ÁRVORES DE DECISÃO

PROFA. MARIANA KLEINA

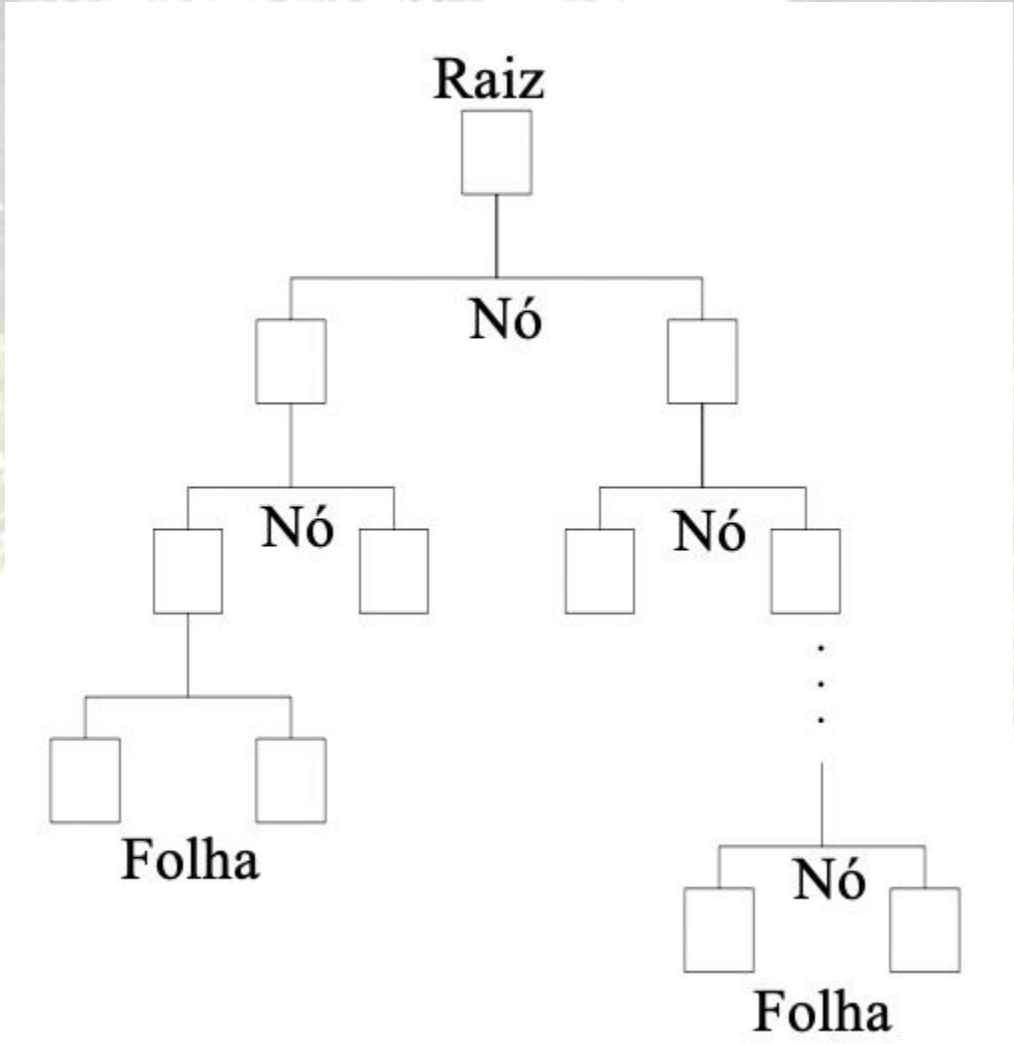
DEFINIÇÃO

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.

É um método de aprendizado de máquina supervisionado.

