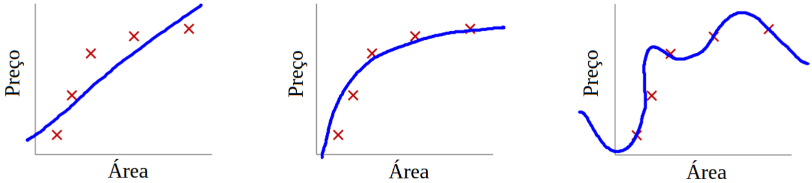
Aula 29 - Teoria do Aprendizado (Parte I)

João B. Florindo

Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas - Brasil
jbflorindo@ime.unicamp.br

Outline

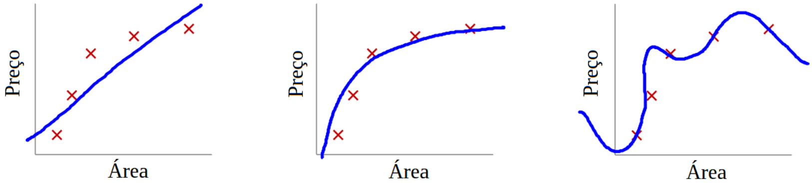


$\theta_0 + \theta_1 x$ — Alto viés (underfit)

$\theta_0 + \theta_1 x + \theta_2 x^2$ — "Ideal"

$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$ — Alta variância (overfit)

- **Erro de generalização**: erro esperado do modelo quando aplicado a exemplos que não foram vistos no conjunto de treino.
- Os modelos à esquerda e à direita possuem ambos um alto erro de generalização, mas as causas são totalmente diferentes.



$\theta_0 + \theta_1 x$ — Alto viés (underfit)

$\theta_0 + \theta_1 x + \theta_2 x^2$ — "Ideal"

$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$ — Alta variância (overfit)

- **Erro de generalização**: erro esperado do modelo quando aplicado a exemplos que não foram vistos no conjunto de treino.
- Os modelos à esquerda e à direita possuem ambos um alto erro de generalização, mas as causas são totalmente diferentes.

- O primeiro sofre de um **viés** alto (*underfit*).
- **Viés** é o erro de generalização esperado quando temos um conjunto de treino muito grande (teoricamente infinito).
- Já o segundo tem alta **variância**.
- Alto risco de se ajustar a padrões específicos de um conjunto pequeno de treino, mas que não expressam a relação ampla entre x e y.
- *Trade-off*: frequentemente, modelo muito "simples" (poucos parâmetros) tem viés alto e variância baixa.
- Já os muito complexos (muitos parâmetros) têm variância alta e viés baixo.

- O primeiro sofre de um **viés** alto (*underfit*).
- **Viés** é o erro de generalização esperado quando temos um conjunto de treino muito grande (teoricamente infinito).
- Já o segundo tem alta **variância**.
- Alto risco de se ajustar a padrões específicos de um conjunto pequeno de treino, mas que não expressam a relação ampla entre x e y.
- *Trade-off*: frequentemente, modelo muito "simples" (poucos parâmetros) tem viés alto e variância baixa.
- Já os muito complexos (muitos parâmetros) têm variância alta e viés baixo.

- O primeiro sofre de um **viés** alto (*underfit*).
- **Viés** é o erro de generalização esperado quando temos um conjunto de treino muito grande (teoricamente infinito).
- Já o segundo tem alta **variância**.
- Alto risco de se ajustar a padrões específicos de um conjunto pequeno de treino, mas que não expressam a relação ampla entre x e y.
- *Trade-off*: frequentemente, modelo muito "simples" (poucos parâmetros) tem viés alto e variância baixa.
- Já os muito complexos (muitos parâmetros) têm variância alta e viés baixo.

