UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
Programa de Engenharia Elétrica (COPPE/UFRJ)
COE 782 - Introdução ao aprendizado de máquina
Professor: Marcos Vinicius Santos Lima

Aluna: Vivian de Carvalho Rodrigues - DRE: 125228569
Programa de Engenharia Civil (COPPE/UFRJ)

22 de maio de 2025

Resumo

Relatório da resolução dos exercícios da Lista 1 (Capítulo 1 - Bishop [1])

1 Exercícios do livro texto

1.1 Exercício 1.1

Dada a função abaixo(eq. 1.2 de [1]):

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y\left(x_n, \mathbf{w}\right) - t_n\}^2 \tag{1}$$

onde (eq.1.1 [1])

$$y(x_n, \mathbf{w}) = w_0 + w_1 x_n + w_2 x_n^2 + \cdots + w_M x_n^M = \sum_{j=0}^{M} w_j x_n^j \tag{2}$$

Os coeficientes $\mathbf{w} = \{w_i\}$ que minimizam a função erro

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}\left[\frac{1}{2}\sum_{n=1}^{N}\{y\left(x_n, \mathbf{w}\right) - t_n\}^2\right] = \frac{1}{2}\sum_{n=1}^{N}\frac{\partial}{\partial \mathbf{w}}\{y\left(x_n, \mathbf{w}\right) - t_n\}^2 =$$

$$= \frac{1}{2}\sum_{n=1}^{N} 2\{y\left(x_n, \mathbf{w}\right) - t_n\}\frac{\partial}{\partial \mathbf{w}}\{y\left(x_n, \mathbf{w}\right) - t_n\} = \sum_{n=1}^{N}\{y\left(x_n, \mathbf{w}\right) - t_n\}\frac{\partial}{\partial \mathbf{w}}\sum_{j=0}^{M} w_j x_n^j =$$

$$= \sum_{n=1}^{N}\{\sum_{j=0}^{M} w_j x_n^j - t_n\}x_n^i$$

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \sum_{n=1}^{N}\{\sum_{j=0}^{M} w_j x_n^{i+j} - t_n x_n^i\} = 0$$

$$\sum_{j=0}^{M} w_j \sum_{n=1}^{N}(x_n)^{i+j} = \sum_{n=1}^{N} t_n (x_n)^i$$

Portanto,

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

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

1.2 Exercício 1.2

Dada a função abaixo(eq. 1.4 de [1]):

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda_r}{2}\|\mathbf{w}\|^2 \tag{3}$$

Onde

$$\|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w} = w_0^2 + w_1^2 + w_2^2 + \cdots + w_M^2 = \sum_{j=0}^{M} w_j^2 \tag{4}$$

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}\left[\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2\right] + \frac{\partial}{\partial \mathbf{w}}\left[\frac{\lambda_r}{2}\|\mathbf{w}\|^2\right]$$

Foi verificado no exercício anterior que, a derivada com relação aos coeficientes $\mathbf{w}$, da primeira parcela da expressão acima é dada por:

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

Resolvendo a derivada da segunda parcela

$$\frac{\lambda}{2}\frac{\partial}{\partial \mathbf{w}}\|\mathbf{w}\|^2 = \frac{\lambda}{2}\frac{\partial}{\partial \mathbf{w}} = \sum_{j=0}^{M} w_j^2 = \frac{\lambda}{2}\sum_{j=0}^{M} 2w_j = \lambda\sum_{j=0}^{M} w_j$$

Então

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

Rearranjando os termos da expressão

$$\sum_{j=0}^{M} w_j \sum_{n=1}^{N}(x_n)^{i+j} + \lambda\sum_{j=0}^{M} w_j = \sum_{n=1}^{N} t_n (x_n)^i$$

$$\left\{\sum_{n=1}^{N}(x_n)^{i+j} + \lambda\right\}\sum_{j=0}^{M} w_j = \sum_{n=1}^{N} t_n (x_n)^i$$

Portanto

$$\left\{\sum_{j=0}^{M} A_{ij} + \lambda \mathbf{I}_{ij}\right\} w_j = T_i$$

1.3 Exercício 1.5

Desenvolvendo a eq.1.38 do [1]

$$var[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \tag{5}$$

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

Portanto

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

1.4 Exercício 1.6

A definição da eq. 1.41 de [1] é

$$\begin{aligned} cov[x,y] &= \mathbb{E}_{x,y}[\{x-\mathbb{E}[x]\}\{y-\mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy]-\mathbb{E}[x]\mathbb{E}[y] \end{aligned} \tag{6}$$

O valor esperado de $f(x)$ é dada pela eq.1.33 de [1]:

$$\mathbb{E}[f] = \sum_n p(x)f(x)$$

Que é uma extensão do valor esperado de x (definição 4.1.1 de [2]):

$$\mathbb{E}[x] = \sum_n p(x)x$$

A partir da eq.2.2.1 de [2] , se 2 eventos são independentes

$$p(x,y) = p(x)p(y)$$

Utilizando o Teorema 4.2.6 de [2] e a definição de valor esperado considerando que x e y são variáveis aleatórias independentes:

$$\mathbb{E}[x,y] = \sum_x \sum_y p(x,y)xy = \sum_x xp(x)\sum_y yp(y) = \mathbb{E}[x]\mathbb{E}[y]$$

1.5 Exercício 1.7

A distribuição gaussiana é definida por (eq.1.46 de [1]):

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{7}$$

que é governada por dois parâmetros: μ, chamado de **média**, e σ^2, chamado de **variância**. A raiz quadrada da variância, representada por σ, é chamada de **desvio padrão** e, o inverso da variância, $\beta = 1/\sigma^2$, é chamado de **precisão**.

A transformação de coordenadas cartesianas é definida por

$$x = r\cos\theta$$

$$y = r\sin\theta$$

Que satisfaz a relação trigonométrica $x^2+y^2 = r^2$.

