



UNIVERSIDADE
FEDERAL DO CEARÁ

DEPARTAMENTO
DE COMPUTAÇÃO

Aprendizagem de Máquina Probabilística

César Lincoln Cavalcante Mattos

2024

Agenda

- ① Revisão probabilidade e estatística
- ② Algumas distribuições de probabilidade
- ③ Aprendizagem de Máquina Probabilística
- ④ Tópicos adicionais
- ⑤ Referências

Revisão probabilidade e estatística

- Interesse em estudar a **incerteza** e medir a **chance** de eventos.

Revisão probabilidade e estatística

- Interesse em estudar a **incerteza** e medir a **chance** de eventos.

Probabilidade × Estatística (\times Aprendizagem de Máquina?)

- **Probabilidade:** modelo/regras → dados; previsões.
- **Estatística:** dados → modelo/regras; análise do que ocorreu.

Revisão probabilidade e estatística

- Interesse em estudar a **incerteza** e medir a **chance** de eventos.

Probabilidade × Estatística (\times Aprendizagem de Máquina?)

- **Probabilidade:** modelo/regras \rightarrow dados; previsões.
- **Estatística:** dados \rightarrow modelo/regras; análise do que ocorreu.

Variável aleatória

- Uma “variável” cujos possíveis valores são resultados de um fenômeno aleatório.
- Mapeia os resultados de um processo imprevisível em quantidades numéricas.
- **Classificação:** Contínua/Discreta; Univariada/Multivariada.

Revisão probabilidade e estatística

Interpretação frequentista

- Considera a **frequência relativa** de ocorrência dos eventos.
- A probabilidade é vista como a frequência relativa no **limite de infinitas observações**.
- Representa parâmetros por **valores determinísticos**.

Revisão probabilidade e estatística

Interpretação frequentista

- Considera a **frequência relativa** de ocorrência dos eventos.
- A probabilidade é vista como a frequência relativa no **limite de infinitas observações**.
- Representa parâmetros por **valores determinísticos**.

Interpretação Bayesiana

- A probabilidade **quantifica a incerteza** sobre um evento.
- A observação de dados **atualiza o grau da incerteza** anterior.
- Representa parâmetros por **distribuições de probabilidade**.

Probabilidade

- **Experimento:** Jogar uma moeda justa \rightarrow variável aleatória E .
- **Espaço amostral?**

Probabilidade

- **Experimento:** Jogar uma moeda justa \rightarrow variável aleatória E .
- **Espaço amostral?** $\Omega = \{K, C\} \rightarrow$ cara (K) ou coroa (C).

Probabilidade

- **Experimento:** Jogar uma moeda justa \rightarrow variável aleatória E .
- **Espaço amostral?** $\Omega = \{K, C\} \rightarrow$ cara (K) ou coroa (C).
- **Evento?**

Probabilidade

- **Experimento:** Jogar uma moeda justa \rightarrow variável aleatória E .
- **Espaço amostral?** $\Omega = \{K, C\} \rightarrow$ cara (K) ou coroa (C).
- **Evento?** $e = \{K\} \subset \Omega$

Probabilidade

- **Experimento:** Jogar uma moeda justa \rightarrow variável aleatória E .
- **Espaço amostral?** $\Omega = \{K, C\} \rightarrow$ cara (K) ou coroa (C).
- **Evento?** $e = \{K\} \subset \Omega$
- **Probabilidade:** $P(E = e) = N_e/N_\Omega$, em que N_e e N_Ω são os números de elementos em e e Ω .

Probabilidade

- **Experimento:** Jogar uma moeda justa \rightarrow variável aleatória E .
- **Espaço amostral?** $\Omega = \{K, C\} \rightarrow$ cara (K) ou coroa (C).
- **Evento?** $e = \{K\} \subset \Omega$
- **Probabilidade:** $P(E = e) = N_e/N_\Omega$, em que N_e e N_Ω são os números de elementos em e e Ω .
- **Note que:**
 - $\rightarrow 0 \leq P(E = e) \leq 1$
 - $\rightarrow \sum_i P(E = e_i) = 1$
 - $\rightarrow P(E \neq e) = 1 - P(E = e)$

Probabilidade

Função de massa de probabilidade

- Especifica a probabilidade de uma **v.a. discreta** X tomar um valor $x \in \Omega$.
- Denotada por $P(X = x)$, em que $\sum_i P(X = x_i) = 1$.

Probabilidade

Função de massa de probabilidade

- Especifica a probabilidade de uma **v.a. discreta** X tomar um valor $x \in \Omega$.
- Denotada por $P(X = x)$, em que $\sum_i P(X = x_i) = 1$.

Função de densidade de probabilidade

- Uma função $f : \mathbb{R}^D \rightarrow \mathbb{R}_{\geq 0}$ tal que

$$\forall \mathbf{x} \in \mathbb{R}^D : f(\mathbf{x}) \geq 0 \text{ e } \int_{\mathbb{R}^D} f(\mathbf{x}) d\mathbf{x} = 1.$$

- Seja X uma **v.a. contínua** e o intervalo $[a, b]$, $a, b \in \mathbb{R}$:

$$P(a \leq X \leq b) = \int_a^b f(x) dx = \int_a^b p(x) dx.$$

Probabilidade

Função de distribuição cumulativa

- Dada uma v.a. $X = [X_1, \dots, X_D]^\top$ com possíveis valores $\mathbf{x} = [x_1, \dots, x_D]^\top$:

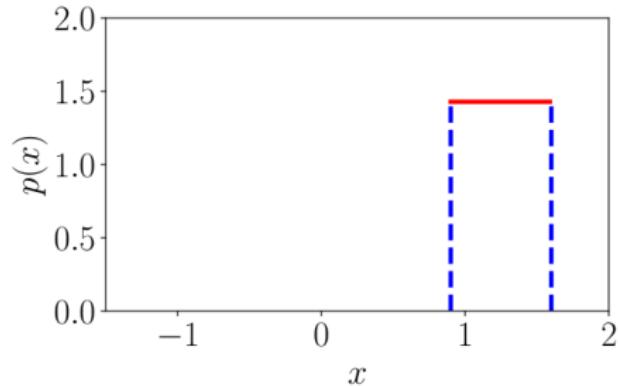
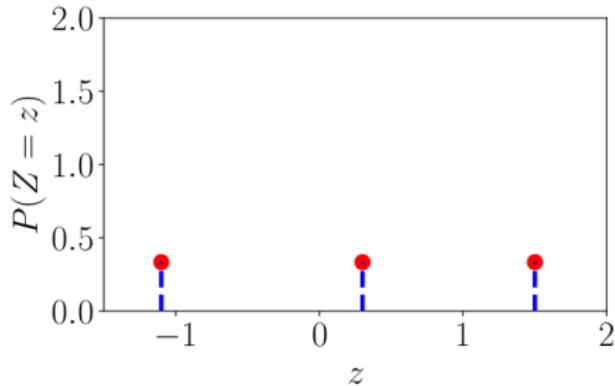
$$F_X(\mathbf{x}) = P(X_1 \leq x_1, \dots, X_D \leq x_D).$$

- Para uma v.a. **contínua**, temos ainda:

$$F_X(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} p(x'_1, \dots, x'_D) dx'_1 \cdots dx'_D.$$

Probabilidade

- **Observação:** Em distribuições discretas, os valores individuais devem estar em $[0, 1]$. Em densidades contínuas, podem haver valores > 1 .



Probabilidade

Regra da soma

- Calcula distribuições **marginais** a partir de **conjuntas**:

$$P(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}} P(\mathbf{x}, \mathbf{y}), \quad p(\mathbf{x}) = \int_{\mathcal{Y}} p(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

- Também chamada de **propriedade da marginalização**.

Probabilidade

Regra da soma

- Calcula distribuições **marginais** a partir de **conjuntas**:

$$P(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}} P(\mathbf{x}, \mathbf{y}), \quad p(\mathbf{x}) = \int_{\mathcal{Y}} p(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

- Também chamada de **propriedade da marginalização**.

Regra do produto

- Calcula distribuições **conjuntas** a partir de **condicionais e marginais**:

$$P(\mathbf{x}, \mathbf{y}) = P(\mathbf{y}|\mathbf{x})P(\mathbf{x}), \quad p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}).$$

- Permite a **fatoração** de distribuições conjuntas.

Probabilidade

Regra da soma

- Calcula distribuições **marginais** a partir de **conjuntas**:

$$P(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}} P(\mathbf{x}, \mathbf{y}), \quad p(\mathbf{x}) = \int_{\mathcal{Y}} p(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

- Também chamada de **propriedade da marginalização**.

