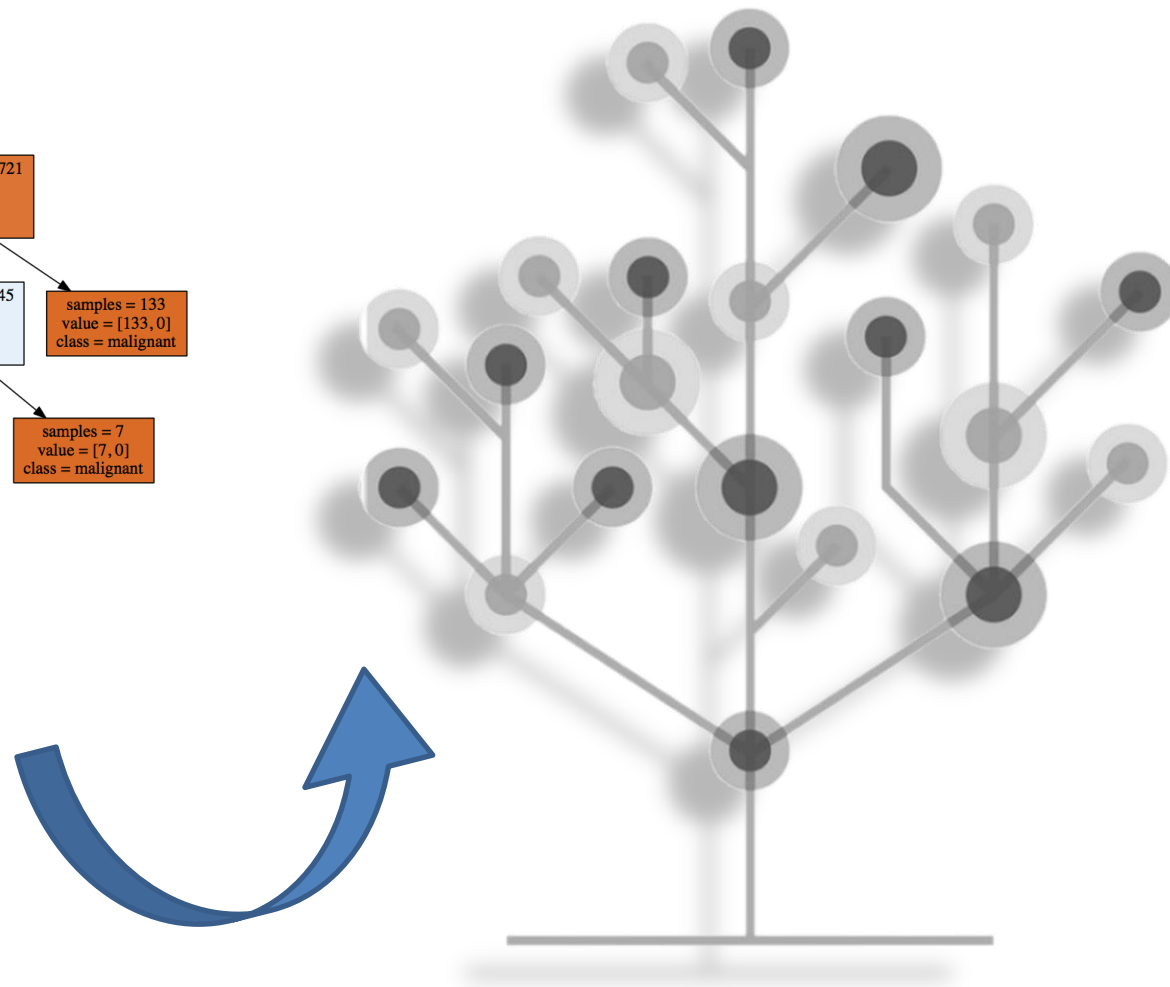
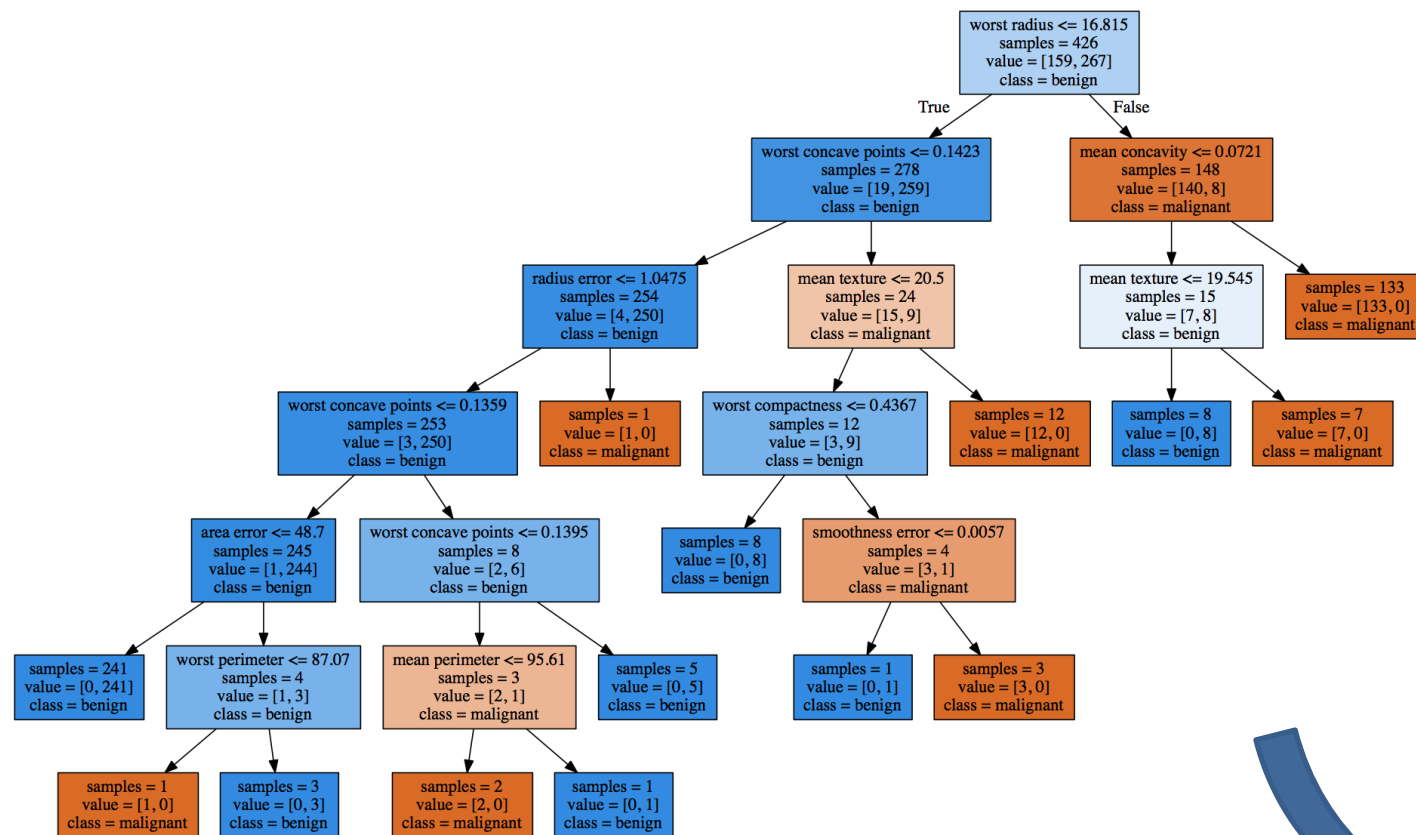




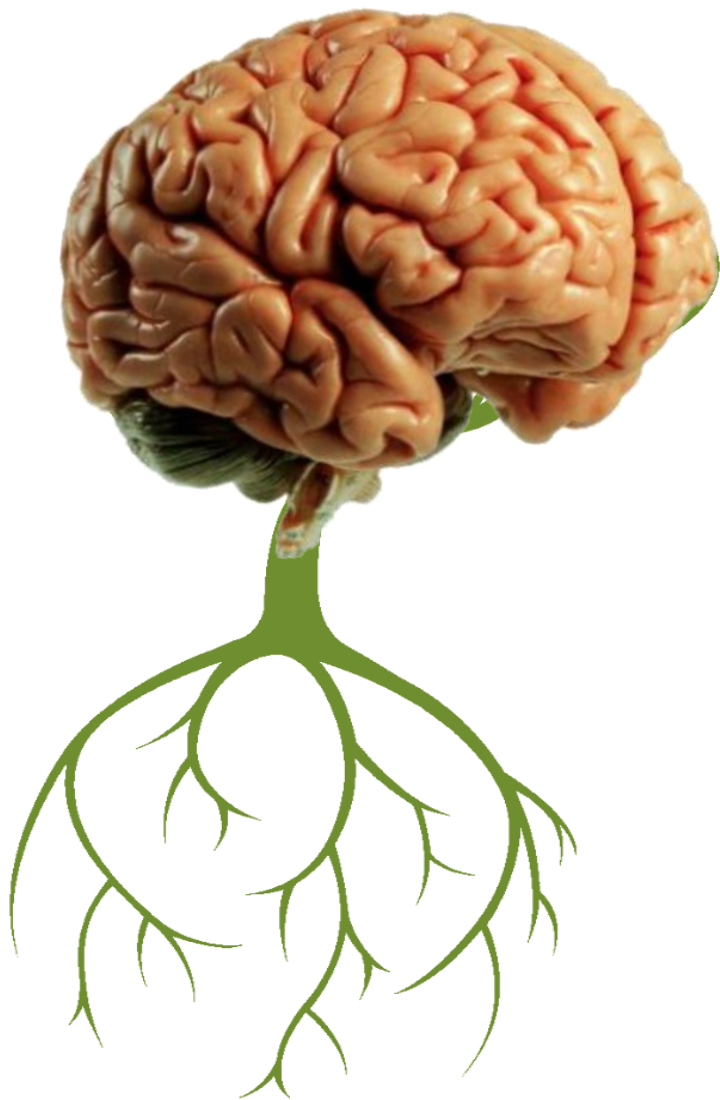
Reconhecimento de padrões e aprendizagem computacional