Se

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

e

$$\frac{\partial(x,y)}{\partial(r,\theta)} = r$$

Então

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

Fazendo mudança de variáveis $r^2 = u$:

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

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

Desta forma realizando a integral da distribuição gaussiana utilizando $y = x-\mu$:

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

1.6 Exercício 1.8

Primeira parte do exercício:
Relembrando, a distribuição gaussiana é definida por (eq.1.46 de [1])

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{8}$$

O valor esperado de funções de x sob a distribuição gaussiana, por definição é dado por:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2) x dx \tag{9}$$

Desta forma

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

Realizando a troca de variáveis

$$x - \mu = u$$
$$\frac{du}{dx} = 1$$

A integral passa a ser escrita como

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

Note que, conforme foi observado no exercício anterior:

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

e a segunda parcela é uma integral de função ímpar de $-\infty$ até $+\infty$ e o termo se anula:

$$\frac{1}{(2\pi\sigma^2)^{1/2}} \int_{-\infty}^{\infty} \exp\left\{-\frac{u^2}{2\sigma^2}\right\} u du = 0$$

Portanto, confirma-se a eq.1.49 de [1]

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

Segunda parte do exercício:
Diferenciando os dois lados da condição de normalização abaixo (eq.1.48 do [1]):

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2) dx = 1 \tag{10}$$

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

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

Para realizar a derivação com relação a σ^2 considera-se as seguintes funções auxiliares $f(\sigma^2)$ e $g(\sigma^2)$:

$$f(\sigma^2) = \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\frac{df(\sigma^2)}{d\sigma^2} = \frac{(x-\mu)^2}{2\sigma^4} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

e

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

$$\frac{dg(\sigma^2)}{d\sigma^2} = \frac{1}{2}(2\pi\sigma^2)^{-1/2}2\pi = (2\pi\sigma^2)^{-1/2}\pi$$

Desta forma:

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

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

Que pode ser reescrito como

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

Assim

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

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

Portanto, o valor esperado de funções de x^2 sob a distribuição gaussiana, por definição é dado por:

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)x^2 dx = \mu^2 + \sigma^2 \tag{11}$$

Finalmente

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

1.7 Exercício 1.9

Gaussiana Univariada
A distribuição gaussiana uni variada é definida por (eq.1.46 de [1])

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{12}$$

A **moda** de uma variável aleatória contínua é o valor em que a função densidade de probabilidade (f.d.p.) atinge o seu máximo — ou seja, o valor mais provável ou mais "frequente"dentro do possível intervalo contínuo.

Para isto, basta derivar a distribuição com relação à variável aleatória x e igualar a zero.

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

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

Portanto

$$x = \mu$$

Gaussiana Multivariada

Conforme apresentado na eq.1.52 de [1], a distribuição gaussiana definida sobre um vetor D-dimensional $\mathbf{x}$ de variáveis contínuas, é dada por

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\boldsymbol{\Sigma}^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{13}$$

onde o vetor D-dimensional $\boldsymbol{\mu}$ é chamado de média, a matriz $\boldsymbol{\Sigma}$, $D \times D$, é chamada de covariância e, $|\boldsymbol{\Sigma}|$ é o determinante de $\boldsymbol{\Sigma}$

Será realizada a derivada de $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ com relação a $\mathbf{x}$.

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

$$\frac{1}{(2\pi)^{D/2}}\frac{1}{\boldsymbol{\Sigma}^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\left[-\frac{1}{2}\frac{\partial}{\partial \mathbf{x}}\left\{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\right]$$

Para resolver a questão, serão utililzadas três propriedades derivadas de matrizes do Apêndice C de [1]:

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$$

$$\frac{\partial}{\partial x}(\mathbf{AB}) = \frac{\partial \mathbf{A}}{\partial x}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial x}$$

$$\frac{\partial}{\partial x}(\mathbf{x}^T\mathbf{a}) = \frac{\partial}{\partial x}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}$$

Desta forma

$$\frac{\partial}{\partial \mathbf{x}}\left\{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} = \frac{\partial}{\partial \mathbf{x}}\left\{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\right\}(\mathbf{x}-\boldsymbol{\mu}) + (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}-\boldsymbol{\mu}) =$$

$$= \frac{\partial}{\partial \mathbf{x}}\left\{(\boldsymbol{\Sigma}^{-1})^T(\mathbf{x}-\boldsymbol{\mu})\right\}(\mathbf{x}-\boldsymbol{\mu}) + (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{I} = \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} = 2(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

Assim,

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

Portanto

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

1.8 Exercício 1.10

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

Para resolução deste exercício serão, também, utilizadas as seguintes definições:

1. Dada uma variável aleatória $f(x)$ e sua função densidade de probabilidade, o valor esperado desta variável é dada por

$$\mathbb{E}[f] = \int p(x)f(x)dx$$

e

$$\int p(x)dx = 1$$

2. A definição de variância da eq.1.38 de [1]

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

3. O Teorema 4.3.5 de [2] ilustra uma parte da resolução deste exercício no seguinte trecho

$$\mathbb{E}[(x - \mathbb{E}[x])(z - \mathbb{E}[z])] = \mathbb{E}[(xz - x\mathbb{E}[z] - z\mathbb{E}[x]) + \mathbb{E}[x]\mathbb{E}[z]]$$

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

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

Resolução

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

$$\mathbb{E}[x+z] = \int\int xp(x)p(z)dxdz + \int\int zp(x)p(z)dxdz = \int p(z)\left(\int xp(x)dx\right)dz + \int p(z)z\left(\int p(x)dx\right)dz$$

Utilizando a definição 1 pode-se concluir que

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

Para a variância, a partir da definição 2 e 3:

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

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

$$var[x+z] = var[x] + var[z]$$

1.9 Exercício 1.11

A eq. 1.54 de [1] é

$$\ln p(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{14}$$

Calculando a sua derivada com relação ao parâmetro μ

$$\frac{\partial}{\partial\mu}\ln p(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}2(x_n-\mu)(-1) = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) = 0$$

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

$$\sum_{n=1}^{N}x_n - N\mu = 0$$

Portanto

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

Fazendo o cálculo da derivada o logaritmo a verossimilhança com relação à σ^2

$$\frac{\partial}{\partial\mu}\ln p(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2\sigma^2} = 0$$

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

Portanto

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

1.10 Exercício 1.13

A solução da máxima verossimilhança para a variância (eq.1.56 de[1]) é

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 \tag{15}$$

A partir das definições abaixo

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)x dx = \mu = \mathbb{E}[x] \tag{16}$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)x^2 dx = \mu^2 + \sigma^2 = \mathbb{E}[x^2] \tag{17}$$

Então

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

1.11 Exercício 1.21

Para a resolução deste exercício serão avaliadas utilizadas as seguintes propriedades

1. Se $a \leq b$ então

$$\sqrt{a} \leq \sqrt{b}$$

$$\sqrt{a}\sqrt{a} \leq \sqrt{a}\sqrt{b}$$

$$a \leq (ab)^{1/2}$$

2. Regra do produto

$$p(X,Y) = p(Y|X)p(X)$$

Em problemas de classificação da variável aleatória $\mathbf{x}$

$$p(\mathbf{x},\mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$$

Assim, um erro ocorre quando um vetor de entrada pertencente à classe $\mathcal{C}_1$ é atribuído à classe $\mathcal{C}_2$ ou vice-versa. A probabilidade de isso ocorrer é dada por

$$\begin{aligned} p(mistake) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x},\mathcal{C}_2)dx + \int_{\mathcal{R}_2} p(\mathbf{x},\mathcal{C}_1)dx \end{aligned} \tag{18}$$

Pela regra do produto

$$p(mistake) = \int_{\mathcal{R}_1} p(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})dx + \int_{\mathcal{R}_2} p(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})dx$$

