02. Modelo PAC generalizado y predictores lineales

1 Problemas

Problema (3.4)**.** Cada hipótesis es determinada por la variable x_i, independientemente de si x_i, $\overline{x}_i$ o ninguno aparece en la conjunción correspondiente. Como $|H| = 3^d + 1$, H es PAC aprendible y

$$m_H(\epsilon, \delta) \leq \frac{d \ln 3 + \ln(1/\delta)}{\epsilon} \tag{1}$$

Considere a $h_0 = x_1 \wedge \overline{x}_1 \wedge \cdots \wedge x_d \wedge \overline{x}_d$ como la hipótesis cuyo valor de verdad tiene mayor posiblidades de ser falso. Sea $\{(c^i, y^i)\}_{i\in[m]}$ un conjunto de entrenamiento i.i.d. El algoritmo no toma en cuenta ejemplos que son falsos. Para cada ejemplo verdadero c_i, remueva de h_i los índices que c_i no posea. El algoritmo retorna h_m, en particular, h_i le asigna verdadero a cada $c_1, \ldots, c_i$. Similarmente, h_i clasifica correcctamente los elementos falsos c_i, por lo que h_m es un ERM. Este algoritmo procede en tiempo lineal por cada dimensión d, por lo que su complejidad está acotada por $\mathcal{O}(md)$.

Problema (3.7)**.** Sea $x \in X$ y p_x la probabilidad condicional de anotar positivamente a x. Por lo tanto,

$$\mathbb{P}[f_D(X) \neq y | X = x] = \min\{p_x,\ 1 - p_x\} \tag{2}$$

Sea el clasificador $g : X \to \{0, 1\}$, entonces

$$\begin{aligned}
\mathbb{P}[g(X) \neq Y | X = x] &= \mathbb{P}[g(X) = 0 | X = x]\mathbb{P}[Y = 1 | X = x] \\
&\quad + \mathbb{P}[g(X) = 1 | X = x]\mathbb{P}[Y = 0 | X = x] \\
&= \mathbb{P}[g(X) = 0 | X = x]p_x + \mathbb{P}[g(X) = 1 | X = x](1 - p_x) \\
&\geq \mathbb{P}[g(X) = 0 | X = x] \min\{p_x, 1 - p_x\} \\
&\quad + \mathbb{P}[g(X) = 1 | X = x] \min\{p_x, 1 - p_x\} \\
&= \min\{p_x, 1 - p_x\}
\end{aligned}$$

Por la ley de la expectancia total,

$$\begin{aligned}
L_D(f_D) &= \mathbb{E}_{(x,y)\sim D}[1_{[f_D(x)\neq y]}] \\
&= \mathbb{E}_{x\sim D_X}[p_x] \\
&\leq \mathbb{E}_{x\sim D_x}[\mathbb{E}_{y\sim D_Y|x}[1_{[g(x)\neq y]}|X = x]] \\
&= L_D(g)
\end{aligned}$$

Problema (4.1)**.** Suponga (1), para todo $\epsilon, \delta \in (0, 1)$, distribución D sobre $X \times \{0, 1\}$ y m. Sea $\epsilon' > 0$ y $\epsilon = \min\{\frac{1}{2}, \frac{\epsilon'}{2}\}$. Fije $m' = m_H(\epsilon, \epsilon)$. Entonces, para $m \geq m'$, dado que la pérdida

está acotada por 1,

$$\begin{aligned}\mathbb{E}_{S\sim D^m}[L_D(A(S))] &\leq 1 \\ &= \mathbb{P}_{S\sim D^m}[L_D(A(S)) > \frac{\epsilon'}{2}] \cdot 1 + \mathbb{P}_{S\sim D^m}[L_D(A(S)) \leq \frac{\epsilon'}{2}] \cdot \frac{\epsilon'}{2} \\ &\leq \mathbb{P}[L_D(A(S)) > \epsilon] + \frac{\epsilon'}{2} \\ &\leq \epsilon'\end{aligned}$$

Por el otro lado, suponga (2). Para $\epsilon, \delta \in (0,1)$, existe m' entero tal que para todo $m \geq m'$, $\mathbb{E}_{S\sim D^m}[L_D(A(S))] \leq \epsilon\delta$. Por aplicación directa de la desigualdad de Markov,

$$\mathbb{P}_{S\sim D^m}[L_D(A(S)) > \epsilon] \leq \frac{\mathbb{E}_{S\sim D^m}[L_D(A(S))]}{\epsilon} = \frac{\epsilon\delta}{\epsilon} = \delta \tag{3}$$

Problema (9.3)**.** Para $d = m$ y $x_i = e_i,\ i \in [m]$. Si $sign(0) = -1$, para $i \in [d]$, sea y_i la anotación de x_i. Si el algoritmo de perceptrón maneja a $w^{(t)}$ en cada iteración, note que

$$w_i = \sum_{j<i} e_j, \quad i \in [d] \tag{4}$$

Por lo tanto, $\langle w^{(i)}, x_i \rangle = 0$. Entonces, los $x_1, \ldots, x_d$ están mal clasificados. Por lo tanto, el vector $w^* = (1, \ldots, 1)$ satisface los requerimientos.

