Nome: Luís Felipe de Melo Costa Silva
Número USP: 9297961

Lista de Exercícios 2 - MAC0460

Exercício 1

a) Vamos fazer $\mathbf{w^T}(t) \cdot \mathbf{x}(t) = y^*(t)$. Como $\mathbf{x}(t)$ está classificado incorretamente, temos que os sinais de $y(t)$ e $y^*(t)$ são diferentes, ou seja, ou $y(t) = 1$ e $y^*(t) = -1$ ou $y(t) = -1$ e $y^*(t) = 1$. Portanto, $y(t) \cdot y^*(t)$ é sempre -1 e então, $y(t) \cdot y^*(t) < 0$

b) Usando (1.3):

$$\begin{aligned}
y(t)\mathbf{w^T}(t+1)\mathbf{x}(t) > y(t)\mathbf{w^T}(t)\mathbf{x}(t) \\
y(t)[\mathbf{w^T}(t) + y(t)\mathbf{x}(t)]\mathbf{x}(t) > y(t)\mathbf{w^T}(t)\mathbf{x}(t) \\
y(t)[\mathbf{w^T}(t) + y(t)\mathbf{x}(t)]\mathbf{x}(t) > y(t)\mathbf{w^T}(t)\mathbf{x}(t) \\
y(t)\mathbf{w^T}(t)\mathbf{x}(t) + y(t)[y(t)\mathbf{x}(t)]\mathbf{x}(t) > y(t)\mathbf{w^T}(t)\mathbf{x}(t) \\
y(t)\mathbf{w^T}(t)\mathbf{x}(t) + y(t)[y(t)\mathbf{x}(t)]\mathbf{x}(t) > y(t)\mathbf{w^T}(t)\mathbf{x}(t) \\
y(t)^2\mathbf{x}(t)^2 > 0
\end{aligned}$$

$\square$

c) Quando temos uma classificação incorreta, temos que $y(t)\mathbf{w^T}(t)\mathbf{x}(t) < 0$. A próxima atualização nos trará um valor $y(t)\mathbf{w^T}(t+1)\mathbf{x}(t)$, que é maior do que o anterior, como provado no item acima. Isso nos deixa mais perto de um valor positivo, que é quando $\mathbf{x}(t)$ estará classificado corretamente.

Exercício 2

a) A hipótese $h(\mathbf{x})$ erra nos índices ímpares de $\mathbf{x}$, ou seja, em aproximadamente $\frac{M}{2}$ índices. Logo, $E_{\text{off}}(h, f) \cong \frac{1}{M} \cdot \frac{M}{2} \cong \frac{1}{2}$.

b) Para cada $\mathbf{x}_k, k \in \{1, ..., N+M\}, y$ pode assumir dois valores, -1 e $+1$, portanto, teremos 2^{N+M} funções distintas. Fixando um N, para garantir um conjunto $\mathcal{D}$ sem ruídos, teremos 2^M funções distintas.

c) O valor $E_{\text{off}} = \frac{k}{M}$ mostra que temos k classificados incorretamente entre M. O número de funções em que isso acontece é $\binom{M}{k}$.

d) Como o número de funções em que $E_{\text{off}}(h, f)$ é $\binom{M}{k}$, temos que:

$$\mathbb{E}_f[E_{\text{off}}(h,f)] = \frac{1}{2^M}\sum_{k=0}^{M}\binom{M}{k}\cdot k = \frac{2^{M-1}M}{2^M} = \frac{M}{2}$$

e) No item anterior calculamos a esperança de E_{off} em relação a f. O que muda é a hipótese $A(\mathcal{D})$, e ela não influi no valor da esperança. Logo, a igualdade é válida.

