Fundamentos de Probabilidade

1. Motivação

Grosso modo, pode-se dizer que a noção de probabilidade é central no âmbito do aprendizado de máquina porque informação tem a ver com incerteza. Num mundo de certezas plenamente cognoscíveis, de que forma poderia haver comunicação ou aprendizado?

Cabe a nós, portanto, recordar alguns conceitos fundamentais da teoria de probabilidade. Também analisaremos uma construção dela derivada, a chamada teoria da informação (IT, do inglês *information theory*).

1.1. Alguns Conceitos

Consideremos um *experimento aleatório*, ou seja, um experimento cujo resultado não pode ser determinado *a priori* com certeza absoluta. Ocupemo-nos de algumas definições:

- O *i-ésimo* ***resultado (outcome)*** de um experimento será denotado por ξ_i. Por exemplo, no lançamento de uma moeda, podemos ter ξ_1 = *'cara'* e ξ_2 = *'coroa'*. No lançamento de um dado, podemos ter ξ_1 = *'face com um ponto'*, ξ_2 = *'face com dois pontos'* etc.
- Um **evento** E é um conjunto de resultados de um experimento. Considerando o exemplo da moeda do item anterior, um evento possível seria E = $\{\xi_1\}$. Esse seria um *evento simples*, formado por um único resultado. No exemplo do dado do item anterior, um evento possível seria E = $\{\xi_1, \xi_3, \xi_5\}$. Nesse caso, podemos associar o evento à seguinte ideia: o valor associado à face corresponde a

um número ímpar. Também se pode definir o evento impossível como sendo E = ∅ (conjunto vazio).

- O conjunto formado por todos os possíveis resultados de um experimento é chamado de **espaço amostral**.

A partir dessas definições, podemos enunciar a base axiomática da teoria (KOLMOGOROV, 2018). Faremos isso segundo uma estrutura de três axiomas:

- Seja A um evento. Necessariamente temos $P(A) \geq 0$, sendo $P(A)$ a probabilidade a ele associada.
- Seja S o espaço amostral, acima definido. Temos $P(S) = 1$.
- Sejam A e B dois eventos disjuntos ($A \cap B = \emptyset$). Nesse caso, $P(A \cup B) = P(A) + P(B)$.

Algumas consequências desses axiomas são relativamente diretas (KAY, 2006):

Seja A^C o complemento de A (com respeito a S).

Por definição, $A^C \cap A = \emptyset$ e $A^C \cup A = S$. Usando o terceiro axioma, temos que $P\left(A^C \cup A\right) = P(A) + P\left(A^C\right) = 1$. Portanto: $P(A^C) = 1 - P(A)$.

Se tratarmos $\emptyset$ e S como complementares, usando o resultado que deduzimos, temos que $P(\emptyset) = 1 - P(S) = 0$.

Para dois eventos quaisquer A e B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

1.2. Probabilidade Condicional e Independência Estatística

O conceito de probabilidade condicional será muito importante para nós. Estabeleçamos uma definição básica:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Costuma-se dizer que $P(A|B)$ é a probabilidade condicional de A dado B. Por uma questão de simplicidade, podemos escrever, numa veia booleana, $P(A \cap B)$ como $P(AB)$. Naturalmente, vale também:

$$P(B|A) = \frac{P(AB)}{P(A)}$$

Usando essas duas equações e notando que $P(AB)$ é igual em ambas, temos:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Esse resultado é conhecido como regra (ou teorema) de Bayes.

Parece válido que tomemos um pouco de ar antes de seguir em frente. Primeiro, reflitamos sobre a probabilidade condicional. Condicionar um evento à ocorrência de outro permite que se identifique algum tipo de dependência entre ambos. Por exemplo, consideremos qual seria a probabilidade de uma pessoa qualquer do planeta falar português. Agora, consideremos qual seria a probabilidade de uma

pessoa falar português dado que ela nasceu no Brasil. Claramente, espera-se que haja uma diferença significativa. Isso se explica pelo fato de que há uma dependência entre "falar português" e "nascer no Brasil" (o local de nascimento influencia o eventual domínio de um idioma).