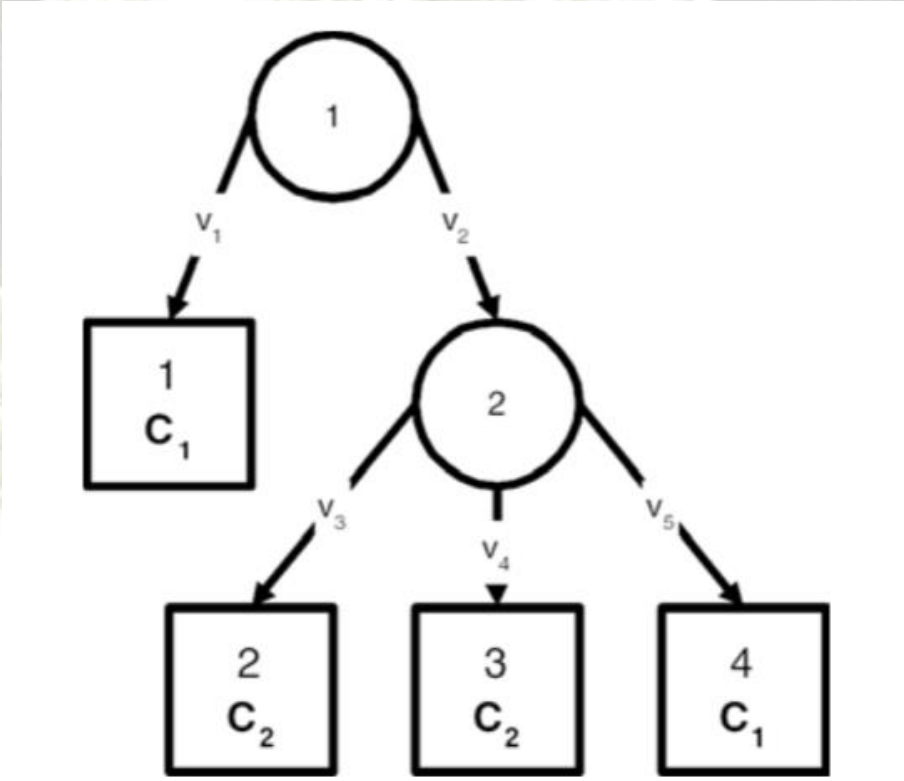
ARQUITETURA

O método consiste na contínua subdivisão de um espaço amostral (chamado de raiz) em classes menores por meio de testes (chamados de nós) feitos para subdividir – em dois subespaços para manter uma maior homogeneidade na divisão - esse espaço amostral em classes até que se tenha um subconjunto homogêneo o suficiente para ser classificado como uma mesma classe, criando, assim, um nó terminal (chamado de folha).



ARQUITETURA

A figura ilustra um árvore de decisão não binária: o nó 2 possui três ramos. Neste tipo de árvore, um teste realizado em um nó resulta na divisão de dois ou mais conjuntos disjuntos que cobrem todas as possibilidades, isto é, todo novo caso deve pertencer a um dos subconjuntos disjuntos.



CONSTRUÇÃO DE UMA ÁRVORE DE DECISÃO

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.

Exemplo: Gosto de

Por indução

Gosto de esportes

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

CONSTRUÇÃO DE 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.

CRIAÇÃO DE UM NÓ

Para definir qual o melhor critério dentre todos os possíveis é feito o cálculo do ganho de informação, que consiste na análise da homogeneidade das subclasses criadas, escolhendo, assim o critério que traga um maior ganho de informação ($Ganho$):

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

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

CLASSIFICAÇÃO DE UM NÓ TERMINAL

A classificação do nó pode ser feita de duas maneiras:

- Quando não é mais possível fazer a subdivisão por se ter um espaço amostral muito pequeno, ou;
- Quando o erro percentual calculado naquela classe é suficientemente menor que o erro calculado na raiz.

EXEMPLO 1

Suponha que deseja-se saber se haverá ou não um jogo de golfe de acordo com as condições climáticas:

Atributo	Possíveis valores
céu	sol, nublado, chuva
temperatura	alta, baixa, suave
umidade	alta, normal
vento	sim, não

EXEMPLO 1

O conjunto de treinamento é:

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
3	nublado	alta	alta	não	joga
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
7	nublado	baixa	normal	sim	joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
10	chuva	suave	normal	não	joga
11	sol	suave	normal	sim	joga
12	nublado	suave	alta	sim	joga
13	nublado	alta	normal	não	joga
14	chuva	suave	alta	sim	não joga

9 possibilidades para ocorrer jogo e 5 para não ocorrer.

EXEMPLO 1

O objetivo do cálculo da entropia está na classificação *booleana* (jogar golfe × não jogar golfe), em que há 14 exemplos, 9 positivos e 5 negativos, ou seja, $T = [9+, 5-]$.

