UNIVERSIDAD CENTROCCIDENTAL
"LISANDRO ALVARADO"
Decanato de Ciencias y Tecnología
Licenciatura en Ciencias Matemáticas

"DESIGUALDADES DE CONCENTRACIÓN Y EL PROBLEMA DE CLASIFICACIÓN BINARIO"

Trabajo Especial de Grado presentado por:

Br. Emely K. Escobar A.

Como requisito final

para obtener el título de Licenciada

en Ciencias Matemáticas.

Area de Conocimiento: Probabilidad y Estadística.

Tutor: Msc. Jhonny Escalona.

Barquisimeto - Venezuela
Diciembre, 2015

Universidad centroccidental
"Lisandro Alvarado"
Decanato de Ciencias y Tecnología
Licenciatura en Ciencias Matemáticas

ACTA
TRABAJO ESPECIAL DE GRADO

Los suscritos miembros del Jurado designado por el Jefe del Departamento de Matemáticas del Decanato de Ciencias y Tecnología de la Universidad Centroccidental Lisandro Alvarado", para examinar y dictar el veredicto sobre el Trabajo Especial de Grado titulado:

"DESIGUALDADES DE CONCENTRACIÓN Y EL PROBLEMA DE CLASIFICACIÓN BINARIO"

presentado por la ciudadana Br. EMELY ESCOBAR titular de la Cédula de Identidad No. 19.436.863. Con el propósito de cumplir con el requisito académico final para el otorgamiento del título de Licenciada en Ciencias Matemáticas.

Luego de realizada la Defensa y en los términos que imponen los Lineamientos para el Trabajo Especial de Grado de la Liecenciatura en Ciencias Matemáticas, se procedió a discutirlo con el interesado habiéndose emitido el veredicto que a continuación se expresa

[1] _____

Con una calificación de _____ puntos.

En fe de lo expuesto firmamos la presente Acta en la Ciudad de Barquisimeto a los _____ días del mes de _____ de _____

_____	_____
TUTOR	FIRMA
_____	_____
PRINCIPAL	FIRMA
_____	_____
PRINCIPAL	FIRMA

OBSERVACIONES:

_____

_____

_____

[1] Aprobado ó Reprobado

Dedicado a mis padres, por su amor incondicional.
Mis razones de ser.

Agradecimientos

Primero que nada quiero agradecer a Dios por darme la oportunidad de vivir este momento, por estar a mi lado en el transcurso de este largo trayecto. Gracias a la Virgensita porque siempre intercedió por mi ante Dios. Le doy gracias por poner en mi camino a cada una de las personas que conocí y están presentes hasta este día.

Además quiero agradecerle a mis padres, Alcides y Deisi por siempre confiar en mi y saber que era capaz de llegar hasta aquí, hasta cuando yo lo dude. Por tener siempre una palabra de apoyo y orgullo desde el momento que decidí comenzar con esto. Gracias por todos sus sacrificios y recompensas. Gracias por todo lo que necesite desde el dinero para el desayuno y las copias como un fuerte abrazo en los mejores y peores momentos.

A mis hermanos, Javier y Tiffany porque con sus risas, burlas por estar loca y estudiar matemáticas, junto con sus travesuras han sabido apoyarme. Siempre preguntandome cómo vas, cuándo presentas, cómo saliste, han mostrado su cariño incondicional.

A Joelviz porque ha sido cómplice y parte importante de cada una de las cosas que me ha tocado vivir, porque es testigo de mis alegrias y mis tristezas en todo este tiempo. Siempre con una palabra bonita, haciéndome reír, ayudándome en las cosas que no entendía, ha sabido comprenderme y aceptar mi carácter. Gracias amor, por simplemente ser como eres conmigo.

Quiero agradecer a toda la familia encuentrista, porque en mucho tiempo fueron y siguen siendo instrumentos de Dios, porque viví muchos momentos bonitos y aprendi a creer y confiar en él, que me ha ayudado en todo este caminar. También quiero agradecer a mis amigos del colegio, que ahora puedo decir que son como mis hermanos, gracias a Mila, Pato, Dari, Banga y Gabo porque siempre que estamos full hemos sacado un tiempito para hablar y reírnos, siendo la mejor terapia para seguir adelante.

No puedo dejar de agradecer a cada una de las personas que veía a diario, y no tan diario, que siempre me ayudaron cuando algo estaba muy complicado, que me explicaban. Personas que desde que comenzamos nos tocó amanecer estudiando,

pasando alegrias y rabias por los examenes. Personas que fueron parte de estudio los sábados y domingo en AsoEM. Pero que mejor manera de culminar el semestre o luego de presentar un parcial que siendo companẽros de viajes y salidas. Gracias Andrei, Pegao, Emily, Mariana, Anais, Jonathan, Ruth, Karla, Andres, Iliana, Jessica, Evelyn, Diana, Genesis, Celismar, Dellys, Patricio, Gustavo, Kleyver. Gracias a Todos los miembros activos de Asoem en general, porque el convivir a diario nos ha permitido ser más que amigos.

Por último agradezco al Prof. Jhonny, mi tutor, porque sin él no fuese podido culminar este trabajo, porque ha sido un excelente apoyo, siempre pudo guiarme y orientarme para poder culminarlo, gracias porque las asesorias siempre fueron un rato agradable, dispuesto a aportar su conocimiento en lo que necesitaba.

Gracias a todos las personas que quizás se me está pasando mencionar, el que me conoce sabe lo despistada que soy. Pero de verdad, un millón de gracias porque sé que mi Felicidad, también es de ustedes.

DESIGUALDADES DE CONCENTRACIÓN Y EL PROBLEMA DE CLASIFICACIÓN BINARIO

Br. Emely Escobar

Resumen

Las desigualdades de concentración proporcionan cotas para la probabilidad de la desviación entre funciones de variables aleatorias independientes y su valor esperado. En la última década se han introducido nuevas herramientas, lo que ha hecho posible introducir desigualdades simples y poderosas. Estas desigualdades son el corazón del análisis matemático en varios problemas de la teoría de aprendizaje haciendo posible derivar nuevos y eficientes algoritmos. En este trabajo, se estudió las herramientas básicas relacionadas con éstas desigualdades aplicadas al problema de clasificación binario.

Índice general

Introducción

El reconocimiento de patrones según Seijas [9], es el estudio de cómo las máquinas pueden observar el ambiente o entorno, aprender a distinguir patrones de interes a partir de la experiencia, y tomar decisiones razonables con respecto a las categorías a las que pertenecen dichos patrones. Es importante destacar, que el mejor reconocedor de patrones conocido hasta ahora es el ser humano, aunque no se sabe a ciencia cierta cúal es el proceso mediante el cual los humanos realizamos esta tarea.

Es allí, donde enmarcados en la rama de la inteligencia artificial, se intenta simular el mejor reconocedor de patrones; para ello se considera que dado un patrón, la clasificación puede ser resuelta de dos maneras: una donde en el proceso de aprendizaje se posee información de salida y otra donde no, estos son llamados, aprendizaje supervisado y aprendizaje no supervisado.

Gallardo [2] se refiere al aprendizaje no supervisado, cuando nos encontramos con los problemas en los que no se dispone de ningún tipo de información de salida sobre los datos, pero se desea organizar los datos de alguna manera para mejorar su comprensión. y el aprendizaje supervisado es cuando se refiere a todas aquellas aplicaciones o procesos en los que se dispone de información tanto de los valores de entrada del sistema como de los valores de salida deseados.

Por lo tanto, apoyados en Lugosi [7], en nuestro trabajo se definirá una función $g : \mathbb{R}^d \to \{0, 1\}$, llamado clasificador, para modelar el problema de aprendizaje. Introduciremos un ajuste probabilístico, (X, Y) un par de variables aleatorias con valores en $\mathbb{R}^d \times \{0, 1\}$, donde al observar el valor de X se tiene que averiguar el valor de Y, mediante g. Estamos interesados cuando Y toma valores en el conjunto $\{0, 1\}$, lo que denominaremos problemas de clasificación binario.

Además, en éste modelo se supone que se tiene acceso a una muestra de pares (X_i, Y_i) con $1 \leq i \leq n$, observados con anterioridad. Luego, un clasificador g_n es construido con los datos, que no es más que una sucesión de pares independientes e identicamente distribuidos con la misma distribución de (X, Y). El proceso de construcción de g_n es llamado aprendizaje supervisado.

Notemos que el desempeño de g_n se mide por su probabilidad de error, L_n,

la cual es una variable aleatoria pues depende de los datos. El valor de $\mathbb{E}L_n$ dado por $\mathbb{P}\{g_n(X) \neq Y\}$ es de interes principal pues proporciona información útil, especialmente si la variable aleatoria L_n esta concentrada alrededor de su media con alta probabilidad.

Es por esto, que pretendremos estudiar las desigualdades de concentración más conocidas que nos serán de gran apoyo para el estudio de la desigualdad de Vapnik-Chervonenkis [10], una herramienta importante que nos permitirá obtener una cota superior de la esperanza de la desviación máxima entre $\mu_n(A)$ y $\mu(A)$. Donde $\mu(A)$ representa $\mathbb{P}\{X_1 \in A\}$ y $\mu_n(A)$ la expresión $\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[X_i \in A]}$. De tal manera que se pueda tener éxito en la minimización del error empírico del clasificador g_n.