Árvores de decisão

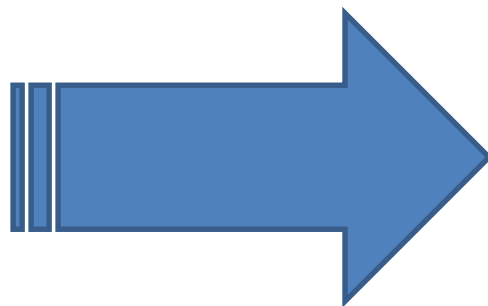
Árvores de decisão



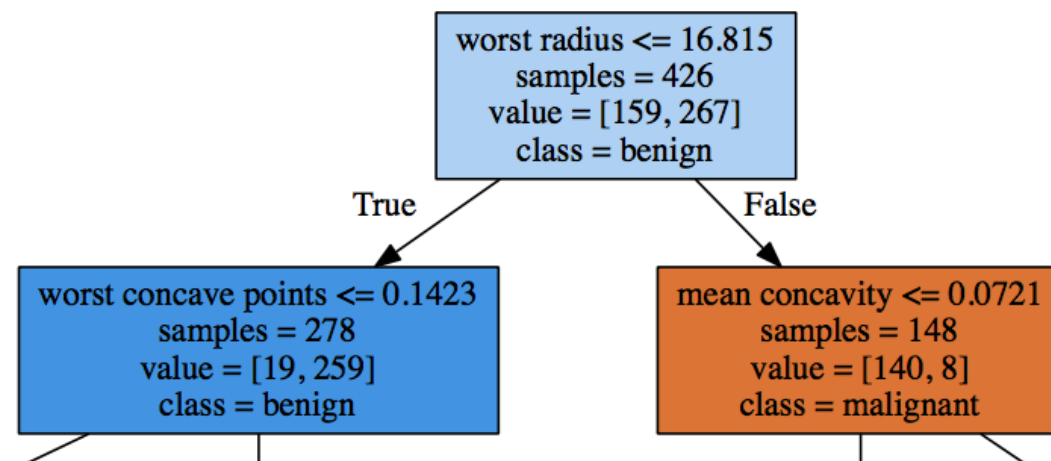
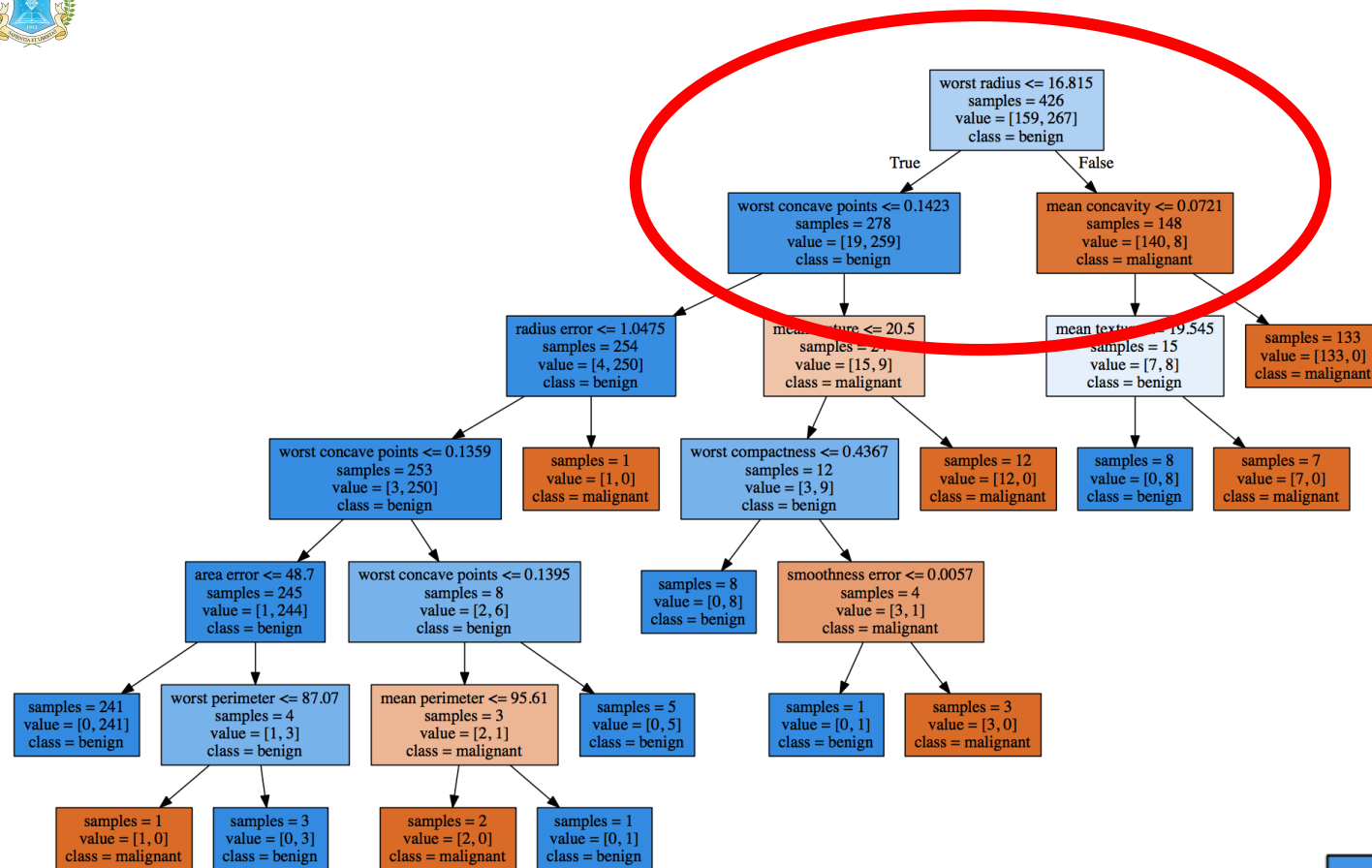
Árvores de decisão



**Mas como será
que ela pensa?**



Árvores de decisão





Árvores de decisão

Definição

Quais palavras devemos colocar aqui?

- | | |
|---|---|
| - | - |
| - | - |
| - | - |
| - | - |

Forme sua definição agora:



Árvores de decisão

Definições informais

- É um algoritmo de aprendizagem de máquina supervisionado que se baseia na ideia de divisão dos dados em grupos homogêneos, podem ser utilizadas em um cenário de classificação ou regressão.
- Uma árvore de decisão é uma ferramenta de **suporte à tomada de decisão** que usa um gráfico no formato de árvore e demonstra visualmente as condições e as probabilidades para se chegar a resultados.



Árvores de decisão

Definição

- É um algoritmo de aprendizagem supervisionada que possui uma estrutura hierárquica onde cada **nó interno** representa um teste em um atributo, cada **ramo** representa uma saída do teste, e cada **nó folha** (ou nó terminal) representa o rótulo de uma classe.

Árvores de decisão

Terminologias





Árvores de decisão

Por quê as árvores de decisão são tão populares?

- Sua construção não exige conhecimento sobre o domínio, ou determinação (calibração) de parâmetros;
- O conhecimento construído na árvore é altamente acessível;
- Os passos de indução e classificação são rápidos;
- Em geral, têm uma boa acurácia. (?)

Árvores de decisão

Entendendo o processo de construção

- O processo de aprendizagem da estrutura de uma árvore de decisão é conhecido com **indução** ou regras.
- **Indução** é o processo de raciocínio sobre um dado conjunto de fatos para princípios gerais ou regras.