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

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

EXEMPLO 1

Ganho de informação – **CÉU**

O atributo céu pode assumir 3 valores (sol, nublado e chuva).

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

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

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

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

Logo,

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

$$= 0{,}940 - \left(\frac{5}{14}\right)\cdot 0{,}97094 - \left(\frac{4}{14}\right)\cdot 0 - \left(\frac{5}{14}\right)\cdot 0{,}97094 = 0{,}2464$$

EXEMPLO 1

Ganho de informação – **TEMPERATURA**

O atributo temperatura pode assumir 3 valores (alta, suave e baixa).

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

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

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

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

Logo,

$$Ganho(\text{info}(T), temperatura) = 0{,}940 - \left(\frac{5}{14}\right)\cdot \text{info}(alta) - \left(\frac{4}{14}\right)\cdot \text{info}(suave) - \left(\frac{5}{14}\right)\cdot \text{info}(baixa)$$

$$= 0{,}940 - \left(\frac{5}{14}\right)\cdot 0{,}97094 - \left(\frac{4}{14}\right)\cdot 0{,}811 - \left(\frac{5}{14}\right)\cdot 0{,}97094 \quad = \boxed{0{,}015}$$

EXEMPLO 1

Ganho de informação – **UMIDADE**

O atributo umidade pode assumir 2 valores (alta e baixa).

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

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

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

Logo,

$$Ganho(\text{info}(T), umidade) = 0{,}940 - \left(\frac{7}{14}\right)\cdot \text{info}(alta) - \left(\frac{7}{14}\right)\cdot \text{info}(baixa)$$

$$= 0{,}940 - \left(\frac{7}{14}\right)\cdot 0{,}985228 - \left(\frac{7}{14}\right)\cdot 0{,}591672 \quad = \boxed{0{,}151}$$

EXEMPLO 1

Ganho de informação – **VENTO**

O atributo vento pode assumir 2 valores (sim e não).

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

$$\text{info}(sim) = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) = 1$$

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

Logo,

$$Ganho(\text{info}(T), vento) = 0{,}940 - \left(\frac{8}{14}\right)\cdot\text{info}(sim) - \left(\frac{6}{14}\right)\cdot\text{info}(não)$$

$$= 0{,}940 - \left(\frac{6}{14}\right)\cdot 1 - \left(\frac{8}{14}\right)\cdot 0{,}811278 \quad = \boxed{0{,}047841}$$

EXEMPLO 1

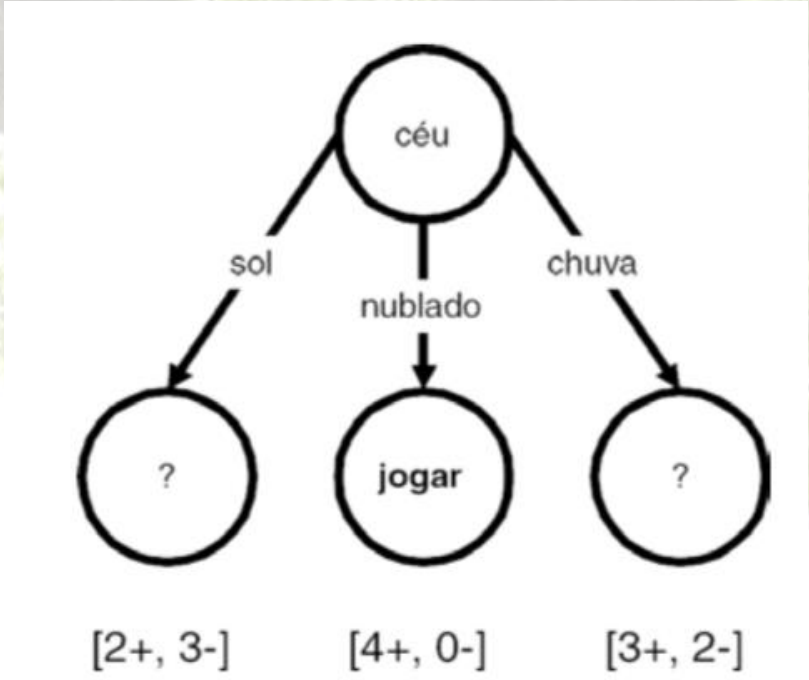
Escolhe-se o atributo (céu) de maior ganho de informação para ser o nó raiz da árvore.

$$\boldsymbol{Ganho(info(T), céu) = 0,2464}$$

$$Ganho(info(T), temperatura) = 0{,}015$$

$$Ganho(info(T), umidade) = 0{,}151$$

$$Ganho(info(T), vento) = 0{,}047841$$



Os ramos sol e chuva ainda estão indefinidos, e o processo deve continuar no próximo nível da árvore.

