p)}{n}}$$

entonces con probabilidad al menos $1 - \delta$

$$Z_{M^*} \leq \lambda_0 M^*.$$

Además al igual que se hicimos en la sección previa si tomamos

$$\tilde{\beta} := t\hat{\beta} + (1 - t)\beta^*,$$

con

$$t := \frac{M^*}{M^* + \|\hat{\beta} - \beta^*\|_1}$$

entonces por el mismo argumento de convexidad que se utiliza en dicha prueba se tiene de (4.6) que

$$\varepsilon(\tilde{\beta}) + \lambda\|\tilde{\beta}\|_1 \leq Z_{M^*} + \varepsilon(\beta^*) + \lambda\|\beta^*\|_1 \leq \lambda_0 M^* + \varepsilon(\beta^*) + \lambda\|\beta^*\|_1,$$

y usando la notación β_S definida en (2.15) y sus propiedades obtenemos

$$\varepsilon(\tilde{\beta}) + \lambda\|\tilde{\beta}_{S_*^C}\|_1 \leq err^* + \varepsilon(\beta^*) + \lambda\|\tilde{\beta}_{S_*} - \beta^*\|_1 \leq 2err^* + \lambda\|\tilde{\beta}_{S_*} - \beta^*\|_1. \quad (4.9)$$

Consideramos ahora dos casos distintos: **Caso 1:**

$$\lambda\|\tilde{\beta}_{S_*} - \beta^*\|_1 \leq err^*,$$

por tanto de (4.9)

$$\varepsilon(\tilde{\beta}) + \lambda\|\tilde{\beta}_{S_*^c}\|_1 \leq 3err^*$$

y por el caso actual

$$\varepsilon(\tilde{\beta}) + \lambda\|\tilde{\beta} - \beta^*\|_1 \leq 4err^* \quad (4.10)$$

de donde si se escoge $\lambda \geq 8\lambda_0$ se deduce fácilmente

$$\|\tilde{\beta} - \beta^*\|_1 \leq 4\frac{err^*}{\lambda} = 4\frac{\lambda_0}{\lambda}M^* \leq \frac{M^*}{2}.$$

Caso 2:

$$\lambda\|\tilde{\beta}_{S_*} - \beta^*\|_1 \geq err^*,$$

por tanto de (4.9)

$$\lambda\|\tilde{\beta}_{S_*^C}\|_1 \leq 2err^* + \lambda\|\tilde{\beta}_{S_*} - \beta^*_{S^*}\|_1 \leq 3\lambda\|\tilde{\beta}_{S_*} - \beta^*\|_1.$$

Y de la condición de compatibilidad

$$\|\tilde{\beta}_{S_*} - \beta^*_{S^*}\|_1 \leq \sqrt{s}_*\|f_{\tilde{\beta}} - f_{\beta^*}\|/\phi_*$$

por lo que también de (4.9)

$$\varepsilon(\tilde{\beta}) + \lambda\|\tilde{\beta}_{S_*}\|_1 + \lambda\|\tilde{\beta}_{S_*} - \beta^*_{S^*}\|_1 \leq err^* + \varepsilon(\beta^*) + 2\lambda\sqrt{s}_*\|f_{\tilde{\beta}} - f_{\beta^*}\|/\phi_*.$$

Por la definición de conjugada convexa para todo $u, v > 0$

$$uv \leq G^*(u) + G(v),$$

se tiene

$$2\lambda\sqrt{s}_*\|f_{\tilde{\beta}} - f_{\beta^*}\|/\phi^* \leq \frac{1}{2}\left(4\lambda\sqrt{s}_*/\phi^*\|f_{\tilde{\beta}} - f^0\| + 4\lambda\sqrt{s}_*/\phi^*\|f_{\beta^*} - f^0\|\right)$$
$$\leq \frac{1}{2}\left(G^*\left(4\lambda\sqrt{s}_*/\phi^*\right) + G\left(\|f_{\tilde{\beta}} - f^0\|\right) + G^*\left(4\lambda\sqrt{s}_*/\phi^*\right) + G\left(\|f_{\beta^*} - f^0\|\right)\right),$$

y de la condición sobre el margen

$$2\lambda\sqrt{s}_*\|f_{\tilde{\beta}} - f_{\beta^*}\|/\phi^* \leq G^*\left(\frac{4\lambda\sqrt{s_*}}{\phi_*}\right) + \varepsilon(\tilde{\beta})/2 + \varepsilon(\beta^*)/2.$$