Por indução → Gosto de esportes

- A indução **busca padrões** em informações disponíveis com o propósito de inferir conclusões racionais



Árvores de decisão

Construindo uma árvore de decisão

O processo de criação de uma árvore de decisão é composto por três etapas:

1. Adotar um critério para a criação de um nó;
2. Classificação de um nó como terminal ou não terminal;
3. Geração de um conjunto de árvores podadas.



Árvores de decisão

Criação de um nó

- A decisão de fazer as divisões dos nós afeta muito a precisão de uma árvore.
- Os critérios de decisão são diferentes para árvores de classificação e de regressão.
- A árvore de decisão divide os nós em todas as variáveis disponíveis e seleciona a divisão que resulta em sub-nós mais homogêneos.
- A seleção do algoritmo também é baseada no tipo de variáveis de destino.



Árvores de decisão

Algoritmos para criação de nós

- Vejamos os quatro algoritmos mais usados na árvore de decisão:

- **Índice Gini**

O índice Gini diz que se selecionarmos aleatoriamente dois itens de uma população, então ambos devem ser da mesma classe e a probabilidade disto é 1 se a população for pura. **O CART (Árvore de Classificação e Regressão) usa o método Gini para criar divisões binárias**

- Qui-Quadrado
- Redução na Variância
- Ganho de informação



Árvores de decisão

Algoritmos para criação de nós

- Vejamos os quatro algoritmos mais usados na árvore de decisão:

- Índice Gini
- **Qui-Quadrado**

É um algoritmo para descobrir a significância estatística entre as diferenças dos sub-nós e do nó pai. O qui-quadrado é medido pela soma dos quadrados das diferenças entre as frequências observadas e esperadas da variável alvo.

Gera a árvore chamada de CHAID (Chi-square Automatic Interaction Detector)

- Redução na Variância
- Ganho de informação



Árvores de decisão

Algoritmos para criação de nós

- Vejamos os quatro algoritmos mais usados na árvore de decisão:

- Índice Gini
- Qui-Quadrado
- **Redução na Variância**

A redução na variância é um algoritmo usado para **variáveis alvo contínuas** (problemas de regressão). Este algoritmo utiliza a fórmula padrão de variância para escolher a melhor divisão. A divisão com menor variância é selecionada como critério para dividir a população.

- Ganho de informação



Árvores de decisão

Algoritmos para criação de nós

- Vejamos os quatro algoritmos mais usados na árvore de decisão:

- Índice Gini
- Qui-Quadrado
- Redução na Variância

- **Ganho de informação**

A teoria da informação dá uma medida para definir esse grau de desorganização em sistema – entropia.

Se a amostra for completamente homogênea, então a entropia é zero.

Se a amostra for dividida em partes iguais (50% – 50%), então terá entropia de um.



Árvores de decisão

Exemplo 1

- Suponha que deseja-se saber se haverá ou não trabalho de campo de acordo com as condições climáticas:

Atributo	Possíveis valores
CÉU	Sol, Nublado, Chuva
TEMPERATURA	Alta, Suave, Baixa
UMIDADE	Alta, Normal
VENTO	Sim, Não



Árvores de decisão

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
1	Sol	Alta	Alta	Não	Não ir
2	Sol	Alta	Alta	Sim	Não ir
3	Nublado	Alta	Alta	Não	Ir ao campo
4	Chuva	Alta	Alta	Não	Ir ao campo
5	Chuva	Baixa	Normal	Não	Ir ao campo
6	Chuva	Baixa	Normal	Sim	Não ir
7	Nublado	Baixa	Normal	Sim	Ir ao campo
8	Sol	Suave	Alta	Não	Não ir
9	Sol	Baixa	Normal	Não	Ir ao campo
10	Chuva	Suave	Normal	Não	Ir ao campo
11	Sol	Suave	Normal	Sim	Ir ao campo
12	Nublado	Suave	Alta	Sim	Ir ao campo
13	Nublado	Alta	Normal	Não	Ir ao campo
14	Chuva	Suave	Alta	Sim	Não ir

Possibilidades:

- **9** Ir ao campo
- **5** Não ir

Árvores de decisão

Entropia

- O cálculo do ganho de informação consiste na análise da homogeneidade das subclasses criadas, escolhendo assim, o critério que traga um maior ganho de informação (*Ganho*):

$$Ganho = info(T) - \sum_{t=1}^m \frac{|T_t|}{|T|} * info(T_t)$$

$$info(T) = - \sum_{j=1}^k \frac{freq(C_j, T)}{|T|} * \log_2 \left(\frac{freq(C_j, T)}{|T|} \right) \quad (\text{Entropia})$$

Onde:

- *freq (C_j,T)*: é o número de amostras *T* subdivididas no subespaço *C_j*;
- *T* : é o número total de amostras;
- *k*: é o número de classes existentes;
- *m*: é o número de subespaços criados na divisão de *T*.

Árvores de decisão

Exemplo 1

- O objetivo do cálculo da entropia está na classificação booleana
(Ir ao campo × Não ir ao campo)
- Há 14 exemplos, 9 positivos e 5 negativos, ou seja, $T = [9+, 5 -]$.

$$\begin{aligned}\text{info}(T) &= -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \dots - p_n \log_2 p_n \\ &= -\left(\frac{9}{14}\right) \log_2 \left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2 \left(\frac{5}{14}\right) \\ &= 0,940\end{aligned}$$

- Após calcular a entropia do sistema, busca-se **qual atributo** possui melhor ganho de informação.

Árvores de decisão

Exemplo 1

- **Ganho de informação – CÉU**
- O atributo céu pode assumir 3 valores (Sol, Nublado e Chuva).

$$T_{\text{sol}} = [2+, 3-], T_{\text{nublado}} = [4+, 0-] \text{ e } T_{\text{chuva}} = [3+, 2-]$$

$$\text{info}(\text{sol}) = -\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) = 0,97094$$

$$\text{info}(\text{nublado}) = -\left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right) = 0$$

$$\text{info}(\text{chuva}) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) = 0,97094$$

Logo,

$$\begin{aligned} \text{Ganho}(\text{info}(T), \text{céu}) &= 0,940 - \left(\frac{5}{14}\right) \cdot \text{info}(\text{sol}) - \left(\frac{4}{14}\right) \cdot \text{info}(\text{nublado}) - \left(\frac{5}{14}\right) \cdot \text{info}(\text{chuva}) \\ &= 0,940 - \left(\frac{5}{14}\right) \cdot 0,97094 - \left(\frac{4}{14}\right) \cdot 0 - \left(\frac{5}{14}\right) \cdot 0,97094 = 0,2464 \end{aligned}$$

Árvores de decisão

Exemplo 1

- **Ganho de informação – TEMPERATURA**
- O atributo **TEMP.** pode assumir 3 valores (Alta, Suave e Baixa).

$$T_{\text{alta}} = [3+, 2-], T_{\text{suave}} = [3+, 1-] \text{ e } T_{\text{baixa}} = [3+, 2-]$$

$$\text{info}(\text{alta}) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) = 0,97094 \quad \text{info}(\text{suave}) = -\left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) = 0,811$$

$$\text{info}(\text{baixa}) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) = 0,97094$$

Logo,

$$\begin{aligned} \text{Ganho}(\text{info}(T), \text{temperatura}) &= 0,940 - \left(\frac{5}{14}\right) \cdot \text{info}(\text{alta}) - \left(\frac{4}{14}\right) \cdot \text{info}(\text{suave}) - \left(\frac{5}{14}\right) \cdot \text{info}(\text{baixa}) \\ &= 0,940 - \left(\frac{5}{14}\right) \cdot 0,97094 - \left(\frac{4}{14}\right) \cdot 0,811 - \left(\frac{5}{14}\right) \cdot 0,97094 = \boxed{0,015} \end{aligned}$$



Árvores de decisão

Exemplo 1

- **Ganho de informação – UMIDADE**
- O atributo **UMID.** pode assumir 2 valores (Alta e Baixa).

$$T_{\text{alta}} = [3+, 4-] \text{ e } T_{\text{baixa}} = [6+, 1-]$$

$$\text{info}(\text{alta}) = -\left(\frac{3}{7}\right)\log_2\left(\frac{3}{7}\right) - \left(\frac{4}{7}\right)\log_2\left(\frac{4}{7}\right) = 0,985228 \quad \text{info}(\text{baixa}) = -\left(\frac{6}{7}\right)\log_2\left(\frac{6}{7}\right) - \left(\frac{1}{7}\right)\log_2\left(\frac{1}{7}\right) = 0,591672$$

Logo,

$$\begin{aligned} \text{Ganho}(\text{info}(T), \text{umidade}) &= 0,940 - \left(\frac{7}{14}\right) \cdot \text{info}(\text{alta}) - \left(\frac{7}{14}\right) \cdot \text{info}(\text{baixa}) \\ &= 0,940 - \left(\frac{7}{14}\right) \cdot 0,985228 - \left(\frac{7}{14}\right) \cdot 0,591672 \quad \boxed{= 0,151} \end{aligned}$$



Árvores de decisão

Exemplo 1

- **Ganho de informação – VENTO**
- O atributo **VENTO** pode assumir 2 valores (Sim e Não).

