Árvore de Decisão

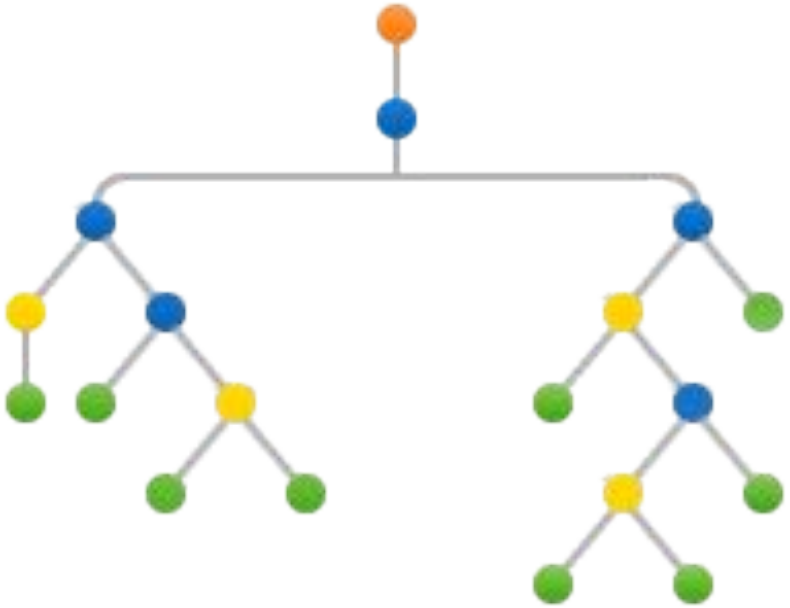
Disciplina: Mineração de Dados

Prof. Braian Varjão

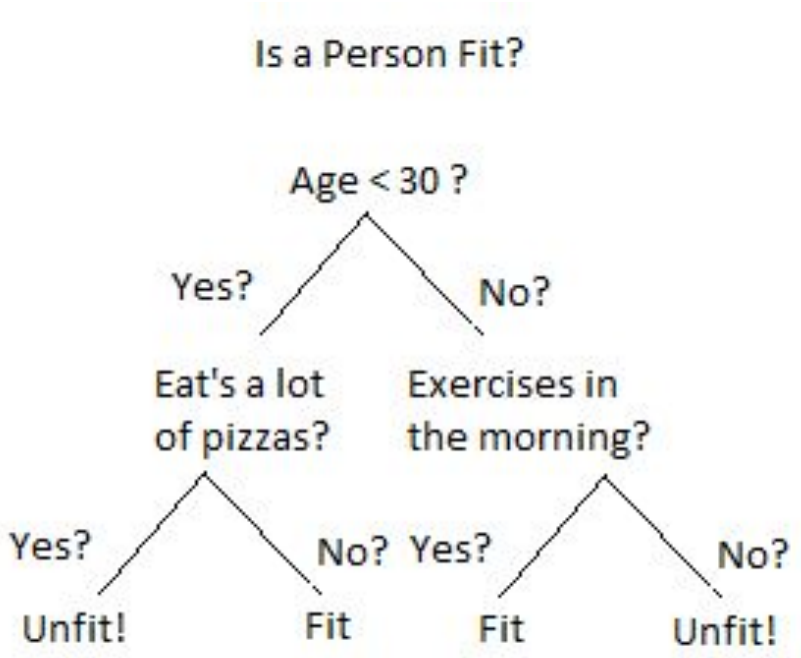
Árvore de Decisão

Árvore de decisão

Algoritmos de árvores de decisão adquirem conhecimento simbólico a partir dos dados de treinamento.