Por lo que poniendo todo en conjunto se deduce

$$\varepsilon(\tilde{\beta})/2 + \lambda\|\tilde{\beta} - \beta^*\|_1 \leq err^* + 3\varepsilon(\beta^*)/2 + G^*\left(\frac{4\lambda\sqrt{s_*}}{\phi_*}\right) = 2err^* = 2\lambda_0 M^*, \tag{4.11}$$

de donde

$$\|\tilde{\beta} - \beta^*\|_1 \leq 2\lambda_0 M^*/\lambda \leq M^*/4 \leq M^*/2.$$

En ambos casos se ha obtenido

$$\|\tilde{\beta} - \beta^*\|_1 \leq M^*/2,$$

lo cual permite concluir

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

Ahora repitiendo los argumentos previos reemplazando $\tilde{\beta}$ por $\hat{\beta}$ se llega bien a (4.10) o bien a (4.11), de donde recordando que

$$\lambda_0 = \frac{4L}{\delta}\sqrt{\frac{2\log(2p)}{n}}$$

y que $\lambda \geq 8\lambda_0$, se deduce el resultado deseado. □

Corolario 4.2.2 *Bajo las mismas condiciones del Teorema 4.2.1 si además se satisface la condición sobre el margen para una función cuadrática $G(u) = cu^2$ se tiene con probabilidad al menos $1 - \delta$*

$$\varepsilon(f_{\hat{\beta}}) \leq \min_{\beta: S_\beta \in \Psi}\left(6\varepsilon\left(f_\beta\right) + \frac{2^{15}L^2\log(2p)s_\beta}{c\sigma^2\phi^2_{S_\beta}n}\right),$$

porque el convexo conjugado de G es $G^*(v) = v^2/(4c)$.

Se observa que la función de pérdida logística $l_{log}(f(X), Y)$ con etiquetas $Y \in \{-1, 1\}$ es Lipschitz con respecto la primera componente con constante

$$L = 1/\log(2),$$

por ser

$$\frac{\partial l_{log}(a, Y)}{\partial a} = \frac{1}{(1 + e^a)\log(2)} \leq \frac{1}{\log(2)}.$$

Puesto que además se ha probado previamente que bajo ciertas condiciones tomando la función de pérdida logística se satisface la condición sobre el margen cuadrático se concreta el siguiente resultado.

Corolario 4.2.3 *Sea Ψ una colección de conjuntos que cumplan la condición de compatibilidad y*

$$\beta^* = \min_{\beta: S_\beta \in \Psi} \left(6\varepsilon\left(f_\beta\right) + \frac{2^{15} L^2 K^2 \log(2p) s_\beta}{t(1-t)\phi^2_{S_\beta} \delta^2 n} \right)$$

Asumiendo que $||f_{\beta^*} - f_{\beta^0}|| \leq \mu$ y que $||f_\beta - f_{\beta^0}|| \leq \mu$ para todo β con $\|\beta - \beta^*\|_1 \leq M^*$, y que para algún t

$$t \leq \pi(x) \leq 1 - t \ \forall x,$$

donde $\pi(x)$ está definida en (4.1).

Si además las columnas de X tienen norma menor que 1, entonces se cumple con probabilidad al menos $1 - \delta$

$$\varepsilon(f_{\hat{\beta}}) \leq \min_{\beta: S_\beta \in \Psi} \left(6\varepsilon\left(f_\beta\right) + \frac{2^{15} l^2 K^2 \log(2p) s_\beta}{t(1-t)\phi^2_{S_\beta} \delta^2 n} \right).$$

Demostración. Ya se ha probado previamente, después de definir la condición sobre el margen que bajo las condiciones del Corolario la función de pérdida logística satisface la condición sobre el margen cuadrático con constante $t(1-t)$.