Note que na região $\mathcal{R}_1$, $p(\mathbf{x},\mathcal{C}_2) \leq p(\mathbf{x},\mathcal{C}_1)$. Assim como $p(\mathbf{x},\mathcal{C}_1) \leq p(\mathbf{x},\mathcal{C}_2)$ na região $\mathcal{R}_2$. Portanto, utilizando a definição 1:

$$p(mistake) \leq \int_{\mathcal{R}_1} \{p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_2|\mathbf{x})\}^{1/2} p(\mathbf{x})dx + \int_{\mathcal{R}_2} \{p(\mathcal{C}_1|\mathbf{x})p(\mathcal{C}_2|\mathbf{x})\}^{1/2} p(\mathbf{x})dx = \int_{\mathcal{R}} \{p(\mathcal{C}_1|\mathbf{x})p(\mathbf{x})p(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})\}^{1/2} dx$$

$$p(mistake) \leq \int_{\mathcal{R}} \{p(\mathbf{x},\mathcal{C}_1)p(\mathbf{x},\mathcal{C}_2)\}^{1/2} dx$$

1.12 Exercício 1.22

A regra de decisão que minimiza a perda esperada é aquela que atribui cada novo $\mathbf{x}$ à classe j para a qual a quantidade

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) \tag{19}$$

é mínima.

Fazendo $L_{kj} = 1 - \mathbf{I}_{kj}$, onde $\mathbf{I}_{kj}$ são os elementos da matriz identidade:

$$min\left\{\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})\right\}$$

$$min\left\{\sum_k (1 - \mathbf{I}_{kj}) p(\mathcal{C}_k|\mathbf{x})\right\}$$

$$min\left\{\sum_k p(\mathcal{C}_k|\mathbf{x}) - \sum_k \mathbf{I}_{kj} p(\mathcal{C}_k|\mathbf{x})\right\}$$

$$min\left\{1 - p(\mathcal{C}_k|\mathbf{x})\right\} \equiv max\left\{p(\mathcal{C}_k|\mathbf{x})\right\}$$

1.13 Exercício 1.23

A partir da equação 1.81 de [1]

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) \tag{20}$$

Sabendo que $p(\mathcal{C}_k, \mathbf{x}) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$, então

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_k L_{kj} p(\mathbf{x}|\mathcal{C}_k) p(\mathcal{C}_k)$$

Existe uma compensação direta entre as probabilidades a priori e a matriz de perdas. Isso significa que ajustar a probabilidade a priori de uma classe,$p(\mathcal{C}_k)$, afeta a forma como as perdas influenciam a tomada de decisão.

1.14 Exercício 1.25

Em problemas de regressão, uma escolha comum de função de custo é a perda quadrática dada por $L(t, y(\mathbf{x})) = \{y(\mathbf{x}) - t\}^2$. Nesse caso, a perda esperada pode ser escrita como

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt \tag{21}$$

para 1 exemplo de treino. Já para múltiplos *targets* e equação pode ser generalizada para

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - \mathbf{t}\}^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \tag{22}$$

Utilizando a equação D-8 (apêndice D) de cálculo de variações

$$\frac{\partial G}{\partial y} - \frac{\partial}{\partial x}\left(\frac{\partial G}{\partial y'}\right) = 0 \tag{23}$$

Assim,

$$\frac{\partial \mathbb{E}[L]}{\partial y} - \frac{\partial}{\partial x}\left(\frac{\partial \mathbb{E}[L]}{\partial y'}\right) = 0$$

$$\frac{\partial \mathbb{E}[L]}{\partial y} = \frac{\partial}{\partial x}\left(\frac{\partial \mathbb{E}[L]}{\partial y'}\right)$$

Como o valor esperado da perda depende apenas de $y(\mathbf{x})$ e $\mathbf{t}$

$$\frac{\partial \mathbb{E}[L(y(\mathbf{x}), \mathbf{t})]}{\partial y'} = 0$$

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

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

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

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

$$y(\mathbf{x}) = \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} = \mathbb{E}_t[\mathbf{t}|\mathbf{x}]$$

Portanto,

$$y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t) dt}{p(\mathbf{x})} = \int t p(t|\mathbf{x}) dt = \mathbb{E}_t[t|\mathbf{x}] \tag{24}$$

1.15 Exercício 1.31

A partir das propriedades da divergência de *Kullback-Leibler*, vemos que $\mathbf{I}(\mathbf{x}, \mathbf{y}) \geq 0$, com igualdade se, e somente se, $\mathbf{x}$ e $\mathbf{y}$ forem independentes. Utilizando as regras da soma e do produto da probabilidade, vemos que a informação mútua está relacionada à entropia condicional por meio da eq. 1.121 de [1]:

$$\mathbf{I}[\mathbf{x}, \mathbf{y}] = \mathbf{H}[\mathbf{y}] - \mathbf{H}[\mathbf{y}|\mathbf{x}] \tag{25}$$

Onde $\mathbf{I}[\mathbf{x}, \mathbf{y}]$ é a **informação mútua** entre as variáveis $\mathbf{x}$ e $\mathbf{y}$. Já $\mathbf{H}$ é a **entropia** de uma variável aleatória.

A partir da eq. 1.112

$$\mathbf{H}[\mathbf{x}, \mathbf{y}] = \mathbf{H}[\mathbf{y}|\mathbf{x}] + \mathbf{H}[\mathbf{x}] \tag{26}$$

Então

$$\mathbf{H}[\mathbf{x}, \mathbf{y}] = \mathbf{H}[\mathbf{y}] - \mathbf{I}[\mathbf{x}, \mathbf{y}] + \mathbf{H}[\mathbf{x}]$$

$$\mathbf{H}[\mathbf{x}, \mathbf{y}] \leq \mathbf{H}[\mathbf{x}] + \mathbf{H}[\mathbf{y}]$$

Demonstração da independência estatística

A eq.1.104 de [1] define entropia como

$$\mathbf{H}[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \tag{27}$$

Então

$$\mathbf{H}[\mathbf{x}, \mathbf{y}] = \int \int p(\mathbf{x}, \mathbf{y}) \ln \{p(\mathbf{x}, \mathbf{y})\} \, d\mathbf{x} d\mathbf{y}$$

$$\mathbf{H}[\mathbf{x}, \mathbf{y}] = \int \int p(\mathbf{x}) p(\mathbf{y}) \ln \{p(\mathbf{x}) p(\mathbf{y})\} \, d\mathbf{x} d\mathbf{y}$$

$$\mathbf{H}[\mathbf{x}, \mathbf{y}] = \int \int p(\mathbf{x}) p(\mathbf{y}) \{\ln p(\mathbf{x}) + \ln p(\mathbf{y})\} \, d\mathbf{x} d\mathbf{y} = \int \int p(\mathbf{y}) p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} d\mathbf{y} + \int \int p(\mathbf{x}) p(\mathbf{y}) \ln p(\mathbf{y}) d\mathbf{y} d\mathbf{x}$$

Assim,

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

$$\mathbf{H}[\mathbf{y}] = \mathbf{H}[\mathbf{y}|\mathbf{x}]$$

Consequentemente, a partir da eq 1.121

$$\mathbf{I}[\mathbf{x}, \mathbf{y}] = 0$$

Portanto as variáveis $\mathbf{x}$ e $\mathbf{y}$ são independentes.