2 Trabajo numérico

1. Una vez implementado el algoritmo del perceptrón, se considera al predictor verdadero de la Figura 1

2. Se generan puntos aleatorios en $\mathbb{R}^2$ acotando una región y tomando en cuenta la capacidad de procesamiento del compilador. Para este problema se eligió $[-10^5, 10^5]^2$. Luego de producir 100 conjuntos de entrenamiento S_m para $m = 1, \ldots, 1000$, se calcula el promedio de $R = \max_i \|x_i\|$. En la Figura 2 se aprecia cómo R escala con el tamaño de la muestra con un comportamiento asintótico con la cota superior de las observaciones, 10^5.

3. Con un tiempo de ejecución de 9 horas, la cantidad de pasos necesarios para que el algoritmo del perceptrón encontrara w^* como función de m se observa en la Figura 3.

4. Para $m = 100$, la secuencia de actualizaciones de $w^{(t)}$ predictores se observa en el video producido y guardado como *actualizaciones.mp4*.

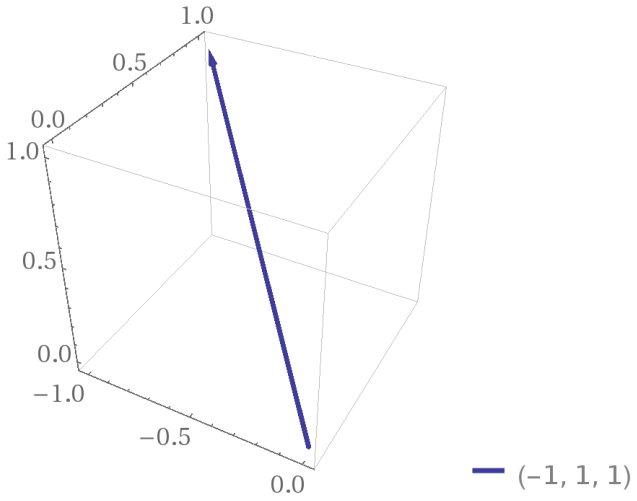

Figure 1: Predictor verdadero

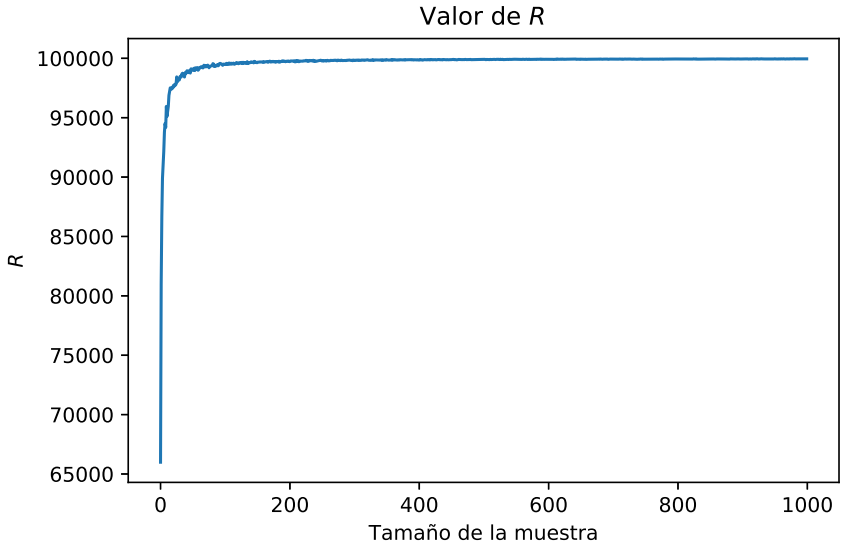

Figure 2: $R = \max_i \|x_i\|$

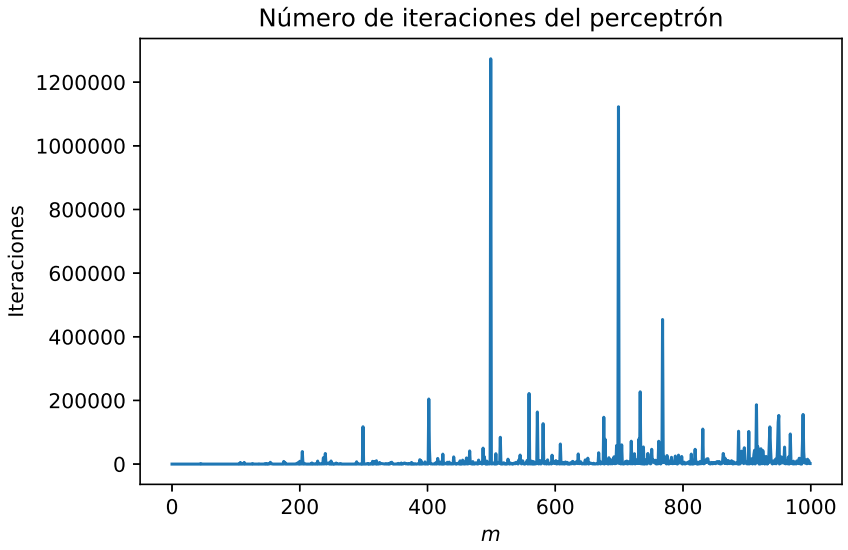

Figure 3: Convergencia del algoritmo del perceptrón