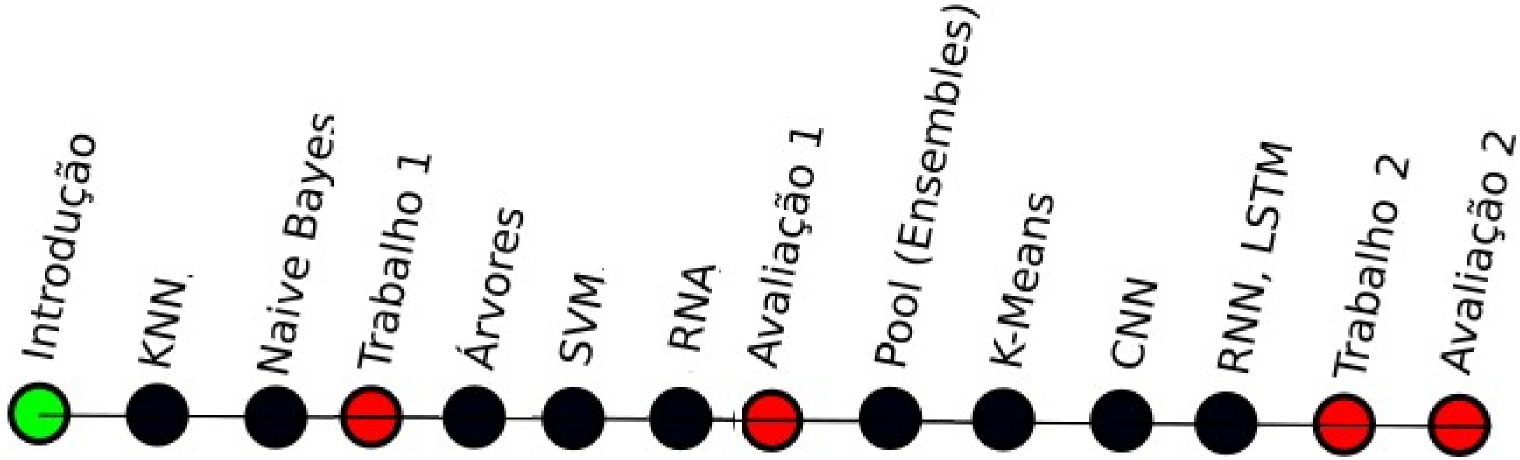
Árvores de Decisão

Prof. André Gustavo Hochuli

gustavo.hochuli@pucpr.br
aghochuli@ppgia.pucpr.br
github.com/andrehochuli/teaching

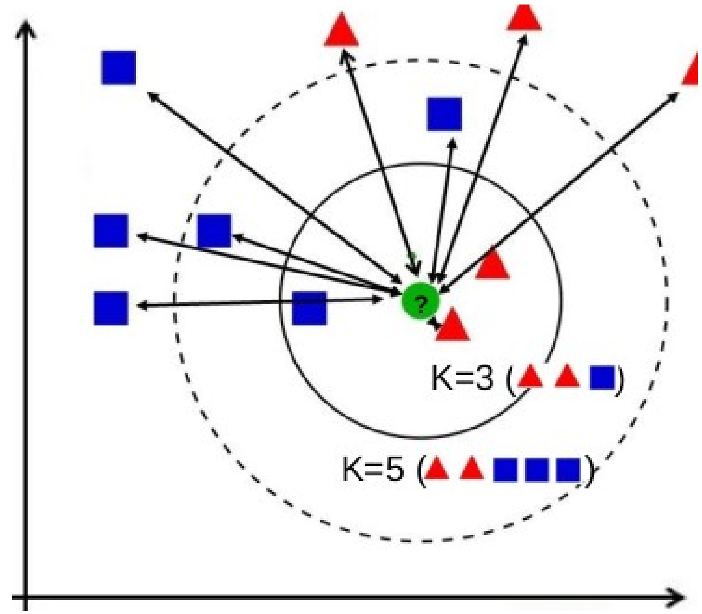
Plano de Aula

- Discussões Iniciais
- Árvores de Decisão
  - Entropia
  - Ganho de Informação
- Exercícios



