MAE 0580/ MAC 6926 - Lista 5

31 de outubro de 2017

Seja $\mathcal{X} = \mathbb{R}^d$ o espaço de *estímulos* ou *entradas* e $\mathcal{Y} = \mathbb{R}$ o espaço de *respostas*. Um elemento $x \in \mathcal{X}$ é um vetor $x = (x(1), \ldots, x(d))$.

Sejam X e Y respectivamente um vetor e uma variável aleatórias definidos sobre o mesmo espaço de probabilidades e assumindo valores respectivamente em $\mathcal{X}$ e $\mathcal{Y}$. No que segue vamos sempre supor que a variável Y e as componentes do vetor $\mathcal{X}$ sejam de quadrado integrável, isto é, $\mathbb{E}[Y^2] < +\infty$ e $\mathbb{E}[X_i^2] < +\infty$, para todo $i = 1, \ldots, d$.

Dada uma função $f : \mathcal{X} \to \mathcal{Y}$, definimos o *risco* associado a f (também chamado de *perda quadrática média* como

$$R(f) = \mathbb{E}\{[f(X) - Y]^2\}$$

Dada uma amostra $(X_1, Y_1), \ldots, (X_n, Y_n)$ de pares de v.a. independentes e identicamente distribuidos, tendo a mesma distribuição do par (X, Y), e uma função $f : \mathcal{X} \to \mathcal{Y}$ definimos o *risco empírico* $R_n(f)$ como

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} [f(X_i) - Y_i]^2$$

Queremos modelar a relação entre as variáveis Xe Y através da classe $\mathcal{L} = \mathcal{L}(\mathcal{X}, \mathcal{Y})$ assim definida

$$\mathcal{L} = \{f : \mathcal{X} \to \mathcal{Y} : f(x) = w_0 + \sum_{i=1}^{d} w_i x_i, \text{ onde } w = (w_1, \ldots, w_d) \in \mathbb{R}^d\}$$

A isso chama-se *regressão linear*.

1. No caso $d = 1$, isto é $\mathcal{X} = \mathbb{R}$, calcule os valores $\tilde{w_0}$ e $\tilde{w_1}$, tais que a função $\tilde{f}(x) = \tilde{w_0} + \tilde{w_1} x$ minimiza o risco $R(f)$ na classe $\mathcal{F}$.

   Mais explicitamente, verifique que

   $$\tilde{w_0} = \mathbb{E}(Y) - \tilde{w_1} \mathbb{E}(X)$$

   e

   $$\tilde{w_1} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X^2}$$

   onde $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$ e $\sigma_X^2 = \mathrm{Cov}(X, X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$.

2. Calcule $\tilde{f}$ quando $Y = a + bX + \xi$, sendo ξ uma variável aleatória de quadrado integrável independente de X e com $\mathbb{E}(\xi) = 0$.

3. Seja $f : \mathcal{X} \to \mathcal{Y}$ e definimos $Y = f(X)$ e $\tilde{Y} = f(X) + \xi$ onde ξ é uma variável aleatória assumindo valores em $\mathbb{R}$, independente de X, com distribuição normal e variância 1, isto é

   $$\mathbb{P}(\xi \leq t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{r^2}{2}} dr$$

   Mostre que $\mathbb{E}\{[f(X) - Y]^2\} = \mathbb{E}\{[f(X) - \tilde{Y}]^2\}$.

4. Seja $(X_1, Y_1), \ldots, (X_n, Y_n)$ variáveis aleatórias independentes e identicamente distribuídas. No caso $d = 1$ (isto é $\mathcal{X} = \mathcal{Y} = \mathbb{R}$), seja $f(X) = a + bX$ uma função genérica em $\mathcal{L}$ com $a, b \in \mathbb{R}$. Calcule os valores $\hat{a}$ e $\hat{b}$ definindo $\hat{f}(X) = \hat{a} + \hat{b}X$, tal que $\hat{f} = \arg\min\{R_n(f) : f \in \mathcal{L}\}$.

5. Dadas as amostras

$$
\begin{aligned}
x_1 &= 10, y_1 = 21,62 \\
x_2 &= 6, y_2 = 13,96 \\
x_3 &= 1, y_3 = 3,83 \\
x_4 &= 2, y_4 = 6,29 \\
x_5 &= 4, y_5 = 10,02 \\
x_6 &= 2, y_6 = 6,10 \\
x_7 &= 3, y_7 = 8,15 \\
x_8 &= 7, y_8 = 16,24 \\
x_9 &= 5, y_9 = 11,58 \\
x_{10} &= 3, y_{10} = 8,41
\end{aligned}
$$

Encontre a função $\hat{f}$ pertencente a classe $\mathcal{L}$ que minimiza R_n no caso $d = 1$ (isto é $\mathcal{X} = \mathcal{Y} = \mathbb{R}$). Desenhe em um gráfico os valores de y_i e $\hat{f}(x_i)$.