Numa perspectiva mais matemática, consideremos o lançamento de um dado honesto. O espaço amostral é composto por seis eventos elementares: S = {um, dois, três, quatro, cinco, seis}, e a probabilidade de ocorrência de qualquer um deles é 1/6. Suponha que consideremos a seguinte questão: qual é a probabilidade de o dado mostrar a face "dois"? A resposta é 1/6. Agora, suponha que consideremos qual seria a probabilidade de o dado mostrar a face "dois" sabendo que o resultado do lançamento corresponde a um número par. Nesse caso, podemos pensar: se é par, tem de ser "dois", "quatro" ou "seis". Como esses casos são equiprováveis, a resposta deve ser 1/3.

Associemos, então, o evento A à ocorrência de "dois" e B à ocorrência de "um número par". Temos que $P(AB) = 1/6$ e $P(B) = 1/2$. Portanto,

$$P(A|B) = 1/6 \div 1/2 = 1/3.$$

Uma situação importante surge quando $P(A|B) = P(A)$. Isso significa que a ocorrência do evento B não afeta a probabilidade de ocorrência do evento A. Usando a definição de probabilidade condicional, nesse caso, percebemos que:

$$P(AB) = P(A)P(B)$$

Quando isso ocorre, os eventos A e B são ditos **independentes**.

1.3. Variáveis Aleatórias

Em muitos casos de interesse, a ocorrência de fenômenos aleatórios se dá no contexto de valores numéricos. Consideremos, a título de exemplo, levantamentos

estatísticos, junto a uma população, de grandezas como idade, renda, altura, peso etc. Pense ainda em noções como "renda média" ou "pirâmide etária".

O formalismo para lidar com valores numéricos em probabilidade dá destaque ao conceito de **variável aleatória**. Uma variável aleatória X é, basicamente, uma função que mapeia resultados de um experimento aleatório em valores numéricos. Caso a imagem de X seja finita ou contável, tem-se uma variável aleatória **discreta**. Caso a imagem de X seja o conjunto dos reais, tem-se uma variável aleatória **contínua**.

Com o mapeamento acima descrito, os valores numéricos passam a se vincular a uma medida de probabilidade. Uma primeira forma de apresentar essa conexão é através da função de distribuição cumulativa (CDF, do inglês *cumulative distribution function*) $F_X(x)$:

$$F_X(x) = P(X \leq x)$$

Em outras palavras, a CDF nos informa a probabilidade de uma variável aleatória X assumir valores menores que um determinado x. Se considerarmos o lançamento de um dado honesto e atribuirmos números naturais às suas faces, temos:

$F_X(1) = 1/6; F_X(2) = 2/6 = 1/3; \; F_X(3) = 3/6 = 1/2; \; F_X(4) = 4/6 = 2/3; \; F_X(5) = 5/6; F_X(6) = 6/6 = 1; F_X(7) = 6/6 = 1$ etc.

Os valores-limite da CDF são 0 ("em" $-\infty$) e 1 ("em" $+\infty$). A CDF nunca decresce, uma vez que sempre "acumula probabilidade".

Quando se lida com variáveis aleatórias discretas, é possível atribuir diretamente valores de probabilidade aos valores numéricos. Nesse caso, define-se uma função de massa de probabilidade (PMF, do inglês *probability mass function*):

$$P_X(x) = P(X = x)$$

Voltando ao nosso exemplo do dado honesto, essa função teria a forma da Fig. 1.

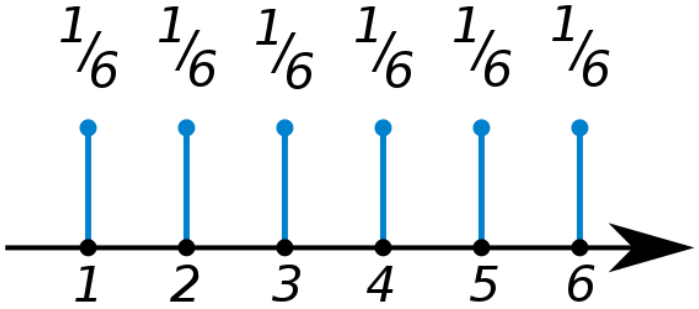


Figura 1 – Função de Massa de Probabilidade (Dado).

