

**Universidade Federal do Rio de Janeiro
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia**



Departamento de Engenharia Elétrica
COE782 - Introdução ao Aprendizado de Máquina
Prof. Dr. Markus Vinícius Santos Lima

Lista 1 de exercícios

Luiz Henrique Souza Caldas
email: lhscaldas@cos.ufrj.br

16 de maio de 2024

Exercícios do Bishop

1. Exercício 1.1

Considere a função de erro da soma dos quadrados dada por $E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, w) - t_n\}^2$, na qual a função $y(x, \mathbf{w})$ é dada pelo polinômio $y(x, \mathbf{w}) = \sum_{j=0}^M w_j x^j$. Mostre que os coeficientes $\mathbf{w} = \{w_i\}$ que minimizam essa função de erro são dados pela solução do seguinte conjunto de equações lineares:

$$\sum_{j=0}^M A_{ij} w_j = T_i,$$

onde

$$A_{ij} = \sum_{n=1}^N (x_n)^{i+j} \quad \text{e} \quad T_i = \sum_{n=1}^N (x_n)^i t_n.$$

Aqui os sufixos i e j denotam os índices de um componente, onde $(x)^i$ denota x elevado a i -ésima potência.

Solução:

$$\frac{\partial E(\mathbf{w})}{\partial w_i} = \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\} \frac{\partial y(x_n, \mathbf{w})}{\partial w_i}$$

$$\text{Se } y(x, \mathbf{w}) = \sum_{j=0}^M w_j x^j,$$

$$\frac{\partial y(x_n, \mathbf{w})}{\partial w_i} = x_n^i$$

$$\frac{\partial E(\mathbf{w})}{\partial w_i} = \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\} x_n^i = 0$$

$$\sum_{n=1}^N y(x_n, \mathbf{w}) x_n^i = \sum_{n=1}^N t_n x_n^i$$

$$\sum_{n=1}^N \left(\sum_{j=0}^M w_j x_n^j \right) x_n^i = \sum_{n=1}^N \underbrace{\left(\sum_{j=0}^M x_n^{j+i} \right)}_{A_{ij}} w_j = \underbrace{\sum_{n=1}^N t_n x_n^i}_{T_i}$$

$$\sum_{j=0}^M A_{ij} w_j = T_i$$

2. Exercício 1.2

Escreva o conjunto de equações lineares acopladas, análogo a $\sum_{j=0}^M A_{ij} w_j = T_i$, satisfeitas pelos coeficientes w_i que minimizam a função de erro da soma dos quadrados regularizada dada por $\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, w) - t_n\}^2 - \frac{\lambda}{2} \|\mathbf{w}\|^2$.

Solução:

$$\frac{\partial \tilde{E}(\mathbf{w})}{\partial w_i} = \frac{1}{2} \sum_{n=1}^N \left\{ \sum_{j=0}^M w_j x_n^j - t_n \right\} x_n^i - \lambda w_i = 0$$

$$\sum_{j=0}^M A_{ij} w_j + \lambda w_i = T_i$$

$$\sum_{j=0}^M A_{ij} w_j + \sum_{j=0}^M \delta_{ij} \lambda w_j = T_i \quad \text{onde} \quad \delta_{ij} = \begin{cases} 0 & , \text{ se } i \neq j \\ 1 & , \text{ se } i = j \end{cases}$$

$$\sum_{j=0}^M (A_{ij} + \delta_{ij}\lambda)w_j = T_i$$

3. Exercício 1.5

Usando a definição $var[f(x)] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$ mostre que $var[f(x)]$ satisfaz $var[f(x)] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$.

Solução:

$$\begin{aligned}\mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] &= \mathbb{E}[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\ \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[\mathbb{E}[f(x)]^2] \\ \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]^2 + \mathbb{E}[f(x)]^2 \\ \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2\end{aligned}$$

4. Exercício 1.6

Mostre que se duas variáveis x e y são independentes, então a covariância entre elas é zero.