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:

Céu = sol

$T_1 = \{1, 2, 8, 9, 11\}$

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
11	sol	suave	normal	sim	joga

EXEMPLO 1

Céu = nublado

$T_2 = \{3, 7, 12, 13\}$

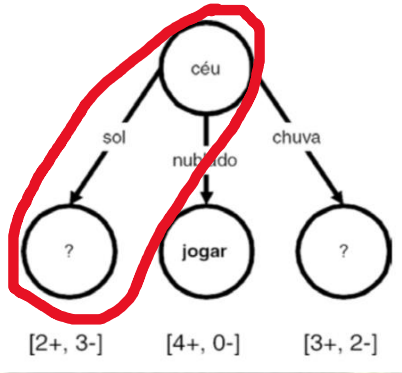
Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
3	nublado	alta	alta	não	joga
7	nublado	baixa	normal	sim	joga
12	nublado	suave	alta	sim	joga
13	nublado	alta	normal	não	joga

Céu = chuva

$T_3 = \{4, 5, 6, 10, 14\}$

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
10	chuva	suave	normal	não	joga
14	chuva	suave	alta	sim	não joga

EXEMPLO 1



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

Ganho de informação – **TEMPERATURA**

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

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

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

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

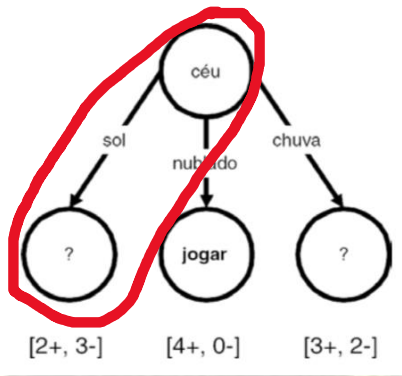
Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
11	sol	suave	normal	sim	joga

Logo, $$Ganho(\text{info}(sol), temperatura) = 0{,}97094 - \left(\frac{2}{5}\right)\cdot\text{info}(alta) - \left(\frac{2}{5}\right)\cdot\text{info}(suave) - \left(\frac{1}{5}\right)\cdot\text{info}(baixa)$$

$$= 0{,}97094 - \left(\frac{2}{5}\right)\cdot 0 - \left(\frac{2}{5}\right)\cdot 1 - \left(\frac{1}{5}\right)\cdot 0$$

$$= \boxed{0.57094}$$

EXEMPLO 1



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

Ganho de informação – **UMIDADE**

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

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

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

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
11	sol	suave	normal	sim	joga

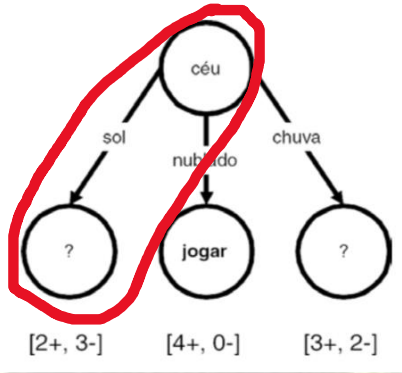
Logo,

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

$$= 0{,}97094 - \left(\frac{3}{5}\right)\cdot 0 - \left(\frac{2}{5}\right)\cdot 0$$

$$= \boxed{0{,}97094}$$

EXEMPLO 1



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

Ganho de informação – **VENTO**

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

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

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

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
11	sol	suave	normal	sim	joga