Regra do produto

- Calcula distribuições **conjuntas** a partir de **condicionais e marginais**:

$$P(\mathbf{x}, \mathbf{y}) = P(\mathbf{y}|\mathbf{x})P(\mathbf{x}), \quad p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}).$$

- Permite a **fatoração** de distribuições conjuntas.

Observação: As regras podem ser combinadas:

$$P(\mathbf{y}) = \sum_{\mathbf{x} \in \mathcal{X}} P(\mathbf{y}|\mathbf{x})P(\mathbf{x}), \quad p(\mathbf{y}) = \int_{\mathcal{X}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) d\mathbf{x}.$$

Probabilidade

Regra de Bayes (versão discreta)

- **Relaciona distribuições condicionais** diferentes:

se $P(X, Y) = P(X|Y)P(Y)$ e

$P(X, Y) = P(Y|X)P(X)$, logo:

$$P(X|Y)P(Y) = P(Y|X)P(X)$$

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} = \frac{P(Y|X)P(X)}{\sum_{X'} P(Y|X')P(X')}$$

- Podemos **calcular probabilidades condicionais desconhecidas** a partir de quantidades disponíveis.
- Exemplo de uso: diagnóstico médico.

Probabilidade

Regra de Bayes (versão contínua)

- Relaciona distribuições condicionais diferentes:

se $p(x, y) = p(x|y)p(y)$ e

$p(x, y) = p(y|x)p(x)$, logo:

$$p(x|y)p(y) = p(y|x)p(x)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int_{\mathcal{X}} p(y|x)p(x)dx}$$

- Podemos calcular probabilidades condicionais desconhecidas a partir de quantidades disponíveis.
- Exemplo de uso: estimação de variáveis *latentes* (não-observadas) contínuas.

Probabilidade

- Cada componente da Regra de Bayes possui uma interpretação clara. Considerando o caso contínuo multivariado:

$$\underbrace{p(\mathbf{x}|\mathbf{y})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y}|\mathbf{x})}^{\text{verossimilhança priori}} \overbrace{p(\mathbf{x})}^{\text{prior}}}{\underbrace{p(\mathbf{y})}_{\text{evidência}}}$$

- **Priori**: Incerteza sobre \mathbf{x} antes da observação de \mathbf{y} .
- **Verossimilhança (likelihood)**: Descreve a relação entre \mathbf{y} e \mathbf{x} .
- **Posteriori**: Incerteza sobre \mathbf{x} após a observação de \mathbf{y} .
- **Evidência (verossimilhança marginal)**: Normaliza a posteriori e funciona como uma verossimilhança esperada:

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

Regra de Bayes - Exemplo

- Aleatoriamente, você escolhe um dos baús abaixo para abrir.
- Aleatoriamente, você pega uma moeda do baú escolhido.
- Se a moeda coletada é de ouro, qual a probabilidade de você ter escolhido o baú A?



Regra de Bayes - Exemplo

- A : Evento de escolher o baú A.
- B : Evento de escolher o baú B.
- G : Evento de pegar uma moeda de ouro.

Regra de Bayes - Exemplo

- A : Evento de escolher o baú A.
- B : Evento de escolher o baú B.
- G : Evento de pegar uma moeda de ouro.
- Aplicando a Regra de Bayes:

$$\begin{aligned} P(A|G) &= \frac{P(G|A)P(A)}{P(G)} = \frac{P(G|A)P(A)}{P(G|A)P(A) + P(G|B)P(B)} \\ &= \frac{1 \times 0.5}{1 \times 0.5 + 0.5 \times 0.5} = \frac{2}{3} \end{aligned}$$



Regra de Bayes - Exemplo 2

- Você fez um teste para verificar se possui COVID-19.

Regra de Bayes - Exemplo 2

- Você fez um teste para verificar se possui COVID-19.
- Na sua região, 1 a cada 10 pessoas contrai a doença (**prevalência**).

Regra de Bayes - Exemplo 2

- Você fez um teste para verificar se possui COVID-19.
- Na sua região, 1 a cada 10 pessoas contrai a doença (**prevalência**).
- O teste feito possui as seguintes características:
 - Se você tem a doença, o teste possui probabilidade de 87.5% de ser positivo (**sensibilidade ou taxa de verdadeiros positivos**).

Regra de Bayes - Exemplo 2

- Você fez um teste para verificar se possui COVID-19.
- Na sua região, 1 a cada 10 pessoas contrai a doença (**prevalência**).
- O teste feito possui as seguintes características:
 - Se você tem a doença, o teste possui probabilidade de 87.5% de ser positivo (**sensibilidade ou taxa de verdadeiros positivos**).
 - Se você não tem a doença, o teste possui probabilidade de 97.5% de ser negativo (**especificidade ou taxa de verdadeiros negativos**).

Regra de Bayes - Exemplo 2

- Você fez um teste para verificar se possui COVID-19.
- Na sua região, 1 a cada 10 pessoas contrai a doença (**prevalência**).
- O teste feito possui as seguintes características:
 - Se você tem a doença, o teste possui probabilidade de 87.5% de ser positivo (**sensibilidade ou taxa de verdadeiros positivos**).
 - Se você não tem a doença, o teste possui probabilidade de 97.5% de ser negativo (**especificidade ou taxa de verdadeiros negativos**).
- Seu teste deu positivo. Qual será aproximadamente a probabilidade de você ter a doença?

Regra de Bayes - Exemplo 2

- $X = 0$: Evento em que você não está doente.
- $X = 1$: Evento em que você está doente.
- $Y = 0$: Evento em que o teste deu negativo.
- $Y = 1$: Evento em que o teste deu positivo.

Regra de Bayes - Exemplo 2

- $X = 0$: Evento em que você não está doente.
- $X = 1$: Evento em que você está doente.
- $Y = 0$: Evento em que o teste deu negativo.
- $Y = 1$: Evento em que o teste deu positivo.
- Queremos encontrar $P(X = 1|Y = 1)$.

Regra de Bayes - Exemplo 2

- $X = 0$: Evento em que você não está doente.
- $X = 1$: Evento em que você está doente.
- $Y = 0$: Evento em que o teste deu negativo.
- $Y = 1$: Evento em que o teste deu positivo.
- Queremos encontrar $P(X = 1|Y = 1)$.
- Pela Regra de Bayes:

$$P(X = 1|Y = 1) = \frac{P(Y = 1|X = 1)P(X = 1)}{P(Y = 1)}$$

$$\rightarrow P(X = 1) = 1/10 = 0.1$$

$$\rightarrow P(Y = 1|X = 1) = 0.875$$

$$\rightarrow P(Y = 1) = ?$$

Regra de Bayes - Exemplo 2

- A probabilidade $P(Y = 1)$ pode ser calculada usando outras condicionais:

$$\begin{aligned}P(Y = 1) &= P(Y = 1|X = 1)P(X = 1) + P(Y = 1|X = 0)P(X = 0) \\&= 0.875 \times 0.1 + (1 - 0.975) \times (1 - 0.1) \\&= 0.875 \times 0.1 + 0.025 \times 0.9 \\&= 0.11\end{aligned}$$

Regra de Bayes - Exemplo 2

- A probabilidade $P(Y = 1)$ pode ser calculada usando outras condicionais:

$$\begin{aligned}P(Y = 1) &= P(Y = 1|X = 1)P(X = 1) + P(Y = 1|X = 0)P(X = 0) \\&= 0.875 \times 0.1 + (1 - 0.975) \times (1 - 0.1) \\&= 0.875 \times 0.1 + 0.025 \times 0.9 \\&= 0.11\end{aligned}$$

- Voltando para a Regra de Bayes:

$$\begin{aligned}P(X = 1|Y = 1) &= \frac{P(Y = 1|X = 1)P(X = 1)}{P(Y = 1)} \\&= \frac{0.875 \times 0.1}{0.11} \approx 0.795\end{aligned}$$

≈ 79.5% de chance de estar doente, dado que o teste foi positivo.

Regra de Bayes - Exemplo 2

- Seu teste deu negativo. Qual será aproximadamente a probabilidade de você ter a doença?

Regra de Bayes - Exemplo 2

- Seu teste deu negativo. Qual será aproximadamente a probabilidade de você ter a doença?
- Queremos encontrar $P(X = 1 | Y = 0)$.