1.16 Exercício 1.33

A entropia condicional de variáveis discretas $\mathbf{x}$ e $\mathbf{y}$ é

$$\mathbf{H}[y|x] = -\sum_i \sum_j p(x_j, y_i) \ln p(y_i|x_j) \tag{28}$$

Utilizando a regra do produto e igualando $\mathbf{H}[y|x]$ a zero

$$\mathbf{H}[y|x] = -\sum_i \sum_j p(y_i|x_j) p(x_j) \ln p(y_i|x_j) = 0$$

Se $p(x) > 0$, resta analisar $p \ln p$:

$$\begin{cases} -p(y_i|x_j) \ln p(y_i|x_j) = 0 & p(y_i|x_j) = 0 \\ -p(y_i|x_j) \ln p(y_i|x_j) = 0 & p(y_i|x_j) = 1 \end{cases} \tag{29}$$

Como

$$\sum_i \sum_j p(y_i|x_j) = 1$$

Existe um $p(y_i|x_i) = 1$ e, o restante é zero.

1.17 Exercício 1.37

A partir das seguintes definições do livro texto do [1]

1. Equação 1.111

$$\mathbf{H}[\mathbf{y}|\mathbf{x}] = -\int \int p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{y} d\mathbf{x} \tag{30}$$

2. Equação 1.104

$$\mathbf{H}[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \tag{31}$$

3. Marginalização da probabilidade conjunta de duas variáveis aleatórias $\mathbf{x}$ e $\mathbf{y}$.

$$\int p(\mathbf{x}, \mathbf{y}) dy = p(\mathbf{x}) \tag{32}$$

$$\mathbf{H}[\mathbf{x}, \mathbf{y}] = -\int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = -\int \int p(\mathbf{x}, \mathbf{y}) \ln \{p(\mathbf{y}|\mathbf{x}) p(\mathbf{x})\} d\mathbf{x} d\mathbf{y}$$

$$\mathbf{H}[\mathbf{x}, \mathbf{y}] = -\int \int p(\mathbf{x}, \mathbf{y}) \{\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})\} d\mathbf{x} d\mathbf{y} = -\int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} - \int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{x}) d\mathbf{x} d\mathbf{y}$$

$$\mathbf{H}[\mathbf{x}, \mathbf{y}] = -\int \int p(\mathbf{x}, \mathbf{y}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{x} d\mathbf{y} - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} = \mathbf{H}[\mathbf{y}|\mathbf{x}] + \mathbf{H}[\mathbf{x}]$$

Tabela 1: Distribuição das probabilidades de x e y.

	$y=0$	$y=1$	$p(x)$
$x=0$	$1/3$	$1/3$	$2/3$
$x=1$	0	$1/3$	$1/3$
$p(y)$	$1/3$	$2/3$	

Tabela 2: Condicional $p(x|y) = p(x,y)/p(y)$.

	$y=0$	$y=1$
$x=0$	1	$1/2$
$x=1$	0	$1/2$

1.18 Exercício 1.39

a)

$$H[x] = -\sum_i p(x_i)\ln p(x_i) = -\frac{2}{3}\ln\frac{2}{3} - \frac{1}{3}\ln\frac{1}{3} = -\frac{2}{3}(\ln 2 - \ln 3) - \frac{1}{3}\ln(\ln 1 - \ln 3) = -\frac{2}{3}\ln 2 + \frac{2}{3}\ln 3 + \frac{1}{3}\ln 3$$

$$H[x] = \ln 3 - \frac{2}{3}\ln 2$$

b)

$$H[y] = -\frac{1}{3}\ln\frac{1}{3} - \frac{2}{3}\ln\frac{2}{3} = H[x]$$

c)

$$H[y|x] = -\sum_i\sum_j p(x_j, y_i)\ln p(y_i|x_j) = -\frac{1}{3}\ln 1 - \frac{1}{3}\ln\frac{1}{2} - 0\ln 0 - \frac{1}{3}\ln\frac{1}{2} = -\frac{2}{3}(\ln 1 - \ln 2) = \frac{2}{3}\ln 2$$

d) $H[x|y] = H[y|x]$

e)

$$H[x,y] = H[y|x] + H[x] = \frac{2}{3}\ln 2 + \ln 3 - \frac{2}{3}\ln 2 = \ln 3$$

f)

$$I[x,y] = H[x] - H[x|y] = \ln 3 - \frac{2}{3}\ln 2 - \frac{2}{3}\ln 2 = \ln 3 - \frac{4}{3}\ln 2$$

1.19 Exercício 1.41

A partir da definição da eq.1.120 de [1] e da regra do produto $p(\mathbf{x},\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$

$$I[\mathbf{x},\mathbf{y}] = -\int\int p(\mathbf{x},\mathbf{y})\ln\left\{\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right\}d\mathbf{x}d\mathbf{y} \tag{33}$$

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

$$I[\mathbf{x},\mathbf{y}] = -\int\int p(\mathbf{x},\mathbf{y})\left[\ln p(\mathbf{x}) + \ln p(\mathbf{y}) - \ln p(\mathbf{x},\mathbf{y})\right]d\mathbf{x}d\mathbf{y} = -\int\int p(\mathbf{x},\mathbf{y})\left\{\ln p(\mathbf{x}) + \ln p(\mathbf{y}) - \ln\left[p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\right]\right\}d\mathbf{x}d\mathbf{y}$$

$$I[\mathbf{x},\mathbf{y}] = -\int\int p(\mathbf{x},\mathbf{y})\left\{\ln p(\mathbf{x}) + \ln p(\mathbf{y}) - \ln p(\mathbf{y}|\mathbf{x}) - \ln p(\mathbf{x})\right\}d\mathbf{x}d\mathbf{y} =$$

$$= -\int\int p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y})d\mathbf{x}d\mathbf{y} + \int\int p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y}$$

$$I[\mathbf{x},\mathbf{y}] = -\int p(\mathbf{y})\ln p(\mathbf{y})d\mathbf{y} + \int\int p(\mathbf{x},\mathbf{y})\ln p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y} = H[\mathbf{y}] + H[\mathbf{y}|\mathbf{x}]$$

Tabela 3: Condicional $p(y|x) = p(x,y)/p(x)$.