- O primeiro sofre de um **viés** alto (*underfit*).
- **Viés** é o erro de generalização esperado quando temos um conjunto de treino muito grande (teoricamente infinito).
- Já o segundo tem alta **variância**.
- Alto risco de se ajustar a padrões específicos de um conjunto pequeno de treino, mas que não expressam a relação ampla entre x e y.
- *Trade-off*: frequentemente, modelo muito "simples" (poucos parâmetros) tem viés alto e variância baixa.
- Já os muito complexos (muitos parâmetros) têm variância alta e viés baixo.

- O primeiro sofre de um **viés** alto (*underfit*).
- **Viés** é o erro de generalização esperado quando temos um conjunto de treino muito grande (teoricamente infinito).
- Já o segundo tem alta **variância**.
- Alto risco de se ajustar a padrões específicos de um conjunto pequeno de treino, mas que não expressam a relação ampla entre x e y.
- *Trade-off*: frequentemente, modelo muito "simples" (poucos parâmetros) tem viés alto e variância baixa.
- Já os muito complexos (muitos parâmetros) têm variância alta e viés baixo.

- O primeiro sofre de um **viés** alto (*underfit*).
- **Viés** é o erro de generalização esperado quando temos um conjunto de treino muito grande (teoricamente infinito).
- Já o segundo tem alta **variância**.
- Alto risco de se ajustar a padrões específicos de um conjunto pequeno de treino, mas que não expressam a relação ampla entre x e y.
- *Trade-off*: frequentemente, modelo muito "simples" (poucos parâmetros) tem viés alto e variância baixa.
- Já os muito complexos (muitos parâmetros) têm variância alta e viés baixo.

Outline

Limitante da união (Desigualdade de Boole)

Lema. Sejam $A_1, A_2, \cdots, A_k$ eventos distintos (podendo não ser independentes). Então:

$$P(A_1 \cup \cdots \cup A_k) \leq P(A_1) + \cdots + P(A_k).$$

Desigualdade de Hoeffding (limitante de Chernoff)

Lema. Sejam $Z_1, \cdots, Z_n$ n variáveis aleatórias independentes e identicamente distribuídas (iid) amostradas a partir de uma distribuição Bernoulli(ϕ), i.e., $P(Z_i = 1) = \phi$ e $P(Z_i = 0) = 1 - \phi$.
Seja $\hat{\phi} = (1/n)\sum_{i=1}^{n} Z_i$ a média destas variáveis aleatórias e seja $\gamma > 0$ fixado. Então

$$P(|\phi - \hat{\phi}| > \gamma) \leq 2\exp(-2\gamma^2 n).$$

- O segundo lema diz que se tomarmos $\hat{\phi}$ (média de n variáveis aleatórias Bernoulli(ϕ)) como estimador de ϕ, então a probabilidade de estarmos distantes do valor verdadeiro é pequena se n for grande.
- Ou ainda: se temos uma moeda enviesada com probabilidade ϕ para "cara" e lançamos n vezes, então a fração de "caras" vai ser uma boa estimativa para ϕ com alta probabilidade se n for grande.
- Estes dois lemas serão base para vários resultados aqui!

- O segundo lema diz que se tomarmos $\hat{\phi}$ (média de n variáveis aleatórias Bernoulli(ϕ)) como estimador de ϕ, então a probabilidade de estarmos distantes do valor verdadeiro é pequena se n for grande.

- Ou ainda: se temos uma moeda enviesada com probabilidade ϕ para "cara" e lançamos n vezes, então a fração de "caras" vai ser uma boa estimativa para ϕ com alta probabilidade se n for grande.

- Estes dois lemas serão base para vários resultados aqui!

- O segundo lema diz que se tomarmos $\hat{\phi}$ (média de *n* variáveis aleatórias Bernoulli(ϕ)) como estimador de ϕ, então a probabilidade de estarmos distantes do valor verdadeiro é pequena se *n* for grande.
- Ou ainda: se temos uma moeda enviesada com probabilidade ϕ para "cara" e lançamos *n* vezes, então a fração de "caras" vai ser uma boa estimativa para ϕ com alta probabilidade se *n* for grande.
- Estes dois lemas serão base para vários resultados aqui!