Exercício 3

a) Vamos minimizar $E_{\text{in}}(h)$ com a derivada (achando o mínimo global).

$$\begin{gathered} E'_{\text{in}}(h) = \sum_{n=1}^{N} 2(h-y_n) = 0 \\ 2Nh - \sum_{n=1}^{N} 2y_n = 0 \\ Nh = \sum_{n=1}^{N} y_n \\ h = \frac{1}{N}\sum_{n=1}^{N} y_n \end{gathered}$$

b) A função módulo não é diferenciável. Por isso, para achar o ponto de mínimo, vamos quebrar nossa função em duas aqui. Para tal, seja um valor k tal que $y_k \leq h \leq y+k+1$. Logo:

$$E_{\text{in}}(h) = \sum_{n=1}^{k}(h-y_n) + \sum_{n=k+1}^{N}(y_n-h)$$

E então:

$$\begin{gathered} E'_{\text{in}}(h) = \sum_{n=1}^{k} 1 + \sum_{n=k+1}^{N} -1 = 0 \\ \sum_{n=1}^{k} 1 = \sum_{n=k+1}^{N} 1 \\ k = N-k \\ k = \frac{N}{2} \end{gathered}$$

c) O E_{in} do primeiro item será bastante afetado, pois ele calcula a média, que tenderá a infinito, junto com ϵ. Já o E_{in} do segundo item não será afetado. Isso acontece pois ele utiliza a mediana, e ela é pouco ou nada afetada pelos *outliers*.

Exercício 4

Queremos que $\epsilon(M,N,\delta) = \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}} \leq 0.05$. Teremos:

$$\sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}} \leq 0.05$$
$$\frac{1}{2N}\ln\frac{2M}{\delta} \leq 0.05^2$$
$$\frac{1}{2N} \leq \frac{0.05^2}{\ln\frac{2M}{\delta}}$$
$$2N \geq \frac{\ln\frac{2M}{\delta}}{0.05^2}$$
$$N \geq \frac{\ln\frac{2M}{\delta}}{2\cdot 0.05^2}$$

Como $\delta = 0.03$:

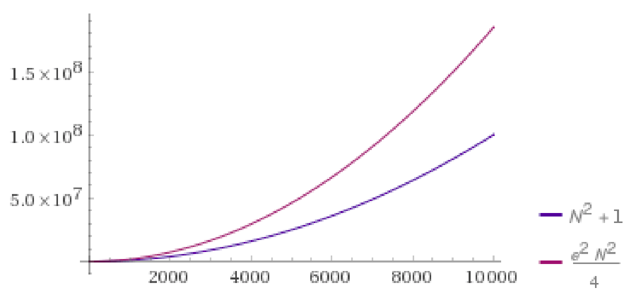
a) Para $M = 1$, $N \geq \frac{\ln\frac{2}{0.03}}{2\cdot 0.05^2} \cong 840$

b) Para $M = 100$, $N \geq \frac{\ln\frac{200}{0.03}}{2\cdot 0.05^2} \cong 1761$

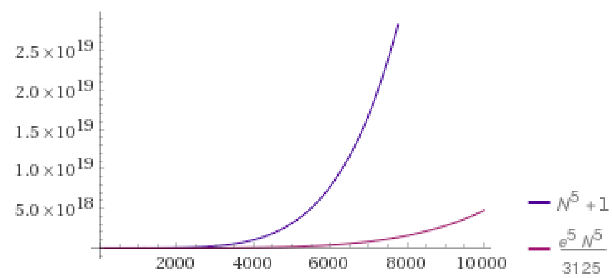
c) Para $M = 10000$, $N \geq \frac{\ln\frac{20000}{0.03}}{2\cdot 0.05^2} \cong 2683$

Exercício 5

Do problema 2.5, temos que $m_{\mathcal{H}}(N) \leq N^{d_{vc}} + 1$. Do problema 2.6, temos que $m_{\mathcal{H}}(N) \leq \left(\frac{eN}{d_{vc}}\right)^{d_{vc}}$. Portanto, teremos os seguintes gráficos:



(a) $d_{vc} = 2$



(b) $d_{vc} = 5$

Como queremos o limiar mais restritivo, quando $d_{vc} = 2$, o limite do Problema 2.5 é mais interessante. Analogamente, para $d_{vc} = 5$, o melhor é o do Problema 2.6.