	$y=0$	$y=1$
$x=0$	$1/2$	$1/2$
$x=1$	0	1

2 Exercícios extras do livro texto

2.1 E1

Utilizando as seguintes definições:

- A eq.1.1 do [1]:

$$y(x_n, \mathbf{w}) = w_0 + w_1 x_n + w_2 x_n^2 + \cdots + w_M x_n^M = \sum_{j=0}^{M} w_j x_n^j \tag{34}$$

- A função objetivo abaixo (eq. 1.2 de [1]):

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left\{ y\left(x_n, \mathbf{w}\right) - t_n \right\}^2 \tag{35}$$

- A parâmetro M é a quantidade de coeficientes a serem ajustados durante o treinamento.
- O N é definido como a quantidade de dados de treinamento x_n (ou $\mathbf{x}_n$) e t_n;

Reescrevendo $y(x_n, \mathbf{w})$:

$$y_n(x_n, \mathbf{w}) = \begin{bmatrix} x^0 & x^1 & x^2 & \cdots & x^M \end{bmatrix}^T_{1\times M} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_M \end{bmatrix}_{M\times 1}$$

Generalizando o problema:

$$\mathbf{A}\mathbf{w} = \mathbf{y}(\mathbf{x}, \mathbf{w})$$

$$\begin{bmatrix} x_0^0 & x_0^1 & x_0^2 & \cdots & x_0^M \\ x_1^0 & x_1^1 & x_1^2 & \cdots & x_1^M \\ x_2^0 & x_2^1 & x_2^2 & \cdots & x_2^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_N^0 & x_N^1 & x_N^2 & \cdots & x_N^M \end{bmatrix}_{N\times M} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_M \end{bmatrix}_{M\times 1} = \begin{bmatrix} y_0(x_0, \mathbf{w}) \\ y_1(x_1, \mathbf{w}) \\ y_2(x_2, \mathbf{w}) \\ \vdots \\ y_N(x_N, \mathbf{w}) \end{bmatrix}_{N\times 1} = \mathbf{y}(\mathbf{x}, \mathbf{w})$$

$$E(\mathbf{w}) = \left\| \begin{bmatrix} y_0(x_0, \mathbf{w}) \\ y_1(x_1, \mathbf{w}) \\ y_2(x_2, \mathbf{w}) \\ \vdots \\ y_N(x_N, \mathbf{w}) \end{bmatrix}_{N\times 1} - \begin{bmatrix} t_0 \\ t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}_{N\times 1} \right\|_2^2 \frac{1}{2} = \left\| \mathbf{y} - \mathbf{t} \right\|_2^2 \frac{1}{2}$$

Ao realizar a minimização da função de custo $E(\mathbf{w})$, ou seja, derivando com relação ao $\mathbf{w}$ e igualando a zero, obtém-se o seguinte desenvolvimento e resultado:

$$\frac{\partial}{\partial \mathbf{w}} E(\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}} \frac{1}{2} \left\| \mathbf{A}\mathbf{w} - \mathbf{t} \right\|_2^2$$

$$\frac{\partial}{\partial \mathbf{w}} E(\mathbf{w}) = \frac{1}{2} 2 \left(\mathbf{A}\mathbf{w} - \mathbf{t} \right)^T \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{A}\mathbf{w} - \mathbf{t} \right)^T = \left(\mathbf{A}\mathbf{w} - \mathbf{t} \right)^T \frac{\partial}{\partial \mathbf{w}} \mathbf{A}\mathbf{w}$$

$$\frac{\partial}{\partial \mathbf{w}} E(\mathbf{w}) = \left(\mathbf{A}\mathbf{w} - \mathbf{t} \right)^T \mathbf{A}\mathbf{I} = 0$$

$$\left(\mathbf{w}^T\mathbf{A}^T - \mathbf{t}^T\right)\mathbf{A} = 0$$

$$\mathbf{w}^T\mathbf{A}^T\mathbf{A} - \mathbf{t}^T\mathbf{A} = 0$$

$$(\mathbf{A}^T\mathbf{A})^T\mathbf{w} = \mathbf{A}^T\mathbf{t}$$

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

Em reconhecimento de reconhecimento de padrões, a utilização da notação vetorial e matricial - por exemplo, $\mathbf{y} = A\mathbf{w}$, onde A representa a matriz de características e $\mathbf{w}$ o vetor de coeficientes (pesos) - oferece vantagens em relação à representação por somatórios explícitos porque esta forma compacta, não apenas melhora a legibilidade e abstração dos modelos, mas também permite o uso direto de ferramentas da álgebra linear, como decomposições matriciais, regularização e projeções em subespaços. Além disso, essa notação se alinha naturalmente com implementações computacionais vetorizadas, exploradas em bibliotecas otimizadas como NumPy, TensorFlow ou PyTorch. Em tarefas como classificação, regressão ou redução de dimensionalidade (como PCA e LDA), a manipulação matricial é essencial para expressar relações entre variáveis, aplicar transformações lineares e derivar soluções fechadas para critérios de otimização.

2.2 E2

Utilizando as seguintes definições:

- A eq.1.1 do [1]:

$$y(x_n, \mathbf{w}) = w_0 + w_1 x_n + w_2 x_n^2 + \cdots + w_M x_n^M = \sum_{j=0}^{M} w_j x_n^j \tag{36}$$

- A função objetivo abaixo (eq. 1.4 de [1]):

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{y\left(x_n, \mathbf{w}\right) - t_n\right\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{37}$$

$$\frac{\partial}{\partial \mathbf{w}}E(\mathbf{w}) = \left(\mathbf{A}\mathbf{w} - \mathbf{t}\right)^T\mathbf{A}\mathbf{I} + \frac{\lambda}{2}\frac{\partial}{\partial \mathbf{w}}\|\mathbf{w}\|^2 = \left(\mathbf{A}\mathbf{w} - \mathbf{t}\right)^T\mathbf{A}\mathbf{I} + \frac{\lambda}{2}2\mathbf{w}^T\frac{\partial}{\partial \mathbf{w}}\mathbf{w} = 0$$

$$\frac{\partial}{\partial \mathbf{w}}E(\mathbf{w}) = \left(\mathbf{A}\mathbf{w} - \mathbf{t}\right)^T\mathbf{A}\mathbf{I} + \lambda\mathbf{w}^T\mathbf{I}$$

$$\mathbf{w}^T\mathbf{A}^T\mathbf{A} - \mathbf{t}^T\mathbf{A} + \lambda\mathbf{w}^T = 0$$

$$(\mathbf{A}^T\mathbf{A})^T\mathbf{w} - \mathbf{A}^T\mathbf{t} + \lambda\mathbf{w} = 0$$

$$\mathbf{A}^T\mathbf{A}\mathbf{w} + \lambda\mathbf{w} = \mathbf{A}^T\mathbf{t}$$

$$(\mathbf{A}^T\mathbf{A} + \lambda)\mathbf{w} = \mathbf{A}^T\mathbf{t}$$

$$\mathbf{w} = (\mathbf{A}^T\mathbf{A} + \lambda)^{-1}\mathbf{A}^T\mathbf{t}$$

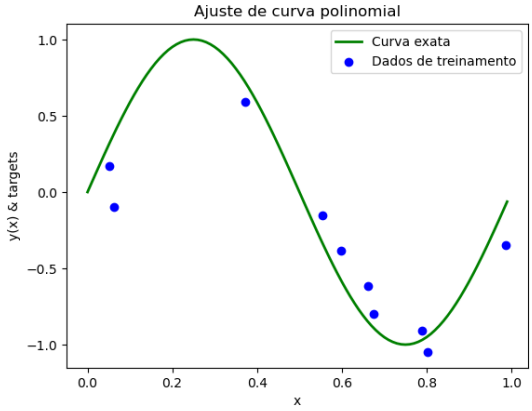

Figura 1: Dados sintéticos.

