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

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

9 de junho de 2025

**Resumo**

Relatório da resolução dos exercícios da Lista 2 (Capítulo 2 - Bishop [1])

# 1 Exercícios do livro texto

## 1.1 Exercício 2.1

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

1. Dada a eq. 2.2 de [1] reproduzida abaixo:

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{1}$$

   Onde a variável aleatória **binária** $x \in \{0, 1\}$ e, também, $p(x = 1|\mu) = \mu$ e $p(x = 0|\mu) = 1 - \mu$.

2. A equação 1.33 do [1]

$$\mathbb{E}[f] = \sum_x p(x)f(x) \tag{2}$$

3. A equação 1.39 da [1]

$$var[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \tag{3}$$

4. A partir da definição de variância, também contida na resolução do exercício 1.8 da lista 1 ([2])

$$\mathbb{E}[(x-\mu)^2] = \sigma^2 \tag{4}$$

**Resolução do exercício:**

- Propriedade (2.257):

$$\sum_{x=0}^{1} p(x|\mu) = p(x = 0|\mu) + p(x = 1|\mu) = 1 - \mu + \mu = 1$$

- Propriedade (2.258):

$$\mathbb{E}[x] = \sum_{x=0}^{1} xp(x|\mu) = 0p(x = 0|\mu) + 1p(x = 1|\mu) = \mu$$

- Propriedade (2.259):

$$var[x] = \sum_{x=0}^{1}(x-\mu)^2 p(x|\mu) = \sum_{x=0}^{1}(x^2 - 2x\mu + \mu^2)p(x|\mu) = \mu^2 p(x=0|\mu) + (1-2\mu+\mu^2)p(x=1|\mu)$$

$$var[x] = \mu^2(1-\mu) + (1-2\mu+\mu^2)\mu = \mu^2 - \mu^3 + \mu - 2\mu^2 + \mu^3 = \mu - \mu^2 = \mu(1-\mu)$$

A partir da definição de entropia abaixo (eq. 1.98 de [1]):

$$H[x] = -\sum_{x} p(x)\ln p(x) \tag{5}$$

Então,

$$H[x] = -\sum_{x=0}^{1}\mu^x(1-\mu)^{1-x}\ln\left\{\mu^x(1-\mu)^{1-x}\right\} = -\sum_{x=0}^{1}\mu^x(1-\mu)^{1-x}\left[\ln\left\{\mu^x\right\} + \ln\left\{(1-\mu)^{1-x}\right\}\right]$$

$$H[x] = -\sum_{x=0}^{1}\mu^x(1-\mu)^{1-x}\left\{x\ln\mu + (1-x)\ln(1-\mu)\right\}$$

Portanto,

$$H[x] = -(1-\mu)\ln(1-\mu) - \mu\ln\mu$$

## 1.2 Exercício 2.2

A partir das definições

1. Dada a eq. 2.2 de [1] reproduzida abaixo:

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{6}$$

   Onde a variável aleatória **binária** $x \in \{0,1\}$ e, também, $p(x=1|\mu) = \mu$ e $p(x=0|\mu) = 1-\mu$.

2. Equação 2.261 do [1]:

$$p(x|\mu) = \left(\frac{1-\mu}{2}\right)^{(1-x)/2}\left(\frac{1+\mu}{2}\right)^{(1+x)/2} \tag{7}$$

   Onde a variável aleatória **binária** $x \in \{-1,1\}$ e, também, $\mu \in [-1,1]$.

**Resolução do exercício:**

(i) Verificação da distribuição normalizada

$$\sum_{x=-1}^{1} p(x|\mu) = p(x=-1|\mu)+p(x=1|\mu) = \left(\frac{1-\mu}{2}\right)^1\left(\frac{1+\mu}{2}\right)^0 + \left(\frac{1-\mu}{2}\right)^0\left(\frac{1+\mu}{2}\right)^1 = \frac{1-\mu}{2}+\frac{1+\mu}{2} = 1$$

(ii) Verificação da média

$$\sum_{x=-1}^{1} xp(x|\mu) = (-1)\left(\frac{1-\mu}{2}\right) + (1)\left(\frac{1+\mu}{2}\right) = \frac{-1+\mu}{2} + \frac{1+\mu}{2} = \mu$$

