

01. Introducción al modelo de aprendizaje PAC

1 Problemas

Problema 1. La clase de los rectángulos alineados en el plano es

$$\mathcal{H}_{\text{rec}}^2 = \{h_{(a_1, b_1, a_2, b_2)} : a_1 \leq b_1, \quad a_2 \leq b_2\}$$

1. Sea A el algoritmo que retorna el rectángulo más pequeño que contiene a todas las observaciones positivas en la muestra de entrenamiento. Si este algoritmo culmina en el rectángulo h_A , note que este no puede contener observaciones negativas. Por lo tanto, si h_A fuese más pequeño, por lo menos una observación positiva sería marcada como negativa, incrementando el error empírico. Si h_A fuese más grande, puede atrapar a una observación negativa, clasificándola como positiva e incrementando $L_{D,f}(h_A)$. Por lo tanto, A es un ERM.
2. Dado que R^* contiene a todos los puntos positivos, entonces $R(S) \subseteq R^*$. Considere el evento $L_{D,f}(h_S) > \epsilon$ como una falla para el algoritmo. Sea $S|_x = (x_1, \dots, x_m)$ instancias del conjunto de entrenamiento. Buscamos acotar la probabilidad en la que S no contiene un ejemplo de R_i , i.e. $D^m(\{S|_x : L_{(D,f)}(h_S) > \epsilon\})$. Sea M el conjunto de muestras que falla en pertenecer a algún R_i y sea \mathcal{H}_B el conjunto de malas hipótesis. Por lo tanto,

$$D^m(\{S|_x : L_{(D,f)}(h_S) > \epsilon\}) \leq D^m(M) \leq \sum_{h \in \mathcal{H}_B} D^m(\{S|_x : L_S(h) = 0\})$$

Por el Lema 2.2, para algún R_i ,

$$D^m(M) \leq \sum_{i=1}^4 D^m(M_{R_i}) \leq \sum_{i=1}^4 |\mathcal{H}| e^{-\epsilon m} = 4|\mathcal{H}| e^{-\epsilon m}$$

Sin embargo, por cada R_i de debe tomar en cuenta un valor de m , obteniendo así el resultado por el Corolario 2.3. Con una probabilidad de, por lo menos $1 - \delta$, si $m \geq 4 \ln(4/\delta)/\epsilon$, entonces A elige una clase de hipótesis con un error de, a lo sumo, ϵ .

3. En \mathbb{R}^d , el argumento anterior es válido si considera a la misma cantidad de R_i como caras de un hipercubo en tal dimensión d . Por lo tanto, si A recibe una muestra de entrenamiento de tamaño $m \geq 2^d \ln(2^d/\delta)/\epsilon$, entonces con probabilidad de, por lo menos, $1 - \delta$, la hipótesis tiene un error de, a lo sumo, ϵ .
4. Debe ser un algoritmo polinomial por la construcción de las posibles combinaciones en el inciso 2 del Trabajo Numérico.

2 Trabajo numérico

El espacio de hipótesis consiste de rectángulos dentro de \mathcal{X} , cuya esquina inferior izquierda se puede desplazar 0.01 unidades en cada dirección. Es decir, cada hipótesis se representa por un rectángulo, el cual está totalmente determinado por las coordenadas de la esquina inferior izquierda (h_1, h_1) y de la esquina superior derecha (h_2, h_2) , dado que ambas recorren la diagonal de \mathcal{X} . Por lo tanto

$$\mathcal{H} = \{[h_1, h_2] \times [h_1, h_2] : 0 \leq h_1 < h_2 \leq 2, \quad h_1, h_2 \in \{0.01[200]\}\} \quad (1)$$

1. La expresión (1) representa una parametrización del espacio de hipótesis.
2. Note que la cantidad de formas de elegir h_1 y h_2 es

$$|\mathcal{H}| = \frac{200 \cdot 201}{2} = 20,100$$

3. Si la muestra S es de tamaño m , la probabilidad de extraer un ejemplo en la clase $h_S = [h_1, h_2] \times [h_1, h_2]$ es de