$$T_{\text{sim}} = [3+, 3-], T_{\text{não}} = [6+, 2-]$$

$$\text{info}(\text{sim}) = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) = 1 \quad \text{info}(\text{não}) = -\left(\frac{6}{8}\right)\log_2\left(\frac{6}{8}\right) - \left(\frac{2}{8}\right)\log_2\left(\frac{2}{8}\right) = 0,811278$$

Logo,

$$\begin{aligned} \text{Ganho}(\text{info}(T), \text{vento}) &= 0,940 - \left(\frac{8}{14}\right) \cdot \text{info}(\text{sim}) - \left(\frac{6}{14}\right) \cdot \text{info}(\text{não}) \\ &= 0,940 - \left(\frac{6}{14}\right) \cdot 1 - \left(\frac{8}{14}\right) \cdot 0,811278 = \boxed{0,047841} \end{aligned}$$

Árvores de decisão

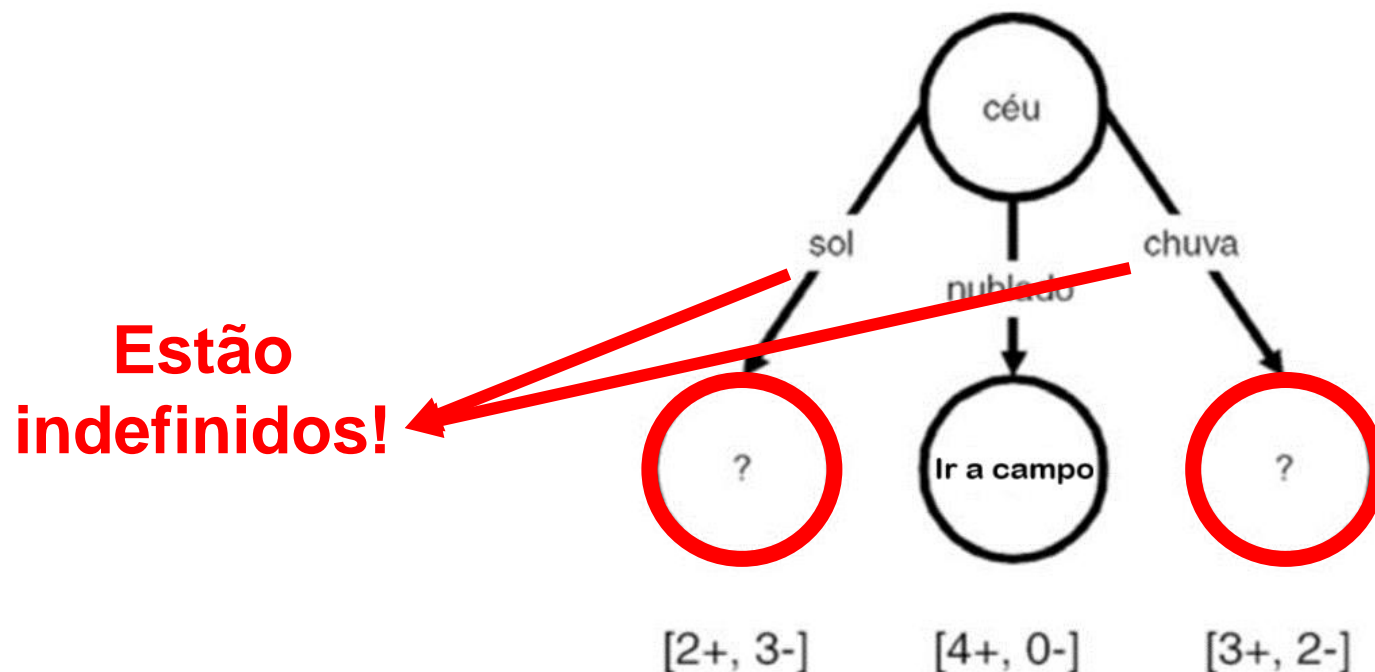
Exemplo 1

- *Ganho info T*, Céu = **0,246**
- *Ganho info T*, Umid. = 0,151

Ganho info T, Temp. = 0,015

Ganho info T, Vento = 0,047

Maior ganho de informação para ser o nó raiz da árvore?



O processo deve continuar no próximo nível da árvore.

Árvores de decisão

Exemplo 1

- As amostras do conjunto de treinamento T são divididos em subconjuntos de acordo com os valores do atributo céu.
- Derivando em 3 subconjuntos:

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
1	Sol	Alta	Alta	Não	Não ir
2	Sol	Alta	Alta	Sim	Não ir
8	Sol	Suave	Alta	Não	Não ir
9	Sol	Baixa	Normal	Não	Ir ao campo
11	Sol	Suave	Normal	Sim	Ir ao campo

Árvores de decisão

Exemplo 1

- As amostras do conjunto de treinamento T são divididos em subconjuntos de acordo com os valores do atributo céu.
- Derivando em 3 subconjuntos:

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
3	Nublado	Alta	Alta	Não	Ir ao campo
7	Nublado	Baixa	Normal	Sim	Ir ao campo
12	Nublado	Suave	Alta	Sim	Ir ao campo
13	Nublado	Alta	Normal	Não	Ir ao campo

Árvores de decisão

Exemplo 1

- As amostras do conjunto de treinamento T são divididos em subconjuntos de acordo com os valores do atributo céu.
- Derivando em 3 subconjuntos:

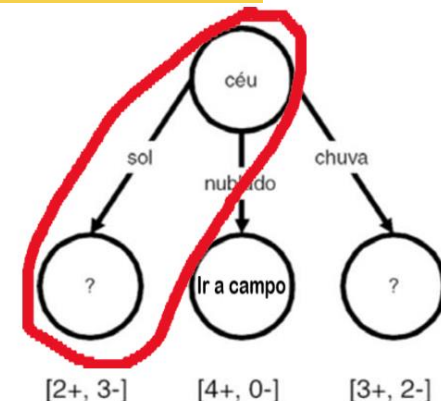
Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
4	Chuva	Alta	Alta	Não	Ir ao campo
5	Chuva	Baixa	Normal	Não	Ir ao campo
6	Chuva	Baixa	Normal	Sim	Não ir
10	Chuva	Suave	Normal	Não	Ir ao campo
14	Chuva	Suave	Alta	Sim	Não ir

Árvores de decisão

Exemplo 1

- **Céu = Sol** (processo de indução para este ramo da árvore)
- **Ganho de informação – TEMPERATURA:**

$$T_{alta} = [2+, 0-], T_{suave} = [1+, 1-] \text{ e } T_{baixa} = [0+, 1-]$$



$$\text{info}(alta) = -\left(\frac{2}{2}\right) \log_2 \left(\frac{2}{2}\right) = 0$$

$$\text{info}(suave) = -\left(\frac{1}{2}\right) \log_2 \left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) \log_2 \left(\frac{1}{2}\right) = 1$$

$$\text{info}(baixa) = -\left(\frac{1}{1}\right) \log_2 \left(\frac{1}{1}\right) - \left(\frac{1}{1}\right) \log_2 \left(\frac{1}{1}\right) = 0$$

Logo,

$$\begin{aligned} \text{Ganho}(\text{info}(\text{sol}), \text{temperatura}) &= 0,97094 - \left(\frac{2}{5}\right) \cdot \text{info}(alta) - \left(\frac{2}{5}\right) \cdot \text{info}(suave) - \left(\frac{1}{5}\right) \cdot \text{info}(baixa) \\ &= 0,97094 - \left(\frac{2}{5}\right) \cdot 0 - \left(\frac{2}{5}\right) \cdot 1 - \left(\frac{1}{5}\right) \cdot 0 = \boxed{0.57094} \end{aligned}$$

Árvores de decisão

Exemplo 1

- **Céu = Sol** (processo de indução para este ramo da árvore)
- **Ganho de informação – UMIDADE:**

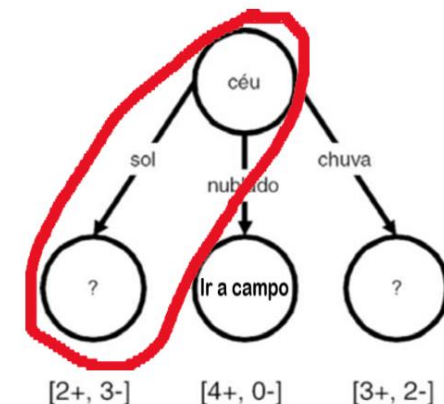
$$T_{\text{alta}} = [3+, 0-] \text{ e } T_{\text{baixa}} = [0+, 2-]$$

$$\text{info}(\text{alta}) = -\left(\frac{3}{3}\right) \log_2 \left(\frac{3}{3}\right) = 0$$

$$\text{info}(\text{baixa}) = -\left(\frac{2}{2}\right) \log_2 \left(\frac{2}{2}\right) = 0$$