(iii) Verificação da Variância

$$\sum_{x=-1}^{1}(x-\mu)^2 p(x|\mu) = (-1-\mu)^2\left(\frac{1-\mu}{2}\right)+(1-\mu)^2\left(\frac{1+\mu}{2}\right) = (1+2\mu+\mu^2)\left(\frac{1-\mu}{2}\right)+(1-2\mu+\mu^2)\left(\frac{1+\mu}{2}\right)$$

$$\sum_{x=-1}^{1}(x-\mu)^2 p(x|\mu) = \frac{1+2\mu+\mu^2-\mu-2\mu^2-\mu^3}{2}+\frac{1-2\mu+\mu^2+\mu-2\mu^2+\mu^3}{2} = 1-\mu^2$$

(iv) Verificação da Entropia

$$H[x] = -\sum_{x=-1}^{1} p(x)\ln p(x) =$$

$$= -\sum_{x=-1}^{1}\left(\frac{1-\mu}{2}\right)^{(1-x)/2}\left(\frac{1+\mu}{2}\right)^{(1+x)/2}\left\{\left(\frac{1-x}{2}\right)\ln\left(\frac{1-\mu}{2}\right)+\left(\frac{1+x}{2}\right)\ln\left(\frac{1+\mu}{2}\right)\right\}$$

$$= \left(\frac{1-\mu}{2}\right)\ln\left(\frac{1-\mu}{2}\right)+\left(\frac{1+\mu}{2}\right)\ln\left(\frac{1+\mu}{2}\right)$$

## 1.3 Exercício 2.8

Relembrando as definições:

- Equação 1.34 do [1]

$$\mathbb{E}[f(x)] = \int p(x)f(x)dx \tag{8}$$

  Considerando $f(x)$ uma variável aleatória (da mesma forma $\mathbb{E}[x] = \int p(x)xdx$ se a VA é $x$).

- Baseado na definição da eq. 1.37 do [1], o valor esperado condicional é

$$\mathbb{E}_x[x|y] = \int p(x|y)xdx$$

  onde $x$ e $y$ são variáveis aleatórias contínuas (da mesma forma $\mathbb{E}_y[f(y)] = \int p(y)f(y)dy$).

- A regra do produto $p(x,y) = p(y|x)p(x) = p(x|y)p(y)$

- Equação 1.40 de [1]

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

- Variância Condicional

$$var_x[x|y] = \mathbb{E}_x[x^2|y] - \mathbb{E}_x[x|y]^2$$

Assim,

$$\mathbb{E}_y[\mathbb{E}_x[x|y]] = \int p(y)\left[\int p(x|y)xdx\right]dy = \int\int xp(x|y)p(y)dxdy$$

$$\mathbb{E}_y[\mathbb{E}_x[x|y]] = \int\int xp(x,y)dydx = \int xp(x)dx = \mathbb{E}[x]$$

e

$$\mathbb{E}_y[var_x[x|y]] + var_y[\mathbb{E}_x[x|y]] = \mathbb{E}_y[\mathbb{E}_x[x^2|y] - \mathbb{E}_x[x|y]^2] + \mathbb{E}_y[\mathbb{E}_x[x|y]^2] - \mathbb{E}_y[\mathbb{E}_x[x|y]]^2$$

$$\mathbb{E}_y[\mathbb{E}_x[x^2|y]] - \mathbb{E}_y[\mathbb{E}_x[x|y]^2] + \mathbb{E}_y[\mathbb{E}_x[x|y]^2] - \mathbb{E}_y[\mathbb{E}_x[x|y]]^2 = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = var[x]$$

## 1.4 Exercício 2.12

Considerando a distribuição uniforme para a variável aleatória contínua $x$ definida por

$$U(x|a,b) = \frac{1}{b-a}$$

onde $a \leq x \leq b$.

(i) Verificação se a distribuição é normal

$$\int_a^b U(x|a,b)dx = \frac{1}{b-a}\int_a^b 1dx = \left[\frac{x}{b-a}\right]_a^b = \frac{b}{b-a} - \frac{a}{b-a} = 1$$

(ii) Média

$$\mu = \int_a^b xU(x|a,b)dx = \frac{1}{b-a}\int_a^b xdx = \frac{1}{2(b-a)}\left[x^2\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}$$

(iii) Variância

$$\int_a^b (x-\mu)^2 U(x|a,b)dx = \int_a^b (x-\frac{b+a}{2})^2 \frac{1}{b-a}dx = \frac{1}{b-a}\frac{1}{3}\left[\left(x-\frac{b+a}{2}\right)^3\right]_a^b =$$

$$= \frac{1}{3(b-a)}\left[\left(b-\frac{b+a}{2}\right)^3 - \left(a-\frac{b+a}{2}\right)^3\right] = \frac{1}{3(b-a)}\left[\left(\frac{b-a}{2}\right)^3 - \left(\frac{a-b}{2}\right)^3\right] = \frac{2(b-a)^3}{24(b-a)} = \frac{(b-a)^2}{12}$$