Además es claro que las condiciones del Teorema 4.2.1 se satisfacen en este Corolario, en particular

$$\lambda \geq 8\lambda_0.$$

Por ello del Corolario 4.2.2 se deduce el resultado. □

Comentamos aquí que al igual que ocurría en el Teorema 3.2.1, que a pesar de haber trabajado en un escenario más general si suponemos que el modelo lineal es correcto se recuperarían las tasas de convergencia que se daban en el Teorema 2.3.2 tanto para el Corolario 4.2.2 como para el Corolario 4.2.3, aunque esto sea a costa de pagar un precio en forma de constante.

Capítulo 5

Conclusiones

Este último capítulo pone en perspectiva los resultados obtenidos a lo largo de esta memoria con los que se obtuvieron en [4]. Aunque durante este trabajo se ha tratado el problema de regresión en el contexto de la estadística de alta dimensión y en [4] se trató el problema de clasificación binaria en estadística clásica buscaremos comparar ambos paradigmas. Para ello comenzamos introduciendo el paradigma de aprendizaje PAC.

Uno de los primeros paradigmas de aprendibilidad en el contexto del *Machine Learning* es el concepto de *aprendizaje probablemente aproximadamente correcto* (aprendizaje PAC) introducido por Vapnik (ver [19]) formalizado posteriormente por Valiant (ver [18]). En este marco consideraremos que una clase de reglas $\mathcal{H}$ es aprendible PAC si existe un algoritmo tal que con una probabilidad alta escoja una regla casi tan buena como la regla óptima dentro de la clase. Esto se formaliza a continuación.

Definición 5.0.1 *(Aprendizaje PAC) Se dice que una clase de reglas $\mathcal{H}$ es aprendible PAC si existe un algoritmo de aprendizaje A y una función $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ tal que para todo $\epsilon, \delta \in (0,1)$, si $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, entonces para cualquier distribución D, con probabilidad al menos $1-\delta$ sobre la elección de la muestra $S \sim D^m$ se tiene:*

$$R_D(A(S)) \leq \min_{h \in \mathcal{H}} R_D(h) + \epsilon, \tag{5.1}$$

donde $R_D(h) = \mathop{\mathbb{E}}_{(x,y)\sim D}[l_{0,1}(h,x,y)]$ *es el riesgo bajo la distribución* D *para la función de pérdida 0-1, que toma el valor* 0 *si* $h(x) = y$ *y toma el valor* 1 *en caso contrario.*

Destacamos que aquí se busca una regla que sea capaz de aprender bajo cualquier distribución D. Se observa además que a diferencia del resto del trabajo donde se buscan garantías para un tamaño de muestra fijo aquí se busca el tamaño de muestra necesario para obtener las garantías deseadas, en particular $m_{\mathcal{H}}(\epsilon, \delta)$ se conoce como *complejidad muestral* y es el mínimo tamaño de muestra necesario para que con probabilidad al menos $1 - \delta$ se cumpla (5.1), por lo que si $m_{\mathcal{H}}(\epsilon, \delta) < \infty \forall \epsilon, \delta > 0$ entonces la clase $\mathcal{H}$ será aprendible PAC.

Como se prueba en [4] (Corolario 1.3.2) si la clase es aprendible PAC entonces el algoritmo de minimización del riesgo empírico (MRE) es válido para conseguir las garantías necesarias en la Definición 5.0.1. Además también se prueba en ese trabajo que es posible caracterizar las clases de reglas PAC mediante el concepto combinatorio de dimensión de Vapnik-Chervonenkis que mide la complejidad de las clases y sobre el que se puede consultar más en [5]. Esto conduce al resultado que enunciamos a continuación, que se conoce como *Teorema Fundamental del Aprendizaje Estadístico*, y que está probado en su forma más general en [14] (Theorem 6.7).

Teorema 5.0.1 *(Teorema Fundamental del Aprendizaje Estadístico) Dada una clase de reglas* $\mathcal{H}$ *son equivalentes:*

1. *El algoritmo MRE permite aprender de forma PAC en* $\mathcal{H}$.
2. $\mathcal{H}$ *es PAC aprendible.*
3. $\mathcal{H}$ *tiene dimensión-VC finita.*

Además existe una versión cuantitativa del Teorema Fundamental del Aprendizaje (en [14] Theorem 6.8) que nos ofrece cotas de carácter similar a las que se han obtenido a lo largo de la memoria.