Logo,

$$\begin{aligned} \text{Ganho}(\text{info}(s), \text{umidade}) &= 0,97094 - \left(\frac{3}{5}\right) \cdot \text{info}(\text{alta}) - \left(\frac{2}{5}\right) \cdot \text{info}(\text{baixa}) \\ &= 0,97094 - \left(\frac{3}{5}\right) \cdot 0 - \left(\frac{2}{5}\right) \cdot 0 = \boxed{0,97094} \end{aligned}$$



Árvores de decisão

Exemplo 1

- **Céu = Sol** (processo de indução para este ramo da árvore)
- **Ganho de informação – VENTO:**

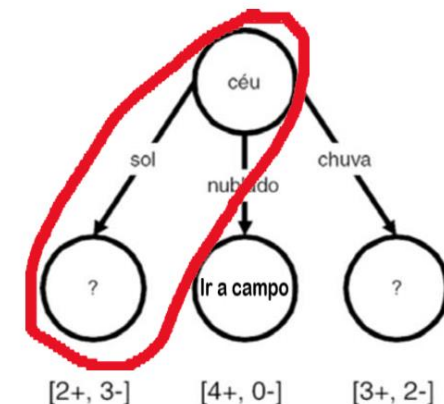
$$T_{\text{sim}} = [1+, 1-], T_{\text{não}} = [2+, 1-]$$

$$\text{info}(\text{sim}) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

$$\text{info}(\text{não}) = -\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) = 0,918295$$

Logo,

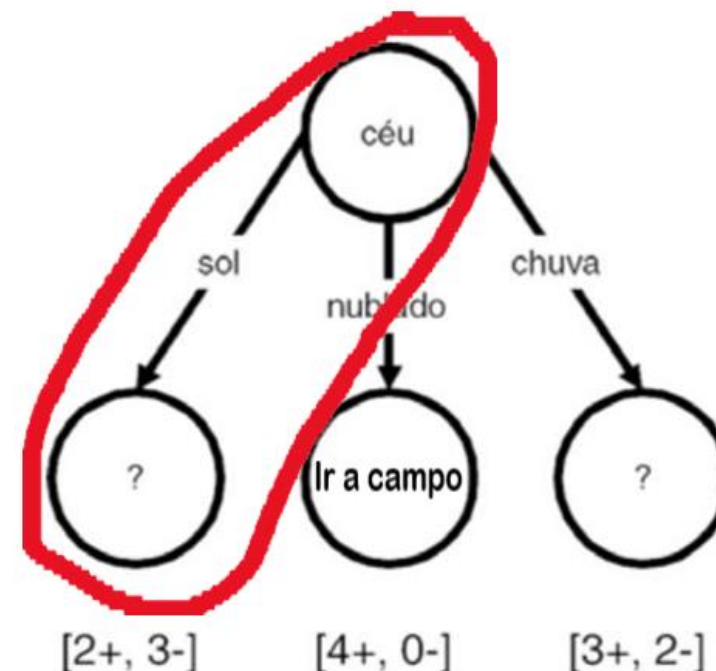
$$\begin{aligned} \text{Ganho}(\text{info}(\text{sol}), \text{vento}) &= 0,97094 - \left(\frac{2}{5}\right) \cdot \text{info}(\text{sim}) - \left(\frac{3}{5}\right) \cdot \text{info}(\text{não}) \\ &= 0,97094 - \left(\frac{2}{5}\right) \cdot 1 - \left(\frac{3}{5}\right) \cdot 0,918295 = \boxed{0.019963} \end{aligned}$$



Árvores de decisão

Exemplo 1

- *Ganho info sol*, Temperatura = 0,57094
- ***Ganho info sol*, Umidade = 0,97094**
- *Ganho info sol*, Vento = 0,019963



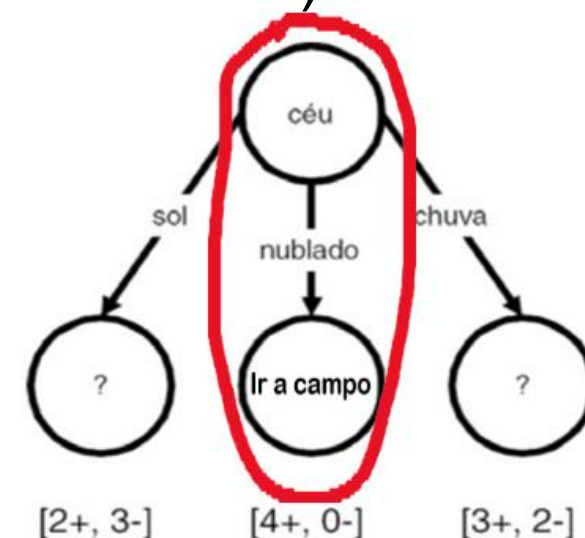
- Examinando os ganhos verifica-se que o atributo com maior ganho de informação é a **umidade**, o qual deve ser o nó seguinte da árvore neste ramo.

Árvores de decisão

Exemplo 1

- **Céu = Nublado** (processo de indução para este ramo da árvore)

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
3	Nublado	Alta	Alta	Não	Ir ao campo
7	Nublado	Baixa	Normal	Sim	Ir ao campo
12	Nublado	Suave	Alta	Sim	Ir ao campo
13	Nublado	Alta	Normal	Não	Ir ao campo



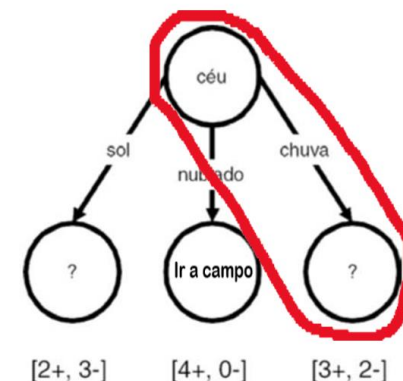
- Observa-se que todas as amostras contidas nesse subconjunto pertencem somente a uma classe (**Ir ao campo**).
- Neste caso, o processo de indução **acaba para este subconjunto** e um nó folha é gerado.

Árvores de decisão

Exemplo 1

- **Céu = Chuva** (processo de indução para este ramo da árvore)
- **Ganho de informação – TEMPERATURA:**

$$T_{\text{alta}} = [0+, 1-], T_{\text{suave}} = [1+, 1-] \text{ e } T_{\text{baixa}} = [1+, 1-]$$



$$\text{info}(\text{alta}) = -\left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right) = 0 \quad \text{info}(\text{suave}) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1 \quad \text{info}(\text{baixa}) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

Logo,

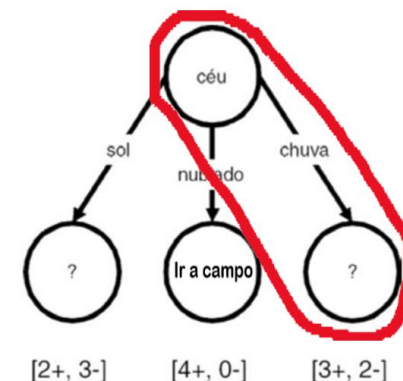
$$\begin{aligned} \text{Ganho}(\text{info}(\text{chuva}), \text{temperatura}) &= 0,97094 - \left(\frac{1}{5}\right) \cdot \text{info}(\text{alta}) - \left(\frac{2}{5}\right) \cdot \text{info}(\text{suave}) - \left(\frac{2}{5}\right) \cdot \text{info}(\text{baixa}) \\ &= 0,97094 - \left(\frac{1}{5}\right) \cdot 0 - \left(\frac{2}{5}\right) \cdot 1 - \left(\frac{2}{5}\right) \cdot 1 = 0,17090 \end{aligned}$$

Árvores de decisão

Exemplo 1

- **Céu = Chuva** (processo de indução para este ramo da árvore)
- **Ganho de informação – UMIDADE:**

$$T_{\text{alta}} = [1+, 1-], \quad T_{\text{normal}} = [2+, 1-]$$



$$\text{info}(\text{alta}) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

$$\text{info}(\text{normal}) = -\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) = 0,9182958$$

Logo,

$$\text{Ganho}(\text{info}(\text{chuva}), \text{umidade}) = 0,97094 - \left(\frac{2}{5}\right) \cdot 1 - \left(\frac{3}{5}\right) \cdot 0,9182958 = 0,019962$$

Árvores de decisão

Exemplo 1

- **Céu = Chuva** (processo de indução para este ramo da árvore)
- **Ganho de informação – VENTO:**

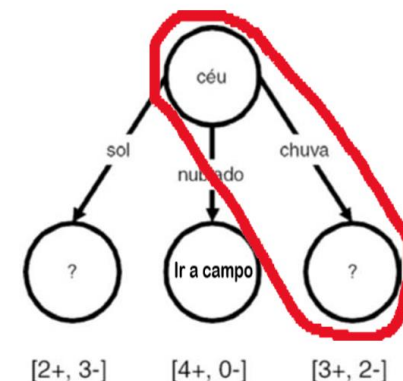
$$T_{\text{sim}} = [2+, 0-], T_{\text{não}} = [0+, 3-]$$

$$\text{info}(\text{sim}) = -\left(\frac{2}{2}\right) \log_2 \left(\frac{2}{2}\right) = 0$$

$$\text{info}(\text{não}) = -\left(\frac{3}{3}\right) \log_2 \left(\frac{3}{3}\right) = 0$$