## 1.5 Exercício 2.13

De acordo com [1], a **entropia relativa** ou divergência de *Kullback-Leibler* é dada pela seguinte expressão (equação 1.113 do [1])

$$KL(p||q) = -\int p(\mathbf{x}) \ln q(\mathbf{x})d\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})d\mathbf{x}\right) = -\int p(\mathbf{x}) \ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} d\mathbf{x}$$

a) Se $q(\mathbf{x}) = p(\mathbf{x})$:

$$KL(p||q) = 0$$

Portanto, não há divergência.

b) Considerando que ambas as pdfs possuem a mesma média ($\boldsymbol{\mu} = \mathbf{m}$):

$$KL(p||q) = -\int p(\mathbf{x}) \ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} d\mathbf{x} = \int p(\mathbf{x}) \left\{\ln q(\mathbf{x}) - \ln p(\mathbf{x})\right\}$$

A partir da definição 2.118 do [1], para $N = 1$:

$$\ln p(\mathbf{x}) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

e

$$\ln q(\mathbf{x}) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{L}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

Então

$$\ln q(\mathbf{x}) - \ln p(\mathbf{x}) = -\frac{1}{2}\ln|\mathbf{L}| + \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

Multiplicando os termos acima por $p(\mathbf{x})$ e realizando a integração:

$$-\frac{1}{2}\ln|\mathbf{L}| + \frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\int p(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu})d\mathbf{x} + \frac{1}{2}\int p(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})d\mathbf{x}$$

Para realizar as integrações acima, serão utilizadas as equações 2.44, 2.50 e 2.56 do [1].

$$(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{L}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \Delta^2 = \sum_{i=1}^{D}\frac{y_i}{\lambda_i}$$

e

$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D}\frac{1}{(2\pi\lambda_j)^{1/2}}\exp\left\{\frac{-y_j^2}{2\lambda_j}\right\}$$

Onde $y_i = \mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu})$ é uma nova coordenada de referência. Assim as integrações acima ficam de seguinte maneira:

$$\int \Delta^2 p(\mathbf{x})d\mathbf{x} = \int \sum_{i=1}^{D}\frac{y_i}{\lambda_i}\prod_{j=1}^{D}\frac{1}{(2\pi\lambda_j)^{1/2}}\exp\left\{\frac{-y_j^2}{2\lambda_j}\right\}dy_j$$

$$\int \Delta^2 p(\mathbf{x})d\mathbf{x} = \sum_{i=1}^{D}\int \frac{y_i}{\lambda_i}\prod_{j=1}^{D}\frac{1}{(2\pi\lambda_j)^{1/2}}\exp\left\{\frac{-y_j^2}{2\lambda_j}\right\}dy_j = \sum_{i=1}^{D}\prod_{j=1}^{D}\frac{1}{(2\pi\lambda_j)^{1/2}}\int \frac{y_i}{\lambda_i}\exp\left\{\frac{-y_j^2}{2\lambda_j}\right\}dy_j$$

Como foi verificado no exercício 1.7 da [2].

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

Então

$$\int \Delta^2 p(\mathbf{x})d\mathbf{x} = \sum_{i=1}^{D} 1 = D$$

Portanto

$$KL(p||q) = -\frac{1}{2}\ln|\mathbf{L}| + \frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{D}{2} + \frac{D}{2}$$

$$KL(p||q) = -\frac{1}{2}\ln|\mathbf{L}| + \frac{1}{2}\ln|\mathbf{\Sigma}|$$

## 1.6 Exercício 2.15

A gaussiana multivariada é dada pela seguinte expressão (eq. 2.43 do [1]):

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{\mathbf{\Sigma}^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{9}$$

E, a partir da eq.1.104 do [1], entropia é calculada como

$$\mathbf{H}[\mathbf{x}] = -\int p(\mathbf{x})\ln p(\mathbf{x})d\mathbf{x} \tag{10}$$

Assim,

$$\mathbf{H}[\mathbf{x}] = -\int p(\mathbf{x})\left\{-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Como foi verificado no exercício 2.13 que

$$\int p(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})d\mathbf{x} = D$$