Discussões Iniciais

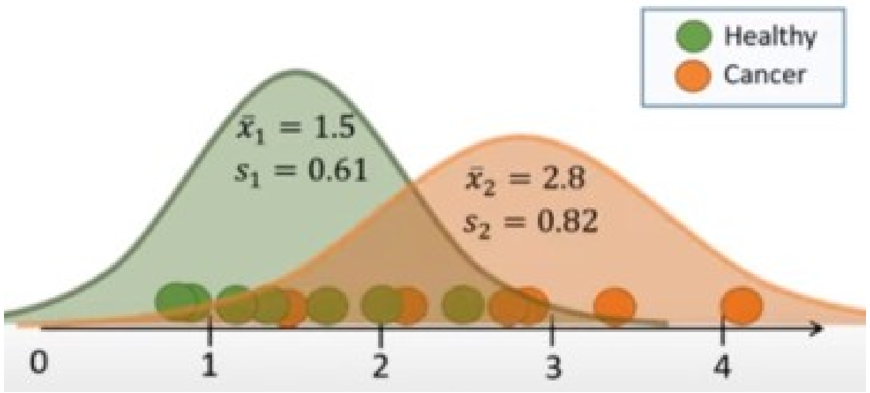
- KNN



- Naive Bayes

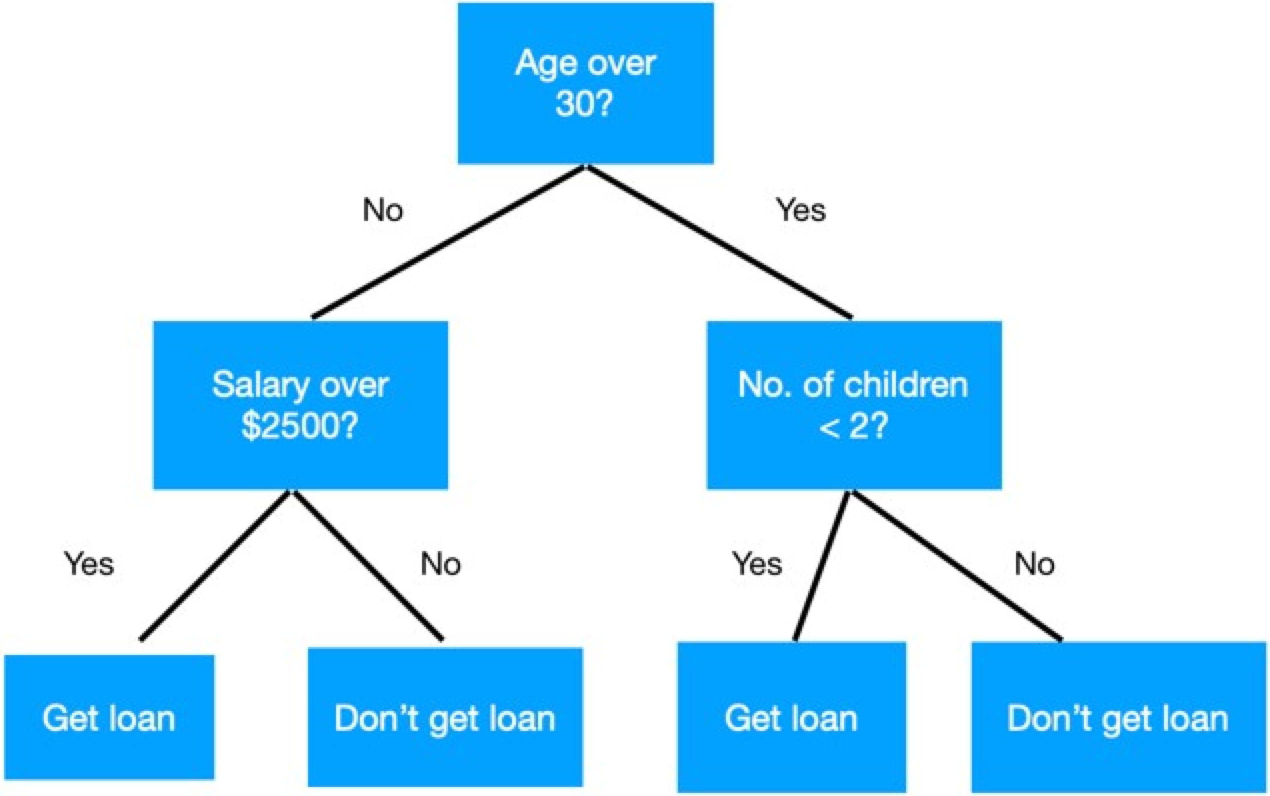
Income	Student	Credit Rating	Buys computer
high	no	fair	YES
medium	no	fair	NO
medium	[illegible]	excellent	NO
high	[illegible]	fair	YES
high	no	fair	NO
medium	yes	fair	YES