Logo,

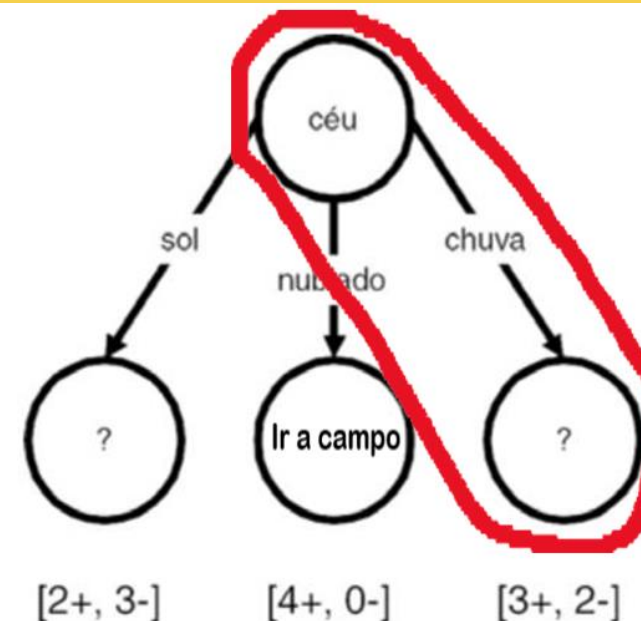
$$\begin{aligned} \text{Ganho}(\text{info}(\text{chuva}), \text{vento}) &= 0,97094 - \left(\frac{2}{5}\right) \cdot \text{info}(\text{sim}) - \left(\frac{3}{5}\right) \cdot \text{info}(\text{não}) \\ &= 0,97094 - \left(\frac{2}{5}\right) \cdot 0 - \left(\frac{3}{5}\right) \cdot 0 \quad \boxed{=0,97094} \end{aligned}$$



Árvores de decisão

Exemplo 1

- *Ganho info chuva, temperatura = 0,17090*
- *Ganho info chuva, umidade = 0,019962*
- ***Ganho info chuva, vento = 0,97094***

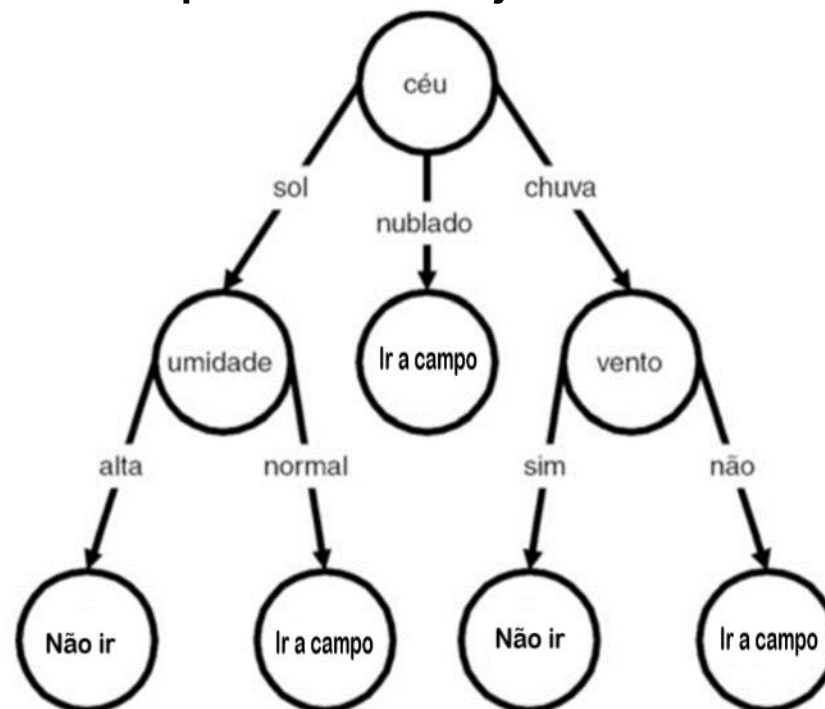


- Examinando os ganhos verifica-se que o atributo com maior ganho de informação é o **vento**, o qual deve ser o nó seguinte na árvore

Árvores de decisão

Exemplo 1

- Árvore de decisão final para o conjunto de treinamento:



E o atributo temperatura?

É irrelevante para a tarefa de classificação, neste caso.



Árvores de decisão

Considerações

- No Exemplo 1, haviam apenas atributos categóricos, e o algoritmo utilizado para a criação da árvore foi o **ID3 (Itemized Dichotomizer 3)**.
- Se os atributos temperatura e umidade do Exemplo 1 fossem atributos contínuos, o ID3 construiria um nó umidade com **14 ramos**, visto que o método cria um ramo para cada valor deste atributo.
- Para dados contínuos, pode-se usar o algoritmo **C4.5** (evolução do ID3).



Árvores de decisão

Exemplo 2

- Suponha que deseja-se saber se haverá ou não trabalho de campo de acordo com as condições climáticas:

Atributo	Possíveis valores
CÉU	Sol, Nublado, Chuva
TEMPERATURA	<i>Dados contínuos</i>
UMIDADE	<i>Dados contínuos</i>
VENTO	Sim, Não



Árvores de decisão

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
1	Sol	85	85	Não	Não ir
2	Sol	80	90	Sim	Não ir
3	Nublado	83	78	Não	Ir ao campo
4	Chuva	70	96	Não	Ir ao campo
5	Chuva	68	80	Não	Ir ao campo
6	Chuva	65	70	Sim	Não ir
7	Nublado	64	65	Sim	Ir ao campo
8	Sol	72	95	Não	Não ir
9	Sol	69	70	Não	Ir ao campo
10	Chuva	75	80	Não	Ir ao campo
11	Sol	75	70	Sim	Ir ao campo
12	Nublado	72	90	Sim	Ir ao campo
13	Nublado	81	75	Não	Ir ao campo
14	Chuva	71	80	Sim	Não ir

Possibilidades:

- **9** Ir ao campo
- **5** Não ir

Árvores de decisão

Atenção

- ✓ Temperatura e umidade agora são valores contínuos e devem ser **analisados de maneira ordenada**.
- ✓ Seja $v = v_1, v_2, \dots, v_n$ o conjunto de valores possíveis para um determinado atributo. Deve-se ordenar v em ordem crescente, ou seja, $v_i \leq v_{i+1}, \forall i$.
- ✓ Para cada $i \in 1, n - 1$, o valor de teste (ponto de partição) será:

$$v = \frac{(v_i + v_{i+1})}{2}$$

- ✓ e os valores dos ramos de partição serão:

$$P_1^v = \{v_j \mid v_j \leq v_p\} \text{ e } P_2^v = \{v_j \mid v_j > v_p\}$$

Árvores de decisão

Exemplo 2

- Para elucidar os conceitos para atributo contínuo, vamos realizar o cálculo do ganho de informação para o atributo contínuo UMIDADE definindo os valores de partição.

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
1	Sol	85	85	Não	Não ir
2	Sol	80	90	Sim	Não ir
8	Sol	72	95	Não	Não ir
9	Sol	69	70	Não	Ir ao campo
11	Sol	75	70	Sim	Ir ao campo

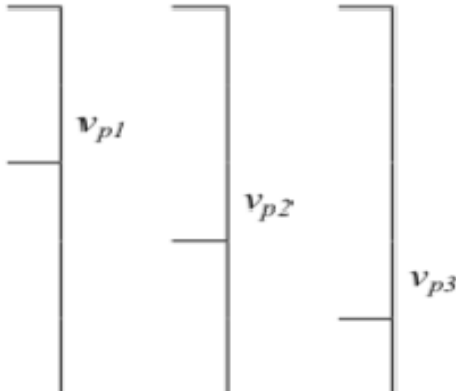
Árvores de decisão

Exemplo 2

- Dispõem-se os exemplos por ordem crescente de umidade e calculam-se os pontos de partição:

Dados	CÉU	TEMP.	UMID.	VENTO	Ir ao campo?
9	Sol	69	70	Não	Ir ao campo
11	Sol	75	70	Sim	Ir ao campo
1	Sol	85	85	Não	Não ir
2	Sol	80	90	Sim	Não ir
8	Sol	72	95	Não	Não ir