Teorema 5.0.2 *(Versión Cuantitativa del Teorema Fundamental del Aprendizaje Estadístico) Sea* $\mathcal{H}$ *una clase de funciones con llegada en* $\{0,1\}$ *con* $\dim VC =$

$d < \infty$, entonces utilizando la función de pérdida 0-1 existen constantes C_1, C_2 que aseguran que $\mathcal{H}$ es aprendible PAC con complejidad muestral

$$C_1 \frac{d + log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + log(1/\delta)}{\epsilon}.$$

Se puede probar que las funciones lineales en $\mathbb{R}^p$ combinadas con la función signo construyen una clase $\mathcal{H}_p$ de funciones con llegada en $\{0, 1\}$ y de $dimVC(\mathcal{H}_p) = p$ (ver [10] para $dimVC(\mathcal{H}_p) \leq p$ y la otra desigualdad es fácil de probar). Y por tanto con un tamaño de muestra n para que de forma uniforme en las distribuciones se cumpla la desigualdad (5.1) con probabilidad al menos $1 - \delta$ para el algoritmo MRE y la clase $\mathcal{H}_p$, el ϵ necesario cumple para cierta constante C_1

$$\epsilon \geq C_1 \frac{p + log(1/\delta)}{n}.$$

Vemos por tanto que para exigir uniformidad en las distribuciones y utilizando la función de pérdida 0-1, se debe tomar un ϵ de orden al menos p/n, que es un orden horroroso en estadística de alta dimensión. En cambio, prescindiendo de la uniformidad en las distribuciones y tomando funciones de pérdida que acoten a la función de pérdida 0-1 (como la función de pérdida hinge o logística), únicamente es necesario un ϵ de orden a lo sumo $\log(p)/n$ como hemos visto por ejemplo en el Teorema 4.2.1. Se destaca por tanto lo irreal que es el paradigma de aprendizaje PAC, para el que además se necesita el algoritmo MRE, que no es computacionalmente eficiente y por tanto no puede ser llevado a la práctica.

Todavía se podrían buscar más argumentos en contra del paradigma de aprendizaje PAC y a favor de la teoría desarrollada en esta memoria, pues por ejemplo como se muestra en el siguiente contraejemplo no existe ningún algoritmo determinista capaz de aprender de forma PAC una clase de regresores lineales si cambiamos la función de pérdida 0-1 por la función de pérdida cuadrática.

Ejemplo 5.0.1 *Se toma el algoritmo determinista A y se procede mediante reducción al absurdo, suponiendo que el algoritmo A junto con la función de complejidad $m : (0, 1)^2 \to \mathbb{N}$ hace al problema de regresión lineal en $\mathbb{R}$ aprendible PAC para la función de pérdida cuadrática.*

Se escogen los valores $\epsilon = 1/100$, $\delta = 1/2$ y un $m \geq m(\epsilon, \delta)$ suficientemente grande. Y se buscan dos distribuciones tales que ningún algoritmo determinista pueda simultáneamente cumplir que con probabilidad al menos $1 - \delta$ el exceso de riesgo de la regla devuelta por el algoritmo $A(S)$ sea menor que delta para ambas distribuciones.

Para ello se define $\mu = \frac{\log(100/99)}{2m}$ *y se toma la distribución* D_1 *siempre devuelve el elemento* $z_1 = (\mu, -1)$ *y la distribución* D_2 *que devuelve con probabilidad* $1 - \mu$ *de nuevo el elemento* $z_1 = (\mu, -1)$ *y con probabilidad* μ *el elemento* $z_2 = (1, 0)$.

Es fácil ver que para ambas distribuciones con al menos un $99\,\%$ *de probabilidad toda la muestra estará formada por el elemento* z_2*, esto es trivial para la primera distribución y para comprobarlo para la segunda basta observar que para un* m *suficientemente grande*

$$(1 - \mu)^m \geq e^{-2\mu m} = 0{,}99.$$

Por ello con probabilidad de al menos $99\,\%$ *el algoritmo determinista* A *devolverá un regresor lineal* $\hat{\beta}$.