A CDF tem uma relação direta com a função de massa de probabilidade. Matematicamente, temos:

$$F_X(x) = \sum_{k=-\infty}^{x} P_X(k)$$

A equação nada mais é que uma descrição eloquente da ideia de acumulação de probabilidade.

No caso de variáveis contínuas, o formalismo é um pouco mais complexo. Uma vez que as possibilidades estão definidas num *continuum*, não é possível mais falar na atribuição direta de massa de probabilidade a cada valor. Fala-se na definição de

uma densidade de probabilidade $f_X(x)$. A relação com a CDF é natural (também uma acumulação):

$$F_X(x) = \int_{-\infty}^{x} f_X(\xi) d\xi$$

A relação inversa também é válida:

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Se desejarmos conhecer $P(a \leq X \leq b)$, podemos fazer:

$$P(a \leq X \leq b) = \int_{a}^{b} f_X(x) dx = F_X(b) - F_X(a)$$

Uma condição crucial "de normalização" é:

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

Um primeiro exemplo de densidade de probabilidade é a densidade gaussiana, talvez a mais célebre de todas. Sua expressão é:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

O valor μ é a média e σ² é a variância (as definições virão mais adiante). A forma dessa densidade é mostrada na Fig. 2. A forma da CDF associada é mostrada na Fig. 3.

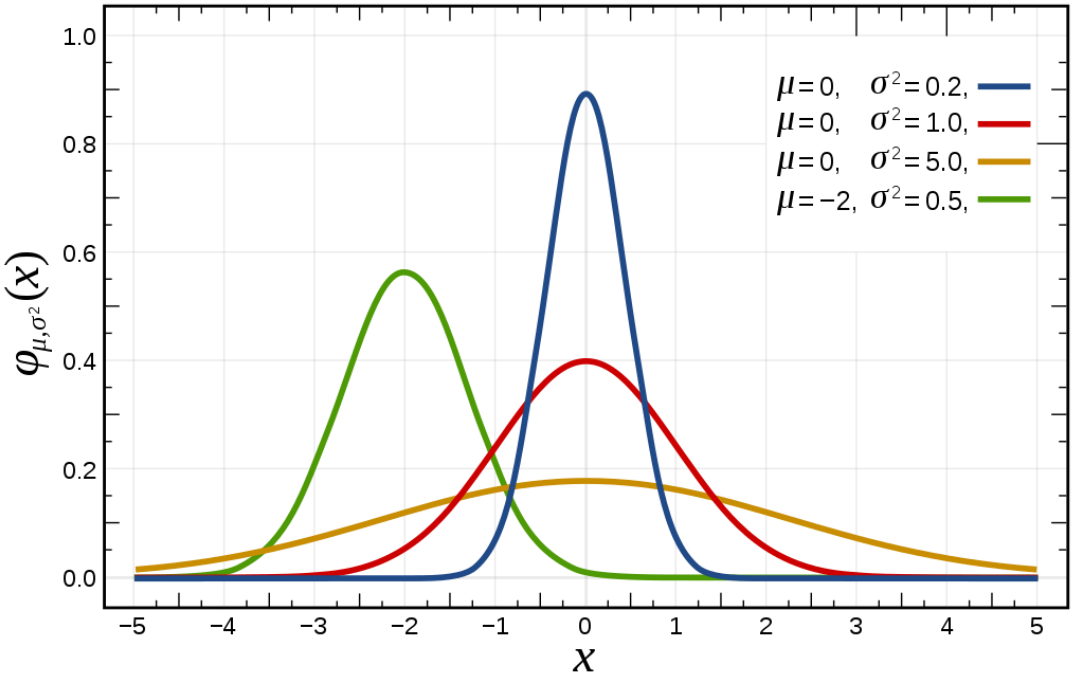


Figura 2 – Exemplos de Densidades Gaussianas.