Então

$$\mathbf{H}[\mathbf{x}] = \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{\Sigma}| + \frac{D}{2}$$

### 1.7 Exercício 2.20

Para a resolução desta questão será utilizada e expressão dos autovetores da matriz de covariância $\mathbf{\Sigma}$ (eq.2.45 do [1])

$$\mathbf{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{11}$$

Para que $\mathbf{\Sigma}$ seja positivo definido, $\lambda_i > 0$. Considerando $\mathbf{a} = \sum_i \alpha_i \mathbf{u}_i$

$$\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} = \mathbf{a}^T\mathbf{\Sigma}\sum_i \alpha_i\mathbf{u}_i = \mathbf{a}^T\sum_i \alpha_i\mathbf{\Sigma}\mathbf{u}_i = \mathbf{a}^T\sum_i \alpha_i\lambda_i\mathbf{u}_i$$

$$\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} = \sum_j \alpha_j\mathbf{u}_j^T\sum_i \alpha_i\lambda_i\mathbf{u}_i = \sum_i \alpha_i^2\lambda_i$$

Desta forma, $\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} > 0$ se $\lambda_i > 0$.

Nota: Os autovetores são escolhidos no formato ortonormal, ou seja (eq. 2.46 do [1]):

$$\mathbf{u}_j^T\mathbf{u}_i = \begin{cases} 1, & \text{se } i = j \\ 0, & \text{se } i \neq 0 \end{cases}$$

## 2 Exercícios computacionais

Os exercício foram realizados na linguagem de programação **Python v3.8.18**, através do **Notebook Jupyter v.6.5.2** (Gerenciador Anaconda).

Neste relatório estão apresentados uma breve discussão sobre os resultados encontrados. O código completo de cada exercício se encontram nos anexos (Impressão do Notebook Jupyter).

### 2.1 Exercício E1 - Inferência Bayesiana Sequencial

Como foi estudado no Cap.2 do [1], a distribuição Beta conjuga bem com a distribuição Binomial. E estas distribuições são utilizadas quando as **variáveis aleatórias discretas binárias**.

Primeiramente, estas distribuições foram estudadas separadamente simulando as figuras 2.1 e 2.2 do [1].

A distribuição Beta está reproduzida na Figura 1 e, a binomial, na Figura 2.

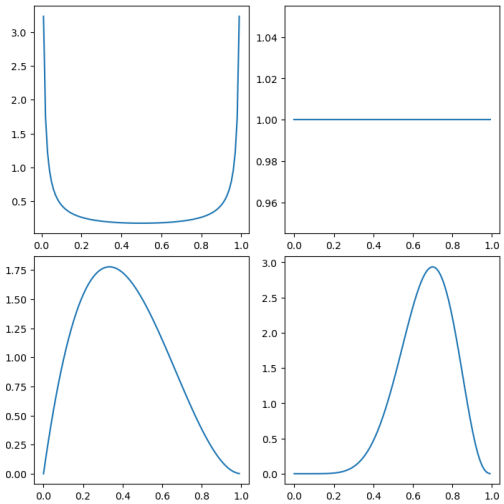

Figura 1: Verificação da distribuição Beta - reprodução da figura 2.2 do [1]

Após verificar cada distribuição, a figura 2.3 do [1] foi reproduzida para uma ensaio de 1 jogada da moeda, tanto para $m = 1$ (cara) como para $m = 0$ (corroa). Aplicou-se a inferência Baysiana considerando a distribuição a **priori** como Beta e **verossimilhança** com a distribuição Binomial.

Neste caso foram considerados $a = b = 2$ e $p(x) = 0.5$. Os resultados estão na Figura 3.

A partir de agora o modelo está pronto e, portanto, serão simuladas 5 jogadas amostradas da distribuição **Bernoulli** com probabilidade $\mu = 0.7$ de cair cara ($m = 1$). As jogadas aleatórias são $\mathbf{t} = \{1, 1, 0, 0, 1\}$.

O experimento foi repetido para o caso $a = b = 2$ e $a = b = 1$. Ambos considerando $p(x) = 0.5$. Os resultados de todas a jogadas estão no anexo E.1 deste exercício.