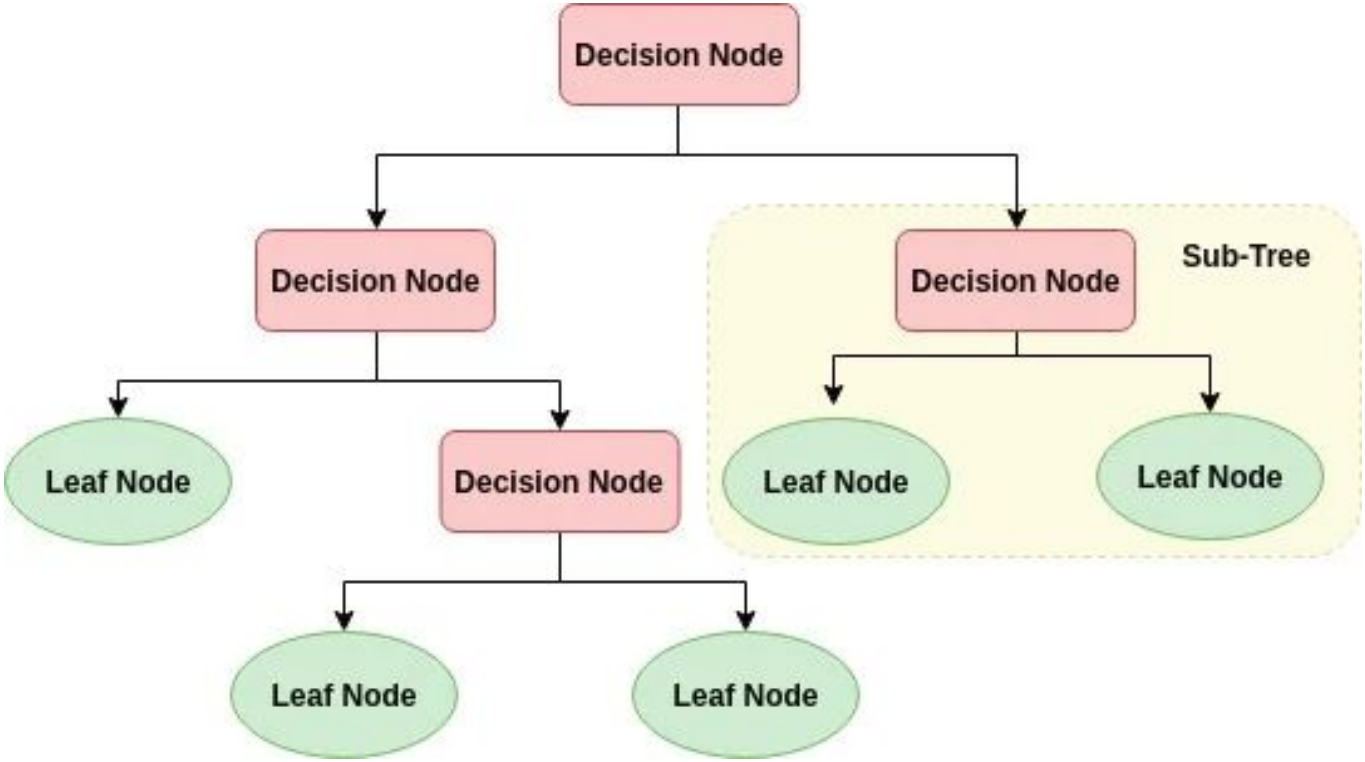
Árvore de decisão




```
IF (age < 30 AND eats_pizza) OR (age >= 30 AND exercices) THEN
    Fit
ELSE
    Unfit!
```