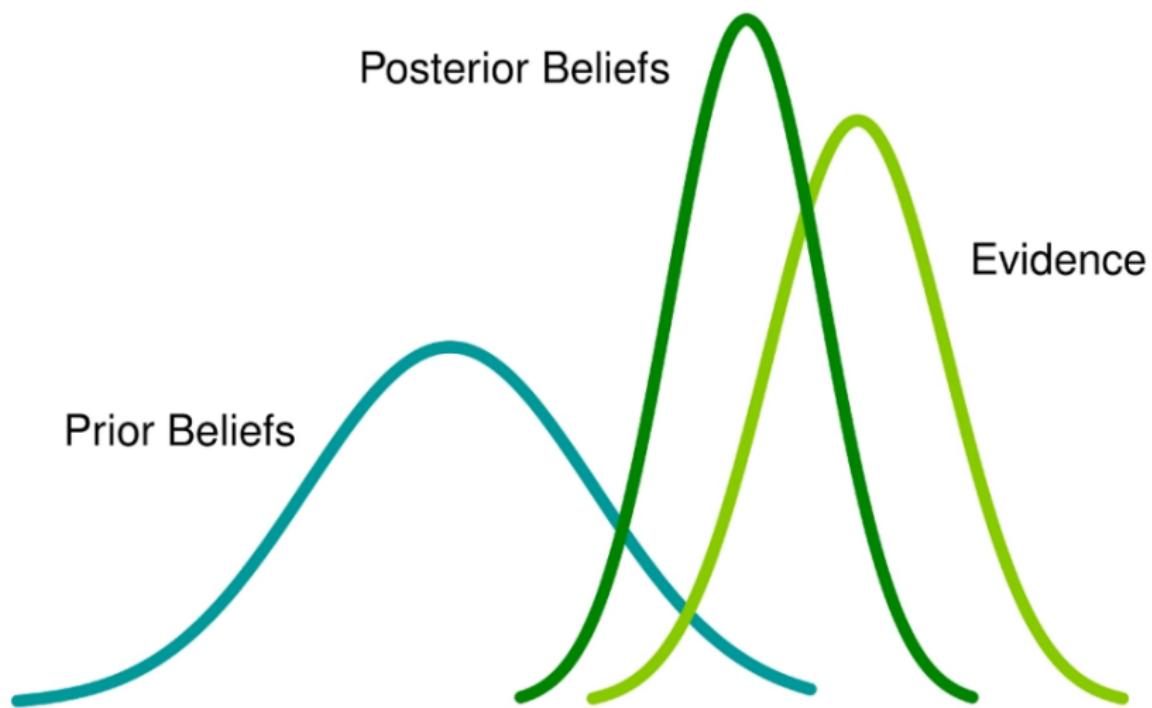
Regra de Bayes - Exemplo 2

- Seu teste deu negativo. Qual será aproximadamente a probabilidade de você ter a doença?
- Queremos encontrar $P(X = 1|Y = 0)$.
- Voltando para a Regra de Bayes:

$$\begin{aligned} P(X = 1|Y = 0) &= \frac{P(Y = 0|X = 1)P(X = 1)}{P(Y = 0)} \\ &= \frac{(1 - 0.875) \times 0.1}{1 - 0.11} \approx 0.014 \end{aligned}$$

$\approx 1.4\%$ de chance de estar doente, dado que o teste foi negativo.

Regra de Bayes - Ilustração



Estatísticas de caracterização

- Estatística populacionais × estatísticas empíricas (ou amostrais).

Estatísticas de caracterização

- Estatística populacionais × estatísticas empíricas (ou amostrais).

Descritores de Tendência Central

- **Média:** Valor esperado da variável aleatória:

$$\mathbb{E}_X[x] = \sum_{x \in \mathcal{X}} xP(x), \quad \mathbb{E}_X[x] = \int_{\mathcal{X}} xp(x)dx$$

- **Caso multivariado:** $\mu = \mathbb{E}_X[\mathbf{x}] = [\mathbb{E}_{X_1}[x_1], \dots, \mathbb{E}_{X_D}[x_D]]^\top$
- **Estimador:** $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$

Estatísticas de caracterização

- Estatística populacionais × estatísticas empíricas (ou amostrais).

Descritores de Tendência Central

- **Média:** Valor esperado da variável aleatória:

$$\mathbb{E}_X[x] = \sum_{x \in \mathcal{X}} xP(x), \quad \mathbb{E}_X[x] = \int_{\mathcal{X}} xp(x)dx$$

- **Caso multivariado:** $\mu = \mathbb{E}_X[\mathbf{x}] = [\mathbb{E}_{X_1}[x_1], \dots, \mathbb{E}_{X_D}[x_D]]^\top$
- **Estimador:** $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$

- **Mediana:** Valor central da variável aleatória discreta ou valor em que $F_X(x) = \int_{-\infty}^x p(x')dx' = 0.5$, para o caso contínuo.
 - **Estimador:** Ordene as observações e escolha o valor central.

Estatísticas de caracterização

- Estatística populacionais × estatísticas empíricas (ou amostrais).

Descritores de Tendência Central

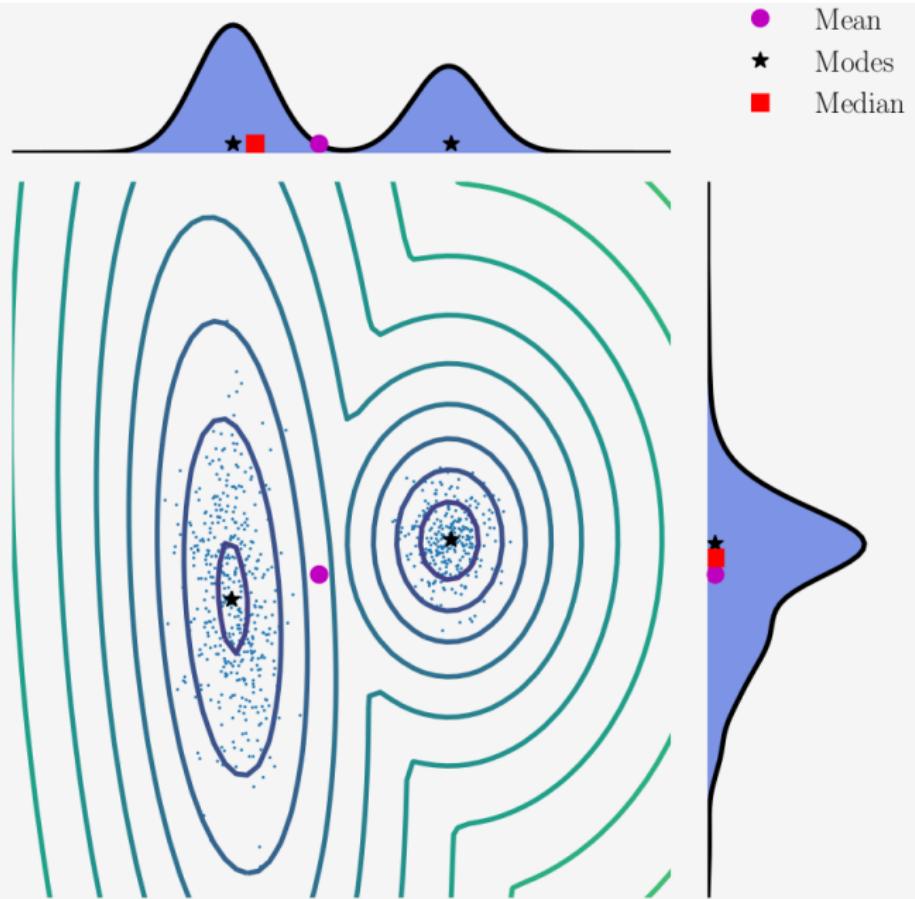
- **Média:** Valor esperado da variável aleatória:

$$\mathbb{E}_X[x] = \sum_{x \in \mathcal{X}} xP(x), \quad \mathbb{E}_X[x] = \int_{\mathcal{X}} xp(x)dx$$

- **Caso multivariado:** $\mu = \mathbb{E}_X[\mathbf{x}] = [\mathbb{E}_{X_1}[x_1], \dots, \mathbb{E}_{X_D}[x_D]]^\top$
- **Estimador:** $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$

- **Mediana:** Valor central da variável aleatória discreta ou valor em que $F_X(x) = \int_{-\infty}^x p(x')dx' = 0.5$, para o caso contínuo.
 - **Estimador:** Ordene as observações e escolha o valor central.
- **Moda:** Valor mais frequente (máximo de $P(x)$ ou $p(x)$).
 - **Estimador:** Escolha a observação mais frequente.