Si este regresor $\hat{\beta}$ *es menor que* $-1/(2\mu)$ *entonces su riesgo para la distribución* D_2 *será*

$$R_{D_2}(\hat{\beta}) \geq \mu(-1/(2\mu))^2 = 1/(4\mu),$$

mientras que el menor riesgo bajo esta distribución es

$$\min_{\beta} R_{D_2}(\beta) \leq R_{D_2}(0) = (1 - \mu)(-1)^2 = 1 - \mu,$$

por lo que el exceso de riesgo asociado a $\hat{\beta} \leq -1/(2\mu)$ *bajo la distribución* D_2 *es*

$$R_{D_2}(\hat{\beta}) - \min_{\beta} R_{D_2}(\beta) \geq 1/(4\mu) - (1 - \mu) > \epsilon,$$

para m *suficientemente grande.*

Si por el contrario este regresor $\hat{\beta}$ *es mayor que* $-1/(2\mu)$ *entonces su riesgo para la distribución* D_1 *será*

$$R_{D_1}(\hat{\beta}) = (-\mu/(2\mu) - (-1))^2 = 1/4,$$

mientras que claramente existe una regla con riesgo 0*, por lo que en este caso el exceso de* $\hat{\beta} > -1/(2\mu)$ *bajo la distribución* D_1 *es mayor que* $1/4$.

Se deduce por tanto que no existe ningún algoritmo determinista que funcione de la forma deseada bajo las dos distribuciones simultáneamente y que por tanto la convexidad de la función de pérdida no es suficiente para garantizar que el problema de regresión es aprendible PAC.

Quedan claras por tanto las ventajas de la teoría desarrollada en esta memoria frente al paradigma de aprendizaje PAC.

Apéndice A

A.1. Cota de Chernoff para normales estándar

Sea Z una variable aleatoria normal estándar

$$\mathbb{P}\left[Z > t\right] \leq e^{-t^2/2},$$

para todo $t > 0$.

Demostración. Vease en primer lugar que para todo $s > 0$ y para todo i

$$\mathbb{E}\left[e^{sX_i}\right] = e^{s^2/2}.$$

$$\begin{aligned}
\mathbb{E}\left[e^{sX_i}\right] &= \int_{-\infty}^{\infty} e^{st} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \\
&= e^{s^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(t-s)^2/2} dt \\
&= e^{s^2/2}
\end{aligned}$$

Además,

$$\mathbb{P}\left[Z > t\right] = \mathbb{P}\left[e^{sZ} > e^{st}\right] \leq \mathbb{E}\left[e^{sZ}\right] e^{-st} = e^{s^2/2 - st}$$

y como esta cota es válida para todo $s > 0$ se toma el valor $s = t$ que la optimiza y proporciona el resultado deseado. □

A.2. Desigualdad de Simetrización

Sea γ una función real y sean $r_1, r_2, ..., r_n$ variables i.i.d. Rademacher independientes de las variables $Z_1, Z_2, ..., Z_n$, entonces para todo $m \geq 1$

$$\mathbb{E}\left(\sup_{\gamma\in\Gamma}\left|\sum_{i=1}^{n}\left\{\gamma\left(Z_i\right)-E\gamma\left(Z_i\right)\right\}\right|^m\right) \leq 2^m\mathbb{E}\left(\sup_{\gamma\in\Gamma}\left|\sum_{i=1}^{n} r_i\gamma\left(Z_i\right)\right|^m\right).$$

A.3. Desigualdad de Contracción

Sean $z_1, z_2, ..., z_n$ elementos no aleatorios del espacio $\mathcal{Z}$, siendo $\mathcal{H}$ una clase de funciones en $\mathcal{Z}$ que toman valores reales y siendo γ_i funciones Lipschitz con constante L.