Exercício 6

Usando a fórmula (2.13) do livro, teremos:

$$
\begin{aligned}
N &\geq \frac{8}{\epsilon^2} \cdot \ln\left(\frac{4((2N)^{d_{vc}}+1)}{\delta}\right) \\
&\geq \frac{8}{(0.05)^2} \cdot \ln\left(\frac{4((2N)^{10}+1)}{0.05}\right) \\
&\geq \frac{8}{0.025} \cdot \ln\left(\frac{4((2N)^{10}+1)}{0.05}\right) \\
&\geq 3200 \cdot \ln(80((2N)^{10}+1)) \\
&\geq 3200 \cdot \ln(80((2N)^{10}+1)) \\
&\geq 3200 \cdot (\ln 80 + \ln((2N)^{10}+1)) \\
&\geq 14000 + 3200 \cdot \ln((2N)^{10}+1)) \\
&\geq 14000 + 3200 \cdot \ln((2N)^{10}) \\
&\geq 14000 + 32000 \cdot \ln(2N) \\
&\geq 14000 + 32000 \cdot (\ln(2) + \ln(N)) \\
&\geq 36000 + 32000 \cdot \ln(N)
\end{aligned}
$$

Fazendo algumas estimativas, chegamos ao valor $N \geq 453000$.

Exercício 7

Temos que $y(x) = f(x) + \epsilon$, onde ϵ é uma v.a. com $\mathbb{E}(\epsilon) = 0$ e $\text{var}(\epsilon) = \mathbb{E}[\epsilon^2] = \sigma^2$. Para auxiliar nos cálculos, vamos definir uma hipótese média $\bar{g}(x) = \mathbb{E}_{\mathcal{D}}[g^{\mathcal{D}}(x)]$.

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})] &= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{X,Y}[(g^{(\mathcal{D})}(x) - y(x))^2]] \\
&= \mathbb{E}_{X,Y}[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(x) - y(x))^2]] \\
&= \mathbb{E}_{X,Y}[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(x) - \bar{g}(x) + \bar{g}(x) - y(x))^2]] \\
&= \mathbb{E}_{X,Y}[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(x) - \bar{g}(x))^2 + (\bar{g}(x) - y(x))^2 + 2 \cdot \overbrace{(g^{(\mathcal{D})}(x) - \bar{g}(x))}^{0} \cdot (\bar{g}(x) - y(x))]] \\
&= \mathbb{E}_{X,Y}[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(x) - \bar{g}(x))^2 + (\bar{g}(x) - y(x))^2]] \\
&= \mathbb{E}_{X,Y}[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(x) - \bar{g}(x))^2] + \mathbb{E}_{\mathcal{D}}[(\bar{g}(x) - y(x))^2]] \\
&= \mathbb{E}_{X,Y}[\mathbb{E}_{\mathcal{D}}[\text{var}(x) + (\bar{g}(x) - y(x))^2]] \\
&= \text{var} + \mathbb{E}_{X,Y}[(\bar{g}(x) - y(x))^2] \\
&= \text{var} + \mathbb{E}_{X,Y}[(\bar{g}(x) - f(x) - \epsilon)^2] \\
&= \text{var} + \mathbb{E}_{X,Y}[(\bar{g}(x))^2 + (f(x))^2 + \epsilon^2 - 2 \cdot \bar{g}(x) \cdot f(x) - 2 \cdot \bar{g}(x) \cdot \epsilon + 2 \cdot f(x) \cdot \epsilon] \\
&= \text{var} + \mathbb{E}_{X,Y}[(\bar{g}(x) - f(x))^2 + \epsilon^2 - 2 \cdot \bar{g}(x) \cdot \epsilon + 2 \cdot f(x) \cdot \epsilon] \\
&= \text{var} + \mathbb{E}_{X,Y}[\text{bias}(x)] + \mathbb{E}_{X,Y}[\epsilon^2] - 2 \cdot \mathbb{E}_{X,Y}[\bar{g}(x) \cdot \epsilon] + 2 \cdot \mathbb{E}_{X,Y}[f(x) \cdot \epsilon] \\
&\overset{ind.}{=} \text{var} + \text{bias} + \mathbb{E}[\epsilon^2] - 2 \cdot \mathbb{E}_X[\bar{g}(x)] \cdot \overbrace{\mathbb{E}[\epsilon]}^{0} + 2 \cdot \mathbb{E}_X[f(x)] \cdot \overbrace{\mathbb{E}[\epsilon]}^{0} = \sigma^2 + \text{bias} + \text{var}
\end{aligned}
$$