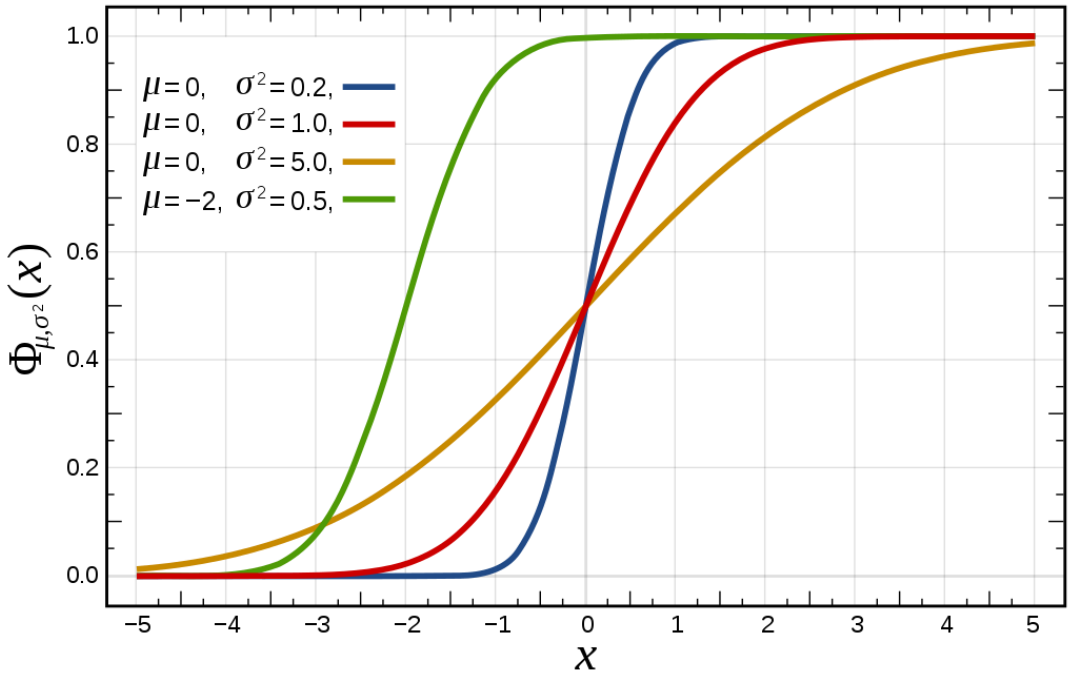


Figura 3 – Exemplos de CDFs Gaussianas.

Outra densidade muito importante é a densidade uniforme (entre dois valores, a e b). Sua definição matemática é:

$$f_X(x) = \begin{cases} 1/(b-a), & a \leq x \leq b \\ 0, & \text{caso contrário} \end{cases}$$

As Figs. 4 e 5 trazem a densidade e a função cumulativa associada.

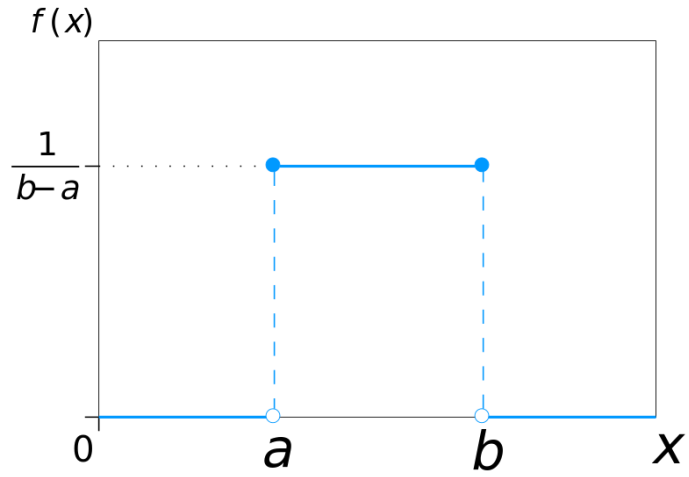


Figura 4 – Densidade Uniforme.