3 Exercícios computacionais

Esta parte da lista de exercícios consiste em replicar o experimento computacional de ajusta de curva polinomial (*Polynomial Curve Fitting*) apresentado no primeiro capítulo do [1].

Este exercício foi realizado na linguagem de programação **Python v3.8.18**, através do **Notebook Jupyter v.6.5.2** (Gerenciador Anaconda).

a) Replicar as figuras 1.4 e 1.6 do [1].

Primeiramente, os dados são gerados de acordo com o Apêndice A do [1](*Synthetic Data*). Os valores de entrada x_n são gerados uniformemente no intervalo $(0, 1)$, e os *targets* correspondentes t_n são obtidos a partir dos valores correspondentes da função $\sin(2\pi x)$, e depois adicionando ruído branco com uma distribuição Gaussiana de desvio padrão $0,3$.

Posteriormente, foi realizado o ajuste de curva polinomial com base nos dados sintéticos para cada grau escolhido, neste caso, conforme o livro (grau 0, 1,3 e 9).

```
n = 3                                                    #quantidade de modelos
M = [1, 3, 9]                                            #grau do polinomio (modelos)

intercept = np.zeros(n)
coef = np.zeros((n, max(M)))
X_new = np.arange(0., 1., 1/N).reshape(N,1)
Y = np.zeros((n,N))

for i in range(n):
poly_features = PolynomialFeatures(degree = M[i], include_bias=False)
X_poly = poly_features.fit_transform(X)
lin_reg = LinearRegression()
lin_reg.fit(X_poly, t)
intercept[i] = lin_reg.intercept_       #vetor de intercept de cada modelo
y_aux = lin_reg.predict(poly_features.fit_transform(X_new))
for j in range(M[i]):
coef[i][j] = lin_reg.coef_[0][j]        #matriz com os coeficientes de cada modelo

for k in range(N):
Y [i][k] =  y_aux[k][0]
```

Assim a figura 1.4 do [1] foi replicada na Figura 2

Assim como foi apresentado no [1], nota-se que para uma amostra com tamanho $N = 10$, os modelos com graus de liberdade menor que 3 apresentam subajuste, ou seja, a complexidade do modelo é menor que a complexidade do problema. Por outro lado, considerando $M = 9$ observa-se que a complexidade do modelo é maior do que a do problema em questão, levando a um super-ajuste, isto é, a curva tende a passar por todos os dados. O modelo aprende também o ruído dos *targets*, fazendo com que o modelo

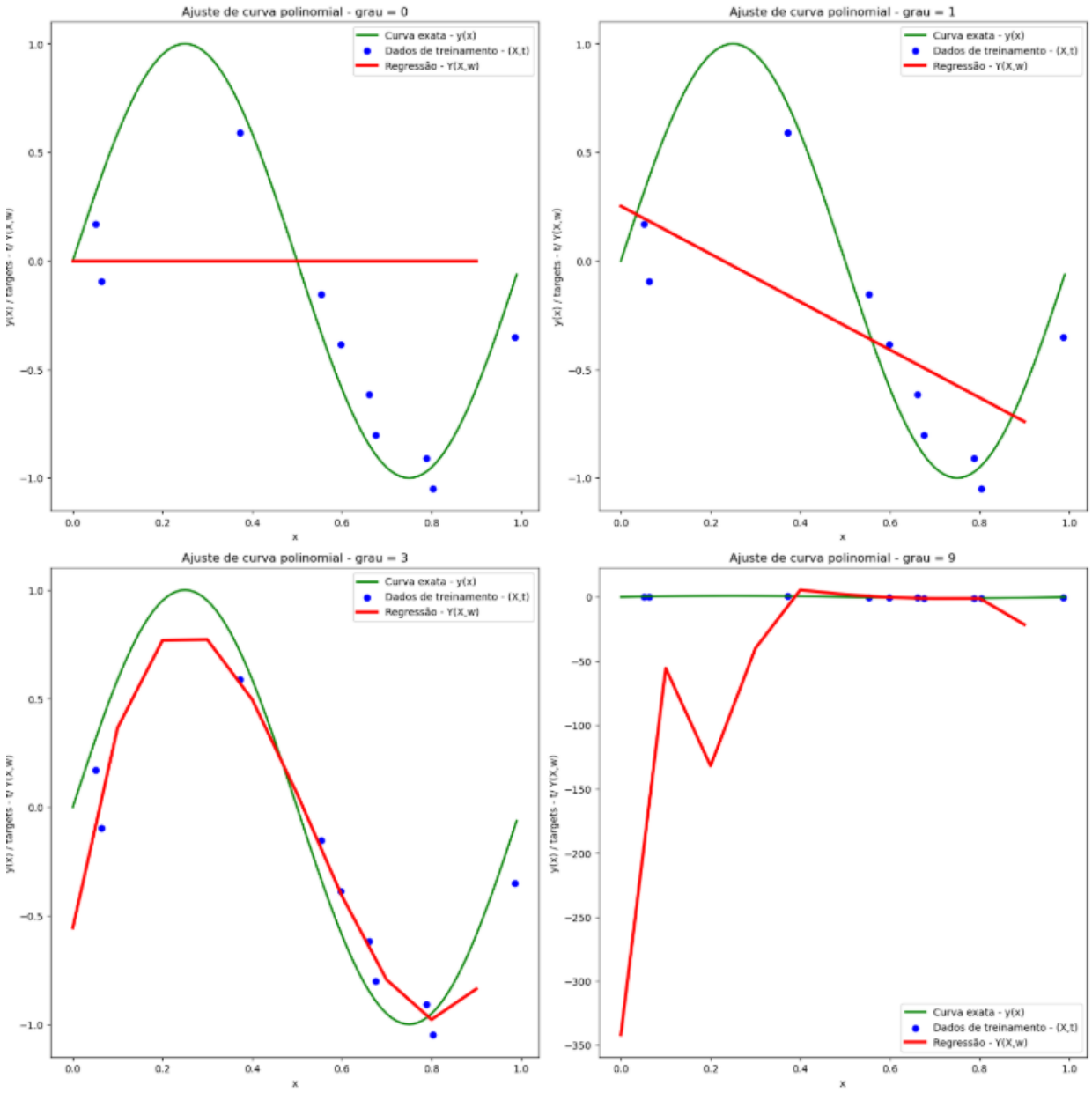

Figura 2: Ajusta da curva polinomial com base nos dados sintéticos para diferentes graus de liberdade.

tenha uma baixo viés, mas com alta variância. Isto faz com que o modelo preditivo tenha um baixa capacidade de generalização.