Atributos Independ.



Árvores de Decisão

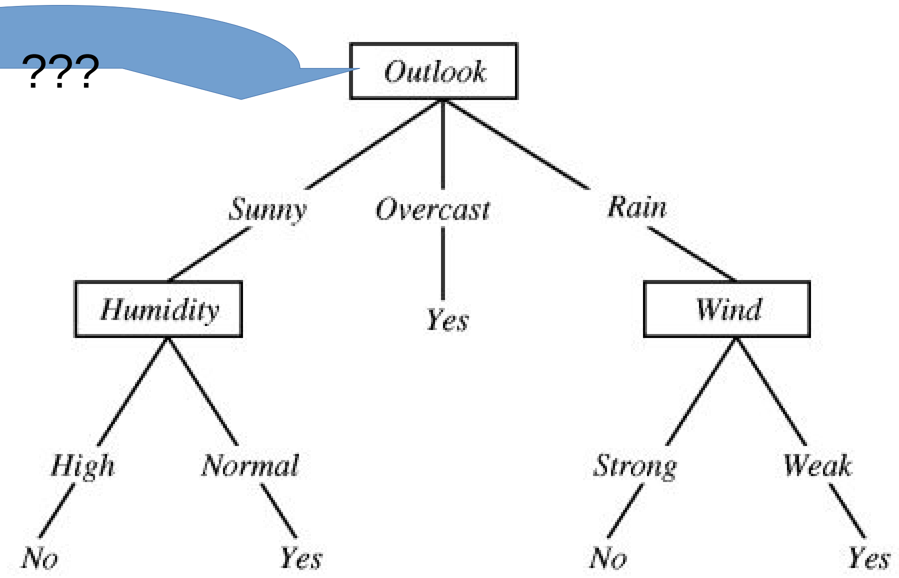
- Estrutura Hierárquica
- Cada nodo é responsável por um nível de decisão



Árvores de Decisão

- Como definir atributos relevantes?
- Como definir a hierarquia entre os atributos?

Day	Outlook	Temp	Humidity	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes



Árvores de Decisão

- Como definir atributos relevantes e limiares?
- Como definir a hierarquia entre os atributos?

  - Entropia

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

  - Ganho de Informação

    - Gain(S, A) = Entropy(S) - ∑[p(Sv) * Entropy(Sv)]

Árvores de Decisão

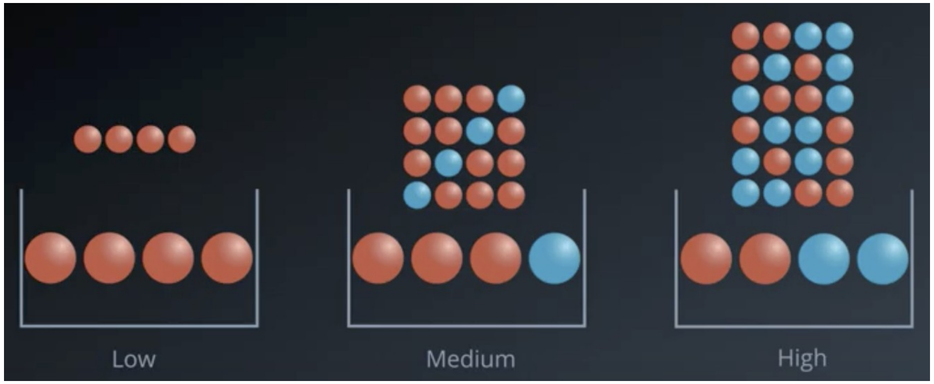
- Entropia: Grau de Incerteza ou Desordem dos **DADOS**