Descritores de Tendência Central - Ilustração



Estatísticas de caracterização

Medidas de Variabilidade

- **Matriz de covariância:** Dispersão em torno da média μ :

$$\begin{aligned}\mathbb{V}_X[\mathbf{x}] &= \text{cov}_X[\mathbf{x}, \mathbf{x}] = \mathbb{E}_X[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] \\ &= \mathbb{E}_X[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}_X[\mathbf{x}]\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbb{E}_X[\mathbf{x}]^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top \\ &= \mathbb{E}_X[\mathbf{x}\mathbf{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top \\ &= \begin{bmatrix} \text{cov}(x_1, x_1) & \text{cov}(x_1, x_2) & \cdots & \text{cov}(x_1, x_D) \\ \text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) & \cdots & \text{cov}(x_2, x_D) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(x_D, x_1) & \text{cov}(x_D, x_2) & \cdots & \text{cov}(x_D, x_D) \end{bmatrix}\end{aligned}$$

→ **Estimador:** $\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top$ ou
 $\hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top$

Estatísticas de caracterização

Medidas de Variabilidade

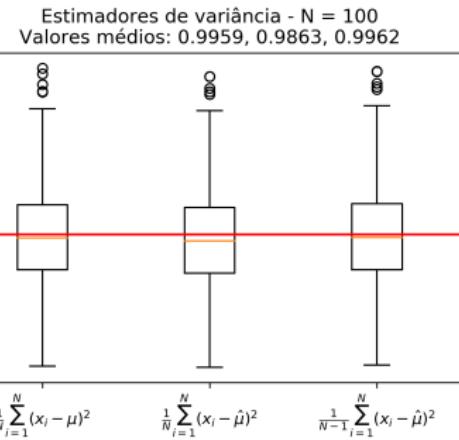
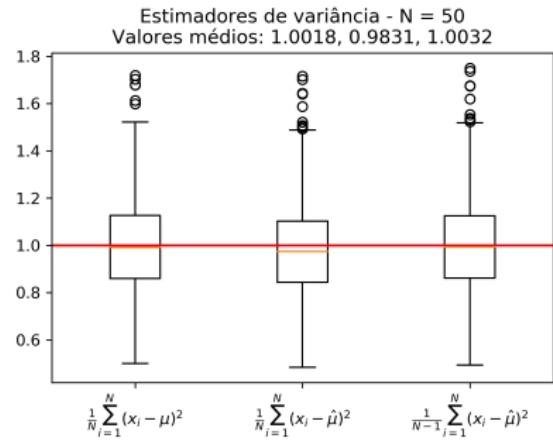
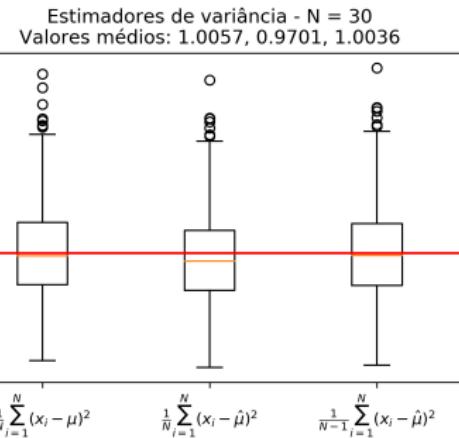
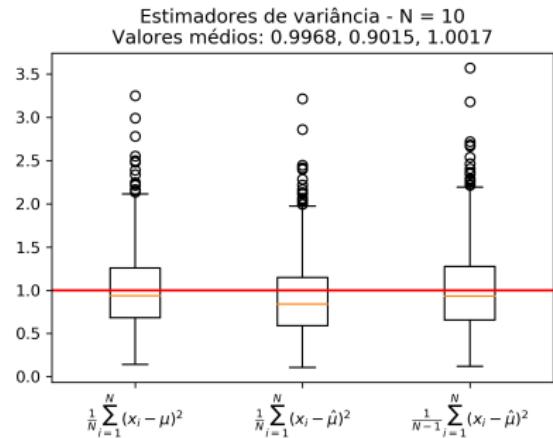
- **Matriz de covariância:** Dispersão em torno da média μ :

$$\begin{aligned}\mathbb{V}_X[\mathbf{x}] &= \text{cov}_X[\mathbf{x}, \mathbf{x}] = \mathbb{E}_X[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] \\ &= \mathbb{E}_X[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}_X[\mathbf{x}]\boldsymbol{\mu}^\top - \boldsymbol{\mu}\mathbb{E}_X[\mathbf{x}]^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top \\ &= \mathbb{E}_X[\mathbf{x}\mathbf{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top \\ &= \begin{bmatrix} \text{cov}(x_1, x_1) & \text{cov}(x_1, x_2) & \cdots & \text{cov}(x_1, x_D) \\ \text{cov}(x_2, x_1) & \text{cov}(x_2, x_2) & \cdots & \text{cov}(x_2, x_D) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(x_D, x_1) & \text{cov}(x_D, x_2) & \cdots & \text{cov}(x_D, x_D) \end{bmatrix}\end{aligned}$$

→ **Estimador:** $\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top$ ou
 $\hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top$

- **Variância:** $\mathbb{V}_X[\mathbf{x}] = \sigma^2 = \mathbb{E}_X[(x - \mu)^2] = \mathbb{E}_X[x^2] - \mu^2$
 - **Estimador:** $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$ ou
 $\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$
 - **Desvio padrão:** σ

Estimadores de variância - Ilustração



Independência estatística

- X e Y são estatisticamente independentes se e somente se:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$$

- Se X e Y são independentes, as propriedades abaixo são válidas:

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

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

$$\mathbb{V}_{X,Y}[\mathbf{x} + \mathbf{y}] = \mathbb{V}_X[\mathbf{x}] + \mathbb{V}_Y[\mathbf{y}],$$

$$\text{cov}_{X,Y}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$$

- **Observação:** X e Y podem ter $\text{cov}_{X,Y}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ e não serem independentes, pois a covariância mede somente dependência linear.

Agenda

- ① Revisão probabilidade e estatística
- ② Algumas distribuições de probabilidade
- ③ Aprendizagem de Máquina Probabilística
- ④ Tópicos adicionais
- ⑤ Referências

Distribuição de Bernoulli

- Distribuição de uma v.a. binária X com estados $x \in \{0, 1\}$ e parâmetro contínuo $q \in [0, 1]$. Exemplo: jogar uma moeda.
- Representada por $\mathcal{B}(x|q)$ ou $\text{Ber}(q)$ em que:

$$P(x|q) = q^x(1 - q)^{1-x}, \quad x \in \{0, 1\},$$

$$\mathbb{E}[x] = q,$$

$$\mathbb{V}[x] = q(1 - q).$$

Distribuição Binomial

- Distribuição de uma v.a. discreta que representa o número de observações 1 em N amostras binárias, considerando o parâmetro $q \in [0, 1]$. Exemplo: número de caras em N arremessos de uma moeda.
- Representada por $\text{Bin}(m|N, q)$ em que:

$$P(m|N, q) = \binom{N}{m} q^m (1 - q)^{N-m}, \quad m \in [0, N],$$

$$\binom{N}{m} = \frac{N!}{(N - m)!m!},$$

$$\mathbb{E}[m] = Nq,$$

$$\mathbb{V}[m] = Nq(1 - q).$$

Distribuição Beta

- Distribuição de uma v.a. contínua no intervalo $[0, 1]$, podendo representar a distribuição sobre o parâmetro q de distribuições de Bernoulli ou Binomial.
- Representada por $\text{Beta}(q|a, b)$ em que:

$$p(q|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} q^{a-1} (1 - q)^{b-1}, \quad q \in [0, 1],$$

$$\mathbb{E}[q] = \frac{a}{a + b},$$

$$\mathbb{V}[q] = \frac{ab}{(a + b)^2(a + b + 1)}.$$

- $a, b > 0$ são hiperparâmetros da distribuição.

Distribuição Beta

- A distribuição beta é **conjugada** das verossimilhanças de Bernoulli e Binomial, pois quando usada como priori de q , resulta em uma posteriori com o mesmo formato da priori.
- Considere:

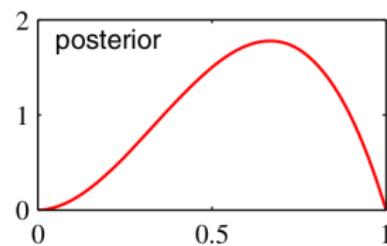
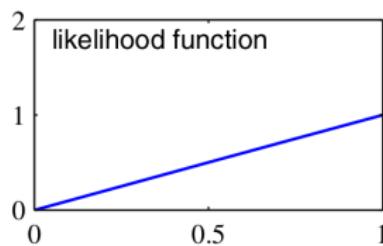
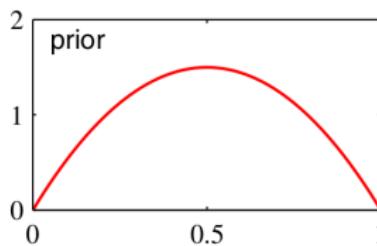
$$p(m|N, q) = \text{Bin}(m|N, q) = \binom{N}{m} q^m (1-q)^{N-m},$$

$$p(q|a, b) = \text{Beta}(q|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} q^{a-1} (1-q)^{b-1}.$$

- A posteriori $p(q|a, b, N, m)$ será:

$$\begin{aligned} p(q|a, b, N, m) &\propto q^{(m+a-1)} (1-q)^{N-m+b-1} \\ &= \text{Beta}(q|m+a, N-m+b). \end{aligned}$$

Distribuição Beta - Ilustração



- Priori (beta, com $a = b = 2$), verossimilhança (binomial, com $N = m = 1$) e posteriori (beta, com $a = 3$ e $b = 2$) para a variável q .
- Note que foi feita somente uma observação ($N = 1$).
- A posteriori encontrada pode ser usada como priori para as próximas observações, resultando em uma **aprendizagem sequencial**.