Capítulo 1

Preliminares

Daremos algunos conceptos necesarios para un mejor entendimiento en el transcurso de la lectura de este trabajo. Comenzaremos apoyándonos en Sheldon Ross, [8] para definir proceso estocástico:

Definición 1.1 *Un proceso estocástico* $\{X(t), t \in T\}$ *es una colección de variables aleatorias. Esto es, para cada* $t \in T$, X_t *es una variable aleatoria. El índice* t *es interpretado, en general, como el tiempo. Nos referimos a* $X(t)$ *como el estado del proceso en el tiempo* t.

En conclusión, un proceso estocástico puede interpretarse como una sucesión de variables aleatorias cuyas características pueden variar a lo largo del tiempo.

Además, Hernandez [5], define martingala como:

Definición 1.2 *Sea* $X = \{X_n : n = 0, 1, \ldots\}$, *y* $Y = \{Y_n : n = 0, 1, \ldots\}$ *dos procesos estocásticos discretos. Decimos que* Y *es una martingala con respecto a la sucesión* X *si para todo* $n \geq 0$, $\mathbb{E}[|Y_n|] < \infty$ *y* $\mathbb{E}[Y_{n+1}|X_n, \ldots, X_0] = Y_n$.

Sin embargo, en este trabajo se utilizará la definición expuesta por Gábor y Lázló en [4], dada de la siguiente manera:

Definición 1.3 *Una sucesión de variables aleatorias* $Z_1, Z_2, \ldots$ *es llamada una martingala si* $\mathbb{E}\{Z_{i+1}|Z_1, \ldots, Z_i\} = Z_i$ *con probabilidad uno para cada* $i > 0$.

Definición 1.4 *Sea* $X_1, X_2, \ldots$ *una sucesión arbitraria de variables aleatorias.* $Z_1, Z_2, \ldots$ *es llamada una martingala con respecto a la sucesión* $X_1, X_2, \ldots$, *si* $\mathbb{E}\{Z_{i+1}|X_1, \ldots, X_i\} = Z_i$ *con probabilidad uno. Para todo* $i > 0$ *y* Z_i, *es una función de* $X_1, \ldots, X_i$.

Obviamente, si $Z_1, Z_2, \ldots$ es una martingala respecto a la sucesión $X_1, X_2, \ldots$ tenemos que $Z_1, Z_2, \ldots$ es una martingala. En efecto,

$$\begin{aligned}\mathbb{E}\{Z_{i+1}|Z_1,\ldots,Z_i\} &= \mathbb{E}\{\mathbb{E}\{Z_{i+1}|X_1,\ldots,X_i\}|Z_1,\ldots,Z_i\}\\ &= \mathbb{E}\{Z_i|Z_1,\ldots,Z_i\}\\ &= Z_i.\end{aligned}$$

Lo anterior, debido a la probabilidad total de la esperanza condicional:

$$\mathbb{E}\{X|Z\} = \mathbb{E}[\mathbb{E}\{X|Y\}|Z].$$

Los ejemplos más importantes de martingalas son las sumas de variables aleatorias independientes con media 0. Sea $U_1, U_2, \ldots$ variables aleatorias independientes con media cero. Entonces las variables aleatorias $S_i = \sum_{j=1}^{i} U_j$ con $i > 0$, forman una martingala. Las martingalas comparten muchas propiedades de sumas de variables independientes.

Definición 1.5 *Una sucesión de variables aleatorias $V_1, V_2, \ldots$ es una sucesión de diferencia de martingala si $\mathbb{E}\{V_{i+1}|V_1,\ldots,V_i\} = 0$ con probabilidad uno.*

Definición 1.6 *Una sucesión de variables aleatorias $V_1, V_2, \ldots$ es una sucesión de diferencia de martingala respecto a la sucesión $X_1, X_2, \ldots$ de variables aleatorias, si para todo $i > 0$, V_i es una función de $X_1, \ldots, X_i$ y*

$$\mathbb{E}\{V_{i+1}|X_1,\ldots,X_i\} = 0.$$

Observamos, que si $V_1, V_2, \ldots$ es una sucesión de diferencia de martingala respecto a la sucesión $X_1, X_2, \ldots$ de variables aleatorias, entonces es una sucesión de diferencias de martingalas.

Además, toda martingala $Z_1, Z_2, \ldots$ conduce naturalmente una sucesión de diferencia de martingala definiendo $V_i = Z_i - Z_{i-1}$, para todo $i > 0$.

Por otra parte, la desigualdad que usaremos frecuentemente en el transcurso de este trabajo y es necesario recordar es la desigualdad de Jensen's, expuesto en [3] como el Corolario siguiente:

Corolario 1.1 *Sea $I \subseteq \mathbb{R}$ un intervalo, y sea $\varphi : I \to \mathbb{R}$ una función convexa y X una variable aleatoria acotada, entonces:*

$$\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X)).$$

Capítulo 2

Modelo de Clasificación y Teoría de Aprendizaje

2.1. Problema de clasificación binario

Lugosi en [7], define el **Reconocimiento de patrones** (ó clasificación de patrones) cuando se trata de aproximar o estimar la clase desconocida de una observación. La **observación** se entiende como la colección de mediciones númericas representada por un vector x de dimensión d y la naturaleza desconocida de la observación es llamada una **clase**, denotada por y, donde toma valores en el conjunto $\{0, 1\}$.

En el reconocimiento de patrones, se define una función. Llamada **Clasificador**, la cual representa una estimación de y dado x. Dada por: $g : \mathbb{R}^d \to \{0, 1\}$. Se habla de un error del clasificador en x si $g(x) \neq y$.

Para modelar el problema de aprendizaje, introduciremos un ajuste probabilístico, sea (X, Y) un par de variables aleatorias con valores en $\mathbb{R}^d \times \{0, 1\}$, donde al observar el valor de X, se tiene que averiguar el valor de Y.

El par (X, Y) de variables aleatorias puede ser descrito de diferentes maneras: Por ejemplo, definamos el par (μ, η) donde μ es la medida de probabilidad para X y η es la regresión de Y en X. De forma más precisa, para un conjunto de medida de Borel $A \subset \mathbb{R}^d$,

$$\mu(A) = \mathbb{P}\{X \in A\},$$

para todo $x \in \mathbb{R}^d$ y como $\eta(x)$ es la regresión de Y en X,

$$\eta(x) = \mathbb{P}\{Y = 1 | X = x\} = \mathbb{E}\{Y | X = x\}.$$

En efecto,

$$\begin{aligned}\mathbb{E}\{Y|X=x\} &= 1.\mathbb{P}\{Y=1|X=x\}+0.\mathbb{P}\{Y=0|X=x\}\\ &= \mathbb{P}\{Y=1|X=x\}\\ &= \eta(x).\end{aligned}$$

Esto es, $\eta(x)$ es la probabilidad condicional que Y es 1 dado $X=x$. La distribución de (X,Y) es determinada por (μ,η). La función η es llamada **Probabilidad Posteriori**.

Por otra parte, toda función $g:\mathbb{R}^d\to\{0,1\}$ define un clasificador. Un error ocurre si $g(X)\neq Y$, y la **probabilidad de error de un clasificador** g viene dado por

$$L(g)=\mathbb{P}\{g(X)\neq Y\}.$$

En el siguiente teorema se muestra que el **Clasificador de Bayes**, dado por

$$g^*(x)=\begin{cases}1 & \text{si} \quad \eta(x)>1/2\\ 0 & \text{otro caso,}\end{cases}$$

minimiza la probabilidad de error de un clasificador.

Teorema 2.1 *Para todo clasificador $g:\mathbb{R}^d\to\{0,1\}$,*

$$\mathbb{P}\{g^*(X)\neq Y\}\leq\mathbb{P}\{g(X)\neq Y\}.$$

Demostración:

$$\begin{aligned}&\mathbb{P}\{g(X)\neq Y|X=x\}\\ &= 1-\mathbb{P}\{g(X)=Y|X=x\}\\ &= 1-(\mathbb{P}\{Y=1,g(X)=1|X=x\}+\mathbb{P}\{Y=0,g(X)=0|X=x\})\\ &= 1-(\mathbb{P}\{g(X)=1|X=x\}\mathbb{P}\{Y=1|X=x\}\\ &+ \mathbb{P}\{g(X)=0|X=x\}\mathbb{P}\{Y=0|X=x\})\\ &= 1-(\mathbb{I}_{\{g(x)=1\}}\mathbb{P}\{Y=1|X=x\}+\mathbb{I}_{\{g(x)=0\}}\mathbb{P}\{Y=0|X=x\})\\ &= 1-(\mathbb{I}_{\{g(x)=1\}}\eta(x)+(1-\eta(x))\mathbb{I}_{\{g(x)=0\}})\end{aligned}$$

Donde $\mathbb{I}_A$ denota la función indicadora del conjunto A.

