Árvores de decisão

Carlos Diego Rodrigues

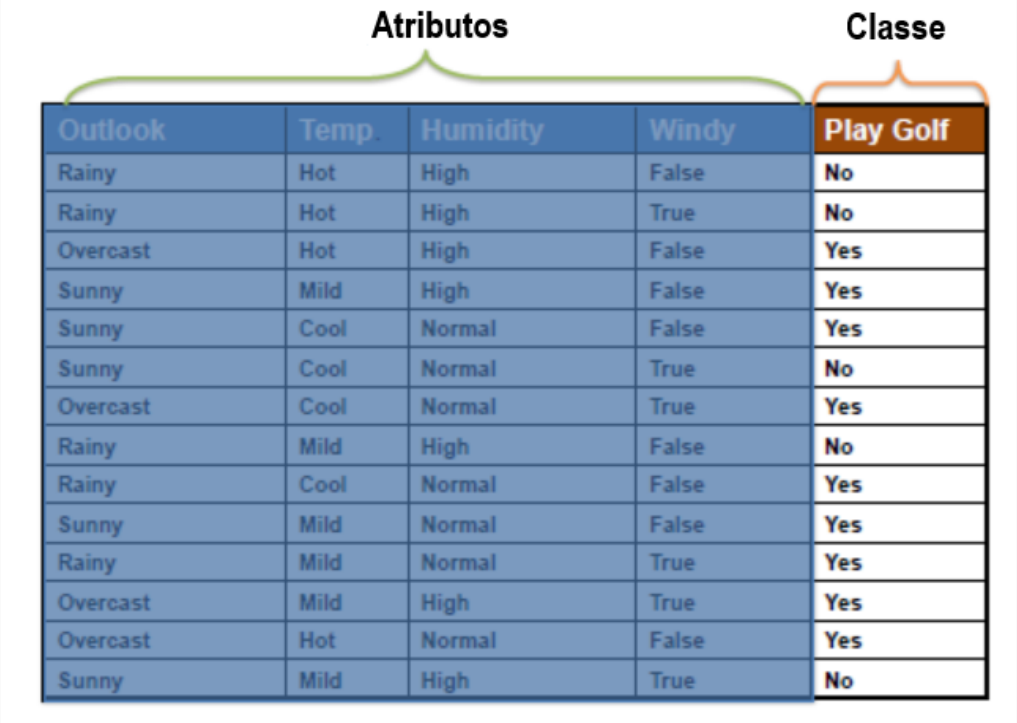
9 de novembro de 2021

Universidade Federal do Ceará

Árvore de decisão

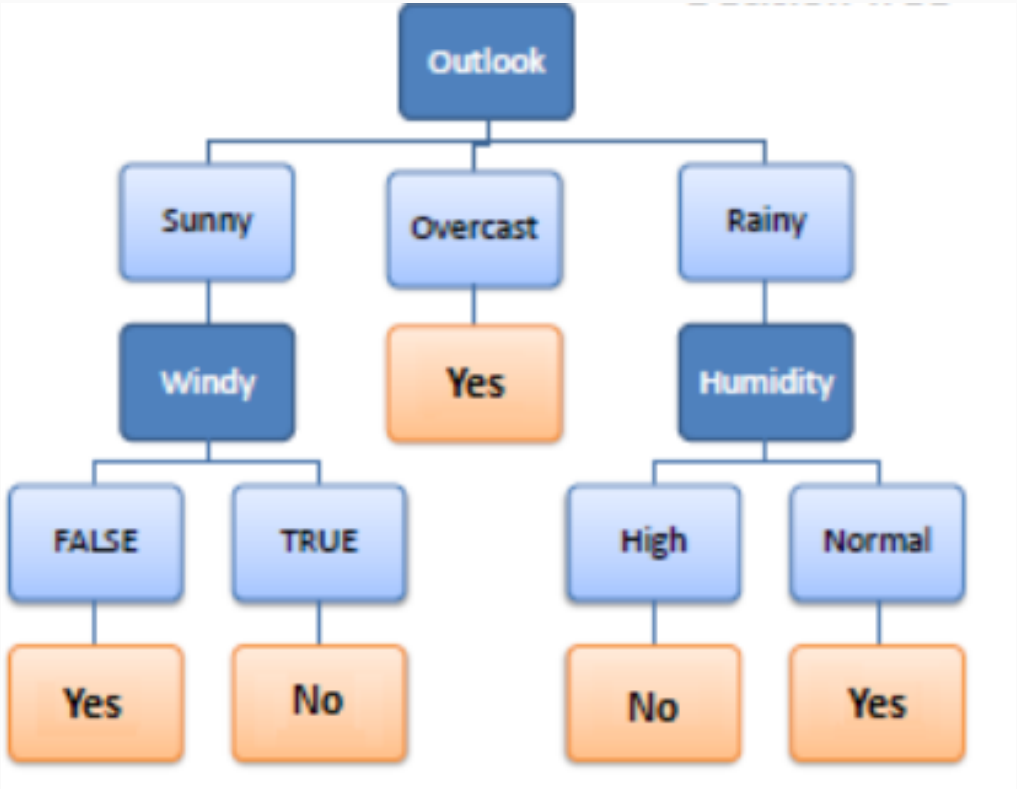
- A construção das árvores de decisão.
- Árvore ID3
- Árvore C4.5

Exemplo de árvore de decisão



Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Exemplo de árvore de decisão



Por que árvores de decisão?