Distribuição Multinoulli (ou categórica)

- Distribuição de uma v.a. discreta X com estados $x \in \{1, \dots, K\}$ e K hiperparâmetros contínuos $0 \leq q_k \leq 1, \sum_k q_k = 1$. Exemplo: jogar um dado.
- Representada por $\text{Cat}(x|\boldsymbol{q})$ em que:

$$P(x = k|\boldsymbol{q}) = q_k,$$

$$P(x|\boldsymbol{q}) = \prod_{k=1}^K q_k^{[x=k]}, \quad x \in \{1, \dots, K\}.$$

Distribuição Multinoulli (ou categórica)

- Distribuição de uma v.a. discreta X com estados $x \in \{1, \dots, K\}$ e K hiperparâmetros contínuos $0 \leq q_k \leq 1, \sum_k q_k = 1$. Exemplo: jogar um dado.
- Representada por $\text{Cat}(x|\boldsymbol{q})$ em que:

$$P(x = k|\boldsymbol{q}) = q_k,$$

$$P(x|\boldsymbol{q}) = \prod_{k=1}^K q_k^{[x=k]}, \quad x \in \{1, \dots, K\}.$$

- Alternativamente, podemos interpretar como a distribuição de um vetor binário $\boldsymbol{x} \in \{0, 1\}^K$ com somente um valor igual a 1:

$$P(\boldsymbol{x}|\boldsymbol{q}) = \prod_{k=1}^K q_k^{x_k}, \quad x_k \in \{0, 1\}, \quad \sum_k x_k = 1,$$

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{q}] = \boldsymbol{q}.$$

Distribuição Multinomial

- Distribuição de uma v.a. discreta que representa o número de observações de cada estado $1 \leq k \leq K$ em N amostras categóricas, considerando $q_k \geq 0, \sum_k q_k = 1$. Exemplo: número de observações de cada face em N arremessos de um dado.
- Representada por $\text{Mult}(m_1, \dots, m_K | N, \mathbf{q})$ ou $\text{Mult}(\mathbf{m} | N, \mathbf{q})$ em que:

$$p(\mathbf{m} | N, \mathbf{q}) = \binom{N}{m_1 \dots m_K} \prod_{k=1}^K q^{m_k}, \quad m_k \in [0, N], \quad \sum_k m_k = 1,$$

$$\binom{N}{m_1 \dots m_K} = \frac{N!}{m_1! \dots m_K!},$$

$$\mathbb{E}[m_k] = Nq_k,$$

$$\mathbb{V}[m_k] = Nq_k(1 - q_k).$$

Distribuição Dirichlet

- Distribuição de uma v.a. contínua multidimensional no intervalo $[0, 1]^K$ cuja soma dos elementos é igual a 1. Pode representar a distribuição sobre o parâmetro \mathbf{q} de distribuições Multinoulli ou Multinomial.
- Representada por $\text{Dir}(\mathbf{q}|\boldsymbol{\alpha})$ em que:

$$p(\mathbf{q}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \prod_{k=1}^K q_k^{\alpha_k - 1}, \quad q_k \in [0, 1],$$

$$\mathbb{E}[q_k] = \frac{\alpha_k}{\sum_k \alpha_k},$$

$$\mathbb{V}[q_k] = \frac{\frac{\alpha_k}{\sum_k \alpha_k} \left(1 - \frac{\alpha_k}{\sum_k \alpha_k}\right)}{\sum_k \alpha_k + 1}.$$

- $\alpha_k > 0$ são hiperparâmetros da distribuição.

Distribuição Dirichlet

- A distribuição Dirichlet é **conjugada** das verossimilhanças Multinoulli e Multinomial, pois quando usada como priori de \mathbf{q} , resulta em uma posteriori com o mesmo formato da priori.
- Considere:

$$p(\mathbf{m}|N, \mathbf{q}) = \binom{N}{m_1 \dots m_K} \prod_{k=1}^K q^{m_k},$$

$$p(\mathbf{q}|\boldsymbol{\alpha}) = \text{Dir}(\mathbf{q}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \prod_{k=1}^K q_k^{\alpha_k - 1}.$$

- A posteriori $p(\mathbf{q}|\boldsymbol{\alpha}, N, \mathbf{m})$ será:

$$\begin{aligned} p(\mathbf{q}|\boldsymbol{\alpha}, N, \mathbf{m}) &\propto \prod_{k=1}^K q_k^{(m_k + \alpha_k - 1)} \\ &= \text{Dir}(\mathbf{q}|\mathbf{m} + \boldsymbol{\alpha}). \end{aligned}$$

Distribuição Gamma

- Distribuição de uma v.a. contínua positiva.
- Representada por $\text{Gamma}(x|a, b)$ em que:

$$p(x|a, b) = \frac{1}{\Gamma(a)} b^a x^{a-1} e^{-bx},$$

$$\mathbb{E}[x] = \frac{a}{b},$$

$$\mathbb{V}[x] = \frac{a}{b^2}.$$

- $a > 0$ e $b > 0$ são hiperparâmetros da distribuição.

Distribuição Gamma inversa

- Distribuição de $\frac{1}{x}$ quando x segue uma distribuição Gamma.
- Representada por $\text{IGamma}(x|a, b)$ em que:

$$p(x|a, b) = \frac{1}{\Gamma(a)} b^a x^{-a-1} e^{-b/x},$$

$$\mathbb{E}[x] = \frac{b}{a-1}, \text{ para } a > 1,$$

$$\mathbb{V}[x] = \frac{b^2}{(a-1)^2(a-2)}, \text{ para } a > 2.$$

- $a > 0$ e $b > 0$ são hiperparâmetros da distribuição.

Distribuição Gaussiana

- Representada por $\mathcal{N}(x|\mu, \sigma^2)$ ou $x \sim \mathcal{N}(\mu, \sigma^2)$ em que:

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

$$\rightarrow \hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\rightarrow \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \text{ ou } \hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

Distribuição Gaussiana

- Representada por $\mathcal{N}(x|\mu, \sigma^2)$ ou $x \sim \mathcal{N}(\mu, \sigma^2)$ em que:

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

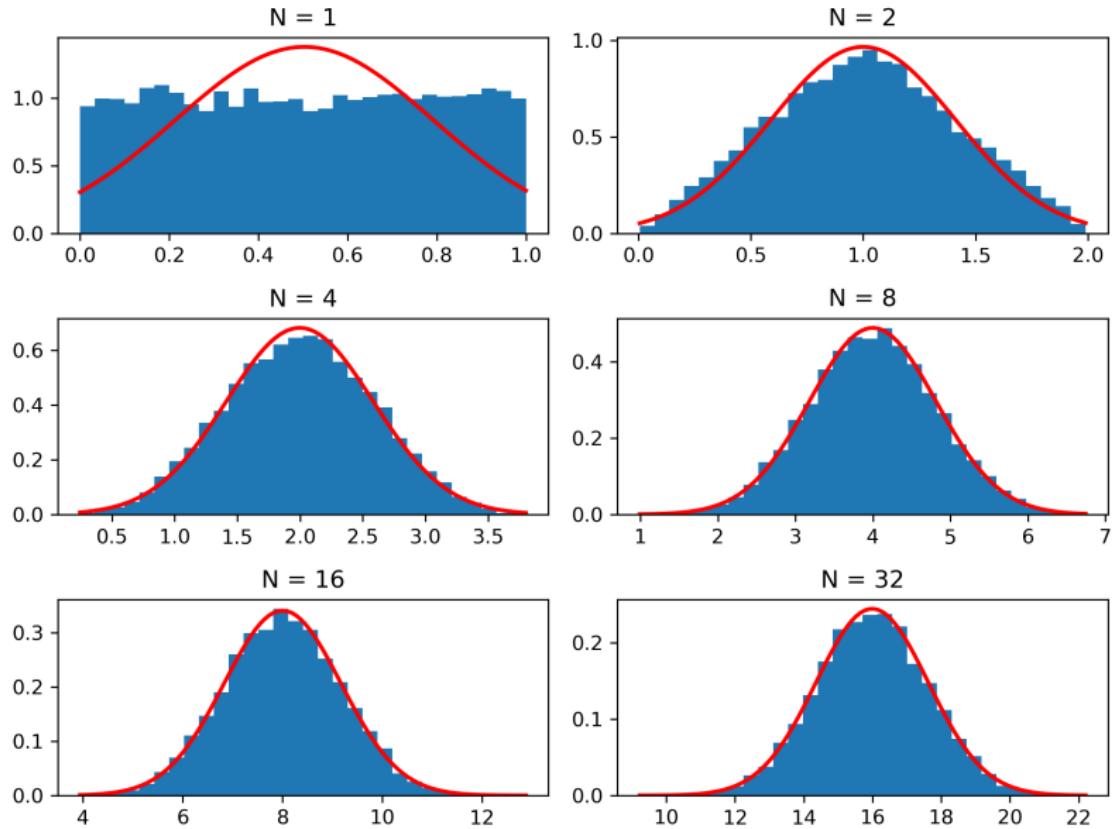
$$\rightarrow \hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\rightarrow \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \text{ ou } \hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