Además, como

$$\begin{aligned}\mathbb{I}_{\{g^*(x)=0\}} &= \mathbb{P}\{g^*(x)=0|X=x\}\\ &= 1-\mathbb{P}\{g^*(x)=1|X=x\}\\ &= 1-\mathbb{I}_{\{g^*(x)=1\}}\end{aligned}\tag{2.1}$$

y de forma análoga,

$$\mathbb{I}_{\{g(x)=0\}} = 1 - \mathbb{I}_{\{g(x)=1\}}. \tag{2.2}$$

Para todo $x \in \mathbb{R}^d$, de (2.1) y (2.2), tenemos:

$$\begin{aligned}
&\mathbb{P}\{g(X) \neq Y | X = x\} - \mathbb{P}\{g^*(X) \neq Y | X = x\} \\
&= -\eta(x)\mathbb{I}_{\{g(x)=1\}} - (1-\eta(x))\mathbb{I}_{\{g(x)=0\}} \\
&+ \eta(x)\mathbb{I}_{\{g^*(x)=1\}} + (1-\eta(x))\mathbb{I}_{\{g^*(x)=0\}} \\
&= \eta(x)(\mathbb{I}_{\{g^*(x)=1\}} - \mathbb{I}_{\{g(x)=1\}}) + (1-\eta(x))(\mathbb{I}_{\{g^*(x)=0\}} - \mathbb{I}_{\{g(x)=0\}}) \\
&= \eta(x)(\mathbb{I}_{\{g^*(x)=1\}} - \mathbb{I}_{\{g(x)=1\}}) + (1-\eta(x))(\mathbb{I}_{\{g(x)=1\}} - \mathbb{I}_{\{g^*(x)=1\}}) \\
&= \eta(x)(\mathbb{I}_{\{g^*(x)=1\}} - \mathbb{I}_{\{g(x)=1\}}) + (\eta(x)-1)(\mathbb{I}_{\{g^*(x)=1\}} - \mathbb{I}_{\{g(x)=1\}}) \\
&= (2\eta(x)-1)(\mathbb{I}_{\{g^*(x)=1\}} - \mathbb{I}_{\{g(x)=1\}}) \\
&\geq 0
\end{aligned}$$

Por lo tanto, para todo clasificador $g : \mathbb{R}^d \to \{0, 1\}$,

$$\mathbb{P}\{g^*(X) \neq Y\} \leq \mathbb{P}\{g(X) \neq Y\}.$$

■

Definimos $L^* = \mathbb{P}\{g^*(X) \neq Y\}$ como **Probabilidad del error de Bayes**, Error de Bayes o Riesgo Bayesiano.

De la demostración anterior podemos notar que

$$L(g) = 1 - \mathbb{E}\{\mathbb{I}_{\{g(X)=1\}}\eta(X) + \mathbb{I}_{\{g(X)=0\}}(1-\eta(X))\}$$

En particular,

$$\begin{aligned}
L(g^*) &= L^* = 1 - \mathbb{E}\{\mathbb{I}_{\{\eta(x)>1/2\}}\eta(X) + \mathbb{I}_{\{\eta(X)\leq 1/2\}}(1-\eta(X))\} \\
&= 1 - \mathbb{E}\{\text{máx}(\eta(X), 1-\eta(X))\} \\
&= \mathbb{E}\{\text{mín}(\eta(X), 1-\eta(X))\}.
\end{aligned}$$

Notemos que en todos estos cálculos g^* depende de la distribución de (X, Y). Si esta distribución es conocida, g^* puede ser calculada. Más aun, si la distribución de (X, Y) es desconocida, tenemos que g^* tambien lo es.

Po lo tanto, para construir un modelo se requiere de una muestra de pares (X_i, Y_i), con $1 \leq i \leq n$, observados con anterioridad. Supondremos que

$$D_n = (X_1, Y_1), ..., (X_n, Y_n),$$

los datos, son pares de variables aleatorias independientes e identicamente distribuidos con la misma distribución de (X, Y).

Un clasificador es construido con la muestra aleatoria $X_1, Y_1, \ldots, X_n, Y_n$ mencionada anteriormente y denotado por g_n. Estimaremos la clasificación de Y mediante la función de clasificación:

$$g_n(X) = g_n(X, D_n).$$

Definición 2.1 *Es llamado aprendizaje supervisado al proceso de construcción de* g_n.

Notemos que el desempeño o el rendimiento de g_n se mide por su probabilidad condicional de error

$$L_n = L(g_n) = \mathbb{P}\{g_n(X) \neq Y | D_n\}.$$

Observemos que L_n es una variable aleatoria pues depende de la muestra D_n. Sería útil saber el valor de la $\mathbb{E}L_n = \mathbb{P}\{g_n(X) \neq Y\}$, especialmente si la variable aleatoria L_n está concentrada alrededor de su media con alta probabilidad; ya que indicaría la calidad del promedio de la muestra D_n. La esperanza $\mathbb{E}L_n$ esta completamente determinada por la distribución del par (X, Y) y el clasificador g_n.

2.2. Minimización del error empírico

Estimar la probabilidad de error L_n de una función de clasificación es de esencial importancia. Para ello, supongamos que se tiene una clase C de clasificadores $g : \mathbb{R}^d \to \{0, 1\}$ y nuestra tarea es encontrar un clasificador con menor probabilidad de error. Ya que desconocemos la distribución subyacente, utilizaremos los datos (la muestra) para estimar la probabilidad de error de los clasificadores en C.

Estamos tentados a escoger un clasificador de C que minimice la estimación de la probabilidad de error sobre la clase. Lo más natural según Gábor y Lázló en [4], para estimar la probabilidad de error $L(g)$, es contar el número de errores que g comete en D_n, es decir el error contable:

$$\widehat{L}_n(g) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}_{\{g(X_j) \neq Y_j\}}.$$

$\widehat{L}_n(g)$ es llamado el **error empírico** o riesgo empírico del clasificador g.

Un buen método busca escoger un clasificador con una probabilidad de error que sea cercana a la mínima probabilidad de error en la clase. Intuitivamente, si podemos estimar la probabilidad de error de los clasificadores en C uniformemente bien, entonces la función de clasificación que minimiza la probabilidad de error estimada,

probablemente tiene una probabilidad de error que es cercana a la mejor en la clase.

De acuerdo a Lugosi [7], se denotará por g_n^* el clasificador que minimiza la probabilidad de error estimada sobre la clase.

$$\widehat{L}_n(g_n^*) \leq \widehat{L}_n(g), \forall\ g \in C. \tag{2.3}$$

Entonces, para la probabilidad de error

$$L(g_n^*) = \mathbb{P}\{g_n^*(X) \neq Y | D_n\}, \tag{2.4}$$

se puede demostrar:

Lema 2.1

$$\begin{aligned} L(g_n^*) - \inf_{g\in C} L(g) &\leq 2\sup_{g\in C}|\widehat{L}_n(g) - L(g)| \\ |\widehat{L}_n(g_n^*) - L(g_n^*)| &\leq \sup_{g\in C}|\widehat{L}_n(g) - L(g)| \end{aligned}$$

Demostración:

La segunda desigualdad es trivial, pues g_n^* minimiza la expresión $\widehat{L}_n(g)$ y $L(g)$ para todo $g \in C$. Por lo tanto, la diferencia de estas expresiones con g_n^* es menor o igual a la diferencia con cualquier otro clasificador $g \in C$ y en particular a su supremo. Por ende, usando este hecho para demostrar la primera desigualdad, tenemos que:

$$\begin{aligned} L(g_n^*) - \inf_{g\in C} L(g) &= L(g_n^*) - \widehat{L}_n(g_n^*) + \widehat{L}_n(g_n^*) - \inf_{g\in C} L(g) \\ &\leq L(g_n^*) - \widehat{L}_n(g_n^*) + \widehat{L}_n(g) - \inf_{g\in C} L(g), \text{por (2.3)} \\ &\leq L(g_n^*) - \widehat{L}_n(g_n^*) + \sup_{g\in C}\widehat{L}_n(g) + \sup_{g\in C}(-L(g)) \\ &= L(g_n^*) - \widehat{L}_n(g_n^*) + \sup_{g\in C}(\widehat{L}_n(g) - L(g)) \\ &\leq |L(g_n^*) - \widehat{L}_n(g_n^*)| + \sup_{g\in C}|\widehat{L}_n(g) - L(g)| \\ &\leq 2\sup_{g\in C}|\widehat{L}_n(g) - L(g)|. \end{aligned}$$

■

Así, vemos que el lema establece cotas superiores para el error de dicho clasificador.

1. Una cota para la distancia entre la probabilidad de error (2.4) y la menor probabilidad de error real en C.

2. Una cota para la distancia entre el error empírico y el error real del clasificador g_n^*.

Ahora bien, siempre que las cotas indiquen que esta cerca de la optima en C, debemos al mismo tiempo tener una buena estimación de la probabilidad de error, y viceversa.

Además, la variable aleatoria $n\widehat{L}_n(g)$ esta distribuida binomialmente con parametros n y $L(g)$. Así, para obtener cotas para el éxito de la minimización del error empírico, necesitamos estudiar desviaciones uniformes de variables aleatorias binomiales de su media. En el siguiente capítulo resumiremos las teorías básicas subyacentes.