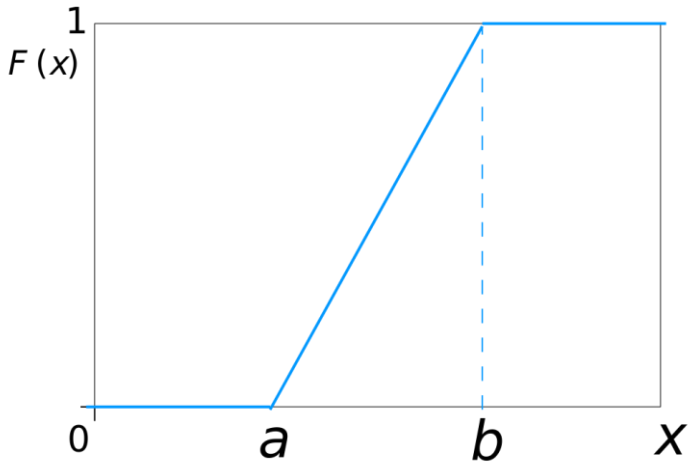


Figura 5 – Função Cumulativa Associada.

1.3.1. Valor Esperado (Esperança Matemática)

O valor esperado de uma variável aleatória é uma "média estatística". Não é uma média obtida de certo número de amostras, mas sim uma média "ideal", "platônica". Essa "idealidade" advém do conhecimento da estrutura probabilística subjacente.

Para variáveis discretas, o valor esperado, denotado pelo operador $E[\cdot]$, é:

$$E[X] = \sum_k x_k \, P_X(x_k)$$

Para variáveis contínuas, temos:

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

O valor esperado de uma função $g(X)$ pode ser obtido de forma direta:

$$E[g(X)] = \sum_k g(x_k) \, P_X(x_k)$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

1.3.2. Momentos

O n-ésimo momento de uma variável aleatória é definido como:

$$m_n = E[X^n]$$

O primeiro momento (m = 1) é o valor esperado da variável, conhecido como média (μ).

Para avaliar a excursão da variável em torno da média, lança-se mão dos *momentos centrais*. O n-ésimo momento central é:

$$c_n = E[(X - \mu)^n]$$

O segundo momento central (n = 2) é conhecido como *variância* (σ^2). O *desvio padrão* é a raiz quadrada da variância (σ).

Os momentos são importantes para a caracterização parcial da estrutura probabilística de uma variável aleatória, e terão papel central nos próximos capítulos do curso.

1.4. Várias Variáveis Aleatórias

Em muitos casos, as variáveis aleatórias devem ser consideradas conjuntamente. A razão é que elas podem apresentar dependência estatística, ou seja, ser mutuamente informativas. Apresentaremos de maneira explícita o caso de duas variáveis aleatórias, mas a extensão para um número superior é direta.

Define-se a função cumulativa da seguinte forma:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y)$$

No caso de variáveis discretas, essa função se relaciona com a função conjunta de massa de probabilidade $P_{XY}(x, y)$ da seguinte forma:

$$F_{XY}(x,y) = \sum_{k=-\infty}^{x} \sum_{m=-\infty}^{y} P_{XY}(k,m)$$

No caso de variáveis contínuas, a relação é análoga:

$$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(\xi,\upsilon) d\xi \, d\upsilon$$

Essa relação significa que:

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$$

Para obter as probabilidades associadas a cada variável, é preciso "realizar uma soma" sobre as demais. No caso de variáveis discretas, um exemplo seria:

$$P_X(x) = \sum_{y} P_{XY}(x,y)$$

Para variáveis contínuas, o exemplo seria:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$$

Também é possível definir funções de massa e de densidade condicionais:

$$P_{Y|X}(Y = y|X = x) = \frac{P_{XY}(x,y)}{P_X(x)}$$

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

Da mesma forma que antes, se $P_{XY}(x,y) = P_X(x)P_Y(y)$ ou $f_{XY}(x,y) = f_X(x)f_Y(y)$, as variáveis são estatisticamente independentes.