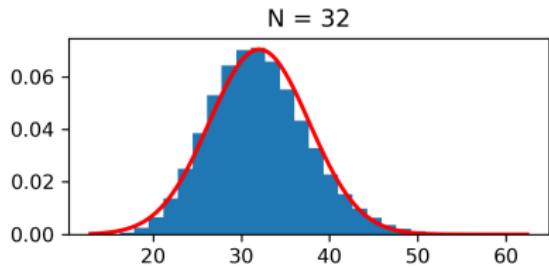
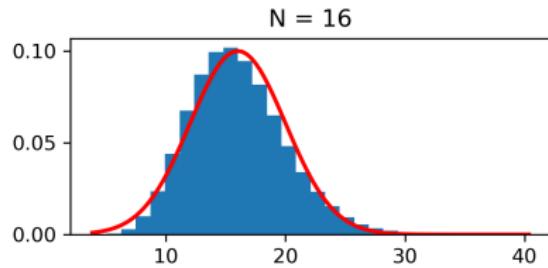
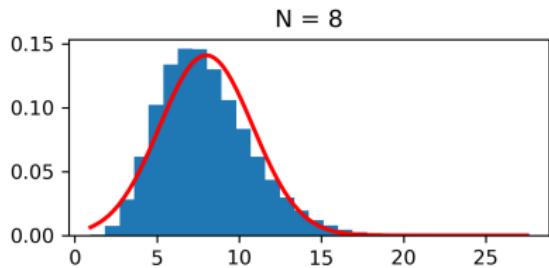
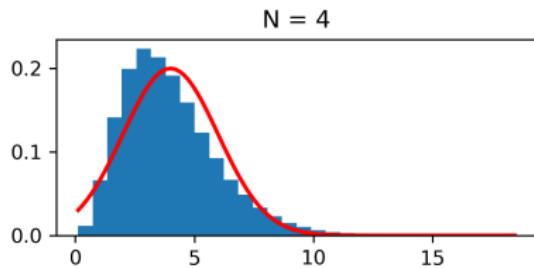
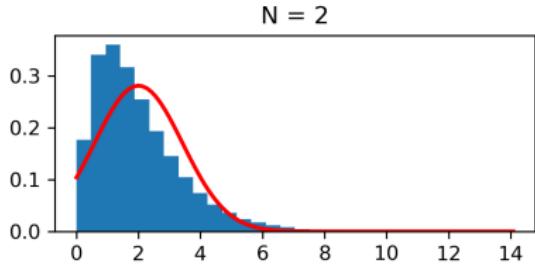
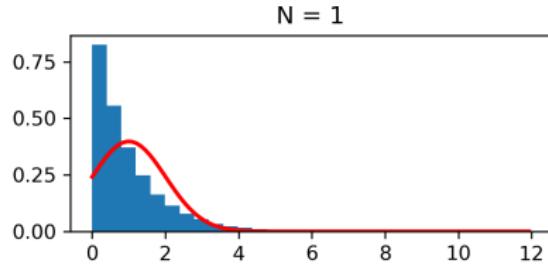
Teorema Central do Limite

Dada uma sequência de N variáveis aleatórias independentes e identicamente distribuídas (i.i.d) com médias e variâncias finitas, a soma dessas variáveis converge para uma distribuição Gaussiana.

Teorema Central do Limite - Ilustração



Teorema Central do Limite - Ilustração



Distribuição Gaussiana Multivariada

- Dado $\mathbf{x} \in \mathbb{R}^D$, temos $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ou $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ em que:

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

$$\rightarrow \hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$$

$$\rightarrow \hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top \text{ ou}$$

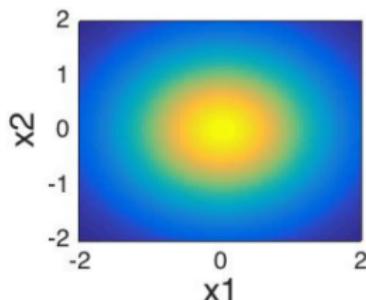
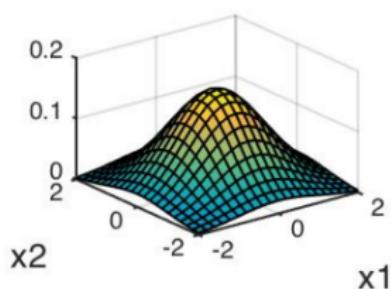
$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top$$

- Note que a quantidade $\Delta = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})}$ corresponde à **distância de Mahalanobis** entre \mathbf{x} e $\boldsymbol{\mu}$.

Distribuição Gaussiana Multivariada

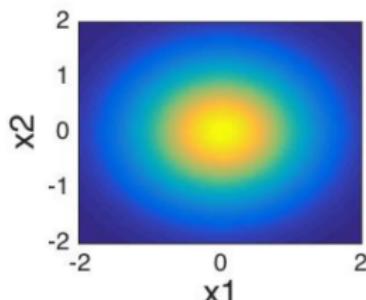
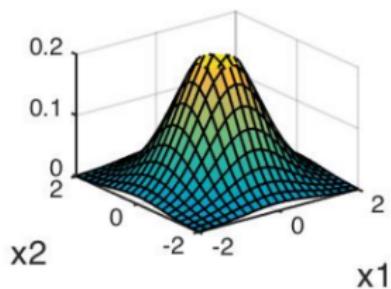
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu = [0 \ 0]^T$$



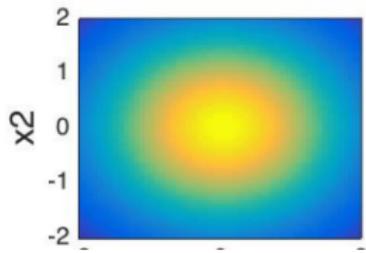
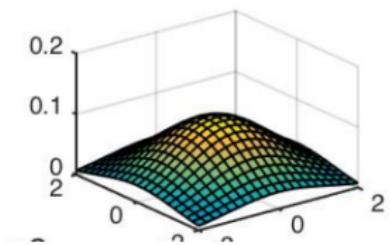
$$\Sigma = \begin{pmatrix} 0.7 & 0 \\ 0 & 0.7 \end{pmatrix}$$

$$\mu = [0 \ 0]^T$$



$$\Sigma = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix}$$

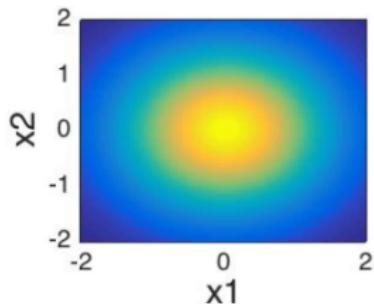
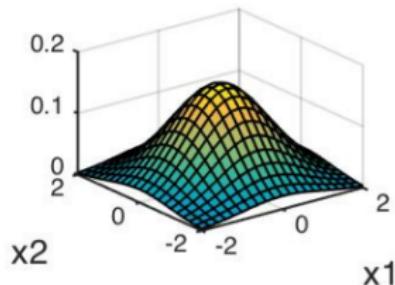
$$\mu = [0 \ 0]^T$$



Distribuição Gaussiana Multivariada

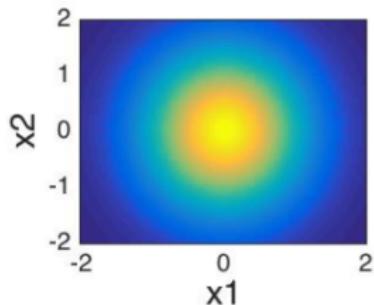
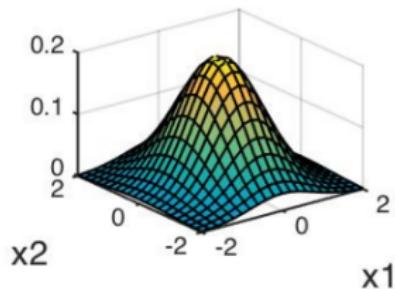
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu = [0 \ 0]^T$$