Capítulo 3

Desigualdades de Concentración

3.1. Desigualdad de Hoeffding

En este capítulo estudiaremos desigualdades que permiten acotar la probabilidad de que los términos en el lado derecho de 2.1 tomen valores grandes. Las desigualdad más simple a utilizar para acotar la diferencia entre una variable aleatoria y su valor esperado es la desigualdad de Chebyshev y apoyandonos en la Desigualdad de Markov podremos demostrar resultados importantes.

Teorema 3.1 (Desigualdad de Markov) *Para toda variable aleatoria X no negativa, y $t > 0$,*

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}X}{t}$$

Demostración:
Sea X una variable aleatoria no-negativa y $t > 0$. Denotemos $\{X \geq t\}$ el evento A y i_A una variable discreta que toma valores 1 y 0 con función de probabilidad $P(A)$ y $1 - P(A)$ respectivamente. Es decir, $P(i_A = 1) = P(A)$ y $P(i_A = 0) = 1 - P(A)$.

Ahora bien, si $X \geq t$ y $t > 0$ entonces $\frac{X}{t} \geq 1$ resultando que, $i_A \leq i_A \frac{X}{t} \leq \frac{X}{t}$ y así, $\mathbb{E}(i_A) \leq \mathbb{E}\frac{X}{t}$, en consecuencia

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}X}{t}$$

.

■

Del teorema anterior, se deduce:

Teorema 3.2 (Desigualdad de Chebyshev) *Para toda variable aleatoria X, y $t >$ o, tenemos que*

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} = \mathbb{P}\{|X - \mathbb{E}X|^2 \geq t^2\} \leq \frac{Var\{X\}}{t^2}.$$

Demostración:

Primero veamos que si X es una variable aleatoria no-negativa para $t, r > 0$ tenemos que $\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}X^r}{t^r}$. En efecto,

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{X^r \geq t^r\}$$

y aplicando la desigualdad de Markov (Teorema 3.1), tenemos:

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{X^r \geq t^r\} \leq \frac{\mathbb{E}X^r}{t^r}$$

En nuestro caso se tiene que $(X - \mathbb{E}X)^2$ es una variable aleatoria no negativa con $r = 2$. Así,

$$\mathbb{P}\{(X - \mathbb{E}X)^2 \geq t^2\} \leq \frac{\mathbb{E}(X - \mathbb{E}X)^2}{t^2}$$

Por lo tanto,

$$\begin{aligned} \mathbb{P}\{|X - \mathbb{E}X| \geq t\} &= \mathbb{P}\{(X - \mathbb{E}X)^2 \geq t^2\} \\ &\leq \frac{\mathbb{E}(X - \mathbb{E}X)^2}{t^2} \\ &= \frac{Var\{X\}}{t^2}. \end{aligned}$$

■

Como ejemplo, podemos sacar provecho a la $\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\}$ con

$$S_n = \sum_{i=1}^{n} X_i,$$

donde $X_1, ..., X_n$ son variables aleatorias independientes reales. Así, por la Desigualdad de Chebyshev y la independencia de las variables aleatorias nos da inmediatamente

$$\mathbb{P}\{|S_n - \mathbb{E}S_n| \geq t\} \leq \frac{Var\{S_n\}}{t^2} = \frac{\sum_{i=1}^{n} Var\{X_i\}}{t^2}$$

Lo anterior es quizás, mejor visto si suponemos que las variables aleatorias X_i con $i = \{1, ..., n\}$ son variables independientes e identicamente distribuidas con distribución de Bernoulli(p), es decir,

$$\mathbb{P}\{X_i = 1\} = 1 - \mathbb{P}\{X_i = 0\} = p.$$

Veamos:

$$\begin{aligned}\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| \geq \epsilon\right\} &\leq \frac{Var\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\}}{\epsilon^2} \\ &\leq \frac{np(1-p)}{n^2\epsilon^2} \\ &= \frac{p(1-p)}{n\epsilon^2}.\end{aligned}$$

Un mejoramiento de esta desigualdad es obtenida por el Método del acotamiento de Chernoff o mejor conocida como **método de Chernoff**. El cual viene dado gracias a la desigualdad de Markov. Si s es un número arbritario positivo, entonces para toda variable aleatoria X y $t > 0$

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

El método de Chernoff consiste en encontrar el valor de $s > 0$ que minimiza la cota superior, o hace la cota superior más pequeña.

En el caso de una suma de variables aleatorias independientes y por propiedades de la esperanza,

$$\begin{aligned}\mathbb{P}\{S_n - \mathbb{E}S_n \geq t\} &\leq \frac{\mathbb{E}e^{s(S_n - \mathbb{E}S_n)}}{e^{st}} \\ &= \frac{\mathbb{E}\left(e^{s(\sum_{i=1}^{n} X_i - \mathbb{E}\sum_{i=1}^{n} X_i)}\right)}{e^{st}} \\ &= \frac{\mathbb{E}\left(e^{s(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}(X_i))}\right)}{e^{st}} \\ &= e^{-st}\mathbb{E}\left(e^{s\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)}\right) \\ &= e^{-st}\mathbb{E}\left(\prod_{i=1}^{n} e^{s(X_i - \mathbb{E}X_i)}\right) \\ &= e^{-st}\prod_{i=1}^{n}\mathbb{E}e^{s(X_i - \mathbb{E}X_i)}.\end{aligned}$$

En consecuencia de lo anterior, el problema de encontrar cotas se reduce a encontrar una cota superior pero para la función generadora de momentos de las variables aleatorias $X_i - \mathbb{E}X_i$. Quizás para acotar variables aleatorias la versión más elegante es la de Hoeffding propuesta en 1963.

Lema 3.1 *Sea X una variable aleatoria con $\mathbb{E}X = 0$, $a \leq X \leq b$. Entonces para todo $s > 0$,*

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

Demostración:

Notemos que como e^x es una función convexa la

$$e^{sx} \leq \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb}, \forall x \in [a,b]$$

y usando la hipótesis que $\mathbb{E}X = 0$ e introduciendo la notación $p = \dfrac{-a}{b-a}$ se sigue que

$$\begin{aligned}
\mathbb{E}e^{sX} &\leq \mathbb{E}\left(\frac{b-X}{b-a}e^{sa}\right) + \mathbb{E}\left(\frac{X-a}{b-a}e^{sb}\right), \forall X \in [a,b] \\
&= \left(\mathbb{E}\frac{b}{b-a} + \mathbb{E}\frac{-X}{b-a}\right)\mathbb{E}e^{sa} + \left(\mathbb{E}\frac{X}{b-a} + \mathbb{E}\frac{-a}{b-a}\right)\mathbb{E}e^{sb} \\
&= \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\
&= \frac{be^{sa} - ae^{sb}}{b-a} \\
&= \frac{e^{sa}(b - ae^{sb}e^{-sa})}{b-a}. \qquad (3.1)
\end{aligned}$$

pero, $e^{-ps(b-a)} = e^{sa} = e^{(-\frac{-a}{b-a})s(b-a)}$ al sustitutir esto en (3.1), tenemos:

$$\begin{aligned}
\frac{e^{-ps(b-a)}(b - ae^{sb}e^{-sa})}{b-a} &= \frac{be^{-ps(b-a)}}{b-a} - \frac{ae^{s(b-a)}e^{-ps(b-a)}}{b-a} \\
&= \frac{b}{b-a}e^{-ps(b-a)} + pe^{s(b-a)}e^{-ps(b-a)}
\end{aligned}$$

sumando y restando convenientemente la expresión: $\dfrac{a}{b-a}e^{-ps(b-a)}$

$$\begin{aligned}
\frac{b}{b-a}e^{-ps(b-a)} + pe^{s(b-a)}e^{-ps(b-a)} &= \frac{b-a}{b-a}e^{-ps(b-a)} - \frac{-a}{b-a}e^{-ps(b-a)} \\
&\quad + pe^{s(b-a)}e^{-ps(b-a)} \\
&= \left(1 - p + pe^{s(b-a)}\right)e^{-ps(b-a)}
\end{aligned}$$

Por lo tanto,

$$\mathbb{E}e^{sX} \leq \left(1 - p + pe^{s(b-a)}\right)e^{-ps(b-a)},$$

Considerando $u = s(b-a)$, y $\phi(u) = -pu + \log(1-p+pe^u)$

$$\begin{aligned} \mathbb{E}e^{sX} &\leq (1-p+pe^u)\, e^{-pu} \\ &= e^{\phi(u)}. \end{aligned}$$

En efecto,

$$\begin{aligned} e^{\phi(u)} &= e^{-pu+\log(1-p+pe^u)} \\ &= e^{-pu} e^{\log(1-p+pe^u)} \\ &= e^{-pu}(1-p+pe^u). \end{aligned}$$

Por lo tanto,

$$\mathbb{E}e^{sX} \leq e^{\phi(u)}.$$

Calculemos ahora la expansión en serie de Taylor de $\phi(u)$
Con $\phi(0) = 0$

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

Por lo tanto,

$$\phi'(0) = 0. \tag{3.2}$$

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

Por otra parte, $\frac{1}{4}$ es el máximo absoluto de la función $\phi''(u)$. Apoyandonos en el criterio de la segunda derivada podemos demostrarlo. Veamos,

Primero busquemos la derivada de $\phi''(u)$ para así poder encontar su valor crítico.