- Apresentação clara do algoritmo de decisão.
- Construção de informação sobre a importância de cada atributo.
- Estrutura altamente eficiente.
- Clareza sobre riscos e recompensas em relação às regras.

Algoritmo ID3

- O algoritmo ID3 constrói uma divisão iterativa dos dados onde em cada nó de decisão ele escolhe um único a ser ramificado.
- São criados ramos para cada valor possível do atributo escolhido.
- O atributo é escolhido a partir dos conceitos de entropia e ganho de informação.
- Os atributos são utilizados conforme possam ser úteis à classificação.

Pseudocódigo

Algorithm 1 ID3

Entrada: A: conjunto de atributos, S: conjunto de instâncias
Saída: R: uma árvore

1: Criar um nó raiz R.
2: **if** $\forall x \in S,\ Classe(x) = 1$ **then**
3: $\quad R \leftarrow 1$
4: **else**
5: $\quad$ **if** $\forall x \in S,\ Classe(x) = 0$ **then**
6: $\quad\quad R \leftarrow 0$
7: $\quad$ **else**
8: $\quad\quad$ **if** $A = \emptyset$ **then**
9: $\quad\quad\quad R \leftarrow$ Resultado mais comum em S.
10: $\quad\quad$ **else**
11: $\quad\quad\quad$ Seja D o atributo com maior *ganho de informação.*
12: $\quad\quad\quad$ **for each** $d_i \in D$ **do**
13: $\quad\quad\quad\quad$ Criar um novo ramo a partir de R com $D = d_i$.
14: $\quad\quad\quad\quad$ Adicionar a subárvore ID3$(A - D,\ S[D = d_i])$

Entropia

- A entropia em uma tabela de um único atributo é definida como:

$$H(S) = \sum_{v \in V(S)} -p_v log_2 p_v$$

- Em uma tabela de dois atributos define-se como:

$$H(S, X) = \sum_{v \in V(X)} \frac{|X_v|}{|X|} H(X_v)$$

- Ganho de informação é definido como:

$$G(S, X) = H(S) - H(S, X)$$

Ganhos de informação da tabela original

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

Algoritmo C4.5

- Melhorias sobre o algoritmo ID3:
  - Tratamento de valores numéricos: discretos e contínuos.
  - Tratamento para valores faltantes ou desconhecidos.
- Poda das árvores de decisão.
- Razão de ganho

Tratamento de valores numéricos.

- A ideia é criar faixas de valores nos quais os valores podem ser inseridos.
- Temos duas maneiras distintas: **sem supervisão** ou **com supervisão**.
- Sem supervisão:
  - Faixas de mesmo tamanho.
  - Faixas de mesma frequência.
  - Quantis e outros métodos.
- Com supervisão:
  - Utilizando-se do conceito de entropia e ganho de informação, tentar estipular os pontos de corte da variável que podem melhor discernir sobre a classe a ser prevista.

Tratamento de valores faltantes.

- Valores faltantes são muito comuns em aplicações práticas.
- Políticas para valores faltantes:
  - Ignorar as instâncias com valores faltantes.
  - Dar um novo valor para indicar a falta de dados.
  - Preencher manualmente a partir de conhecimento prévio.
  - Utilizar uma média (quantitativo) ou moda (qualitativo).
  - Utilizar técnicas de ciência de dados para prever o valor faltante.
- Atualização do ganho de informação: ponderar pela proporção de instâncias em que o valor é conhecido.

$$G(S, X) = F \cdot (H(S) - H(S, X))$$

Poda das árvores de decisão.

- Árvores completas podem sofrer com o fenômeno do *overfitting*.
- Pré-poda: na construção da árvore são feitas avaliações se novos nós devem ser adicionados.
- Pós-poda: após a construção da árvore, nós são removidos para termos uma árvore mais eficiente.

Razão de ganho

- O ganho de informação pode esconder uma armadilha: o número de instâncias afetadas não é contabilizado.
- Então divisões em partes muito pequenas podem parecer boas, mas refletem na realidade apenas particularidades que podem não ter conexão com a generalidade.
- Por isso o autor dos métodos ID3 e C4.5 sugere a utilização de uma medida "padronizada", a taxa de ganho.
- Define-se portanto um fator de taxa chamado *split info* (valor intrínseco):

$$I(S,X) = \sum_{v \in V(X)} \frac{|X_v|}{|X|} log_2 \frac{|X_v|}{|X|}$$

- E a taxa de ganho passa a ser:

$$GR(S,X) = \frac{G(S,X)}{I(S,X)}$$