Solução:

$$\begin{aligned}cov[x, y] &= \mathbb{E}[x, y] - \mathbb{E}[x]\mathbb{E}[y] \\ \mathbb{E}[x, y] &= \sum_x \sum_y p(x, y)xy \\ \text{Se } x \text{ e } y \text{ são independentes, } p(x, y) &= p(x)p(y), \text{ então temos} \\ \mathbb{E}[x, y] &= \sum_x \sum_y p(x)p(y)xy = \sum_x p(x)x \sum_y p(y)y = \mathbb{E}[x]\mathbb{E}[y] \\ cov[x, y] &= \mathbb{E}[x, y] - \mathbb{E}[x]\mathbb{E}[y] = \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[y] \\ cov[x, y] &= 0\end{aligned}$$

5. Exercício 1.7

Neste exercício, provamos a condição de normalização $\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2)dx = 1$ para a distribuição gaussiana univariável. Para fazer isso, considere a integral

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right) dx$$

que podemos avaliar primeiro escrevendo o seu quadrado na forma

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\right) dx dy.$$

Agora faça a transformação de coordenadas cartesianas (x, y) para coordenadas polares (r, θ) e então substitua $u = r^2$. Mostre que, ao realizar as integrais em relação a θ e u , e em seguida, tirar a raiz quadrada de ambos os lados, obtemos

$$I = (2\pi\sigma^2)^{1/2}.$$

Finalmente, use esse resultado para mostrar que a distribuição Gaussiana $\mathcal{N}(x|\mu, \sigma^2)$ é normalizada.

Solução:

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\right) dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}(x^2 + y^2)\right] dx dy$$

Fazendo a transformação de coordenadas cartesianas para polares temos

$$r^2 = x^2 + y^2 \text{ e } dx dy = r dr d\theta$$

Substituindo

$$I^2 = \int_0^{2\pi} \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) r dr d\theta = 2\pi \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}r^2\right) r dr$$

Fazendo $u = r^2$, temos $du = 2r dr$

$$I^2 = \pi \int_0^{\infty} \exp\left(-\frac{1}{2\sigma^2}u\right) du = \pi \left[-2\sigma^2 \exp\left(-\frac{u}{2\sigma^2}\right)\right]_0^{\infty} = 2\pi\sigma^2$$

$$\underline{I = (2\pi\sigma^2)^{1/2} \quad |}$$

Integrando a distribuição Gaussiana temos

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} dx$$

Fazendo $u = x - \mu$ e, conseqüentemente, $du = dx$ temos

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = \frac{1}{(2\pi\sigma^2)^{1/2}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}u^2\right\} du = \frac{I}{(2\pi\sigma^2)^{1/2}} = \frac{(2\pi\sigma^2)^{1/2}}{(2\pi\sigma^2)^{1/2}}$$

$$\underline{\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1 \quad |}$$

6. Exercício 1.8

Usando uma mudança de variáveis, verifique que a distribuição gaussiana univariável dada por $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$ satisfaz $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2)x dx = \mu$. Em seguida, diferenciando ambos os lados da condição de normalização

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1$$

em relação a σ^2 verifique que a Gaussiana satisfaz $\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2)x^2 dx = \mu^2 + \sigma^2$. Finalmente, mostre que $\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$ é verdadeira.

Solução:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \frac{e^{\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}}}{(2\pi\sigma^2)^{1/2}} x dx$$

$$y = x - \mu \rightarrow x = y + \mu \text{ e } dx = dy$$

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \frac{e^{\{-\frac{1}{2\sigma^2}y^2\}}}{(2\pi\sigma^2)^{1/2}} (y + \mu) dy$$

Porém $ye^{-\frac{y^2}{2\sigma^2}}$ é ímpar, então a sua integral no intervalo $(-\infty, \infty)$ é nula

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2\sigma^2}}}{(2\pi\sigma^2)^{1/2}} \cdot \mu dy = \frac{\mu}{(2\pi\sigma^2)^{1/2}} \cdot I \quad (\text{do exercício anterior})$$

$$\mathbb{E}[x] = \frac{\mu}{(2\pi\sigma^2)^{1/2}} \cdot (2\pi\sigma^2)^{\frac{1}{2}}$$

$$\underline{\mathbb{E}[x] = \mu}$$

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} dx = 1$$

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} dx = (2\pi\sigma^2)^{1/2}$$

$$\frac{d}{d\sigma^2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} dx = \frac{d}{d\sigma^2} (2\pi\sigma^2)^{1/2}$$