Pontos de partição



$$v_{p1} = \frac{(70 + 85)}{2} = 77,5$$

$$v_{p2} = \frac{(85 + 90)}{2} = 87,5$$

$$v_{p3} = \frac{(90 + 95)}{2} = 92,5$$

Árvores de decisão

Exemplo 2

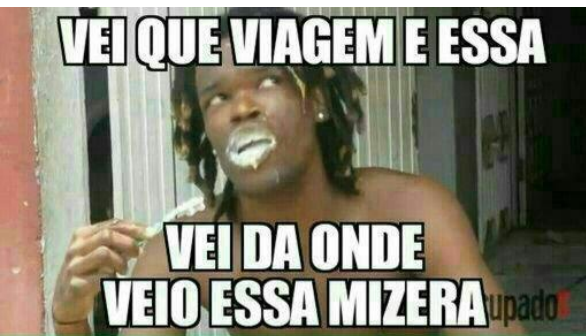
- Cálculo do ganho de informação para V_{p1} :

$$p(\text{Ir a campo} \mid \text{umidade} < 77,5) = 2/2 = 1$$

$$p(\text{Ir a campo} \mid \text{umidade} > 77,5) = 0/3 = 0$$

$$p(\text{Não ir} \mid \text{umidade} < 77,5) = 0/2 = 0$$

$$p(\text{Não ir} \mid \text{umidade} > 77,5) = 3/3 = 1$$



$$\text{info}(\text{umidade} < 77,5) = -1 * \log_2(1) - 0 * \log_2(0) = 0$$

$$\text{info}(\text{umidade} > 77,5) = -0 * \log_2(0) - 1 * \log_2(1) = 0$$

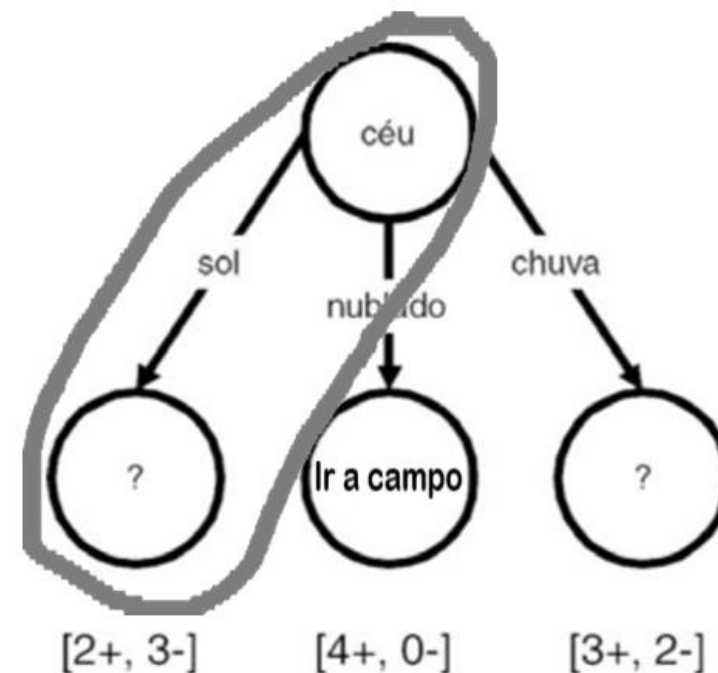
$$\text{info}(\text{umidade}) = \frac{2}{5} * \text{info}(\text{umidade} < 77,5) + \frac{3}{5} * \text{info}(\text{umidade} > 77,5) = 0$$

$$\text{Ganho}(\text{info}(\text{sol}), \text{umidade}) = 0,97094 - \text{info}(\text{umidade}) = 0,97094$$

Árvores de decisão

RELEMBRANDO O EXEMPLO

- *Ganho info sol*, Temperatura = 0,57094
- ***Ganho info sol*, Umidade = 0,97094**
- *Ganho info sol*, Vento = 0,019963



- Examinando os ganhos verifica-se que o atributo com maior ganho de informação é a **umidade**, o qual deve ser o nó seguinte da árvore neste ramo.

Árvores de decisão

Exemplo 2

- Cálculo do ganho de informação para V_{p2} :

$$p(\text{Ir a campo} \mid \text{umidade} < 87,5) = 2/3$$

$$p(\text{Ir a campo} \mid \text{umidade} > 87,5) = 0/2 = 0$$

$$p(\text{Não ir} \mid \text{umidade} < 87,5) = 1/3$$

$$p(\text{Não ir} \mid \text{umidade} > 87,5) = 2/2 = 1$$

$$\text{info}(\text{umidade} < 87,5) = -\left(\frac{2}{3}\right) * \log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) * \log_2\left(\frac{1}{3}\right) = 0,918$$

$$\text{info}(\text{umidade} > 87,5) = -0 * \log_2 0 - 1 * \log_2(1) = 0$$

$$\text{info}(\text{umidade}) = \frac{3}{5} * \text{info}(\text{umidade} < 87,5) + \frac{2}{5} * \text{info}(\text{umidade} > 87,5) = 0,550$$

$$\text{Ganho}(\text{info}(\text{sol}), \text{umidade}) = 0,97094 - \text{info}(\text{umidade}) = 0,420$$

Árvores de decisão

Exemplo 2

- Cálculo do ganho de informação para V_{p3} :

$$p(\text{Ir a campo} \mid \text{umidade} < 92,5) = 2/4 = 1/2$$

$$p(\text{Ir a campo} \mid \text{umidade} > 92,5) = 0/1 = 0$$

$$p(\text{Não ir} \mid \text{umidade} < 92,5) = 2/4 = 1/2$$

$$p(\text{Não ir} \mid \text{umidade} > 92,5) = 1/1 = 1$$

$$\text{info}(\text{umidade} < 92,5) = -\left(\frac{1}{2}\right) * \log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) * \log_2\left(\frac{1}{2}\right) = 1$$

$$\text{info}(\text{umidade} > 92,5) = -0 * \log_2 0 - 1 * \log_2(1) = 0$$

$$\text{info}(\text{umidade}) = \frac{4}{5} * \text{info}(\text{umidade} < 92,5) + \frac{1}{5} * \text{info}(\text{umidade} > 92,5) = 0,8$$

$$\text{Ganho}(\text{info}(\text{sol}), \text{umidade}) = 0,97094 - \text{info}(\text{umidade}) = 0,170$$

Árvores de decisão

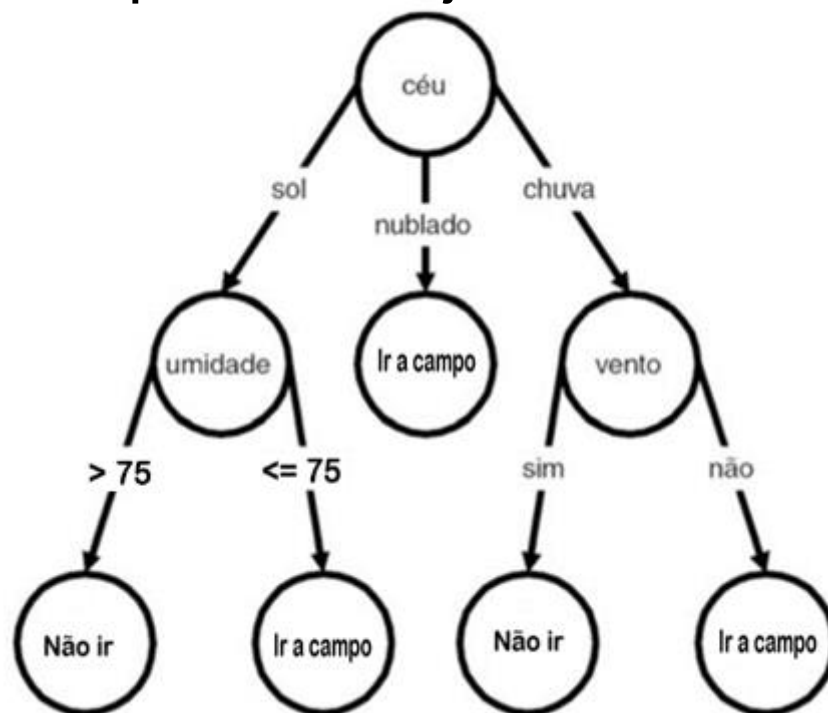
Exemplo 2

- A partição que possui o maior ganho de informação é *vp1*, portanto essa partição será escolhida para o nó teste da árvore.
- O valor de teste nos ramos do atributo umidade pode ser o próprio *vp1* ou utilizar um valor que pertença ao conjunto de valores possíveis da umidade (não ultrapassando o valor da partição).

Árvores de decisão

Exemplo 2

- Árvore de decisão final para o conjunto de treinamento:



E o atributo temperatura?

Continua sendo irrelevante para a tarefa de classificação, neste caso.