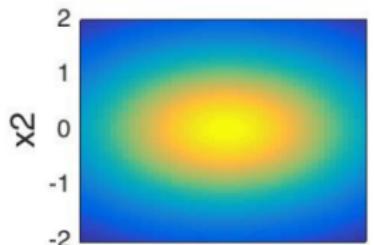
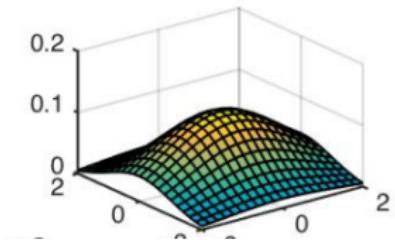
$$\Sigma = \begin{pmatrix} 0.6 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu = [0 \ 0]^T$$



$$\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

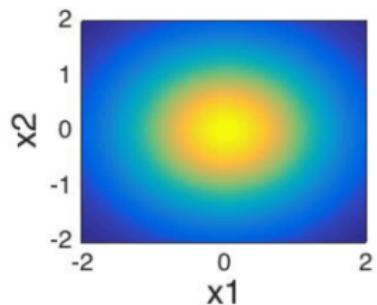
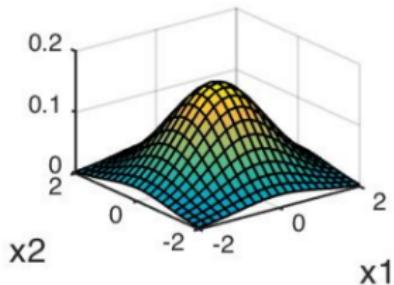
$$\mu = [0 \ 0]^T$$



Distribuição Gaussiana Multivariada

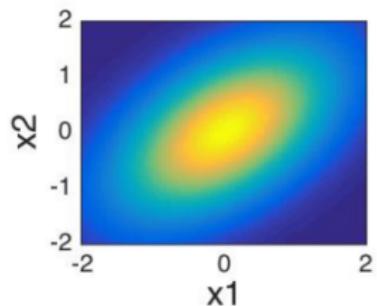
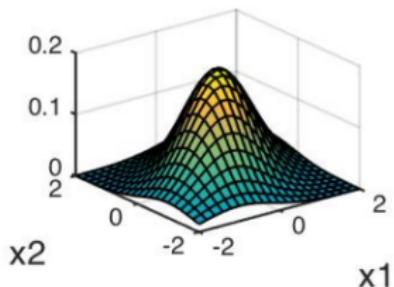
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu = [0 \ 0]^T$$



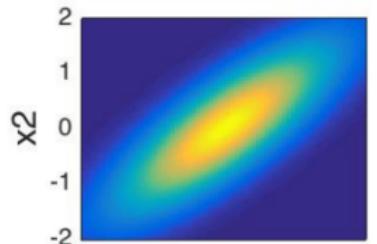
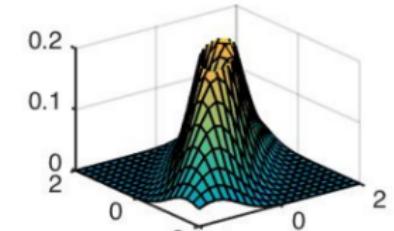
$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

$$\mu = [0 \ 0]^T$$



$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

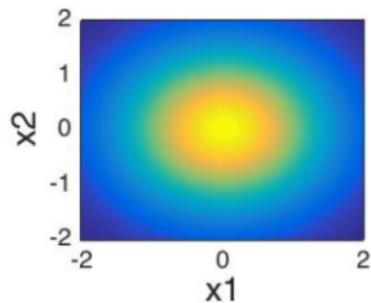
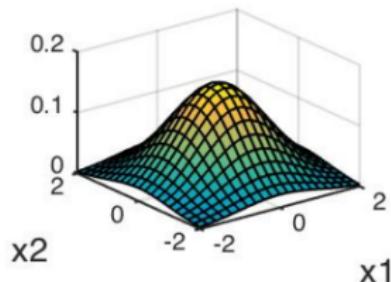
$$\mu = [0 \ 0]^T$$



Distribuição Gaussiana Multivariada

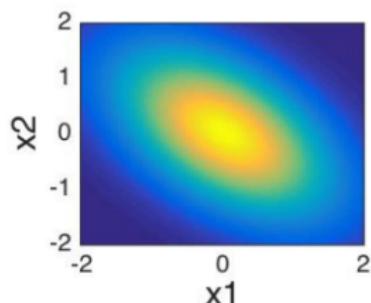
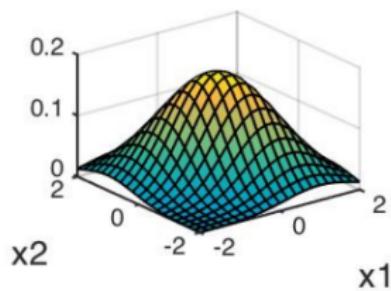
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu = [0 \ 0]^T$$



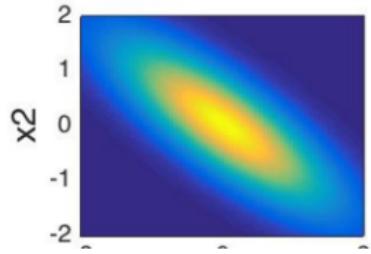
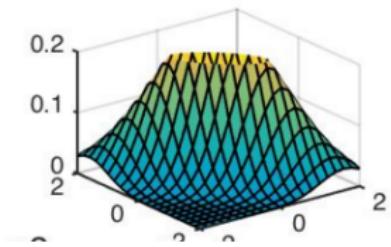
$$\Sigma = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

$$\mu = [0 \ 0]^T$$



$$\Sigma = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$$

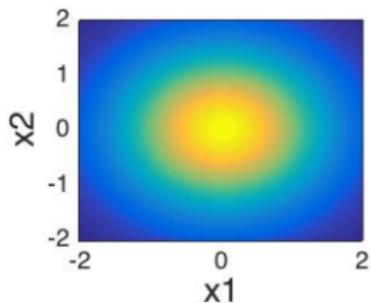
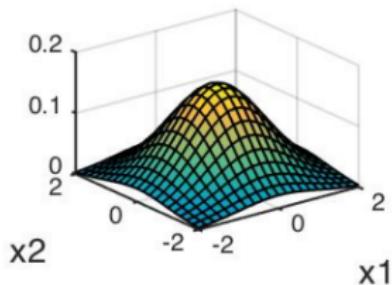
$$\mu = [0 \ 0]^T$$



Distribuição Gaussiana Multivariada

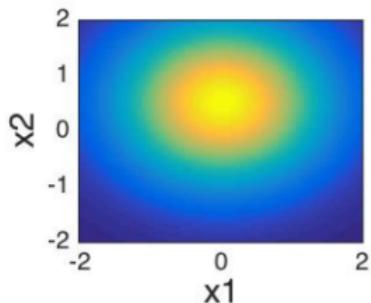
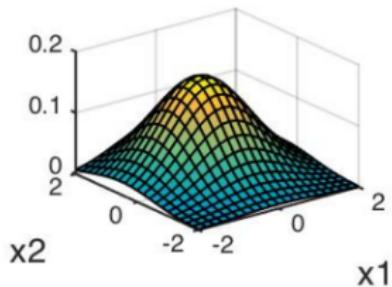
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu = [0 \ 0]^T$$



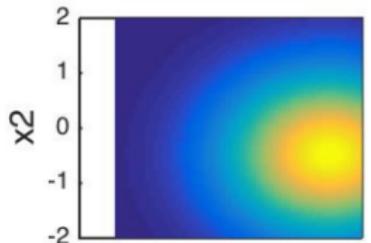
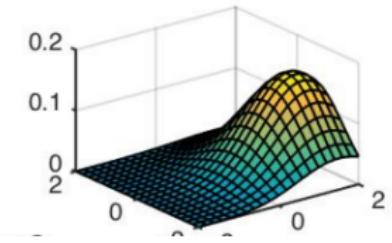
$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu = [0 \quad 0.5]^T$$



$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu = [1.5 \quad -0.5]^T$$



Propriedades da Distribuição Gaussiana

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Propriedades da Distribuição Gaussiana

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Marginalização

A observação de uma coleção maior de variáveis não afeta a distribuição de subconjuntos menores, ou seja:

$$x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}) \text{ e } x_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$$

Propriedades da Distribuição Gaussiana

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Marginalização

A observação de uma coleção maior de variáveis não afeta a distribuição de subconjuntos menores, ou seja:

$$x_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \text{ e } x_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