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \left(\frac{-(x-\mu)^2}{2} \cdot (-1) \cdot \frac{1}{\sigma^4}\right) dx = \sqrt{2\pi} \frac{1}{2} (\sigma^2)^{-1/2}$$

$$\frac{1}{(2\pi\sigma^2)^{1/2}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} (x - \mu)^2 dx = \sigma^2$$

$$\mathbb{E}[(x - \mu)^2] = \sigma^2$$

Porém,

$$\mathbb{E}[(x - \mu)^2] = \mathbb{E}[x^2 - 2x\mu + \mu^2] = \mathbb{E}[x^2] + \mathbb{E}[-2x\mu] + \mathbb{E}[\mu^2]$$

$$\mathbb{E}[(x - \mu)^2] = \mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2$$

$$\mathbb{E}[(x - \mu)^2] = \mathbb{E}[x^2] - 2\mu^2 + \mu^2 = \mathbb{E}[x^2] - \mu^2$$

$$\mathbb{E}[x^2] = \mu^2 + \underbrace{\mathbb{E}[(x - \mu)^2]}_{\sigma^2}$$

$$\underline{\mathbb{E}[x^2] = \mu^2 + \sigma^2}$$

$$var[x] = E[x^2] - E[x]^2 = \mu^2 + \sigma^2 - \mu^2$$

$$\underline{var[x] = \sigma^2}$$

7. Exercício 1.9

Mostre que a moda (ou seja, o máximo) da distribuição Gaussiana

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$

é dado por μ . Da mesma forma, mostre que o modo da distribuição gaussiana multivariada

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^2) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$$

é dado por $\boldsymbol{\mu}$.

Solução:

- Para a univariada temos:

$$\frac{d}{dx} \mathcal{N}(x|\mu, \sigma^2) = \frac{d}{dx} \left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \right) = 0$$

$$\frac{d}{dx} \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \left[\frac{-2(x-\mu)}{2\sigma^2} \right] \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} = \frac{-(x-\mu)}{\sigma^2} \mathcal{N}(x|\mu, \sigma^2) = 0$$

Como $\mathcal{N}(x|\mu, \sigma^2) = 0$ só ocorre para $x = \pm\infty$, o ponto de máximo estará em x tal que $\frac{-(x-\mu)}{\sigma^2} = 0$

$$\underline{x = \mu}$$

- De forma análoga, para a multivariada temos:

$$\frac{d}{d\mathbf{x}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^2) = \frac{d}{d\mathbf{x}} \left(\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\} \right) = \mathbf{0}$$

$$\frac{d}{d\mathbf{x}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^2) = -\frac{1}{2} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^2) \nabla_{\mathbf{x}} \{ (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \} = \mathbf{0}$$

$$\frac{d}{d\mathbf{x}} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^2) = -\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^2) \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{0}$$

$$\mathbf{x} - \boldsymbol{\mu} = \mathbf{0}$$

$$\underline{\mathbf{x} = \boldsymbol{\mu}}$$

8. Exercício 1.10

Suponha que as variáveis x e z são estatisticamente independentes. Mostre que a média e a variância da soma delas satisfazem

$$\mathbb{E}[x + z] = \mathbb{E}[x] + \mathbb{E}[z]$$

$$\text{var}[x + z] = \text{var}[x] + \text{var}[z].$$

Solução:

$$\mathbb{E}[x + z] = \int \int p(x, z)(x + z) dx dz$$

Se x e z são independentes, $p(x, z) = p(x)p(z)$

$$\mathbb{E}[x + z] = \int \int p(x)p(z)(x + z) dx dz$$

$$\mathbb{E}[x + z] = \int \int p(x)p(z)x dx dz + \int \int p(x)p(z)z dx dz$$

$$\mathbb{E}[x + z] = \underbrace{\int p(z) dz}_1 \int p(x)x dx + \underbrace{\int p(x) dx}_1 \int p(z)z dz$$

$$\mathbb{E}[x + z] = \int p(x)x dx + \int p(z)z dz$$

$$\mathbb{E}[x + z] = \mathbb{E}[x] + \mathbb{E}[z]$$

$$\text{var}[x + z] = \mathbb{E}[(x + z - \mathbb{E}[x + z])^2] = \mathbb{E}[(x - \mathbb{E}[x] + z - \mathbb{E}[z])^2]$$