$$\begin{aligned}
\phi'''(u) &= \frac{(p(1-p)e^{-u})'(p+(1-p)e^{-u})^2 - (p(1-p)e^{-u})((p+(1-p)e^{-u})^2)'}{(p+(1-p)e^{-u})^4} \\
&= \frac{(-p(1-p)e^{-u})(p^2+2(1-p)e^{-u}+(1-p)^2e^{-u})}{(p+(1-p)e^{-u})^4} \\
&\quad + \frac{2p(1-p)^2e^{-2u}(p+(1-p)e^{-u})}{(p+(1-p)e^{-u})^4} \\
&= \frac{-p^3(1-p)e^{-u} - \cancel{2p(1-p)^2e^{-2u}} - p(1-p)^3e^{-3u}}{(p+(1-p)e^{-u})^4} \\
&\quad + \frac{\cancel{2p^2(1-p)^2e^{-2u}} + 2p(1-p)^3e^{-3u}}{(p+(1-p)e^{-u})^4} \\
&= \frac{-p^3(1-p)e^{-u} + p(1-p)^3e^{-3u}}{(p+(1-p)e^{-u})^4}.
\end{aligned}$$

Entonces, $\phi'''(u) = 0$ si y solo si $-p^3(1-p)e^{-u} + p(1-p)^3e^{-3u} = 0$

$$\begin{aligned}
&\implies p(1-p)^3e^{-3u} = p^3(1-p)e^{-u} \\
&\implies e^{-2u} = \frac{p^2}{(1-p)^2} \\
&\implies \log\left(e^{-2u}\right) = \log\left(\frac{p^2}{(1-p)^2}\right) \\
&\implies -2u = \log\left(\frac{p}{1-p}\right)^2 \\
&\implies u = -\frac{1}{2}\log\left(\frac{p}{1-p}\right)^2 \\
&\implies u = -\log\left(\left(\frac{p}{1-p}\right)^2\right)^{1/2} \\
&\implies u = -\log\left(\frac{p}{1-p}\right).
\end{aligned}$$

Por lo tanto,

$$\phi'''\left(-\log\left(\frac{p}{1-p}\right)\right) = 0.$$

Ahora bien, la segunda derivada de $\phi''(u)$ viene dada por:

$$\begin{aligned}\phi^{(iv)}(u) &= \frac{(p^3(1-p)e^{-u}-3p(1-p)^3e^{-3u})(p+(1-p)e^{-u})^4}{(p+(1-p)e^{-u})^8} \\ &\quad + \frac{4(-p^3(1-p)e^{-u}+p(1-p)^3e^{-3u}(1-p)e^{-u})}{(p+(1-p)e^{-u})^8} \\ &= \overbrace{p^3(1-p)e^{-u}-3p(1-p)^3e^{-3u}(p+(1-p)e^{-u})^{-4}}^{f_1(u)} \\ &\quad + \overbrace{4(1-p)e^{-u}(-p^3(1-p)e^{-u}+p(1-p)^3e^{-3u})(p+(1-p)e^{-u})^{-5}}^{f_2(u)}\end{aligned}$$

es decir,

$$\phi^{(iv)}(u) = f_1(u) + f_2(u).$$

Por lo tanto,

$$\phi^{(iv)}\left(-\log\left(\frac{p}{1-p}\right)\right) = f_1\left(-\log\left(\frac{p}{1-p}\right)\right) + f_2\left(-\log\left(\frac{p}{1-p}\right)\right). \tag{3.3}$$

y considerando para $a > 0$, que:

$$e^{-a\left(-\log\left(\frac{p}{1-p}\right)\right)} = \frac{p^a}{(1-p)^a},$$

al sustituirlo en (3.3), resulta

$$\begin{aligned}\phi^{(iv)}\left(-\log\left(\frac{p}{1-p}\right)\right) &= (p^4-3p^4)(2p)^{-4} + 4p(-p^4+p^4)(p+p)^{-5} \\ &= -\frac{2p^4}{(2p)^4} \\ &= -\frac{1}{8} \\ &< 0. \end{aligned} \tag{3.4}$$

En conclusión, como $\phi''(u)$ es una función que cumple (3.3), con $\phi^{(iv)}$ definida en $\left(-\log\frac{p}{(1-p)}\right)$ y se cumple (3.4), por el criterio de la segunda derivada $\phi''\left(-\log\frac{p}{(1-p)}\right)$ es un máximo relativo de ϕ''. Y por ser unico, necesariamente es un máximo absoluto.

Es decir,

$$\begin{aligned}\phi''\left(-\log\frac{p}{(1-p)}\right) &= \frac{p(1-p)\frac{p}{(1-p)}}{\left(p+(1-p)\frac{p}{(1-p)}\right)^2}\\ &= \frac{p^2}{4p^2}\\ &= \frac{1}{4},\end{aligned}$$

es un máximo absoluto de ϕ''.

De todo lo anterior tenemos que $\phi(0)=0$, $\phi'(0)=0$ y $\phi''(u)\leq\frac{1}{4}$ y así, la serie de Taylor, para algún $\theta\in[0,u]$, esta dada por

$$\phi(u)=\phi(0)+u\phi'(0)+\frac{u^2}{2}\phi''(\theta)\leq\frac{u^2}{2}\frac{1}{4}=\frac{u^2}{8}=\frac{s^2(b-a)^2}{8}$$

En consecuencia,

$$\mathbb{E}e^{sX}\leq e^{\phi(u)}\leq e^{s^2(b-a)^2/8}.$$

■

Ahora, directamente usaremos éste lema para acotar lo obtenido en el método de Chernoff, resultando de la forma:

$$\begin{aligned}&\mathbb{P}\{S_n-\mathbb{E}S_n\geq\epsilon\}\\ &\quad\leq e^{-s\epsilon}\prod_{i=1}^n\mathbb{E}\{e^{s(X_i-\mathbb{E}X_i)}\}\\ &\quad\leq e^{-s\epsilon}\prod_{i=1}^n\mathbb{E}\{e^{s^2(b_i-a_i)^2/8}\},\text{ por 3.1}\\ &\quad= e^{-s\epsilon}e^{s^2\sum_{i=1}^n(b_i-a_i)^2/8}\\ &\quad= e^{-s\epsilon+s^2\sum_{i=1}^n(b_i-a_i)^2/8},\text{tomando } s=4\epsilon/\sum_{i=1}^n(b_i-a_i)^2\\ &\quad= e^{\left\{\frac{-4\epsilon^2}{\sum_{i=1}^n(b_i-a_i)^2}+\frac{4^2\epsilon^2}{(\sum_{i=1}^n(b_i-a_i)^2)^2}\cdot\frac{\sum_{i=1}^n(b_i-a_i)^2}{8}\right\}}\\ &\quad= e^{\left\{\frac{-4\epsilon^2}{\sum_{i=1}^n(b_i-a_i)^2+}+\frac{2\epsilon^2}{\sum_{i=1}^n(b_i-a_i)^2}\right\}}\\ &\quad= e^{-2\epsilon^2/\sum_{i=1}^n(b_i-a_i)^2}\end{aligned}$$

El resultado anterior es conocido como la desigualdad de Hoeffding. Es decir, todo lo anterior, no es más que la demostración de la primera desigualdad del siguiente teorema.

Teorema 3.3 (Desigualdad de Hoeffding) *Sea $X_1, X_2 \dots X_n$ variables aleatorias independientes acotadas talque X_i pertenezca al intervalo $[a_i, b_i]$ con probabilidad 1. Denote su suma por $S_n = \sum_{i=1}^{n} X_i$. Entonces para todo $\epsilon > 0$ tenemos*

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

y

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

Quedando demostrado la primera desigualdad basta considerar para la segunda que se procede de manera análoga, tomando en cuenta que

$$\mathbb{P}\{S_n - \mathbb{E}S_n \leq -\epsilon\} = \mathbb{P}\{-S_n - (-\mathbb{E}S_n) \geq \epsilon\}.$$

■

Dicho teorema fue demostrado por Chernoff(1952) y Okamoto(1952), para variables aleatorias con distribución binomial.

Esto es, cuando $X_i's$ son variables i.i.d Bernoulli(p), tenemos:

$$\mathbb{P}\{S_n/n - p \geq \epsilon\} = \mathbb{P}\{S_n - np \geq \epsilon\} \leq e^{-2n\epsilon^2}.$$

Podemos combinar esta desigualdad con el lema 2.1 para acotar el rendimiento de la minimización del error empírico en el caso especial que la clase C contenga una cantidad finita de clasificadores. Para ello enunciaremos el siguiente teorema:

Teorema 3.4 *Supongamos que el cardinal de C es acotada por N. Entonces tenemos que para todo $\epsilon > 0$*

$$\mathbb{P}\left\{\sup_{g \in C} |\widehat{L}_n(g) - L(g)| > \epsilon\right\} \leq 2Ne^{-2n\epsilon^2}.$$

Demostración:

Para demostrar este teorema basta usar la desigualdad de Hoeffding, junto con la propiedad que Durrett en [1], señala como $\mathbb{P}(\bigcup A_i) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$, y el hecho que la variable aleatoria $n\widehat{L}_n(g)$ esta distribuida binomialmente con parametros n y $L(g)$.

■

Notemos que la distribución actual de los datos no juega un papel importante en la cota superior resultante.