- Vamos nos restringir à classificação binária $y \in \{0, 1\}$.

- Mas a teoria se generaliza.

- Conjunto de treinamento

$$S = \{(x^{(i)}, y^{(i)}), i = 1, \cdots, n\}.$$

- Cada exemplo $(x^{(i)}, y^{(i)})$ obtido iid de uma distribuição $\mathcal{D}$.

- Definimos o erro de treinamento (**risco empírico** ou **erro empírico** na teoria do aprendizado) para a hipótese h por

$$\hat{\varepsilon}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}.$$

- Vamos nos restringir à classificação binária $y \in \{0, 1\}$.

- Mas a teoria se generaliza.

- Conjunto de treinamento

$$S = \{(x^{(i)}, y^{(i)}), i = 1, \cdots, n\}.$$

- Cada exemplo $(x^{(i)}, y^{(i)})$ obtido iid de uma distribuição $\mathcal{D}$.

- Definimos o erro de treinamento (**risco empírico** ou **erro empírico** na teoria do aprendizado) para a hipótese h por

$$\hat{\varepsilon}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}.$$

- Vamos nos restringir à classificação binária $y \in \{0, 1\}$.
- Mas a teoria se generaliza.
- Conjunto de treinamento

$$S = \{(x^{(i)}, y^{(i)}), i = 1, \cdots, n\}.$$

- Cada exemplo $(x^{(i)}, y^{(i)})$ obtido iid de uma distribuição $\mathcal{D}$.
- Definimos o erro de treinamento (**risco empírico** ou **erro empírico** na teoria do aprendizado) para a hipótese h por

$$\hat{\varepsilon}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}.$$

- Vamos nos restringir à classificação binária $y \in \{0, 1\}$.
- Mas a teoria se generaliza.
- Conjunto de treinamento

$$S = \{(x^{(i)}, y^{(i)}), i = 1, \cdots, n\}.$$

- Cada exemplo $(x^{(i)}, y^{(i)})$ obtido iid de uma distribuição $\mathcal{D}$.
- Definimos o erro de treinamento (**risco empírico** ou **erro empírico** na teoria do aprendizado) para a hipótese h por

$$\hat{\varepsilon}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}.$$

- Vamos nos restringir à classificação binária $y \in \{0, 1\}$.
- Mas a teoria se generaliza.
- Conjunto de treinamento

$$S = \{(x^{(i)}, y^{(i)}), i = 1, \cdots, n\}.$$

- Cada exemplo $(x^{(i)}, y^{(i)})$ obtido iid de uma distribuição $\mathcal{D}$.
- Definimos o erro de treinamento (**risco empírico** ou **erro empírico** na teoria do aprendizado) para a hipótese h por

$$\hat{\varepsilon}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}.$$

- Já o erro de generalização é dado por

$$\varepsilon(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y).$$

- O fato de treino e teste serem obtidos da *mesma* distribuição $\mathcal{D}$ e dos exemplos de treino serem independentes são as mais importantes condições **PAC** (*probably approximately correct*).

- Já o erro de generalização é dado por

$$\varepsilon(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y).$$

- O fato de treino e teste serem obtidos da *mesma* distribuição $\mathcal{D}$ e dos exemplos de treino serem independentes são as mais importantes condições **PAC** (*probably approximately correct*).

- Considere o classificador linear

$$h_\theta(x) = \mathbb{1}\{\theta^T x \geq 0\}$$

.

- A abordagem mais "básica" para obter θ e na qual focaremos é a **minimização do risco empírico** (ERM).

- Minimizar o erro de treinamento:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \hat{\varepsilon}(h_\theta)$$

- A hipótese obtida é $\hat{h} = h_{\hat{\theta}}$.