$$\text{var}[x + z] = \underbrace{\mathbb{E}[(x - \mathbb{E}[x])^2]}_{\text{var}[x]} + \underbrace{\mathbb{E}[(z - \mathbb{E}[z])^2]}_{\text{var}[z]} + \underbrace{\mathbb{E}[2(z - \mathbb{E}[z])(x - \mathbb{E}[x])]}_{-2A}$$

$$A = \int \int (x - \mathbb{E}[x])(z - \mathbb{E}[z])p(x)p(z) dx dz$$

$$A = \int \int xzp(x)p(z) dx dz + \int \int \mathbb{E}[x]\mathbb{E}[z]p(x)p(z) dx dz - \int \int \mathbb{E}[x]zp(x)p(z) dx dz - \int \int \mathbb{E}[z]xp(x)p(z) dx dz$$

$$A = \mathbb{E}[x]\mathbb{E}[z] + \mathbb{E}[x]\mathbb{E}[z] - \mathbb{E}[x]\mathbb{E}[z] - \mathbb{E}[z]\mathbb{E}[x] = 0$$

$$\text{var}[x + z] = \text{var}[x] + \text{var}[z]$$

9. Exercício 1.11

Fazendo as derivadas da função de log-verossimilhança $\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)$ em relação a μ e σ^2 iguais a zero, verifique os resultados $\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n$ e $\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2$.

Solução:

$$\frac{d}{d\mu} \ln p(\mathbf{x}|\mu, \sigma^2) = \frac{d}{d\mu} \left(-\frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \right)$$

$$\frac{1}{2\sigma^2} (-2) \sum_{n=1}^N (x_n - \mu) = 0$$

$$\sum_{n=1}^N \mu = \sum_{n=1}^N x_n$$

$$N\mu = \sum_{n=1}^N x_n$$

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n$$

$$\frac{d}{d\sigma^2} \ln p(\mathbf{x}|\mu, \sigma^2) = \frac{d}{d\sigma^2} \left(-\frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \right)$$

$$-\frac{1}{2} \frac{(-1)}{\sigma^4} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \frac{1}{\sigma^2} = 0$$

$$\frac{N}{\sigma^2} = \frac{1}{\sigma^4} \sum_{n=1}^N (x_n - \mu)^2$$

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2$$

10. Exercício 1.13

Suponha que a variância de uma distribuição Gaussiana seja estimada usando o resultado $\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2$, mas com a estimativa de máxima verossimilhança μ_{ML} substituída pelo valor verdadeiro μ da média. Mostre que esse estimador tem a propriedade de que sua esperança é dada pela verdadeira variância σ^2 .

Solução:

$$\mathbb{E}[\sigma_{ML}^2] = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2\right] = \frac{1}{N} \mathbb{E}\left[\sum_{n=1}^N (x_n^2 - 2x_n\mu_{ML} + \mu_{ML}^2)\right]$$

$$\mathbb{E}[\sigma_{ML}^2] = \frac{1}{N} \left(\mathbb{E}\left[\sum_{n=1}^N x_n^2\right] - 2\mu \mathbb{E}\left[\sum_{n=1}^N x_n\right] + \mathbb{E}\left[\sum_{n=1}^N \mu^2\right] \right)$$

$$\mathbb{E}[\sigma_{ML}^2] = \frac{1}{N} (N(\mu^2 + \sigma^2) - 2\mu N\mu + N\mu^2)$$

$$\mathbb{E}[\sigma_{ML}^2] = \frac{1}{N} (N\mu^2 + N\sigma^2 - 2N\mu^2 + N\mu^2)$$

$$\underline{\mathbb{E}[\sigma_{ML}^2] = \sigma^2}$$

11. Exercício 1.21

Considere dois números não negativos a e b , e mostre que, se $a \leq b$, então $a \leq (ab)^{1/2}$. Use esse resultado para mostrar que, se as regiões de decisão de um problema de classificação de duas classes são escolhidas para minimizar a probabilidade de classificação errada, essa probabilidade satisfará

$$p(\text{erro}) \leq \int \{p(x, C_1)p(x, C_2)\}^{1/2} dx.$$

Solução:

$$a = b \rightarrow a^2 \leq ab \rightarrow \underline{a \leq (ab)^{1/2}}$$

$$p(\text{erro}) = p(x \in R_1, C_2) + p(x \in R_2, C_1)$$

$$p(\text{erro}) = \int_{R_1} p(x, C_2) dx + \int_{R_2} p(x, C_1) dx$$

Em R_1 temos $p(x, C_2) \leq p(x, C_1)$ e em R_2 temos $p(x, C_1) \leq p(x, C_2)$ o que resulta em

$$\begin{cases} \int_{R_1} p(x, C_2) dx \leq \int_{R_1} \{p(x, C_1)p(x, C_2)\}^{1/2} dx \\ \int_{R_2} p(x, C_1) dx \leq \int_{R_2} \{p(x, C_1)p(x, C_2)\}^{1/2} dx \end{cases}$$

Então

$$p(\text{erro}) \leq \int_{R_1} \{p(x, C_1)p(x, C_2)\}^{1/2} dx + \int_{R_2} \{p(x, C_1)p(x, C_2)\}^{1/2} dx$$

Como os integrandos são iguais e as regiões são complementares, podemos unir as duas integrais para todo o domínio

$$\underline{p(\text{erro}) \leq \int \{p(x, C_1)p(x, C_2)\}^{1/2} dx}$$

12. Exercício 1.22

Dada uma matriz de perda com elementos L_{kj} , o risco esperado é minimizado se, para cada x , escolhermos a classe que minimiza $\sum_k L_{kj}p(C_k|x)$. Verifique que, quando a matriz de perda é dada por $L_{kj} = 1 - I_{kj}$, onde I_{kj} são os elementos da matriz identidade, isso reduz ao critério

de escolher a classe que possui a maior probabilidade posterior. Qual é a interpretação dessa forma de matriz de perda?

Solução:

$$\mathbb{E}[L] = \sum_K L_{kj} p(C_k|x)$$

$$\text{Se } L_{kj} = 1 - I_{kj}$$

$$\mathbb{E}[L] = \sum_K (1 - I_{kj}) p(C_k|x)$$

$$\mathbb{E}[L] = \sum_K p(C_k|x) - \sum_K I_{kj} p(C_k|x)$$

$$\mathbb{E}[L] = 1 - \sum_K I_{kj} p(C_k|x)$$

$$\begin{cases} \text{se } k = j \rightarrow \mathbb{E}[L] = 1 - p(C_j|x) \\ \text{se } k \neq j \rightarrow \mathbb{E}[L] = 1 \end{cases}$$

Para minimizar $\mathbb{E}[L]$ devemos escolher a classe C_j com a maior probabilidade $p(C_j|x)$.

A matriz de perdas da forma $L_{kj} = 1 - I_{kj}$ atribui o mesmo peso, de valor 1, para todos os erros de classificação.

13. Exercício 1.23

Derive o critério para minimizar a perda esperada quando há uma matriz de perda geral e probabilidades a priori gerais para as classes.

Solução:

$$\mathbb{E}[L] = \sum_K L_{kj} p(C_k|x)$$

$$\text{Pelo Teorema de Bayes, } p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)}$$

$$\mathbb{E}[L] = \sum_K L_{kj} \frac{p(x|C_k)p(C_k)}{p(x)}$$

$$\underline{\mathbb{E}[L] \propto \sum_K L_{kj} p(x|C_k)p(C_k)} \quad |$$

14. Exercício 1.25

Considere a generalização da função de perda quadrática $\mathbb{E}[L] = \int \int \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt$ para uma única variável alvo t para o caso de múltiplas variáveis alvo descritas pelo vetor \mathbf{t} dado por

$$\mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x}))] = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}.$$

Usando o cálculo das variacional, mostre que a função $\mathbf{y}(\mathbf{x})$ para a qual essa perda esperada é minimizada é dada por $\mathbf{y}(\mathbf{x}) = \mathbb{E}_{\mathbf{t}}[\mathbf{t}|\mathbf{x}]$. Mostre que esse resultado se reduz a $y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t) dt}{p(\mathbf{x})} = \int t p(t|\mathbf{x}) dt = \mathbb{E}_t[t|\mathbf{x}]$ para o caso de uma única variável alvo t .