Logo,

Logo,

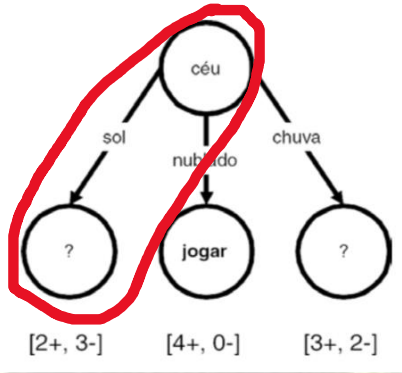
$$Ganho(\text{info}(sol), vento) = 0{,}97094 - \left(\frac{2}{5}\right)\cdot\text{info}(sim) - \left(\frac{3}{5}\right)\cdot\text{info}(não)$$

$$= 0{,}97094 - \left(\frac{2}{5}\right)\cdot 1 - \left(\frac{3}{5}\right)\cdot 0{,}918295$$

$$= 0.019963$$

EXEMPLO 1

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

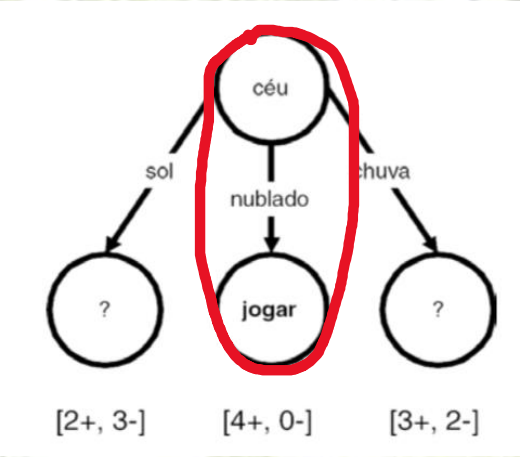


$$Ganho(info(sol), temperatura) = 0{,}57094$$
$$\boldsymbol{Ganho(info(sol), umidade) = 0{,}97094}$$
$$Ganho(info(sol), vento) = 0{,}019963$$

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.

EXEMPLO 1

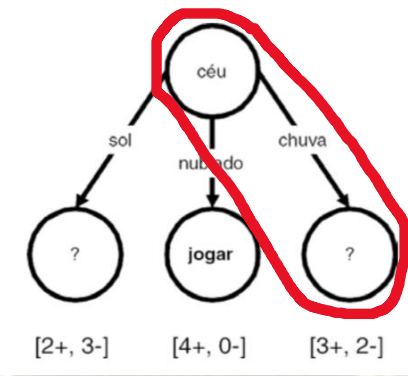
- Céu = nublado (processo de indução para este ramo da árvore)



Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
3	nublado	alta	alta	não	joga
7	nublado	baixa	normal	sim	joga
12	nublado	suave	alta	sim	joga
13	nublado	alta	normal	não	joga

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

EXEMPLO 1



- Céu = chuva (proc. de indução para este ramo da árvore)

Ganho de informação – **TEMPERATURA**

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

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

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

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

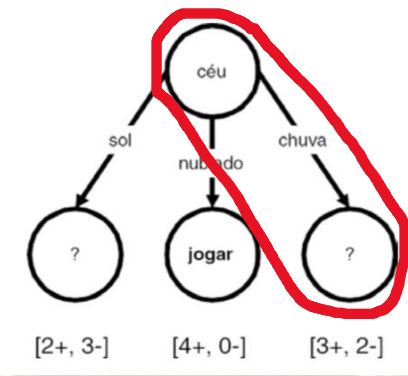
Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
10	chuva	suave	normal	não	joga
14	chuva	suave	alta	sim	não joga

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

$$= 0{,}97094 - \left(\frac{1}{5}\right)\cdot 0 - \left(\frac{2}{5}\right)\cdot 1 - \left(\frac{2}{5}\right)\cdot 1$$

$$= 0.17090$$

EXEMPLO 1



- Céu = chuva (proc. de indução para este ramo da árvore)

Ganho de informação – **UMIDADE**

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

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

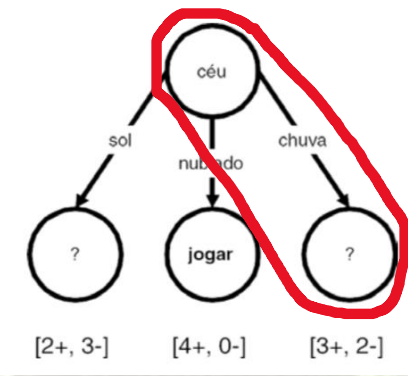
$$\text{info}(normal) = -\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) = 0{,}9182958$$