$$\mathbb{P}_{ex}(h_S) = m \cdot \frac{(h_2 - h_1)^2}{4}$$

4. A continuación se muestra uno de los conjuntos de entrenamiento en la Figura 1.

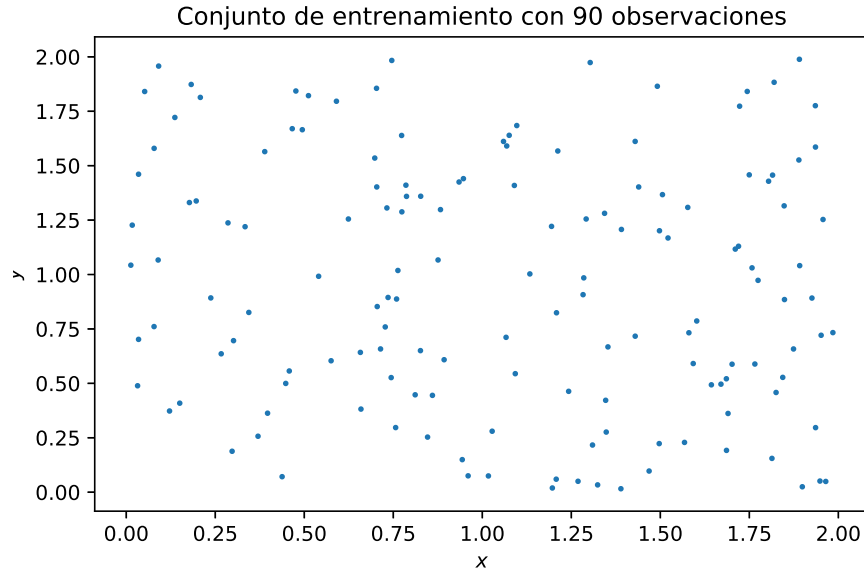


Figure 1: Ejemplo de uno de los conjuntos de entrenamiento

Para la misma cantidad de observaciones, los errores se presentan en la Figura 2. Y el rectángulo h_s que minimiza L_S se muestra en la Figura 3.

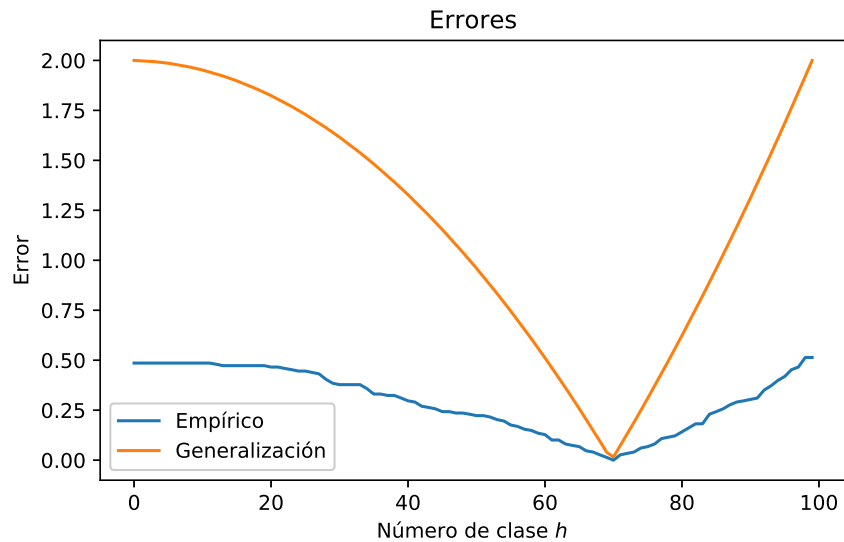


Figure 2: Errores para el dataset de la Figura 1



Figure 3: Clase que minimiza el error de entrenamiento

5. Para cada valor de m , se forman 1000 conjuntos de entrenamiento y se grafica $L_{D,f}$. Para $m = 90$, el histograma de errores se aprecia en la Figura 4.
6. Seleccionando $\epsilon = 0.01$ y $\delta = 0.05$, el número de muestras de tamaño 90 con un error menor a ϵ es 470. Note que el tamaño de las muestras debe ser de, por lo menos, 465, para que con un 95% de confianza, el error sea menor a 0.05.

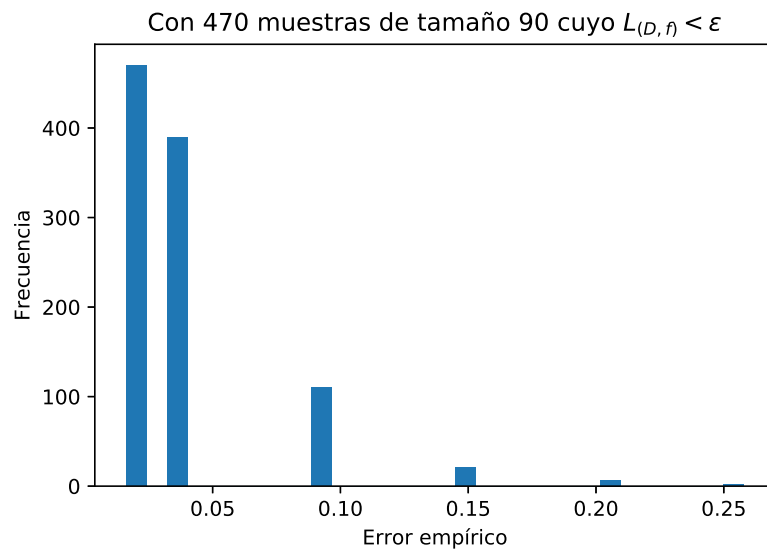


Figure 4: Histograma de errores empíricos