Para tener una idea del tamaño del error, estamos interesados en la esperanza de la desviación máxima

$$\mathbb{E}\sup_{g\in C}|\widehat{L}_n(g) - L(g)|.$$

Para encontrar una cota para esta expresión, podemos usar la desigualdad anterior (Teorema 3.4) y el hecho como se define en [6], que para toda variable aleatoria no negativa X,

$$\mathbb{E}X = \int_0^\infty \mathbb{P}\{X \geq t\}dt,$$

pero, obtendremos resultados más claros combinando el lema 3.1 y el siguiente resultado:

Lema 3.2 *Sea $\sigma > 0$, $n \geq 2$ y sea $Y_1, Y_2 \dots Y_n$ variables aleatorias reales tal que para todo $s > 0$ y $1 \leq i \leq n$, $\mathbb{E}\{e^{sY_i}\} \leq e^{s^2\sigma^2/2}$. Entonces*

$$\mathbb{E}\{\max_{i\leq n} Y_i\} \leq \sigma\sqrt{2\ln n}.$$

Si además, $\mathbb{E}\{e^{s(-Y_i)}\} \leq e^{s^2\sigma^2/2}$, para todo $s > 0$ y $1 \leq i \leq n$, entonces para todo $n \geq 1$,

$$\mathbb{E}\{\max_{i\leq n} |Y_i|\} \leq \sigma\sqrt{2\ln(2n)}.$$

Demostración:

Por la desigualdad de Jensen's, para todo $s > 0$,

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

Por lo tanto,

$$e^{s\mathbb{E}\max_{i\leq n} Y_i} \leq ne^{s^2\sigma^2/2}$$

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

y tomando $s = \sqrt{\frac{2\ln n}{\sigma^2}}$, tenemos:

$$\begin{aligned} E\{\max_{i\leq n} Y_i\} &\leq \frac{\ln n\sigma}{\sqrt{2\ln n}} + \frac{\sqrt{2\ln n}n\sigma^{\not 2}}{2\not\sigma} \\ &= \frac{2\ln n\sigma + 2\ln n\sigma}{2\sqrt{2\ln n}} \\ &= \frac{2\ln n\sigma}{\sqrt{2\ln n}} \\ &= \sigma\sqrt{2\ln n} \end{aligned}$$

Finalmente, para demostrar la segunda desigualdad notemos que hay $2n$ Y_i en el $\max_{i\leq n}|Y_i| = \max(Y_1, -Y_1, \ldots, Y_n, -Y_n)$. Así, basta aplicar la primera desigualdad del lema 3.2 para esta prueba.

■

En conclusión, aplicando este resultado del lema anterior a nuestra variable $\sup_{g\in C}|\widehat{L}_n(g) - L(g)|$, usando el Lema 3.1 para demostrar que cumple las hipótesis y tomando $\sigma = \dfrac{1}{2\sqrt{n}} > 0$, tenemos que:

$$\mathbb{E}\sup_{g\in C}|\widehat{L}_n(g) - L(g)| \leq \sqrt{\frac{\ln(2N)}{2n}}. \tag{3.5}$$

3.2. Desigualdad de diferencias acotadas

En esta sección daremos una extensión de las desigualdades de concentración para funciones de variables aleatorias independientes.

Sea A un conjunto, y $g : A^n \to \mathcal{R}$ una función de n variables. Estudiaremos desigualdades para la diferencia entre $g(X_1, \ldots, X_n)$ y su valor esperado, cuando $X_1, \ldots, X_n$ son variables aleatorias independientes con valores en A. A veces escribiremos g en lugar de $g(X_1, \ldots, X_n)$ siempre que no cause confusión.

Estas desigualdades se obtendrán mediante un refinamiento del método de Chernoff y de su aplicación para la desigualdad de Hoeffding.

Recordemos que si X, Y son variables aleatorias acotadas e independientes, en [1], se cita que :

$$\mathbb{E}\{XY\} = \mathbb{E}\{\mathbb{E}(XY|Y)\} = \mathbb{E}\{Y\mathbb{E}\{X|Y\}\}. \tag{3.6}$$

El primer resultado de esta sección es una mejora de la desigualdad de Efron y Stein (1981) probada por Steele en 1986.

Teorema 3.5 (Desigualdad de Efron-Stein) *Si $X_1^{'}, \ldots, X_n^{'}$ forman una copia independiente de $X_1, \ldots, X_n$, entonces*

$$Var(g(X_1, \ldots, X_n)) \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\{(g(X_1, \ldots, X_n) - g(X_1, \ldots, X_i^{'}, \ldots, X_n))^2\}$$

Demostración:

Introduciremos la notación de $Z = g(X_1, \ldots, X_n)$ para mayor comprensión. y denotemos $V = g - \mathbb{E}g = Z - \mathbb{E}Z$.

Definamos $V_i = \mathbb{E}\{Z|X_1, \ldots, X_i\} - \mathbb{E}\{Z|X_1, \ldots, X_{i-1}\}$ una sucesión de diferencia de martingala respecto a a la sucesión $X_1, X_2, \ldots$
Por lo tanto, $V = \sum_{i=1}^{n} V_i$, se escribe como suma se diferencias de martingala. Veamos:

$$\begin{aligned}
\sum_{i=1}^{n} V_i &= \cancel{\mathbb{E}\{Z|X_1\}} - \mathbb{E}Z + \cancel{\mathbb{E}\{Z|X_1, X_2\}} - \cancel{\mathbb{E}\{Z|X_1\}} \\
&+ \cancel{\mathbb{E}\{Z|X_1, X_2, X_3\}} - \cancel{\mathbb{E}\{Z|X_1, X_2\}} + \ldots - \ldots \\
&+ \mathbb{E}\{Z|X_1, \ldots X_n\} - \cancel{\mathbb{E}\{Z|X_1, \ldots X_{n-1}\}} \\
&= \mathbb{E}\{Z|X_1, \ldots X_n\} - \mathbb{E}\{Z\} \\
&= Z - \mathbb{E}Z = V.
\end{aligned}$$

Luego,

$$\begin{aligned}
Var(Z) &= \mathbb{E}\{(Z - \mathbb{E}Z)^2\} \\
&= \mathbb{E}V^2 \\
&= \mathbb{E}\left(\sum_{i=1}^{n} V_i\right)^2 \\
&= \mathbb{E}\sum_{i=1}^{n} {V_i}^2 + 2\mathbb{E}\sum_{i>j} V_i V_j \\
&= \sum_{i=1}^{n} \mathbb{E}{V_i}^2.
\end{aligned}$$

Pues,

$$\begin{aligned}
\mathbb{E}V_i V_j &= \mathbb{E}\mathbb{E}\{V_i V_j | X_1, \ldots, X_j\} \\
&= \mathbb{E}V_j \mathbb{E}\{V_i | X_1, \ldots, V_j\}\} \\
&= 0.
\end{aligned}$$

Ya que V_i es una sucesión de diferencia de martingala respecto a $X_1, X_2, \ldots$ y como $i > j$ tenemos $\mathbb{E}\{V_i|X_1, \ldots, X_j\} = 0$.

Para acotar la esperanza de ${V_i}^2$, notemos que:

$$\begin{aligned} {V_i}^2 &= (\mathbb{E}\{Z|X_1, \ldots, X_i\} - \mathbb{E}\{Z|X_1, \ldots, X_{i-1}\})^2 \\ &= (\mathbb{E}[\mathbb{E}\{Z|X_1, \ldots, X_n\} - \mathbb{E}\{Z|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\,|X_1, \ldots X_i\}])^2. \end{aligned}$$

y como X^2 es una función convexa, por la desigualdad de Jensen tenemos que:

$$\begin{aligned} {V_i}^2 &\leq \mathbb{E}[(\mathbb{E}\{Z|X_1, \ldots, X_n\} - \mathbb{E}\{Z|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\})^2|X_1, \ldots X_i] \\ &= \mathbb{E}[(Z - \mathbb{E}\{Z|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\})^2|X_1, \ldots X_i]. \end{aligned}$$

Por lo tanto,

$$\begin{aligned} \mathbb{E}{V_i}^2 &\leq \mathbb{E}(Z - \mathbb{E}\{Z|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\})^2 \\ &= \mathbb{E}(Z - \mathbb{E}_i Z)^2. \end{aligned}$$

tomando, $\mathbb{E}_i Z = \mathbb{E}\{Z|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$.

Ahora bien, usando el hecho que $Var(X) = \frac{1}{2}\mathbb{E}\{(X - Y)^2\}$ para X y Y variables aleatorias independiente e identicamente distribuidas, y como $X_i^{'}$ es una copia independiente de X_i, podemos usar esa propiedad para $g(X_1, \ldots, X_i^{'}, \ldots X_n)$ y $g(X_1, \ldots, X_i, \ldots, X_n)$.