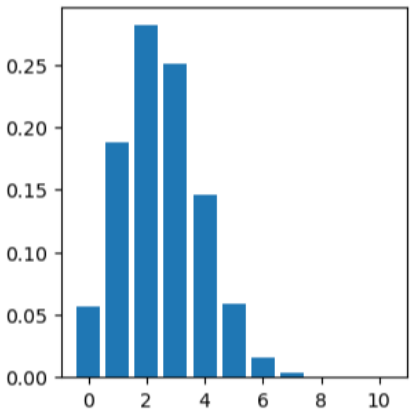


Figura 2: Verificação da distribuição Binomial - reprodução da figura 2.1 do [1]

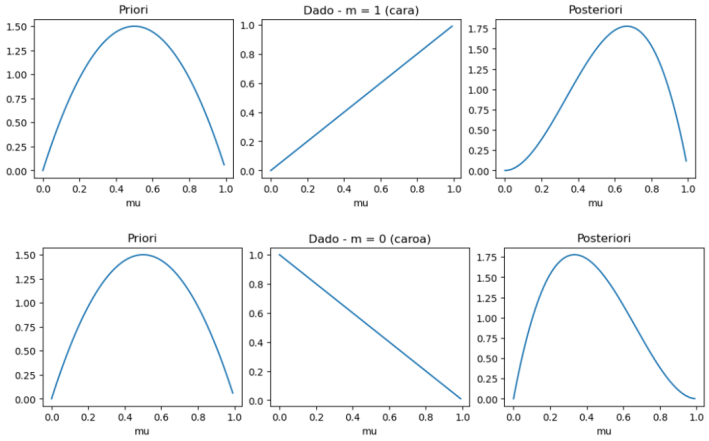


Figura 3: Verificação de 1 jogada (cara ou coroa) - reprodução da figura 2.3 do [1] - $a = b = 2$ e $\mu = 0.5$

Com base na simulação de cinco jogadas de moedas, utilizando uma distribuição Beta como distribuição a priori e a Binomial como verossimilhança, foi possível observar como a informação prévia influencia a construção da distribuição a posteriori, especialmente na primeira jogada, a distribuição a priori exerce forte influência sobre os resultados inferenciais, refletindo-se diretamente na forma da posteriori obtida.

No entanto, à medida que o número de jogadas aumenta, a influência da distribuição a priori diminui gradualmente, fazendo com que distribuições a priori diferentes conduzam a distribuições a posteriori cada vez mais semelhantes. Isso ilustra um dos princípios fundamentais da inferência bayesiana: com dados suficientes, a verossimilhança tende a dominar a influência da priori, levando à convergência da distribuição a posteriori independentemente da escolha inicial da distribuição a priori.

## 2.2 Exercício E2 - Verificação experimental do Teorema Central do Limite

Para este exercício foram consideradas as seguintes distribuições:

- Uniforme (0,1) no intervalo 0 a 1.
- Bernoulli com probabilidade $\mu = 0.4$.

Os detalhes estão apresentados no anexo E.2. O resultado final da vefificação do Teorema Central do Limite se encontra na Figura 4.

Ao somar 500 variáveis aleatórias independentes com distribuições Uniforme e Bernoulli, foi possível observar empiricamente a convergência das distribuições resultantes para uma forma aproximadamente Gaussiana. Esse comportamento é uma ilustração prática do Teorema Central do Limite, que afirma que, sob certas condições, a soma (ou média) de um grande número de variáveis aleatórias independentes tende a seguir uma distribuição Normal, independentemente da distribuição original de cada variável.

O experimento mostra que, tanto variáveis originalmente contínuas (como as Uniformes) quanto discretas (como as Bernoulli) obedecem a esse princípio, desde que sejam independentes e tenham variância finita.

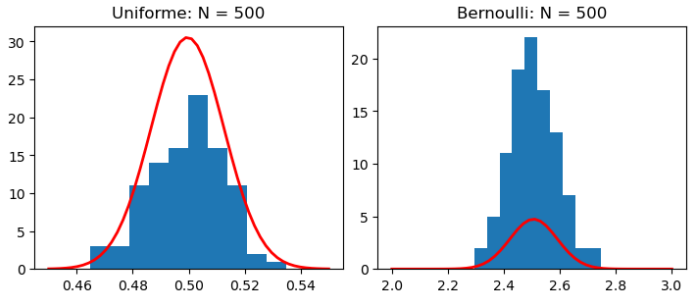


Figura 4: Verificação do Teorema Central do Limite - $N = 500$ variáveis com $n = 100$ registros.

Assim, o exercício reforça a importância do Teorema Central do Limite como ferramenta teórica e prática para justificar o uso da distribuição Normal em diversas aplicações estatísticas.