É importante definir ainda dois momentos, a correlação e a covariância. Ambos indicam uma relação entre variáveis aleatórias, embora esta seja menos contundente que a dependência estatística. A correlação entre variáveis X e Y é:

$$\text{corr}(X,Y) = E[XY].$$

Já a covariância é um momento conjunto central:

$$\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Se $\text{corr}(X, Y) = 0$, as variáveis são ortogonais. Se $\text{cov}(X, Y) = 0$, as variáveis são descorrelacionadas. Sempre que duas variáveis são independentes, elas também são descorrelacionadas. A implicação oposta só vale para variáveis gaussianas.

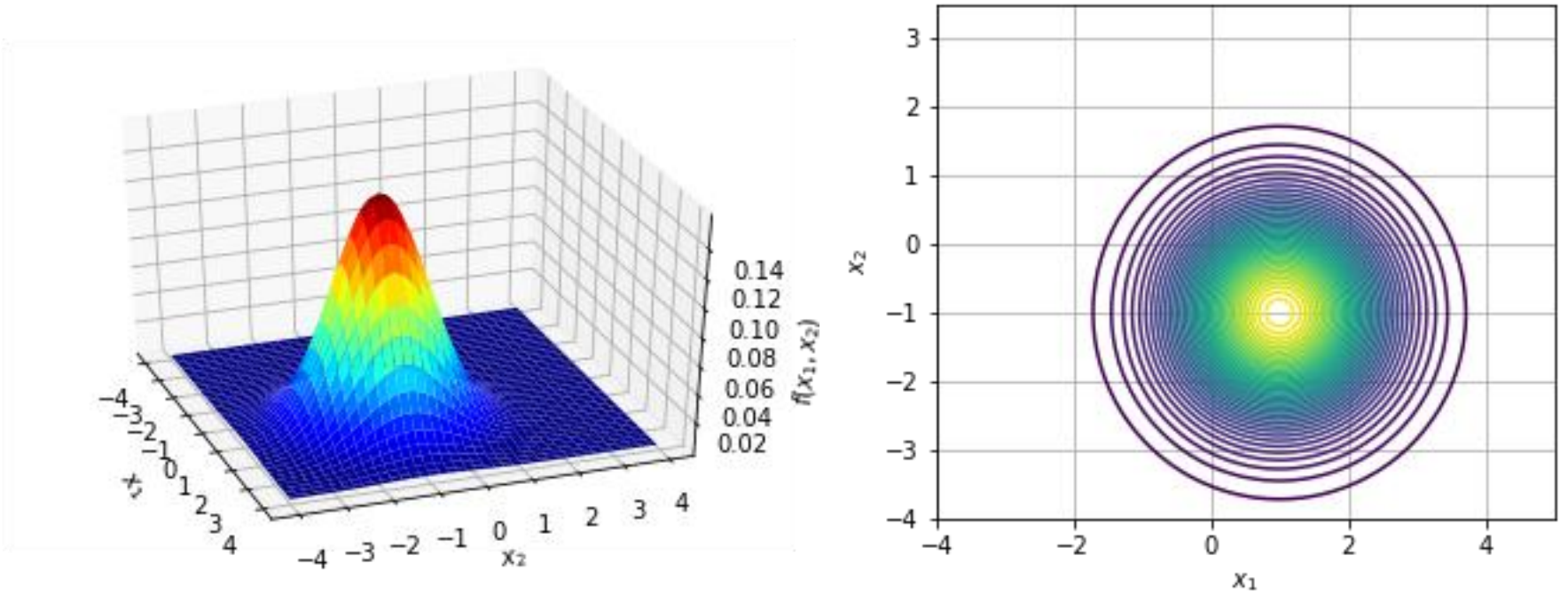
Por fim, é importante apresentar a forma matemática da densidade gaussiana multivariada:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \, exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$

Na equação, $x = [x_1 \ \ldots \ x_k]$ é o vetor de variáveis aleatórias e $\boldsymbol{\Sigma}$ é a matriz de covariância, na qual cada elemento σ_{ij} dessa matriz é dado por $\text{cov}(x_i, x_j)$. O termo $|\boldsymbol{\Sigma}|$ corresponde ao determinante da matriz $\boldsymbol{\Sigma}$.