Solução:

$$\frac{\delta \mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x}))]}{\delta \mathbf{y}(\mathbf{x})} = \frac{\delta}{\delta \mathbf{y}(\mathbf{x})} \left(\int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \right) = 0$$

$$\frac{\delta \mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x}))]}{\delta \mathbf{y}(\mathbf{x})} = \int 2(\mathbf{y}(\mathbf{x}) - \mathbf{t}) p(\mathbf{x}, \mathbf{t}) d\mathbf{t} = 0$$

$$\mathbf{y}(\mathbf{x}) = \frac{\int \mathbf{t} p(\mathbf{t}, \mathbf{x}) d\mathbf{t}}{\int p(\mathbf{t}, \mathbf{x}) d\mathbf{t}}$$

$$\mathbf{y}(\mathbf{x}) = \int \frac{\mathbf{t} p(\mathbf{t}, \mathbf{x}) d\mathbf{t}}{p(\mathbf{x})}$$

$$\mathbf{y}(\mathbf{x}) = \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t}$$

$$\underline{\mathbf{y}(\mathbf{x}) = \mathbb{E}[\mathbf{t}|\mathbf{x}]}$$

Para o caso de um target escalar t , temos

$$\underline{y(x) = \int t p(t|x) dt = \mathbb{E}[t|x]}$$

15. Exercício 1.31

Considere duas variáveis x e y com distribuição conjunta $p(x, y)$. Mostre que a entropia diferencial desse par de variáveis satisfaz

$$H[x, y] \leq H[x] + H[y]$$

com igualdade se, e somente se, x e y forem estatisticamente independentes.

Solução:

$$H[x, y] = - \int \int p(x, y) \ln[p(x, y)] dx dy$$

Se x e y são independentes,

$$H[x, y] = - \int \int p(x) p(y) \ln[p(x) p(y)] dx dy$$

$$H[x, y] = - \int \int p(x) p(y) [\ln p(x) + \ln p(y)] dx dy$$

$$H[x, y] = - \int \int p(x) p(y) \ln p(x) dx dy - \int \int p(x) p(y) \ln p(y) dx dy$$

$$H[x, y] = - \int p(x) \ln p(x) dx - \int p(y) \ln p(y) dy$$

$$\underline{H[x, y] = H[x] + H[y]}$$

Se $H[x, y] = H[y|x] + H[x]$, no caso de x e y serem estatisticamente independentes teremos que $H[y|x] = H[y]$. Calculando a informação mútua temos

$$I[x, y] = H[y] - H[y|x] = 0$$

Porém a informação mútua é dada pela divergência de Kullback-Leibler

$$I[x, y] = - \int \int p(x, y) \ln \left[\frac{p(x)p(y)}{p(x, y)} \right] dydx$$

a qual só será nula se $p(x, y) = p(x)p(y)$, condição que só é satisfeita se x e y forem estatisticamente independentes.

Se elas não forem independentes, $I[x, y] \neq 0$. Nesse caso, fazendo $I[x, y] = C$, sendo C uma constante positiva, temos

$$H[y|x] = H[y] - I[x, y] = H[y] - C$$

$$H[x, y] = H[y|x] + H[x] = H[y] + H[x] - C < H[y] + H[x]$$

Então,

$$\underline{H[x, y] \leq H[x] + H[y]} \quad \text{como se queria demonstrar.}$$

16. Exercício 1.33

Suponha que a entropia condicional $H[y|x]$ entre duas variáveis aleatórias discretas x e y seja zero. Mostre que, para todos os valores de x para os quais $p(x) > 0$, a variável y deve ser uma função de x , em outras palavras, para cada x , há apenas um valor de y tal que $p(y|x) \neq 0$.

Solução:

$$H[x, y] = H[y|x] + H[x]$$

$$\text{Se } H[y|x] = 0$$

$$\underline{H[x, y] = H[x]}$$

Esse resultado significa que a informação necessária para descrever x e y é a mesma informação necessária para descrever apenas x , o que só pode ocorrer se y for totalmente dependente de x , ou seja, $y = f(x)$.

17. Exercício 1.37

Usando a definição $H[y|x] = - \int \int p(y, x) \ln p(y|x) dydx$ junto com a regra do produto da probabilidade, prove o resultado $H[x, y] = H[y|x] + H[x]$.