□

Exercício 8

a) Do enunciado, temos que o conjunto $\mathcal{H}$ possui funções no estilo $ax + b$. Logo, usando a definição de $\bar{g}(x)$, e definindo como K o número total de *data sets*:

$$\bar{g}(x) = \mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(x)] = \mathbb{E}_{\mathcal{D}}[a^{(\mathcal{D})} \cdot x + b^{(\mathcal{D})}] = \mathbb{E}_{\mathcal{D}}[a^{(\mathcal{D})} \cdot x] + \mathbb{E}_{\mathcal{D}}[b^{(\mathcal{D})}]$$
$$= \frac{1}{K}\sum_{k=1}^{K} a_k \cdot x + \frac{1}{K}\sum_{k=1}^{K} b_k = \frac{1}{K}\sum_{k=1}^{K} a_k \cdot x + b_k$$

b) Fixando um número K de iterações, poderíamos fazer, a cada iteração:

- Gerar dois valores, x_1 e x_2, cada um retirado de uma distribuição Uniforme$(-1, 1)$. Com isso, geramos dois pontos, usando a $f(x)$, e temos nosso *data set*.
- Rodar o algoritmo de aprendizado, que resultará em uma $g(x)$.

Teremos K *data sets* e K funções $g(x)$. Com isso, podemos calcular:

- $\bar{g}(x)$, com a média das $g(x)$.
- O *bias*, com o valor $\mathbb{E}_X[(\bar{g}(x) - f(x))^2]$.
- A var, fazendo $\mathbb{E}_X[\mathbb{E}_{(\mathcal{D})}(g^{\mathcal{D}}(x) - \bar{g}(x))^2]]$
- Por último, E_{out}, com a fórmula $\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})] = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_X[(g^{(\mathcal{D})}(x) - f(x))^2]]$

Exercício 10

a) Vamos chamar de $h(x)$ a saída do nosso algoritmo e C a variável que representa o custo. Teremos:

- cost(accept) = $\mathbb{E}(C|h(x) = 1) = g(x) \cdot 0 + (1 - g(x)) \cdot c_a = (1 - g(x)) \cdot c_a$
- cost(reject) = $\mathbb{E}(C|h(x) = -1) = g(x) \cdot c_r + (1 - g(x)) \cdot 0 = g(x) \cdot c_r$

b) A regressão logística gera uma saída num intervalo contínuo $[0, 1]$. Quando usamos esse resultado para uma classificação binária, precisamos de um limiar κ. O custo de aceite é igual ao custo de rejeite nesse ponto, já que o intervalo é contínuo. Então, fazendo $g(x) = \kappa$, teremos:

cost(reject) = cost(accept)
$$(1 - g(x)) \cdot c_a = g(x) \cdot c_r$$
$$(1 - \kappa) \cdot c_a = \kappa \cdot c_r$$
$$c_a - \kappa \cdot c_a = \kappa \cdot c_r$$
$$\kappa = \tfrac{c_a}{c_a + c_r}$$

$\square$

c) Para o exemplo do supermercado, temos que $c_a = 1$ e $c_r = 10$. Portanto, teremos $\kappa = \frac{1}{11}$, que indica um bom intervalo de aceite, já que é o falso negativo que incomoda. Já para a CIA, com $c_a = 1000$ e $c_r = 1$, $\kappa = \frac{1000}{1001}$, e nos dá um intervalo bem restritivo para o aceite, de modo a evitar ao máximo falsos positivos.