Exemplos:

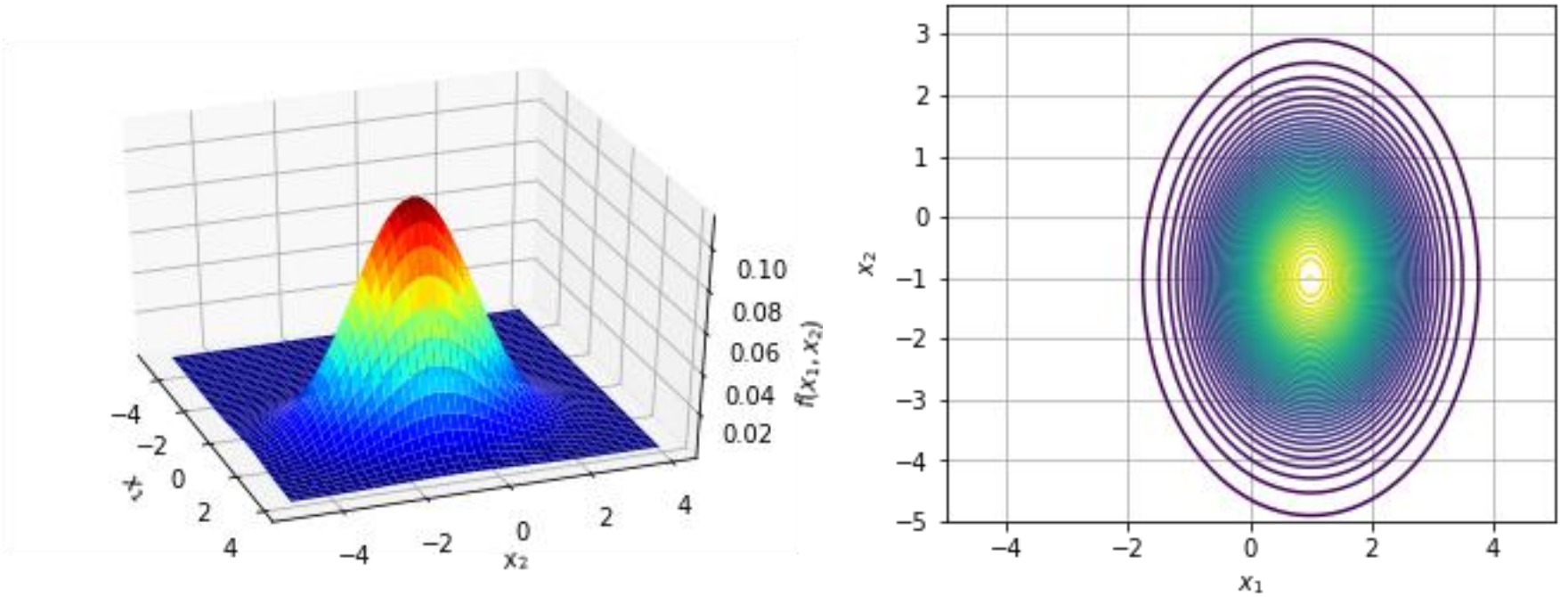
- $k = 2, \boldsymbol{\mu} = [1 \ -1]^{\mathrm{T}}$ e $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.



(a) (b)

Figura 6 – PDF Gaussiana e as respectivas curvas de nível para o caso de variáveis aleatórias independentes e de mesma variância.