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
10	chuva	suave	normal	não	joga
14	chuva	suave	alta	sim	não joga

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

$= 0{,}019962$

EXEMPLO 1



- Céu = chuva (proc. de indução para este ramo da árvore)

Ganho de informação – **VENTO**

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

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

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

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
10	chuva	suave	normal	não	joga
14	chuva	suave	alta	sim	não joga

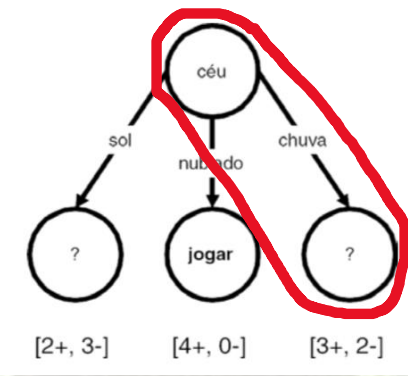
Logo, $Ganho(\text{info}(chuva), vento) = 0{,}97094 - \left(\frac{2}{5}\right)\cdot\text{info}(sim) - \left(\frac{3}{5}\right)\cdot\text{info}(não)$

$$= 0{,}97094 - \left(\frac{2}{5}\right)\cdot 0 - \left(\frac{3}{5}\right)\cdot 0$$

$$= 0.97094$$

EXEMPLO 1

- Céu = chuva (proc. de indução para este ramo da árvore)

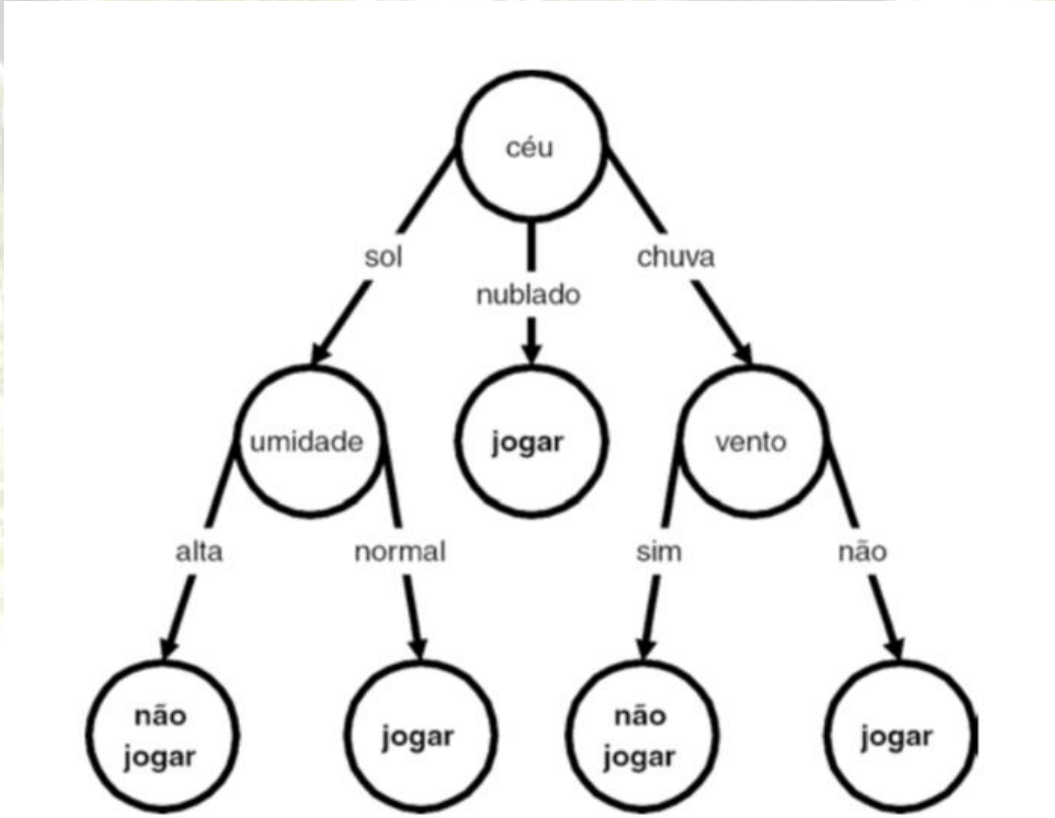


$$Ganho(info(chuva), temperatura) = 0{,}17090$$
$$Ganho(info(chuva), umidade) = 0{,}019962$$
$$\boldsymbol{Ganho(info(chuva), vento) = 0,97094}$$

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.