## 2.3 Exercício E3 -Verificação experimental do *Law of Large Numbers* - LLN

Para a verificação experimental da Lei dos grandes números foram consideradas $N = 500$ variáveis aleatórias com $n = 100$ registros a partir de uma distribuição normal padrão (Gaussiana de média 0 e variância 1).

O código completo encontra-se no anexo E.3.

O resultado para $N = 1$ e $N = 500$ está na Figura 5

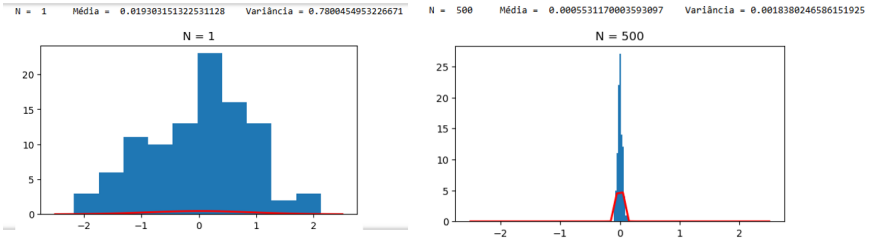


Figura 5: Verificação da Lei dos Grandes Números - Comparativo entre $N = 1$ e $N = 500$ variáveis com $n = 100$ registros.

Observou-se que a média amostral de cada conjunto tende a se concentrar em torno do valor esperado à medida que o número de amostras aumenta. Isso se refletiu no formato do histograma do estimador, que se tornou progressivamente mais estreito conforme $N$ crescia, evidenciando a diminuição da variância e o aumento da precisão da estimativa da média populacional. Esse comportamento está de acordo com o previsto teoricamente, reforçando a compreensão empírica da Lei dos Grandes Números.

## 2.4 Exercício E4 - Estimação de pdf

Os **métodos paramétricos** de estimativa de densidade de probabilidade (pdf) assumem que os dados seguem uma distribuição conhecida (como normal, exponencial, etc.), e a tarefa é estimar os parâmetros dessa distribuição com base nos dados. No entanto, quando não se quer assumir uma forma específica para a distribuição, utilizam-se **métodos não paramétricos**, como o histograma e o método kernel. O histograma é uma técnica simples e intuitiva: ele divide o eixo dos dados em intervalos (ou *bins*) de largura fixa e conta quantas observações caem em cada intervalo. A densidade é então estimada pela frequência relativa de observações em cada *bin* dividida pela largura do *bin*. Apesar de sua simplicidade, o histograma depende fortemente da escolha da largura e da posição dos *bin*, o que pode afetar significativamente a suavidade e a precisão da estimativa.

O método kernel, por sua vez, oferece uma abordagem mais refinada e contínua para estimar a densidade. Em vez de contar observações em intervalos fixos, ele coloca uma função de suavização (chamada de **função kernel**, como a gaussiana) centrada em cada ponto de dado. A estimativa da densidade em qualquer ponto é obtida somando essas funções kernel, ponderadas por um parâmetro de suavização conhecido como largura de banda (ou *bandwidth*). Essa técnica produz curvas de densidade suaves e é menos sensível à escolha da posição dos pontos do que o histograma. No entanto, a escolha adequada da largura de banda é crucial:

valores muito pequenos levam a uma estimativa muito ruidosa, enquanto valores muito grandes podem ocultar características importantes da distribuição dos dados.

Foi gerada uma amostra com $N = 50$ dados cuja a distribuição é dada pela mistura de 2 gaussianas de acordo com a equação 2.188 do [1] (representação da curva verde das figuras 2.24 e 2.25).

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{12}$$

A pdf do modelo gerador dos dados foi estimado utilizando o histograma e o kernel Gaussiano. Foram utilizados os mesmos parâmetros $h$ das figuras reproduzidas do livro([1]). O código completo está no anexo E.4.

Os resultados estão na Figura 6.