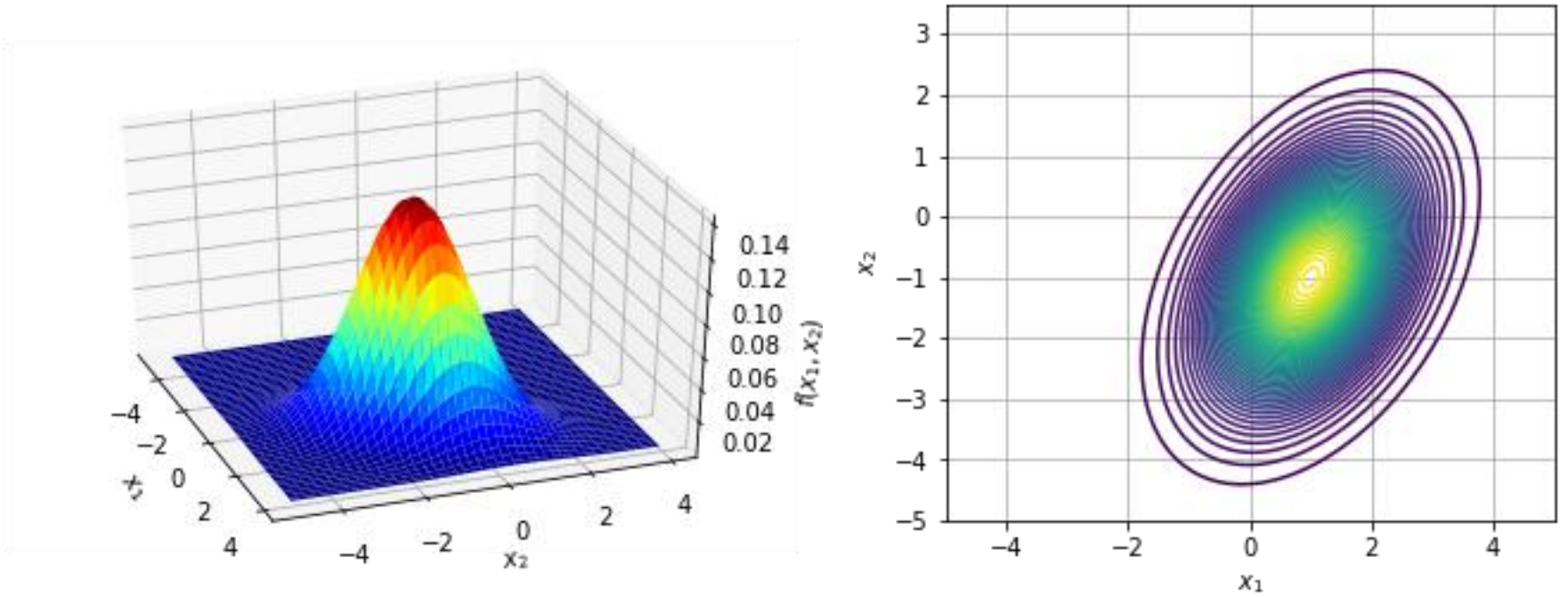
- $k = 2, \boldsymbol{\mu} = [1 \ -1]^{\mathrm{T}}$ e $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$.



(a) (b)

Figura 7 – PDF Gaussiana e as respectivas curvas de nível para o caso de variáveis aleatórias independentes e com variâncias diferentes.

- $k = 2, \boldsymbol{\mu} = [1 \ -1]^{\mathrm{T}}$ e $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0{,}5 \\ 0{,}5 & 1{,}5 \end{bmatrix}$.



(a) (b)

Figura 8 – PDF Gaussiana e as respectivas curvas de nível para o caso de variáveis aleatórias correlacionadas.

2. Referências bibliográficas

KAY, S., **Intuitive Probability and Random Processes Using MATLAB**, Springer, 2006.

KOLMOGOROV, A. N., **Foundations of the Theory of Probability**, Dover, 2018.

SHYNK, J. J., Probability, **Random Variables and Random Processes**, Wiley, 2013.