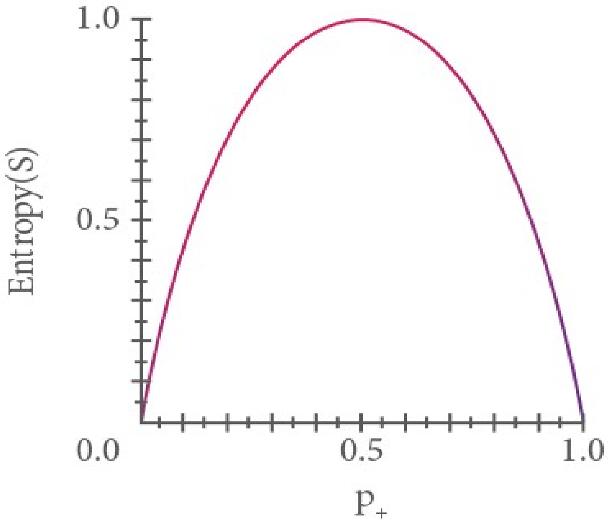
$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

  - S = Dataset
  - C = Número de Classes
  - P_i = proporção da classe ‘i’ no conjunto
  - E = 1 (Entropia Máxima)

Árvores de Decisão

- Calcular a entropia do conjunto abaixo:
  - 50 bolas vermelhas
  - 50 bolas azuis
  - $$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$
    - E = [-p(vermelha) * log2(p(vermelha))] + [- p(azul) * log2(p(azul))]
    - E = [-0.5 * log2(0.5)] + [-0.5 * log2(0.5)]
    - E = [-0.5 * (-1)] + [- 0.5 * (-1)] = 1
  - E para 98 vermelhas e 2 azuis
    - E = [-0.98 * log2(0.98)] + [-0.02 * log2(0.02)]
    - E = 0.141





Árvores de Decisão

- Entropia vs Probabilidade
  - Probabilidade: Chance ou Incerteza relacionada a um **evento**
  - Entropia: Incerteza ou Desordem associada a um conjunto de **dados**.

- Dado 50 bolas vermelhas e 50 bolas azuis, então:
  - Entropia = 1
  - Probabilidade Vermelha = 50%
  - Probabilidade Azul = 50%

- Dado 98 bolas vermelhas e 2 bolas azuis, então:
  - Entropia Conjunto = 0.141
  - Probabilidade Vermelha = 98%
  - Probabilidade Azul = 2%

Árvores de Decisão

- Ganho de Informação:
  - Gain(S, A) = Entropy(S) - ∑[p(Sv) * Entropy(Sv)]

  - ‘A’ é o atributo que está sendo avaliado
  - ‘Sv’ é o subconjunto dos dados que corresponde ao valor v do atributo A
  - ‘p(Sv)’ é a proporção dos valores em ‘Sv’ em relação ao número de valores no conjunto de dados ‘S’
  - ‘Entropy(S)’ e ‘Entropy(Sv)’ são as entropias do conjunto de dados original e dos subconjuntos resultantes

Árvores de Decisão

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Gain(S, A) = Entropy(S) - ∑[p(Sv) * Entropy(Sv)]

Day	Outlook	Temp	Humidity	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- Calculando a entropia da classe "play tennis":
  - Probabilidade de jogar tênis: 9/14 (Yes) , 5/14 (No)
  - Entropia
    - (9/14) * log2(9/14) - (5/14) * log2(5/14) = 0.940

Árvores de Decisão

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Gain(S, A) = Entropy(S) - ∑[p(Sv) * Entropy(Sv)]

Calculando a entropia para cada valor do atributo "humidity":