Outro forma da avaliar que houve super ajuste é a magnitude dos coeficientes $\mathbf{w}$ calculados (ver Figura 3).

Quando há superajuste do modelo, os valores dos coeficientes atingem magnitudes muito alta.

Repetiu-se o experimento considerando o ajuste polinomial com grau 9 ($M = 9$) para dados sintéticos com tamanho de amostra $N = 15$ e $N = 100$ (Replicar a figura 1.6 do [1]). Os resultados estão apresentados nas Figura 4 e Figura 5 respectivamente.

Novamente, quando tamanho de dados de treino $N = 15$ o modelo fica super ajustado (Figura 4). Porém, quando o tamanho dos dados aumentam para $N = 100$, o problema de super ajuste é resolvido (Figura 5).

O superajuste (ou *overfitting*) tende a diminuir quando se aumenta o tamanho dos dados de treino porque o modelo passa a ter uma representação mais ampla e variada do problema que está tentando aprender. Com um conjunto de dados pequeno, o modelo pode facilmente memorizar os exemplos específicos de treinamento, incluindo o ruído e variações irrelevantes, em vez de aprender padrões gerais. Ao aumentar a quantidade de dados, o modelo é exposto a mais exemplos e variações, o que força o

```
In [5]: intercept

Out[5]: array([ 2.52236664e-01, -5.55673277e-01, -3.42036792e+02])

In [6]: coef

Out[6]: array([[-1.10247598e+00,  0.00000000e+00,  0.00000000e+00,
                 0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
                 0.00000000e+00,  0.00000000e+00,  0.00000000e+00],
               [ 1.22178356e+01, -3.20096555e+01,  2.00965413e+01,
                 0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
                 0.00000000e+00,  0.00000000e+00,  0.00000000e+00],
               [ 1.59527194e+04, -2.60191069e+05,  1.88006568e+06,
                -7.36141895e+06,  1.71612980e+07, -2.46000761e+07,
                 2.13181298e+07, -1.02613957e+07,  2.10801994e+06]])
```

Figura 3: Coeficientes ajustados (peso w_n) para os valores de $M = 1, 3$ e 9 respectivamente.

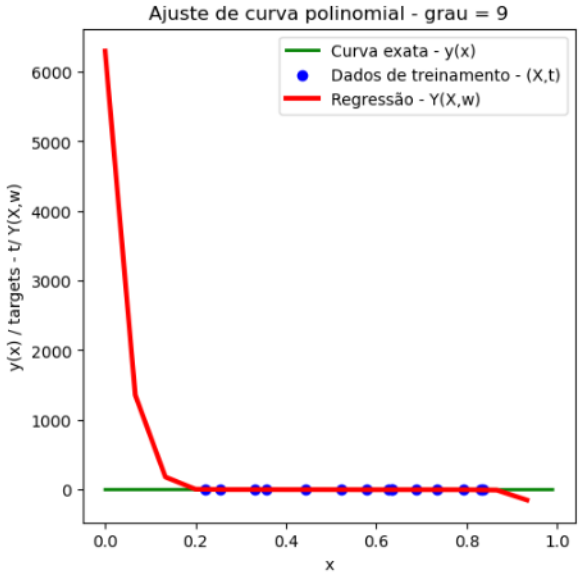

Figura 4: Ajuste da curva polinomial - $M = 9$ e $N = 15$.

aprendizado de características mais robustas e generalizáveis. Isso torna mais difícil que o modelo se ajuste apenas às peculiaridades do conjunto de treinamento, resultando em melhor desempenho em dados novos e não vistos, e, consequentemente, em uma redução do superajuste.

b) Simulação com base de dados sem relevância estatística.

Nesta parte do problema foi gerada uma amostra de tamanho $N = 50$ que não representa a população (ver Figura 6).

Considerando um polinômio de grau $M = 4$, a curva ajustada ficou com o seguinte comportamento (Figura 7)

Ter uma base de dados com relevância estatística é fundamental para garantir que as conclusões obtidas a partir dos dados sejam confiáveis, representativas e generalizáveis para o problema real que se deseja resolver. Uma base estatisticamente relevante possui tamanho suficiente e diversidade adequada para capturar a variabilidade dos dados do mundo real, reduzindo o risco de vieses.

Se a base de dados for pequena ou mal distribuída, o modelo pode aprender padrões incorretos ou não representativos, levando a erros de generalização, superajuste, ou mesmo a decisões injustas em contextos sensíveis, como saúde ou finanças. Além disso, a relevância estatística é essencial para validar hipóteses com testes estatísticos, garantindo que os resultados obtidos não sejam fruto do acaso, mas sim de relações significativas nos dados.

Em casos práticos com dimensão elevada (ou seja, quando o número de variáveis ou características dos dados é muito grande), a relação entre a relevância estatística da base de dados e a qualidade do modelo se torna ainda mais crítica. Esse cenário é conhecido como o problema da maldição da

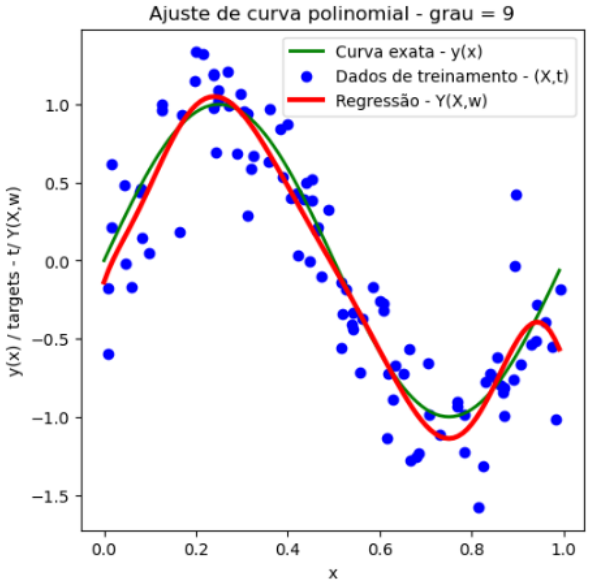

Figura 5: Ajuste da curva polinomial - $M = 9$ e $N = 100$.