Árvores de decisão

Overfitting em árvores de decisão

- Sobreajuste (“Overfitting”) é um dos principais desafios enfrentados ao modelar árvores de decisão.
- Se não forem estabelecidos limites a uma árvore de decisão, ela acabará dando 100% de precisão no conjunto de treinamento, porque no pior caso ela vai acabar fazendo uma folha para cada observação.
- Por isso, prevenir o sobreajuste é fundamental ao se modelar uma árvore de decisão. Isso pode ser feito de duas maneiras:
- **Definindo restrições no tamanho da árvore**
- **Podando a árvore.**



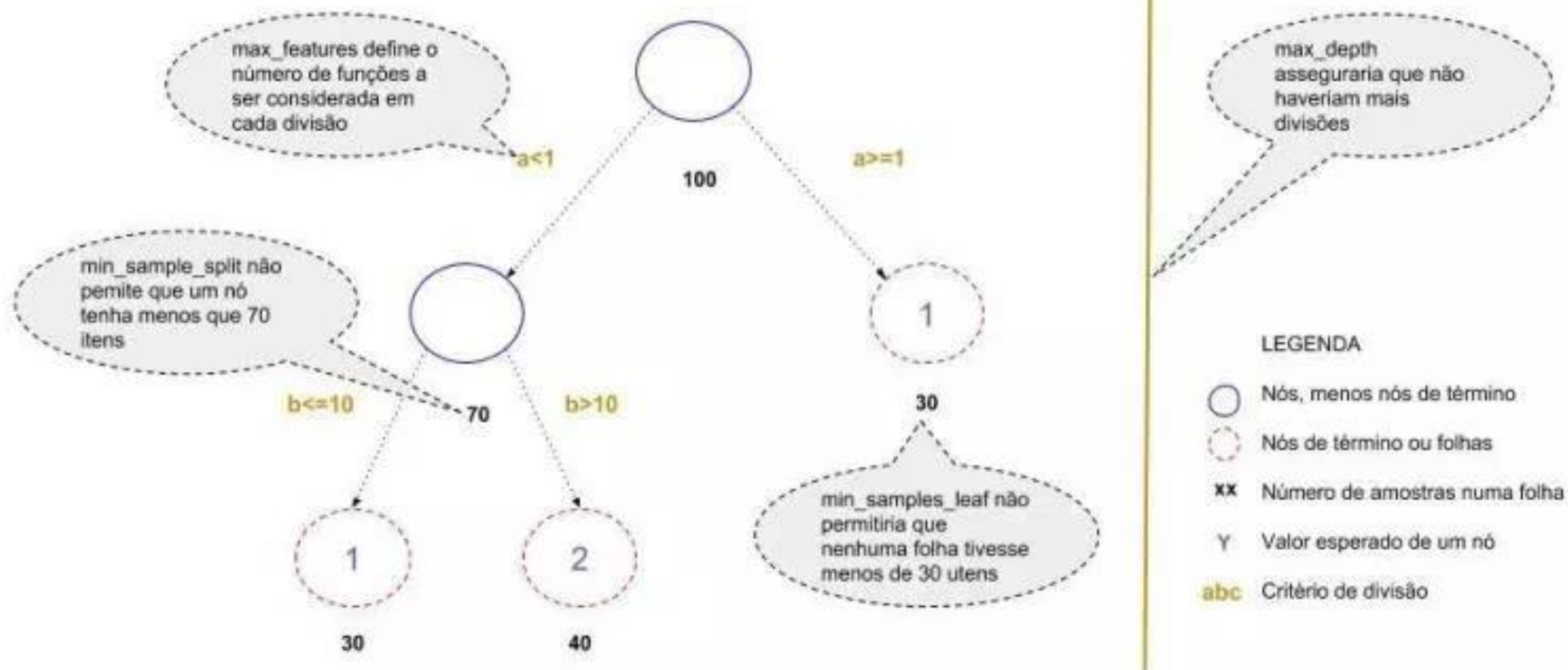
Árvores de decisão

Restrições

- Número mínimo de amostras para uma divisão de nó:
Valores mais elevados impedem o modelo de aprender sobre relações que podem ser muito específicas da amostra selecionada para a árvore.
- Número mínimo de amostras para um nó de término:
Valores mais baixos para problemas de classe mais desbalanceados.
- Profundidade máxima da árvore:
Uma maior profundidade permitirá ao modelo aprender relações cada vez mais específicas.
- Número máximo de nós de término.
- Número máximo de atributos a considerar para uma divisão:
Valores mais altos podem levar a um sobreajuste.

Árvores de decisão

Restrições





Árvores de decisão

Pré-poda

- Regras de parada que previnem a construção daqueles ramos que não parecem melhorar a precisão preditiva da árvore.
 - Todas as observações alcançando um nó pertencem à mesma classe.
 - Todas as observações alcançando um nó têm o mesmo vetor de características (mas não necessariamente pertencem à mesma classe).
 - O número de observações no nó é menor que um certo limiar.
 - O mérito atribuído a todos os possíveis testes que particionam o conjunto de observações no nó é muito baixo.



Árvores de decisão

Pós-poda

- Um dos métodos mais simples é baseado em duas medidas (Bratko, 1984), o erro estático e o erro de backed-up:
 - O erro estático é o número de classificações incorretas considerando que todos os exemplos que chegam nesse nó são classificados usando a classe majoritária da distribuição de classes desse nó.
 - O erro de backed-up é a soma das classificações incorretas de todas as subárvores do nó corrente. Se o erro de backed-up é maior ou igual ao erro estático, então o nó é trocado por uma folha com a classe majoritária do nó.

Árvores de decisão

Pseudo algoritmo de poda

- Para cada nó interno da árvore, é calculada a taxa de erro caso esse nó vire folha (e tudo abaixo dele seja eliminado).
- Em seguida, é calculada a taxa de erro caso não haja a poda.
- Se a diferença entre essas duas taxas de erro for menor que um valor pré-estabelecido, a árvore é podada; caso contrário, não ocorre a poda.
- Esse processo se repete progressivamente, gerando um conjunto de árvores podadas.
- Por fim, para cada uma delas é calculado erro na classificação de um conjunto de dados teste, e a árvore que obtiver o menor erro será a escolhida.



Árvores de decisão

Vantagens

- 1. Flexibilidade:** Árvores de decisão não assumem nenhuma distribuição para os dados. Elas são métodos não paramétricos. O espaço de objetos é dividido em subespaços, e cada subespaço é ajustado com diferentes modelos. Uma árvore de decisão fornece uma cobertura exaustiva do espaço de instâncias. Havendo exemplos suficientes, pode aproximar o erro de Bayes de qualquer função.
- 2. Robustez:** Árvores univariáveis são invariantes a transformações (estritamente) monótonas de variáveis de entrada. Por exemplo, usar X_j , $\log x_j$, ou e^{x_j} como a j -ésima variável de entrada produz árvores com a mesma estrutura. Como uma consequência dessa invariância, a sensibilidade a distribuições com grande cauda e outliers é também reduzida (Friedman, 1999).



Árvores de decisão

Vantagens

3. Seleção de atributos: O processo de construção de uma árvore de decisão seleciona os atributos a usar no modelo de decisão. Essa seleção de atributos produz modelos que tendem a ser bastante robustos contra a adição de atributos irrelevantes e redundantes.

4. Interpretabilidade: Decisões complexas e globais podem ser aproximadas por uma série de decisões mais simples e locais. Todas as decisões são baseadas nos valores dos atributos usados para descrever o problema. Ambos os aspectos contribuem para a popularidade das árvores de decisão.

5. Eficiência: O algoritmo para aprendizado de árvore de decisão é um algoritmo guloso que é construído de cima para baixo (top-down), usando uma estratégia dividir para conquistar sem backtracking. Sua complexidade de tempo é linear com o número de exemplos.



Árvores de decisão

Desvantagens

- 1. Replicação:** O termo refere-se à duplicação de uma sequência de testes em diferentes ramos de uma árvore de decisão, levando a uma representação não concisa, que também tende a ter baixa precisão preditiva. Pagallo e Haussler (1990) argumentam que a replicação é inerente à representação da árvore de decisão.
- 2. Valores ausentes:** Uma árvore de decisão é uma hierarquia de testes. Se o valor de um atributo é desconhecido, isso causa problemas em decidir que ramo seguir. Algoritmos devem empregar mecanismos especiais para abordar falta de valores. Friedman et al. (1996) sustentam que "Cerca de metade do código no CART e 80% dos esforços de programação foram desenvolvidos para falta de valores!".



Árvores de decisão

Desvantagens

3. Atributos contínuos: O gargalo do algoritmo é a presença de atributos contínuos. Nesse caso, uma operação de ordenação é solicitada para cada atributo contínuo de cada nó de decisão. Alguns autores estimam que a operação de ordenação consuma 70% do tempo necessário para induzir uma árvore de decisão em grandes conjuntos de dados com muitos atributos contínuos (Catlett, 1991).

4. Instabilidade: Muitos pesquisadores, especialmente Breiman et al. (1984), Breiman (1996b) e Kohavi e Kunz (1997) apontaram que pequenas variações no conjunto de treinamento podem produzir grandes variações na árvore final. Há uma forte tendência a inferências feitas próximo das folhas serem menos confiáveis que aquelas feitas próximas da raiz.



Árvores de decisão

Métodos ensemble

- Próxima Aula.

Random forest

- Próxima Aula.

!!! SIMBORA !!!