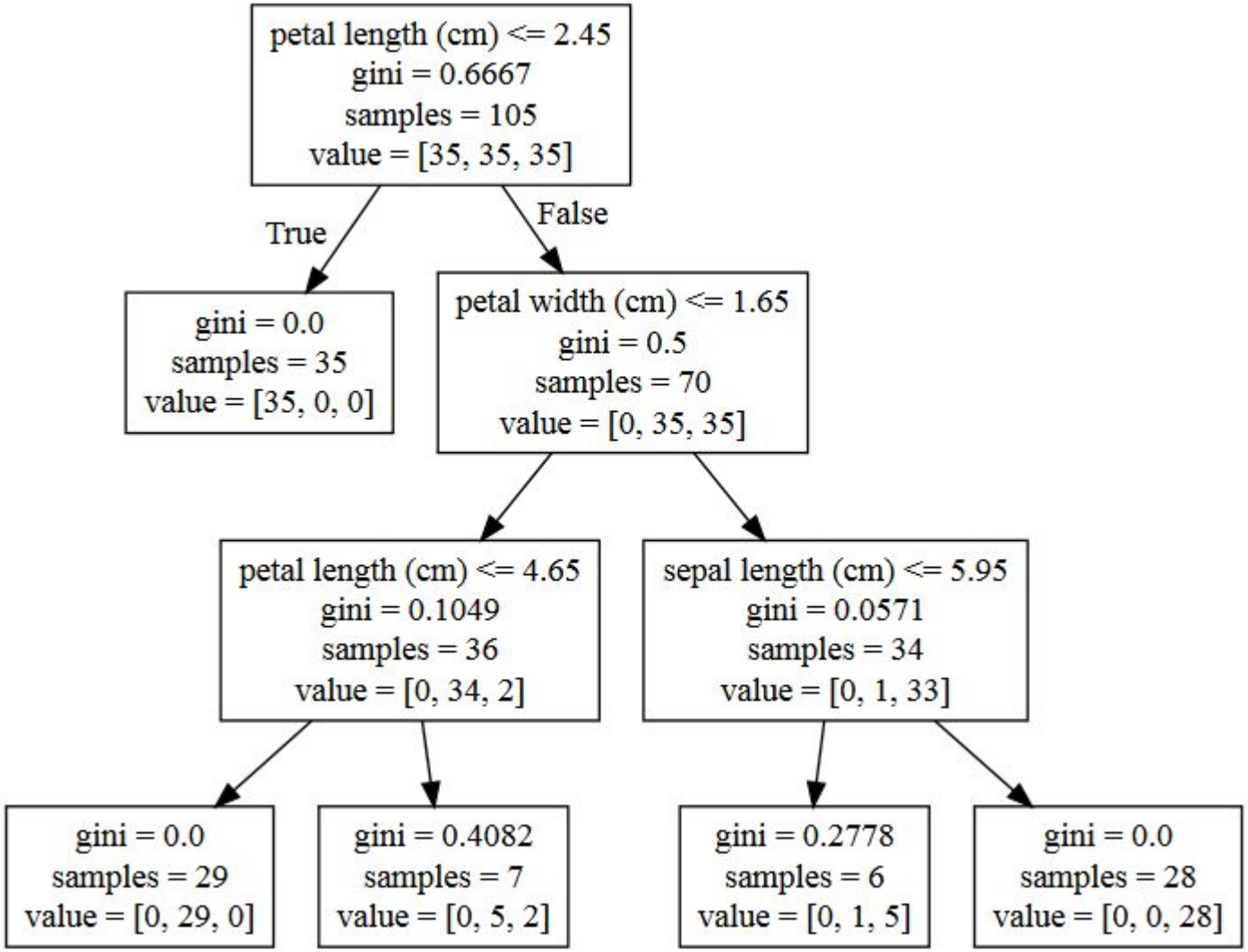
Prof. MSc. Braian Varjão

Representação



Prof. MSc. Braian Varjão

Árvore de decisão com atributos contínuos



Prof. MSc. Braian Varjão

Processo de aprendizado

De modo geral, os algoritmos utilizam uma abordagem de busca gulosa *top-down* no espaço de possíveis árvores de decisão.

Prof. MSc. Braian Varjão

Processo de aprendizado

Como selecionar o nó raiz?

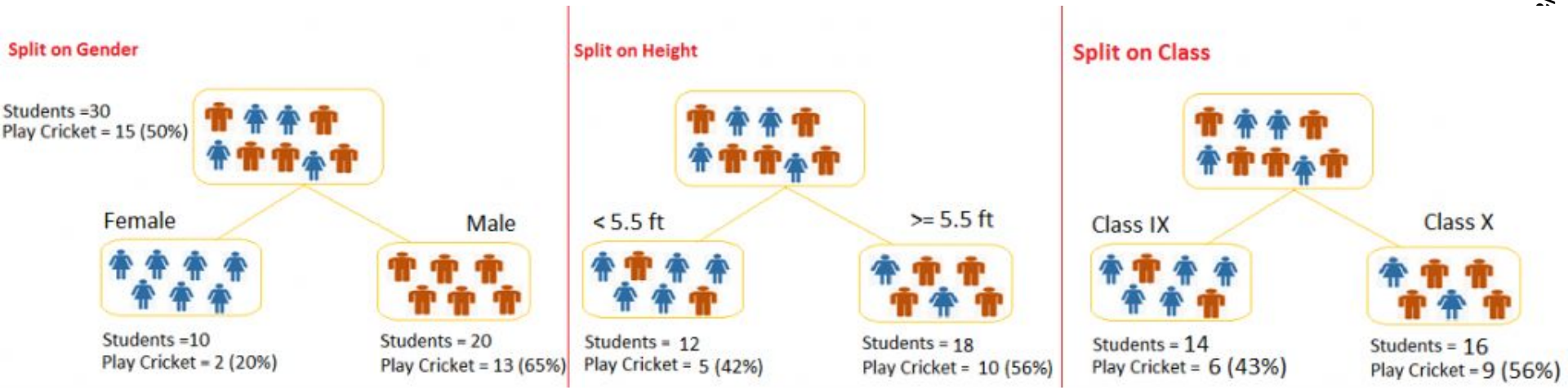
Algumas opções:

1. Atributo selecionado aleatoriamente;
2. O atributo com mais valores possíveis;
3. O atributo com menos valores possíveis;
4. Maior ganho de informação;
5. Maior índice Gini, etc.