Entonces siendo $r_1, r_2, ..., r_n$ variables Rademacher (definidas en 2.12) para cualquier función $f_{\beta^*} : \mathcal{Z} \to \mathbb{R}$

$$\mathbb{E}\left(\sup_{f\in\mathcal{H}}\left|\sum_{i=1}^{n} r_i\left\{\gamma_i\left(f\left(z_i\right)\right)-\gamma_i\left(f^*\left(z_i\right)\right)\right\}\right|\right) \leq 2L\mathbb{E}\left(\sup_{f\in\mathcal{H}}\left|\sum_{i=1}^{n} r_i\left(f\left(z_i\right)-f^*\left(z_i\right)\right)\right|\right)$$

A.4. Desigualdad de momentos de Hoeffding

Sean $Z_1, Z_2, ..., Z_n$ variables aleatorias que toman valores en $\mathcal{Z}$ y $\gamma_1, \gamma_2, ..., \gamma_p$ funciones de dicho espacio en la recta real tales que

$$\mathbb{E}\gamma_j\left(Z_i\right) = 0, \left|\gamma\left(Z_i\right)\right| \leq c_{i,j},$$

para todo $i \leq n, j \leq p$.

Entonces para todo $m \geq 1$ satisfaciendo $p \geq e^{m-1}$, se tiene

$$\mathbb{E}\max_{1\leq j\leq p}\left|\sum_{i=1}^{n}\gamma_j\left(Z_i\right)\right|^m \leq \left[2\log(2p)\right]^{m/2}\max_{1\leq j\leq p}\left[\sum_{i=1}^{n} c_{i,j}^2\right]^{m/2}.$$

Bibliografía

[1] Amaku, M., Coutinho, F. A. B. and Oliveira, L. N., (2020), *Thermodynamic Potentials and Natural Variables*, Revista Brasileira de Ensino de Física, 42, e20190127. Epub December 02.

[2] Bertsekas D., (1995). '*Nonlinear Programming*', Athena Scientific, Belmont, MA.

[3] Bühlmann, P. y van de Geer, S., (2011). '*Statistics for High-Dimensional Data Methods*', Theory and Applications. Springer.

[4] Del Barrio, E. y Del Río Almajano, M. Tereso, (2019), '*Fundamentos y aplicaciones de la teoría de aprendizaje estadístico.*', Universidad de Valladolid, Facultad de Ciencias.

[5] R. M. Dudley,(2014), '*Uniform Central Limit Theorems*', Cambridge University Press.

[6] Lutz Dümbgen, Sara A. van de Geer, Mark Veraar, Jon A. Wellner, 2010, *Nemirovski's Inequalities Revisited*. Am. Math. Mon. 117(2): 138-160

[7] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R., (2004), *Least angle regression (with discussion)*, Annals of Statistics 32 407–451.

[8] Hastie, Trevor and Tibshirani, Robert and Wainwright, Martin, 2015, *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman and Hall/CRC.

[9] Jakubovitz D., Giryes R., Rodrigues M.R.D., (2019) *'Generalization Error in Deep Learning'*, Birkhäuser.

[10] Sham Kakade and Ambuj Tewari, (2008), *VC Dimension and Sauer's Lemma*. CMSC 35900 (Spring 2008) Learning Theory.

[11] Ledoux, M. y Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Verlag, New York

[12] López-Abente G, Núñez O, Pérez-Gómez B, Aragonés N, Pollán M. *La situación del cáncer en España: Informe 2015*. Instituto de Salud Carlos III. Madrid, 2015.

[13] Mallows, C., (1973), *Some Comments on CP*. Technometrics, 15(4), 661-675. doi:10.2307/1267380

[14] S. Shalev-Shwartz y S. Ben-David, (2014), '*Understanding Machine Learning*', Cambridge University Press.

[15] John Shawe-Taylor y Nello Cristianini,(2004), '*Kernel Methods for Pattern Analysis*', Cambridge University Press.

[16] R. Tibshirani, (1996), *'Regression shrinkage and selection via the lasso'*, Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[17] Van Der Vaart, A. y Wellner, J., (1996), Weak Convergence and Empirical Processes. Springer Series in Statistics, Springer-Verlag, New York.

[18] Valiant, L. G., (1984). '*A theory of the learnable*', Communications of the ACM 27, 1134–1142.

[19] Vladimir N. Vapnik,(2004), '*The Nature of Statistical Learning Theory*', Springer-Verlag New York, Inc.