Por lo tanto,

$$\begin{aligned} \mathbb{E}{V_i}^2 &\leq \mathbb{E}(Z - \mathbb{E}_i Z)^2 \\ &= Var(Z) \\ &= \frac{1}{2}\mathbb{E}\{(g(X_1, \ldots, X_n) - g(X_1, \ldots, X_i^{'}, \ldots, X_n))^2\}. \end{aligned}$$

En conclusión, queda demostrado que

$$Var(g(X_1, \ldots, X_n) \leq \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\{(g(X_1, \ldots, X_n) - g(X_1, \ldots, X_i^{'}, \ldots, X_n))^2\}$$

■

Definición 3.1 *Sea $f : A^n \to \mathbb{R}$. Diremos que f satisface la propiedad de diferencias acotadas si existe $c_1, c_2, \ldots, c_n \geq 0$, para todo $1 \leq i \leq n$ y*

$$\sup_{X_1, \ldots, X_n} |f(X_1, \ldots, X_n) - f(X_1, \ldots, X_{i-1}, X_i^{'}, X_{i+1}, \ldots X_n)| \leq c_i.$$

Supongamos que $g : A^n \to \mathbb{R}$ satisface la definición anterior. En otras palabras, supongamos que si se cambia la i-esima variable de g, mientras las otras se mantienen fijas, el valor de la función no cambia mas que en c_i. Entonces de la desigualdad de Efron-Stein tenemos que:

$$\begin{aligned} Var(g) &\leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\{g(X_1,\ldots,X_n) - g(X_1,\ldots,X_i^{'},\ldots X_n)^2\} \\ &\leq \frac{1}{2}\sum_{i=1}^{n} {c_i}^2. \end{aligned}$$

Por lo tanto, para tales funciones es posible probar que la desigualdad de diferencias acotadas, es una extensión poderosa de la desigualdad de Hoeffding.

McDiarmid(1989) prueba esta desigualdad usando técnicas de martingala, que es la demostración que detallaremos aquí. La prueba del Teorema 3.6 usa la siguiente extensión del lema 3.1:

Lema 3.3 *Sea V y Z variables aleatorias tal que $\mathbb{E}\{V|Z\} = 0$ con probabilidad uno. y para alguna función h y constante $c \geq 0$*

$$h(Z) \leq V \leq h(Z) + c.$$

Entonces, para todo $s > 0$, $\mathbb{E}\{e^{sV|Z}\} \leq e^{s^2c^2/8}$.

Teorema 3.6 (Desigualdad de diferencia acotada) *Supongamos que*

$$X_1,\ldots,X_n \in A$$

son independientes, y $g : A^n \to \mathbb{R}$. Sea $c_1, c_2, \ldots, c_n \geq 0$, satisfaciendo que para $1 \leq i \leq n$

$$\sup_{X_1,\ldots,X_n,X_i^{'}\in A} |f(X_1,\ldots,X_n) - f(X_1,\ldots,X_{i-1},X_i^{'},X_{i+1},\ldots X_n)| \leq c_i.$$

Entonces, para todo $t > 0$ tenemos:

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

y

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

Demostración:

Al igual que en la prueba del teorema de Efron-Stein, introduciremos la notación $V = g - \mathbb{E}g$, y definimos

$$V_i = \mathbb{E}\{Z|X_1,\ldots,X_i\} - \mathbb{E}\{Z|X_1,\ldots,X_{i-1}\}.$$

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

Introduciremos las variables aleatorias:

$$H_i(X_1, \ldots, X_i) = \mathbb{E}\{g(X_1, \ldots, X_n)|X_1, \ldots, X_i\}.$$

y denotando la distribución de X_i por F_i para todo $i = 1, \ldots, n$, tenemos que:

$$V_i = H_i(X_1, \ldots, X_i) - \int H_i(X_1, \ldots, X_{i-1}, x)dF_i(x).$$

Además, definimos las variables aleatorias

$$W_i = \sup_u \left(H_i(X_1, \ldots, X_{i-1}, u) - \int H_i(X_1, \ldots, X_{i-1}, x)dF_i(x) \right),$$

y

$$Z_i = \inf_v \left(H_i(X_1, \ldots, X_{i-1}, v) - \int H_i(X_1, \ldots, X_{i-1}, x)dF_i(x) \right).$$

Claramente, $Z_i \leq V_i \leq W_i$ con probabilidad uno, y suponiendo que para todo i, Z_i es una función de $X_1, \ldots, X_{i-1}$ tenemos:

$$\begin{aligned}
W_i - Z_i &= \sup_u \left(H_i(X_1, \ldots, X_{i-1}, u) - \int H_i(X_1, \ldots, X_{i-1}, x)dF_i(x) \right) \\
&- \inf_v H_i(X_1, \ldots, X_{i-1}, v) + \int (H_i(X_1, \ldots, X_{i-1}, x)dF_i(x)) \\
&= \sup_u H_i(X_1, \ldots, X_{i-1}, u) - \inf_v H_i(X_1, \ldots, X_{i-1}, v) \\
&= \sup_u H_i(X_1, \ldots, X_{i-1}, u) + \sup_v H_i(X_1, \ldots, X_{i-1}, v) \\
&= \sup_{u,v} (H_i(X_1, \ldots, X_{i-1}, u) - H_i(X_1, \ldots, X_{i-1}, v)) \\
&\leq c_i.
\end{aligned}$$

Por suponer que se cumple la definición 3.1.

Por lo tanto, $Z_i \leq V_i \leq Z_i + c_i$ con $\mathbb{E}\{V_i|X_1, \ldots, X_{i-1}\} = 0$, es decir, se cumplen las hipotesis del lema 3.3, por ende para todo $i = 1, \ldots, n$

$$\mathbb{E}\{e^{sV_i}|X_1, \ldots, X_{i-1}\} \leq e^{s^2 {c_i}^2/8}.$$

Finalmente, por lo anterior y el método de acotamiento de Chernoff, para todo $s > 0$ tenemos:

$$\begin{aligned}
&\mathbb{P}\{g - \mathbb{E}g \geq t\} \\
&\qquad \leq \quad e^{-st}\mathbb{E}e^{(sg-\mathbb{E}g)} \\
&\qquad = \quad e^{-st}\mathbb{E}e^{s(\sum_{i=1}^{n} V_i)} \\
&\qquad = \quad e^{-st}\mathbb{E}\{e^{sV_n} e^{s\sum_{i=1}^{n-1} V_i}\} \\
&\qquad = \quad e^{-st}\mathbb{E}\{e^{s\sum_{i=1}^{n-1} V_i}\mathbb{E}e^{sV_n}|X_1, \ldots, X_{n-1}\} \\
&\qquad \leq \quad e^{-st}\mathbb{E}\{e^{s\sum_{i=1}^{n-1} V_i} e^{s^2c^2/8}\} \\
&\qquad = \quad e^{-st}e^{s^2c^2/8}\mathbb{E}e^{s\sum_{i=1}^{n-1} V_i}.
\end{aligned}$$

Ahora, repitiendo este procedimiento $n-1$ veces más, resulta:

$$\begin{aligned}
\mathbb{P}\{g - \mathbb{E}g \geq t\} &= e^{-st}e^{s^2c^2/8}e^{s^2{c_{n-1}}^2/8} \cdot \ldots \cdot e^{s^2{c_2}^2/8}e^{s^2{c_1}^2/8} \\
&= e^{-st}e^{s^2(\sum_{i=1}^{n} {c_i}^2/8)},
\end{aligned}$$

y tomando, $s = 4t/\sum_{i=1}^{n} {c_i}^2$ demostramos la primera desigualdad. La segunda se prueba de manera similar.

■

Una importante aplicación de la desigualdad de diferencias acotadas muestra que si $\mathcal{C}$ es alguna de las clases de clasificadores de la forma $g : \mathbb{R}^d \to \{0, 1\}$, y vemos la variable $\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)|$ como una función de n pares de variables aleatorias (X_i, Y_i) , $i = 1, \ldots, n$, inmediatamente veremos que de la propiedad de diferencias acotadas es satisfecho con $c_i = 1/n$, y 3.1 implica inmediatamente que:

$$\mathbb{P}\left\{\left|\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)| - \mathbb{E}\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)|\right| > \epsilon\right\} \leq 2e^{-2n\epsilon^2}$$

Lo interesante de este resultado es que independientemente del tamaño del valor esperado, la variable aleatoria $\sup_{g\in\mathcal{C}} |\widehat{L}_n(g) - L(g)|$ está claramente concentrada alrededor de su media con alta probabilidad. En el siguiente y último capítulo de este trabajo estudiaremos su valor esperado.

Capítulo 4

Desigualdad de Vapnik-Chervonenkis

Recordando de secciones anteriores la expresión 3.4 y 3.5, dicen que para toda clase finita C de clasificadores, y para todo $\epsilon > 0$,

$$\mathbb{P}\left\{\sup_{g\in C}|\widehat{L}_n(g) - L(g)| > \epsilon\right\} \leq 2Ne^{-2n\epsilon^2},$$

y

$$\mathbb{E}\sup_{g\in C}|\widehat{L}_n(g) - L(g)| \leq \sqrt{\frac{\ln(2N)}{2n}}.$$

Esta simple cota puede resultar inútil si el cardinal de la clase C es muy grande o infinito. El propósito de éste capítulo es introducir una teoría que mejore estos casos.