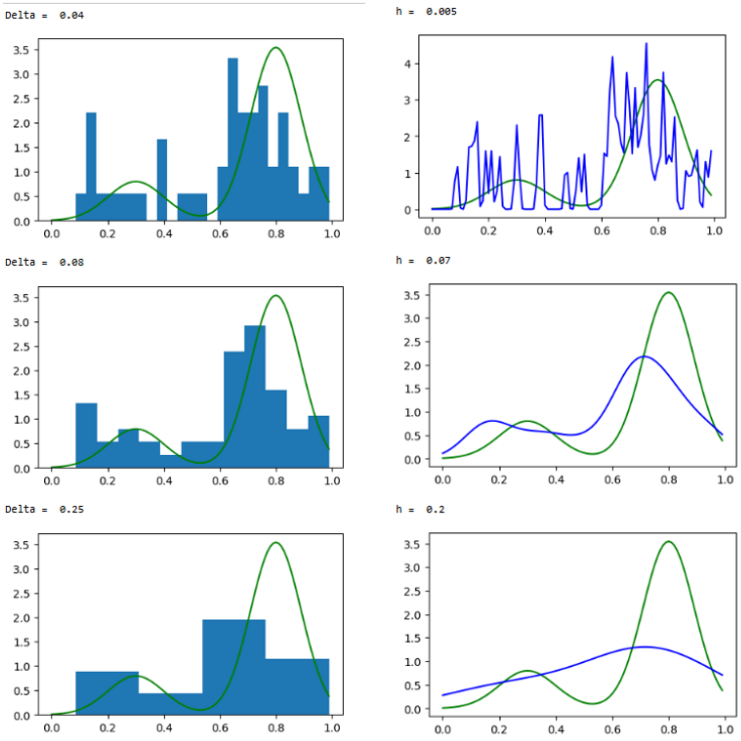


Figura 6: Estimativa de pdf pelo histograma(esquerdo) e pelo método kernel(direito).

## 2.5 Exercício E5 - Classificador K-NN

O KNN (*K-Nearest Neighbors*) é um modelo de classificação baseado em instâncias que não faz suposições explícitas sobre a distribuição dos dados. Ele funciona identificando os $k$ vizinhos mais próximos de um ponto de teste com base em uma métrica de distância (como a Euclidiana) e classificando o ponto com a classe mais comum entre esses vizinhos. O valor de $k$ influencia diretamente o desempenho: valores baixos podem levar a classificações ruidosas, enquanto valores muito altos podem suavizar demais as fronteiras entre classes.

Por ser um algoritmo simples e intuitivo, o KNN é amplamente usado em problemas de reconhecimento de padrões e classificação. No entanto, ele pode ser computacionalmente custoso em grandes conjuntos de dados, já que precisa calcular a distância entre o ponto de teste e todos os pontos do conjunto de treino. Além disso, o desempenho do KNN pode ser sensível à escala dos dados, tornando importante a normalização ou padronização prévia.

Foram consideradas 2 classes $C_1$ em vermelho, com modelo gerador gaussiano ($\mu = -1$ / $\sigma^2 = 1$) e, $C_2$ em azul, com modelo gerador gaussiano ($\mu = 1$ / $\sigma^2 = 1$). Foram geradas 10 observações de cada classe. Os dados de treinamento estão da Figura 7

A seguir foram geradas 4 novas observações aleatórias com média e variância aleatórias, que foram chamados de dados de teste, já que serão classificados com o modeo K-NN treinado a partir dos 20 pontos iniciais.

Os novos dados, ainda não classificados estão representados em com "×"sem definição de cor na Figura 8.

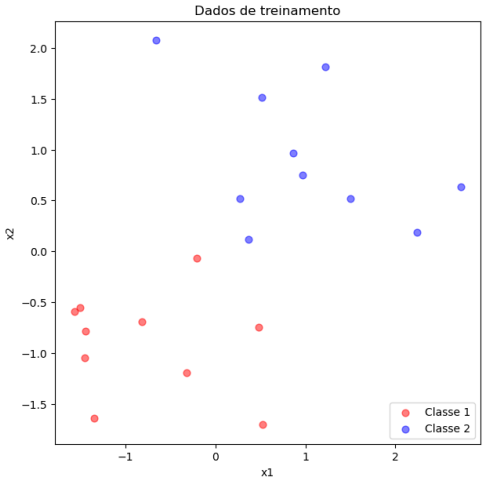


Figura 7: Observações geradas para classificação (10 pontos para cada classe).

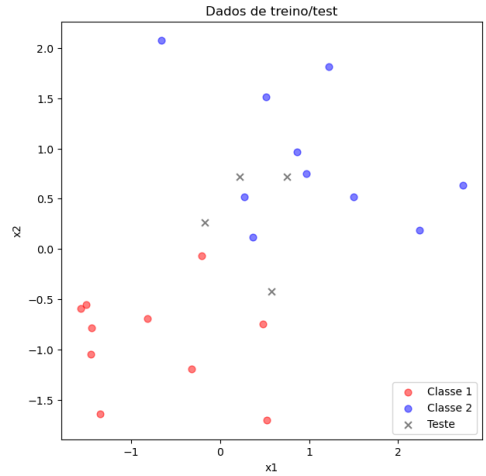


Figura 8: Dados de treino(20 pontos) e 4 dados não classificados de teste.