Condicionamento

Condicionar Gaussianas resulta em uma Gaussiana:

$$p(\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{z}) = \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{z} - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

Propriedades da Distribuição Gaussiana

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Marginalização

A observação de uma coleção maior de variáveis não afeta a distribuição de subconjuntos menores, ou seja:

$$x_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \text{ e } x_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

Condicionamento

Condicionar Gaussianas resulta em uma Gaussiana:

$$p(\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{z}) = \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{z} - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

Linearidade

Uma combinação linear de Gaussianas resulta em uma Gaussiana:

$$p(a\mathbf{x}_1 + b\mathbf{x}_2) = \mathcal{N}(a\boldsymbol{\mu}_1 + b\boldsymbol{\mu}_2, a^2\boldsymbol{\Sigma}_{11} + b^2\boldsymbol{\Sigma}_{22})$$

Família exponencial

- As distribuições revisadas aqui são todas da chamada **família exponencial**, podendo ser escritas no formato:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})),$$

em que \mathbf{x} pode ser escalar ou vetorial, contínuo ou discreto, $\boldsymbol{\eta}$ são os **parâmetros naturais** da distribuição e $h(\cdot)$, $g(\cdot)$, $\mathbf{u}(\cdot)$ são funções.

Família exponencial

- As distribuições revisadas aqui são todas da chamada **família exponencial**, podendo ser escritas no formato:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})),$$

em que \mathbf{x} pode ser escalar ou vetorial, contínuo ou discreto, $\boldsymbol{\eta}$ são os **parâmetros naturais** da distribuição e $h(\cdot)$, $g(\cdot)$, $\mathbf{u}(\cdot)$ são funções.

- Note que $g(\boldsymbol{\eta})$ independe de \mathbf{x} e garante que a distribuição seja normalizada, ou seja:

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} = 1.$$

Família exponencial

- As distribuições revisadas aqui são todas da chamada **família exponencial**, podendo ser escritas no formato:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})),$$

em que \mathbf{x} pode ser escalar ou vetorial, contínuo ou discreto, $\boldsymbol{\eta}$ são os **parâmetros naturais** da distribuição e $h(\cdot)$, $g(\cdot)$, $\mathbf{u}(\cdot)$ são funções.

- Note que $g(\boldsymbol{\eta})$ independe de \mathbf{x} e garante que a distribuição seja normalizada, ou seja:

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})) d\mathbf{x} = 1.$$

- Toda distribuição da família exponencial $p(\mathbf{x}|\boldsymbol{\eta})$ possui uma priori $p(\boldsymbol{\eta})$ **conjugada** à sua função de verossimilhança, tal que a posteriori de $\boldsymbol{\eta}$ possui o mesmo formato da priori.

Agenda

- ① Revisão probabilidade e estatística
- ② Algumas distribuições de probabilidade
- ③ Aprendizagem de Máquina Probabilística
- ④ Tópicos adicionais
- ⑤ Referências

Aprendizagem de Máquina Probabilística

- Quantidades desconhecidas são tratadas como **variáveis aleatórias** com **distribuições de probabilidade**.
- Abordagem adequada para tomar decisões diante de **incertezas**.
- Modelos supervisionados aprendem distribuições condicionais $p(\mathbf{y}|\mathbf{x})$.
- Modelos não-supervisionados aprendem distribuições não-condicionais $p(\mathbf{x})$.

Aprendizagem de Máquina Probabilística

- Quantidades desconhecidas são tratadas como **variáveis aleatórias** com **distribuições de probabilidade**.
- Abordagem adequada para tomar decisões diante de **incertezas**.
- Modelos supervisionados aprendem distribuições condicionais $p(\mathbf{y}|\mathbf{x})$.
- Modelos não-supervisionados aprendem distribuições não-condicionais $p(\mathbf{x})$.

“Quase todo o aprendizado de máquina pode ser visto em termos probabilísticos, tornando o pensamento probabilístico fundamental. É claro que não é a única visão. Mas é por meio dessa visão que podemos conectar o que fazemos às outras ciências computacionais, seja na otimização estocástica, teoria de controle, pesquisa operacional, econometria, teoria da informação, física estatística ou bioestatística. Somente por essa razão, **o domínio do pensamento probabilístico é essencial.**”

(Shakir Mohamed, 2018)

Aprendizagem de Máquina Probabilística

Exemplo paramétrico supervisionado - máxima verossimilhança

Sejam N entradas $\mathbf{x}_i|_{i=1}^N$, agrupadas na matriz \mathbf{X} , N saídas $y_i|_{i=1}^N$, e os parâmetros \mathbf{w} :

$$y = f_{\mathbf{w}}(\mathbf{x}) + \epsilon \text{ [observações ruidosas]},$$

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \mathbf{w}) \text{ [verossimilhança]},$$

$$\hat{\mathbf{w}} = \arg \max p(\mathbf{y}|\mathbf{X}, \mathbf{w}) \text{ [máxima verossimilhança]}$$

$$p(y_*|\mathbf{x}_*, \hat{\mathbf{w}}) \text{ [distribuição preditiva]}.$$

Aprendizagem de Máquina Probabilística

Exemplo paramétrico supervisionado - abordagem Bayesiana

Sejam N entradas $\mathbf{x}_i|_{i=1}^N$, agrupadas na matriz \mathbf{X} , N saídas $y_i|_{i=1}^N$, e os parâmetros \mathbf{w} :

$$y = f_{\mathbf{w}}(\mathbf{x}) + \epsilon \text{ [observações ruidosas]},$$
$$p(\mathbf{w}) \text{ [priori]},$$

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \mathbf{w}) \text{ [verossimilhança]},$$

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w} \text{ [verossimilhança marginal]},$$

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})} \text{ [posteriori via Regra de Bayes]}$$

$$p(y_*|\mathbf{x}_*) = \int p(y_*|\mathbf{x}_*, \mathbf{w})p(\mathbf{w}|\mathbf{X}, \mathbf{y})d\mathbf{w} \text{ [distribuição preditiva a posteriori]}.$$

Aprendizagem de Máquina Probabilística

Exemplo paramétrico não-supervisionado - variáveis latentes

Sejam N observações $\mathbf{x}_i|_{i=1}^N$, agrupadas na matriz \mathbf{X} , N variáveis latentes (não observadas) $\mathbf{z}_i|_{i=1}^N$, agrupadas na matriz \mathbf{Z} , e os parâmetros \mathbf{w} :

$$p(\mathbf{Z}) \text{ [prior]},$$

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{w}) = \prod_{i=1}^N p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{w}) \text{ [verossimilhança]},$$

$$p(\mathbf{Z}|\mathbf{X}, \mathbf{w}) = \frac{p(\mathbf{X}|\mathbf{Z}, \mathbf{w})p(\mathbf{Z})}{p(\mathbf{X}|\mathbf{w})} \text{ [posteriori via Bayes]}$$

$$p(\mathbf{X}|\mathbf{w}) = \frac{p(\mathbf{X}|\mathbf{Z}, \mathbf{w})p(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \mathbf{w})} = \int p(\mathbf{X}|\mathbf{Z}, \mathbf{w})p(\mathbf{Z})d\mathbf{Z} \text{ [verossimilhança marginal]},$$

$$\hat{\mathbf{w}} = \arg \max p(\mathbf{X}|\mathbf{w}) \text{ [máxima verossim. marg.]}$$

$$p(\mathbf{x}_*|\tilde{\mathbf{z}}, \hat{\mathbf{w}}), \text{ em que } \tilde{\mathbf{z}} \sim p(\mathbf{z}) \text{ [geração de dados]}.$$

Agenda

- ① Revisão probabilidade e estatística
- ② Algumas distribuições de probabilidade
- ③ Aprendizagem de Máquina Probabilística
- ④ Tópicos adicionais
- ⑤ Referências

Tópicos adicionais

- Família exponencial e distribuições conjugadas.
- Função/transformação de uma variável aleatória.

Agenda

- ① Revisão probabilidade e estatística
- ② Algumas distribuições de probabilidade
- ③ Aprendizagem de Máquina Probabilística
- ④ Tópicos adicionais
- ⑤ Referências

Referências bibliográficas

- Cap. 6 - DEISENROTH, M. et al. **Mathematics for machine learning**. 2019.
- HINES, W. W., MONTGOMERY, D. C., GOLDSMAN, D. M., e BORROR, C. M., **Probabilidade e Estatística na Engenharia**. LTC, 2006.
- PAPOULIS, A., and PILLAI, S.U., **Probability, Random Variables and Stochastic Processes**. (Electrical & Electronic Engineering Series). McGraw-Hill International, 4th edition, 2002.