EXEMPLO 1

Árvore de decisão final para o conjunto de treinamento:



Observa-se que o atributo temperatura não foi selecionado para fazer parte da árvore (irrelevante para a tarefa de classificação, neste caso).

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).

EXEMPLO 2

Considere agora o seguinte conjunto de dados:

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	85	85	não	não joga
2	sol	80	90	sim	não joga
3	nublado	83	78	não	joga
4	chuva	70	96	não	joga
5	chuva	68	80	não	joga
6	chuva	65	70	sim	não joga
7	nublado	64	65	sim	joga
8	sol	72	95	não	não joga
9	sol	69	70	não	joga
10	chuva	75	80	não	joga
11	sol	75	70	sim	joga
12	nublado	72	90	sim	joga
13	nublado	81	75	não	joga
14	chuva	71	80	sim	não joga

EXEMPLO 2

Temperatura e umidade agora são valores contínuos e devem ser analisados de maneira ordenada.

Seja $v = (v_1, v_2, \ldots, v_n)$ 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 ($i \in [1, n-1]$), o valor de teste (ponto de partição) será:

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

e os valores dos ramos de partição serão:

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

EXEMPLO 2

Para elucidar os conceitos para atributo contínuo, é apresentado o cálculo do ganho de informação para o atributo contínuo umidade definindo os valores de partição.

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	85	85	não	não joga
2	sol	80	90	sim	não joga
8	sol	72	95	não	não joga
9	sol	69	70	não	joga
11	sol	75	70	sim	joga

EXEMPLO 2

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

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
9	sol	69	70	não	joga
11	sol	75	70	sim	joga
1	sol	85	85	não	não joga
2	sol	80	90	sim	não joga
8	sol	72	95	não	não joga

Pontos de partição

v_{p1} v_{p2} v_{p3}

Pontos de partição:

$$v_{p1} = \frac{(70+85)}{2} = 77{,}5 \qquad v_{p2} = \frac{(85+90)}{2} = 87{,}5 \qquad v_{p3} = \frac{(90+95)}{2} = 92{,}5$$

EXEMPLO 2

Cálculo do ganho de informação para cada partição:

(v_{p1})

$$p(jogar \mid umidade < 77{,}5) = \frac{2}{2} = 1$$

$$p(não\ jogar \mid umidade < 77{,}5) = \frac{0}{2} = 0$$

$$p(jogar \mid umidade > 77{,}5) = \frac{0}{3} = 0$$

$$p(não\ jogar \mid umidade > 77{,}5) = \frac{3}{3} = 1$$

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

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

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

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

EXEMPLO 2

Cálculo do ganho de informação para cada partição:

(v_{p2})

$$p(jogar \mid umidade < 87{,}5) = \frac{2}{3}$$

$$p(não\ jogar \mid umidade < 77{,}5) = \frac{1}{3}$$

$$p(jogar \mid umidade > 87{,}5) = \frac{0}{2} = 0$$

$$p(não\ jogar \mid umidade > 87{,}5) = \frac{2}{2} = 1$$

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

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

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

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

EXEMPLO 2

Cálculo do ganho de informação para cada partição:

(v_{p3})