Solução:

$$H[y|x] = - \int \int p(x, y) \ln[p(y|x)] dydx$$

Aplicando a regra do produto

$$H[y|x] = - \int \int p(x, y) \ln \left[\frac{p(x, y)}{p(x)} \right] dy dx$$

$$H[y|x] = - \int \int p(x, y) \ln[p(x, y)] dy dx + \int \int p(x, y) \ln[p(x)] dy dx$$

$$H[y|x] = H[x, y] + \int p(x) \ln[p(x)] dx$$

$$H[y|x] = H[x, y] - H[x]$$

$$H[x, y] = H[y|x] - H[x]$$

18. **Exercício 1.39**

Considere que duas variáveis binárias x e y tendo a distribuição conjunta dada por

x	y	$p(x, y)$
0	0	1/3
0	1	1/3
1	0	0
1	1	1/3

Avalie as quantidades abaixo

$$(a) H[x] \quad (b) H[y] \quad (c) H[y|x] \quad (d) H[x|y] \quad (e) H[x, y] \quad (f) I[x, y].$$

Desenhe um diagrama para mostrar a relação entre essas diversas quantidades.

Solução:

19. **Exercício 1.41**

Usando as regras da soma e do produto de probabilidade, mostre que a informação mútua $I(x, y)$ satisfaz a relação $I[x, y] = H[x] - H[x|y] = H[y] - H[y|x]$.

Solução:

1 Exercícios extras

1.1 E1

Enunciado:

Considere a função custo/objetivo dada na equação $E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, w) - t_n\}^2$ em que y está definido na equação $y(x, \mathbf{w}) = \sum_{j=0}^M w_j x^j$. Mostre que $E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, w) - t_n\}^2$ pode ser escrito como

$$E(\mathbf{w}) = \frac{1}{2} \|\mathbf{y} - \mathbf{t}\|_2^2$$

onde

$$\mathbf{y} = [y(x_1, \mathbf{w}) \dots y(x_N, \mathbf{w})]^T$$

$$\mathbf{t} = [t_1 \dots t_N]^T$$

$$\mathbf{y} = \mathbf{A}\mathbf{w}$$

Determine as dimensões e os elementos (também chamados de entradas) da matriz \mathbf{A} . Mostre que o vetor de coeficientes que minimiza esta função objetivo pode ser escrito como

$$\mathbf{w}^* = \mathbf{A}^\dagger = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{t}.$$

Compare com o exercício 1.1 e note a vantagem de se usar Álgebra Linear para trabalhar com uma notação mais compacta.

Solução:

1.2 E2

Enunciado:

Mesma ideia de E1, porém agora considerando a função objetivo dada na equação $\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, w) - t_n\}^2 - \frac{\lambda}{2} \|\mathbf{w}\|^2$ do livro. Escrevê-la de forma matricial. Encontre o vetor de coeficientes ótimos (em fórmula fechada).

Solução:

1.3 E3 (Exercício Computacional)

Enunciado:

Replique o experimento computacional denominado “Polynomial Curve Fitting” usado diversas vezes no livro texto (veja páginas 4 e 5 do livro, bem como Apêndice A). Faça:

- Replique os resultados da Figura 1.4 e da Figura 1.6 para validar seu código (i.e., ter certeza de que ele está funcionando adequadamente);
- Simule uma base de dados que não tenha relevância estatística, isto é, que seja uma amostra que NÃO representa bem o todo (a população). Verifique alguns resultados experimentais para compreender a importância de ter uma amostra relevante. Explique qual a relação entre o caso simulado e casos práticos envolvendo vetores de dimensão elevada.

Dica: Para a simulação, ao invés de pegar dados igualmente espaçados no intervalo $[0,1]$, você pode forçar com que seus dados sejam amostrados apenas do semiciclo positivo (ou apenas do negativo) do modelo gerador.

- Simule uma base de dados em que 1 dos dados seja outlier. O que ocorre com a curva vermelha, estimativa da curva verde (modelo gerador), neste caso?

Dica: Para a simulação, você pode gerar seus dados de treinamento normalmente, igual feito no item (a), e ao final do processo escolher 1 desses dados pra atribuir um novo valor de target que seja completamente “maluco” (por exemplo, $\text{target} = 10$).

Solução:

Testando