Sea $X_1, \ldots, X_n$ variables aleatorias independientes que toman valores en $\mathcal{R}^d$ con la misma distribución.

$$\mu(A) = \mathbb{P}\{X_1 \in A\} \quad (A \subset \mathbb{R}^d).$$

Se define la distribución empírica como

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{[X_i\in A]}$$

Considere una clase $\mathcal{A}$ de subconjuntos de $\mathbb{R}^d$. Nuestra principal preocupación es el comportamiento de la variable aleatoria $\sup_{A\in\mathcal{A}}|\mu_n(A) - \mu(A)|$. Se vio en el capítulo anterior que una simple consecuencia de la desigualdad de diferencias acotadas es que

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

para todo n y $t > 0$. Esto muestra que para toda clase $\mathcal{A}$, la desviación máxima es claramente concentrada alrededor de su media.

Introduciremos un término nuevo, que nos permitirá dar una cota para

$$\mathbb{E} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|.$$

Definición 4.1 *El Coeficiente de separación VC es una cantidad dada por:*

$$\mathcal{S}_{\mathcal{A}}(n) = \max_{X_1,\ldots,X_n \in \mathcal{R}^d} |\{\{X_1, \ldots, X_n\} \cap A; A \in \mathcal{A}\}|$$

Es decir, $\mathcal{S}_{\mathcal{A}}(n)$ es el número máximo de los diferentes subconjuntos de un conjunto de n puntos el cual puede obtenerse por la intersección con elementos de A.

A partir de esta definición podemos introducir, el teorema principal de éste capítulo, un resultado de Vapnik-Chervonenki.

Teorema 4.1 (Desigualdad de Vapnik-Chervonenkis)

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

Demostración:

Introduciremos $X_1^{'}, \ldots, X_n^{'}$ una copia independiente de $X_1, \ldots, X_n$. Además, definimos n variables aleatoria de signos $\sigma_1, \ldots, \sigma_n$ talque $\mathbb{P}\{\sigma_1 = -1\} = \mathbb{P}\{\sigma_1 = 1\} = 1/2$, independientes de $X_1, X_1^{'}, \ldots, X_n, X_n^{'}$.

Entonces, denotando $\mu_n^{'}(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[X_i^{'} \in A]}$, podemos escribir

$$\begin{aligned}
&\mathbb{E}\left\{ \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu_(A)| \right\} \\
&= \mathbb{E}\left\{ \sup_{A \in \mathcal{A}} |\mathbb{E}\{\mu_n(A) - \mu_n^{'}(A) | X_1, \ldots, X_n\}| \right\} \\
&\leq \mathbb{E}\left\{ \sup_{A \in \mathcal{A}} \mathbb{E}\{\mu_n(A) - \mu_n^{'}(A) | X_1, \ldots, X_n\} \right\} \\
&\quad \text{(por la desigualdad de JensenÂ´s)} \\
&\leq \mathbb{E}\left\{ \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu_n^{'}(A)| \right\} \\
&\quad \text{(ya que} \quad \sup \mathbb{E}(\cdot) \leq \mathbb{E}\sup(\cdot)) \\
&= \frac{1}{n}\mathbb{E}\left\{ \sup_{A \in \mathcal{A}} \left| \sum_{i=1}^{n} \sigma_i \left(\mathbb{I}_{[X_i \in A]} - \mathbb{I}_{[X_i^{'} \in A]} \right) \right| \right\} \\
&= \frac{1}{n}\mathbb{E}\left\{ \mathbb{E}\left\{ \sup_{A \in \mathcal{A}} \left| \sum_{i=1}^{n} \sigma_i \left(\mathbb{I}_{[X_i \in A]} - \mathbb{I}_{[X_i^{'} \in A]} \right) \right| \mid X_1, X_1^{'}, \ldots, X_n, X_n^{'} \right\} \right\} \\
&\quad \text{(ya que , } X_1, X_1^{'}, \ldots, X_n, X_n^{'} \text{son i.i.d).}
\end{aligned}$$

Luego, por la indenpendencia de $\sigma_i{}'s$ del resto de las variables, fijamos los valores de $X_1 = x_1, X_1^{'} = x_1^{'}, \ldots, X_n = x_n X_n^{'} = x_n^{'}$, y estudiamos

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

Ahora bien, denotemos por $\widehat{\mathcal{A}}\subset\mathcal{A}$ una colección de conjuntos tal que cada par de conjuntos en $\widehat{\mathcal{A}}$ tienen diferentes intersecciones con el conjunto $\{x_1, x_1^{'}, \ldots, x_n, x_n^{'}\}$, y toda posible intersección es representada una vez.

Esto es,

$$|\widehat{\mathcal{A}}| \leq |\mathcal{A}| \leq 2n \leq \mathcal{S}_{\mathcal{A}}(2n), \tag{4.1}$$

y

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

Además, observemos que cada $\sigma_i(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i^{'}\in A]})$ tiene media cero y toma valores en $[-1,1]$, cumpliendo así las hipótesis del Lema 3.1 para todo $s>0$, por lo tanto

$$\mathbb{E}e^{s\sum_{i=1}^{n}\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i^{'}\in A]}\right)} = \prod_{i=1}^{n}\mathbb{E}e^{s\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i^{'}\in A]}\right)} \leq e^{ns^2(2)^2/8} = e^{ns^2/2}.$$

Observación 4.1 *Según cita Lugosi en [7],el coeficiente de separación cuando n, m son enteros, cumple la siguiente propiedad:*

$$\mathcal{S}_{\mathcal{A}}(n+m) \leq \mathcal{S}_{\mathcal{A}}(n).\mathcal{S}_{\mathcal{A}}(m)$$

En consecuencia, cuando n y m son iguales, tenemos que $\mathcal{S}_{\mathcal{A}}(2n) \leq (\mathcal{S}_{\mathcal{A}}(n))^2$.

Por lo tanto, de todo lo anterior y como $\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i^{'}\in A]}\right)$ es simétrica, el Lema 3.2 implica inmediatamente que:

$$\begin{aligned}
\mathbb{E}\left\{\max_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{I}_{[x_i\in A]}-\mathbb{I}_{[x_i^{'}\in A]}\right)\right|\right\} &\leq \sqrt{n}\sqrt{2\ln 2n} \\
&\leq \sqrt{2n\ln(\mathcal{S}_{\mathcal{A}}(2n))} \text{ Por 4.1} \\
&\leq \sqrt{2n\ln(\mathcal{S}_{\mathcal{A}}(n))^2}, \text{ de Observación 4.1} \\
&\leq 2\sqrt{n\ln(\mathcal{S}_{\mathcal{A}}(n))}.
\end{aligned}$$

Así,

$$\begin{aligned}\mathbb{E}\left\{\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu_(A)|\right\} &\leq \frac{2}{n}\sqrt{n\ln(\mathcal{S}_{\mathcal{A}}(n))}\\ &= 2\sqrt{\frac{\ln(\mathcal{S}_{\mathcal{A}}(n))}{n}}\\ &\leq 2\sqrt{\frac{\log(2\mathcal{S}_{\mathcal{A}}(n))}{n}} \text{ (Por cambio de base del}\log_e X).\end{aligned}$$

Por lo tanto,

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

■

Por lo tanto, apoyandonos en la teoría de Vapnik-Chervonenkis hemos encontrado una cota del valor esperado de $\sup_{A\in\mathcal{A}}\{|\mu_n(A)-\mu(A)|\}$, en el caso de que N tome valores muy grandes o que el cardinal de C sea infinito.

Conclusiones

En este trabajo hemos estudiado algunas desigualdades de concentración que son de gran utilidad para la teoría de aprendizaje, en particular la teoría de aprendizaje supervisado aplicado al problema de clasificación binario.

Además, se observa lo importante que es implementar la teoría de Vapnik-Chervonenkis pues nos permite encontrar cotas del valor esperado, en el caso que C tenga cardinal infinito, para así, poder obtener una cota para el exito de una buena minimización del error empírico y en consecuencia un excelente clasificador con una mínima probabilidad de error.

Por lo tanto, la elección adecuada del clasificador, requiere del uso de herramientas probabilísticas. Es por esto que el desarrollo de este trabajo permitirá ser base para el estudio de nuevas desigualdades y nuevas cotas, para entender correctamente el funcionamiento de otros métodos de aprendizaje supervisado. Permitiendo ser una herramienta de información para crear o mejorar aplicaciones ya existentes como, diagnosticos médicos, reconocimiento de caracteres, redes neuronales, clasificación de textos, entre otros.

Bibliografía

[1] Rick Durrett. *Probability Theory and Examples.* 4th edition, 2010.

[2] Margarita Gallardo. Aplicación de técnicas de clustering para la mejora del aprendizaje. pages 7–9, 2009.

[3] Marco Gaxiola. Concentración de medidas de probabilidad. page 11, 2012.

[4] Lázló Gyorfi and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. 1996.

[5] Rosalia Hernandez. Una introducción a la teoría de martingalas. page 28, 2002.

[6] Ji Liu. Notes of statistical inequalities. pages 1–12.

[7] Gábor Lugosi. Principles of nonparametric learning. pages 3–20, 2002.

[8] Sheldon Ross. *Introduction to Probability Models.* 6th edition, 1997.

[9] Leticia Seijas. Reconocimiento de patrones utilizando técnicas estadísticas y conexionistas aplicadas a la clasificación de dígitos manuscritos. page 16, 2011.

[10] Vapnik y A. Chervonenkis. Theory of pattern recognition. 1974.