Os novos 4 dados foram classificados com $K = \{1, 2, 3, 5\}$. Todos os detalhes da análise e código estão no anexo E.5. Observou-se que a classificação de um ponto modificou com a mudança do parâmetro $K$. A classificação de um dado no KNN pode mudar significativamente dependendo da quantidade de vizinhos escolhida, ou seja, o valor de $K$. Esse parâmetro define quantos vizinhos mais próximos (do ponto de teste) o algoritmo irá considerar para tomar a decisão de qual classe atribuir. Serão apresentados os resultados que apresentaram impacto na classificação ($K = \{2, 3\}$). Para valores de $k$ superior a 2 o ponto destacado na Figura 9 passou da classe 1 para a classe 2.

Quando $k$ é pequeno (ex: k = 1), o algoritmo considera apenas o vizinho mais próximo. Isso o torna muito sensível ao ruído e a *outliers*, pois uma única amostra próxima — mesmo que esteja isolada ou incorreta — pode determinar a classe do ponto testado. Por isso, ocorre o *overfitting*, já que o modelo se adapta demais às pequenas variações dos dados de treino.

Quando $k$ é grande (ex: k = 15 ou 20), o modelo se torna mais estável e resistente a *outliers*, pois a decisão passa a refletir uma média mais geral do comportamento das classes próximas.

Porém, se $k$ for grande demais, o modelo pode começar a incluir pontos de classes diferentes que estão mais distantes, resultando em *underfitting* — ou seja, uma perda de precisão nas fronteiras entre as classes. Por isso, escolher um $k$ adequado é crucial. Em geral, usa-se validação cruzada para encontrar o melhor valor de $k$ com base no desempenho do modelo nos dados de validação.

Além da escolha do valor de $k$, outro fator que influencia diretamente o resultado da classificação no K-NN é a distribuição das classes no conjunto de dados. Quando há um desequilíbrio entre as classes (por exemplo, uma classe aparece muito mais que as outras), o K-NN pode tender a favorecer a classe majoritária, especialmente com valores maiores de $k$. Para mitigar esse efeito, pode-se usar técnicas como ponderar os votos dos vizinhos pela distância (dando mais peso aos vizinhos mais próximos) ou aplicar métodos de balanceamento, como *undersampling*, *oversampling*. Isso garante que a decisão do K-NN seja mais justa e eficaz, mesmo em contextos com distribuição desigual das classes.

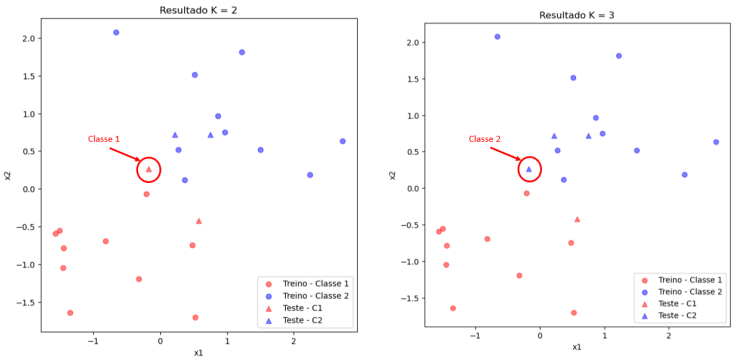


Figura 9: Resultado do classificador K-NN.

Em resumo, o K-NN é um algoritmo simples e eficaz para tarefas de classificação, mas sua performance depende fortemente da escolha adequada do número de vizinhos (k), da escala dos dados e da distribuição das classes. Valores pequenos de $k$ tornam o modelo mais sensível a ruídos, enquanto valores grandes podem comprometer sua capacidade de capturar padrões locais. Além disso, em contextos com classes desbalanceadas, estratégias adicionais como ponderação por distância ou balanceamento de dados podem ser essenciais. Quando bem ajustado, o K-NN pode oferecer bons resultados e servir como uma base comparativa sólida para modelos mais complexos.

# Referências

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

[2] Vivian de Carvalho Rodrigues. Coe782 lista 1. Technical report, COPPE/UFRJ, 2025.