- Considere o classificador linear

$$h_\theta(x) = \mathbb{1}\{\theta^T x \geq 0\}$$

.

- A abordagem mais "básica" para obter θ e na qual focaremos é a **minimização do risco empírico** (ERM).

- Minimizar o erro de treinamento:

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \hat{\varepsilon}(h_\theta)$$

- A hipótese obtida é $\hat{h} = h_{\hat{\theta}}$.

- Considere o classificador linear

$$h_\theta(x) = \mathbb{1}\{\theta^T x \geq 0\}$$

.

- A abordagem mais “básica” para obter θ e na qual focaremos é a **minimização do risco empírico** (ERM).

- Minimizar o erro de treinamento:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \hat{\varepsilon}(h_\theta)$$

- A hipótese obtida é $\hat{h} = h_{\hat{\theta}}$.

- Considere o classificador linear

$$h_\theta(x) = \mathbb{1}\{\theta^T x \geq 0\}$$

.

- A abordagem mais "básica" para obter θ e na qual focaremos é a **minimização do risco empírico** (ERM).

- Minimizar o erro de treinamento:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \hat{\varepsilon}(h_\theta)$$

- A hipótese obtida é $\hat{h} = h_{\hat{\theta}}$.

- Não focamos aqui em hipóteses específicas, mas sim em uma **classe de hipóteses** $\mathcal{H}$.

- Na classificação linear, temos a classe de todos os classificadores com fronteira linear sobre o domínio de entradas $\mathcal{X}$:

$$\mathcal{H} = \{h_\theta : h_\theta(x) = \mathbb{1}\{\theta^T x \geq 0\}, \theta \in \mathbb{R}^{d+1}\}$$

- Similarmente, teríamos a classe das redes neurais, etc.

- O ERM é feito sobre toda a classe $\mathcal{H}$ da qual a hipótese foi retirada:

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\varepsilon}(h).$$

- Não focamos aqui em hipóteses específicas, mas sim em uma **classe de hipóteses** $\mathcal{H}$.
- Na classificação linear, temos a classe de todos os classificadores com fronteira linear sobre o domínio de entradas $\mathcal{X}$:

$$\mathcal{H} = \{h_\theta : h_\theta(x) = \mathbb{1}\{\theta^T x \geq 0\}, \theta \in \mathbb{R}^{d+1}\}$$

- Similarmente, teríamos a classe das redes neurais, etc.
- O ERM é feito sobre toda a classe $\mathcal{H}$ da qual a hipótese foi retirada:

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \hat{\varepsilon}(h).$$

- Não focamos aqui em hipóteses específicas, mas sim em uma **classe de hipóteses** $\mathcal{H}$.

- Na classificação linear, temos a classe de todos os classificadores com fronteira linear sobre o domínio de entradas $\mathcal{X}$:

$$\mathcal{H} = \{h_\theta : h_\theta(x) = \mathbb{1}\{\theta^T x \geq 0\}, \theta \in \mathbb{R}^{d+1}\}$$

- Similarmente, teríamos a classe das redes neurais, etc.

- O ERM é feito sobre toda a classe $\mathcal{H}$ da qual a hipótese foi retirada:

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\varepsilon}(h).$$

- Não focamos aqui em hipóteses específicas, mas sim em uma **classe de hipóteses** $\mathcal{H}$.

- Na classificação linear, temos a classe de todos os classificadores com fronteira linear sobre o domínio de entradas $\mathcal{X}$:

$$\mathcal{H} = \{h_\theta : h_\theta(x) = \mathbb{1}\{\theta^T x \geq 0\}, \theta \in \mathbb{R}^{d+1}\}$$

- Similarmente, teríamos a classe das redes neurais, etc.

- O ERM é feito sobre toda a classe $\mathcal{H}$ da qual a hipótese foi retirada:

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\varepsilon}(h).$$