$$p(jogar \mid umidade < 92{,}5) = \frac{2}{4} = \frac{1}{2}$$

$$p(não\ jogar \mid umidade < 92{,}5) = \frac{2}{4} = \frac{1}{2}$$

$$p(jogar \mid umidade > 92{,}5) = \frac{0}{1} = 0$$

$$p(não\ jogar \mid umidade > 92{,}5) = \frac{1}{1} = 1$$

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

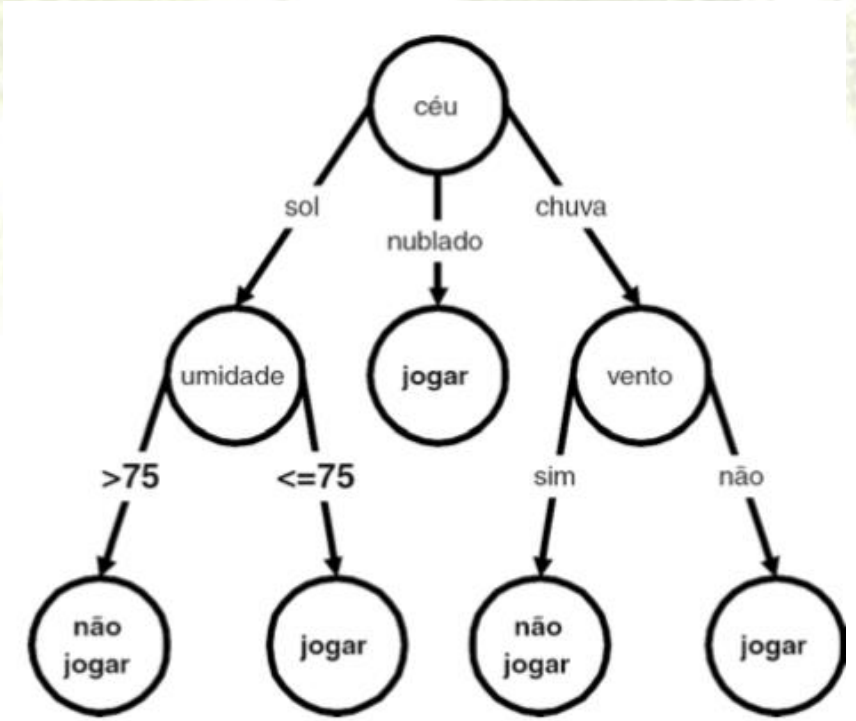
$$\text{info}(umidade > 92{,}5) = -0 * \log_2 0 - 1 * \log_2(1) = 0$$

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

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

EXEMPLO 2

A partição que possui o maior ganho de informação é v_{p1}. Portanto essa partição será escolhida para o nó teste da árvore. O valor de teste nos ramos do atributo umidade pode ser o próprio v_{p1} ou utilizar um valor que pertença ao conjunto de valores possíveis da umidade (não ultrapassando o valor da partição).



MÉTODOS DE PODA

Quando árvores de decisão são construídas, muitos ramos ou sub-árvores podem conter ruídos ou erros. O aprendizado é muito específico ao conjunto de treinamento, não permitindo generalizar para o conjunto de teste (*overfitting*).

Para melhorar o modelo, utilizam-se métodos de poda (*pruning*) na árvore, cujo objetivo é melhorar a taxa de acerto do modelo para novas amostras que não foram utilizadas no treinamento.

MÉTODOS DE PODA

Existem diversas formas de realizar uma poda, e todas elas são classificadas como pré-poda ou pós-poda.

- Pré-poda: realizada durante a construção da árvore. Em um certo momento, se o ganho de informação for menor que um valor pré-estabelecido, então esse nó vira folha.

MÉTODOS DE PODA

- Pós-poda: realizada após a construção da árvore. 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.

MÉTODOS DE PODA

Dentre os métodos de poda existentes, destacam-se: *Cost Complexity Pruning*, *Reduced Error Pruning*, *Minimum Error Pruning* (MEP), *Pessimistic Pruning*, *Error-Based Pruning* (EBP), *Minimum Description Length* (MDL) *Pruning*, *Minimum Message Length* (MML) *Pruning*, *Critical Value Pruning* (CVP), OPT e OPT-2.

OUTROS ALGORITMOS DE INDUÇÃO DE ÁRVORES DE DECISÃO

Além do ID3 e C4.5 apresentados, existem muitos outros algoritmos para construção de árvores de decisão por indução.
São eles: CART, TDIDT, NBTree, ADTree, LMT e BFTree.

REFERÊNCIAS

PAULA, M. B. Indução automática de árvores de decisão. Universidade Federal de Santa Catarina. Dissertação, 2002.

VON ZUBEN, F. J., ATTUX, R. R. F. Árvores de Decisão. UNICAMP. Disponível em: ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia004_1s10/notas_de_aula/topico7_IA004_1s10.pdf

http://www.alanfielding.co.uk/multivar/crt/dt_example_04.htm

Exemplos práticos:
https://www.wrprates.com/o-que-e-arvore-de-decisao-decision-tree-linguagem-r/

https://medium.com/machine-learning-beyond-deep-learning/árvores-de-decisão-3f52f6420b69