- Humidity = High
  - Probabilidade de jogar tênis: 3/7 (Yes), 4/7 (No)
  - Entropia
    - - (3/7) * log2(3/7) - (4/7) * log2(4/7) = 0.985
- Humidity = Normal
  - Probabilidade de jogar tênis: 6/7 (Yes), 1/7 (No)
  - Entropia
    - -(6/7) * log2(6/7) - (1/7) * log2(1/7) = 0.592

- Ganho de informação:

0.940 - [(7/14)*0.985 + (7/14)*0.592] = 0.151

Day	Outlook	Temp	Humidity	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Árvores de Decisão

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Gain(S, A) = Entropy(S) - ∑[p(Sv) * Entropy(Sv)]

- Calculando a entropia para cada valor do atributo "outlook":
  - Outlook = Sunny
    - Probabilidade de jogar tênis: 2/5 (Yes), 3/5 (No)
    - Entropia
      - \- (2/5) * log2(2/5) - (3/5) * log2(3/5) = 0.971
  - Outlook = Overcast
    - Probabilidade de jogar tênis: 4/4 (Yes), 0/4 (No)
    - Entropia = 0 (já que todos jogaram tênis)
  - Outlook = Rainy
    - Probabilidade de jogar tênis: 3/5 (Yes), 2/5 (No)
    - Entropia
      - \- (3/5) * log2(3/5) - (2/5) * log2(2/5) = 0.971

Day	Outlook	Temp	Humidity	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- Ganho de informação = 0.940 - [(5/14)*0.971 + (4/14)*0 + (5/14)*0.971] = 0.247

Árvores de Decisão

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Gain(S, A) = Entropy(S) - ∑[p(Sv) * Entropy(Sv)]

- Resumindo
  - Gain (Tennis, Humidity) = 0.151

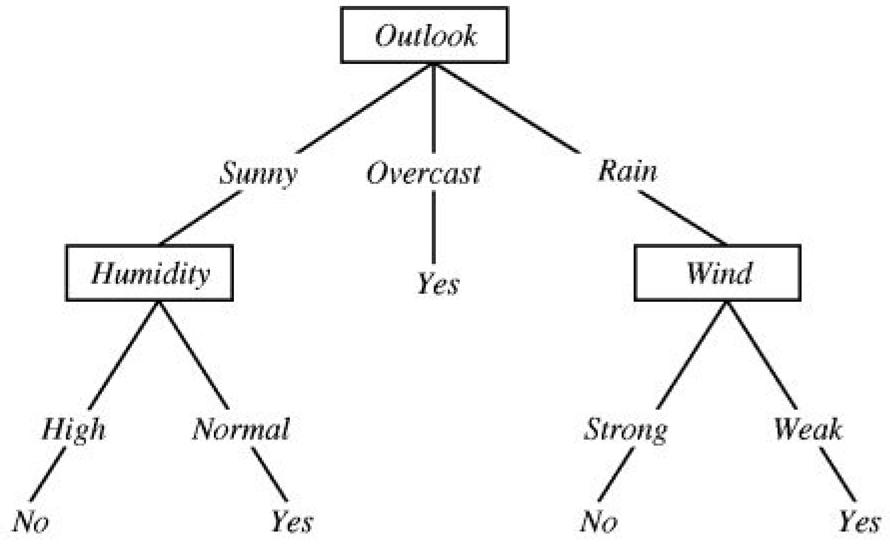
Gain (Tennis, Outlook) = 0.247

Logo, Outlook tem mais ganho de informação

Day	Outlook	Temp	Humidity	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Árvores de Decisão

- Quando parar de construir a árvore?
- De maneira breve:
  - Quando todos os nós folha são puros
  - (nós folha têm dados que pertencem a uma única classe).
  - Quando um determinado critério é atingido (I.E Altura, tempo...)



Árvores de Decisão

- Vantagens:
  - Interpretabilidade
  - Velocidade

- Desvantagens:
  - Overfitting
  - Sensibilidade a dados
  - Dificuldade em capturar relações complexas

Árvores de Decisão

- Let’s Code: Tópico_02_Aprendizado_Supervisionado_Arvores_Decisão.ipynb