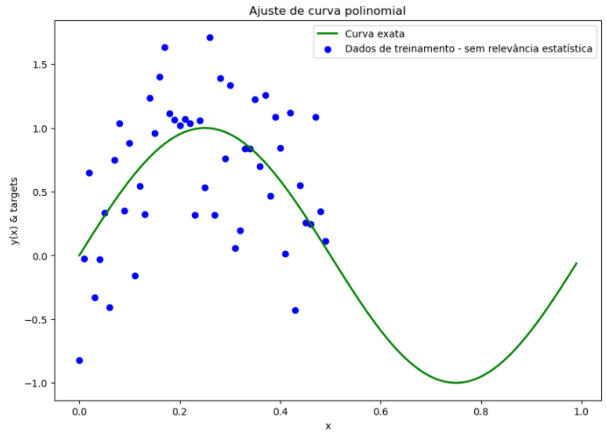

Figura 6: Base de dados sem relevância estatística $N = 50$.

dimensionalidade (*curse of dimensionality*), e afeta diretamente a generalização e o desempenho de modelos de reconhecimento de padrões.

À medida que a dimensão cresce, os dados se tornam mais esparsos no espaço, e é necessário muito mais dados para cobrir esse espaço de forma representativa. Sem uma base estatisticamente relevante e suficientemente grande, os modelos tendem a superajustar ou simplesmente não aprender padrões úteis.

Com muitos atributos, cresce a chance de encontrar relações aparentes entre variáveis e os *targets* que não são verdadeiras, apenas fruto do acaso. Isso pode enganar o modelo, que aprende padrões sem significado real.

Em resumo, o modelo com dados de treinamento que não sejam relevantes para o problema, isto é, em casos de dimensão elevada, em que há atributos com baixa ou nenhuma relação mútua com o os *targets*, podem gerar modelos com baixíssima capacidade de generalização. O modelos final estará errado.

c) Simulação de base de dados com 1 *outlier*.

Um *outlier* é um valor ou observação que se desvia significativamente do padrão geral dos dados, situando-se fora do intervalo onde a maioria dos valores se concentra. Esses pontos atípicos podem surgir por diversos motivos, como erros de medição, variabilidade natural extrema ou eventos raros. Sua presença pode distorcer análises estatísticas, afetar médias, variâncias e influenciar negativamente o desempenho de modelos de aprendizado de máquina, especialmente os mais sensíveis a variações nos dados. Identificar e tratar *outliers* adequadamente é essencial para garantir a robustez e a confiabilidade das conclusões obtidas a partir dos dados.

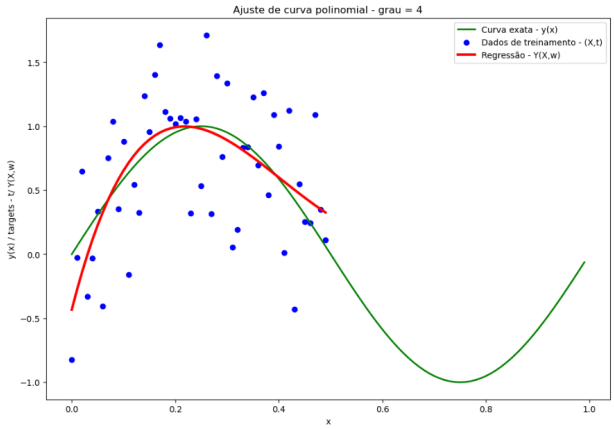

Figura 7: Ajuste polinimial a partir de base de dados sem relevância estatística $N = 50$ e $M = 4$.

Nesta parte da resolução, foi acrescentado um *outlier* nos dados sintéticos (ver Figura 8)

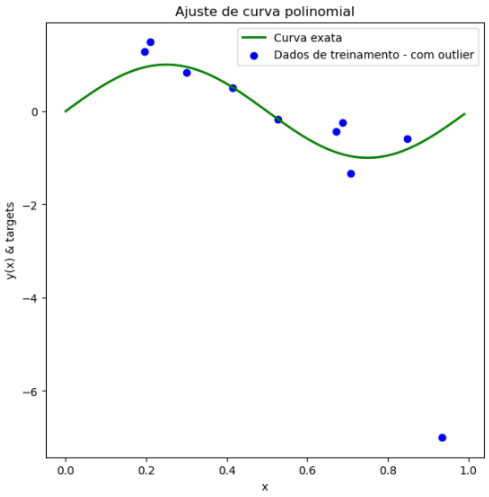

Figura 8: Dados sintéticos com 1 *outlier* ($N = 10$).

Os *outliers* podem ter um impacto significativo e geralmente negativo em ajustes de curvas, especialmente quando se utilizam métodos que minimizam o erro quadrático médio (como a regressão linear clássica). Esses métodos são sensíveis a valores extremos, pois um único *outlier* pode gerar um erro muito alto que domina a função de custo, fazendo com que a curva se ajuste de forma inadequada ao restante dos dados.

O resultado do ajuste polinomial com $M = 3$ está na Figura 9. Um dado errado no conjunto de treinamento, dependendo do tamanho da amostra é capaz de alterar completamente o modelo preditivo, e o modelo final estará errado.

Principais impactos dos outliers em ajustes de curvas:

1. Desvio do modelo: a curva ajustada pode ser "puxada"em direção ao *outlier*, desviando-se dos padrões reais da maioria dos dados. Isso reduz a capacidade de generalização do modelo.
2. Aumento do erro global: O erro quadrático médio aumenta, pois estes pontos geram diferenças muito grandes entre os valores reais e os previstos, afetando negativamente as métricas de desempenho.
3. Perda de interpretabilidade: O modelo ajustado pode se tornar menos intuitivo ou interpretável, refletindo padrões artificiais criados pela influência dos outliers.
4. Necessidade de técnicas robustas: a presença de*outliers* frequentemente exige o uso de métodos de ajuste mais robustos,ou técnicas de pré-processamento

Portanto, *outliers* distorcem o ajuste de curvas ao introduzirem variações não representativas que afetam a forma da curva, o desempenho do modelo e a validade da análise. Por isso, é essencial identificá-los e tratá-los adequadamente durante a modelagem.

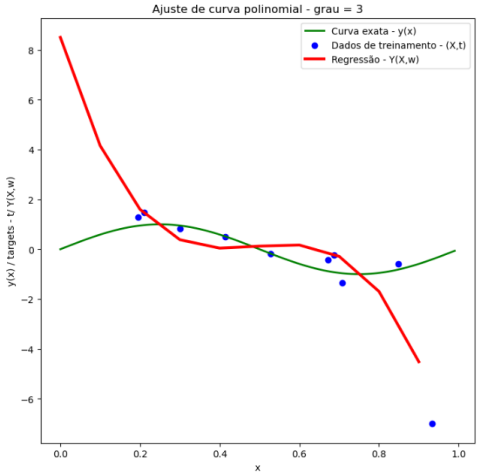

Figura 9: Resultado com 1 *outlier* ($N = 10$ e $M = 3$).

Referências

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, 1 edition, 2006.

[2] Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics*. Pearson Education, Boston, Mass. [u.a.], 4. ed. edition, 2012. Includes bibliographical references and index.