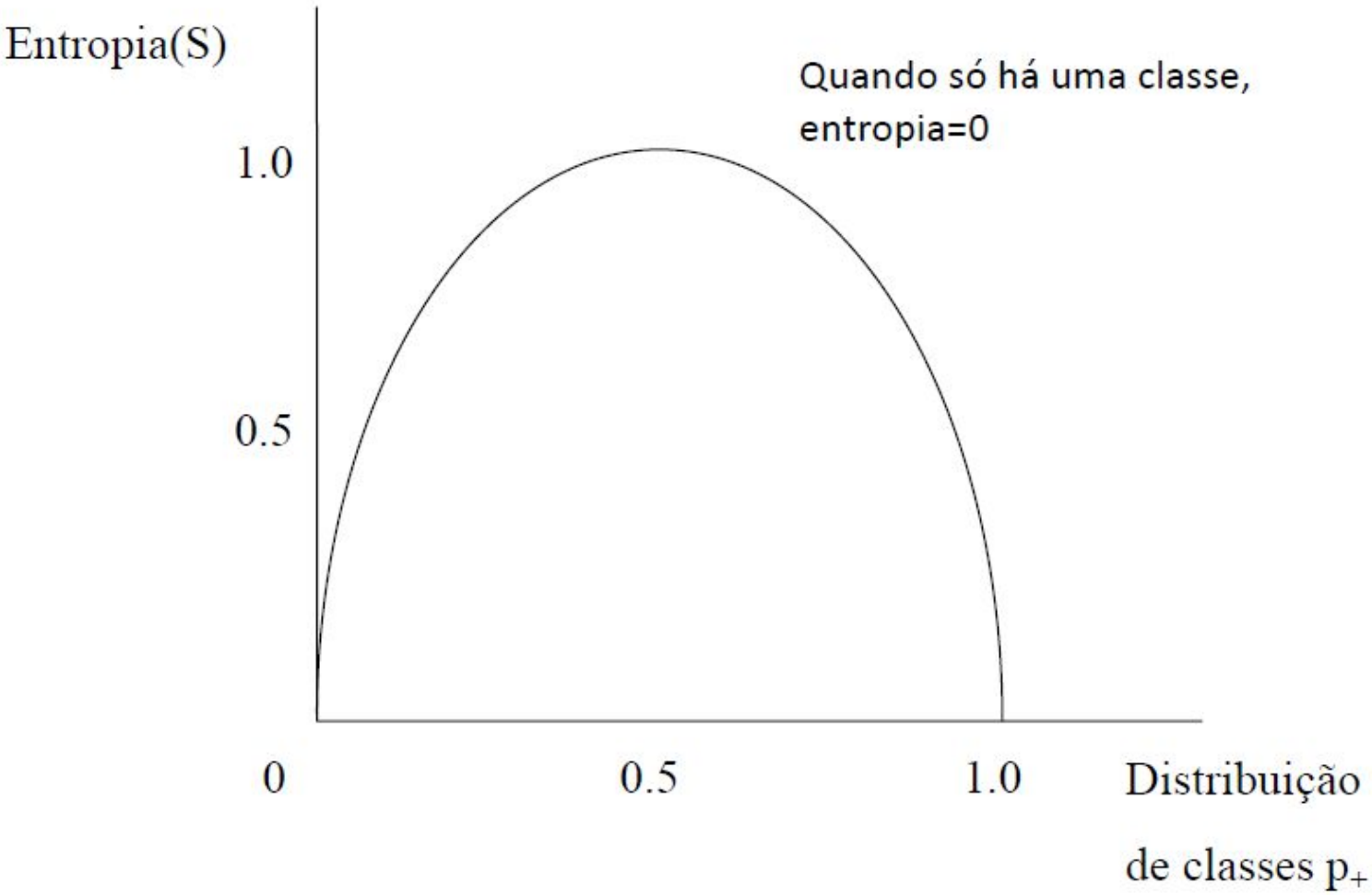
Prof. MSc. Braian Varjão

Algoritmo ID3

O particionamento da base se dá em busca de reduzir a impureza dos dados a cada iteração. Essa impureza é observada por uma medida chamada entropia.



Entropia


Entropia(S)
Quando só há uma classe, entropia=0
1.0
0.5
0
0.5
1.0
Distribuição de classes p_+

Entropia

Cálculo em problemas binários:

- Seja um conjunto de exemplo S
- Proporção de exemplos positivos $p(+)$
- Proporção de exemplos negativos $p(-)$

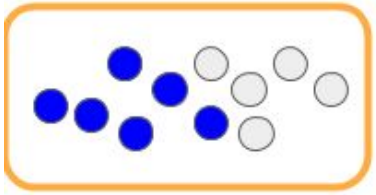
A entropia dessa classificação booleana é:

$$Entropia(S) = -p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$$

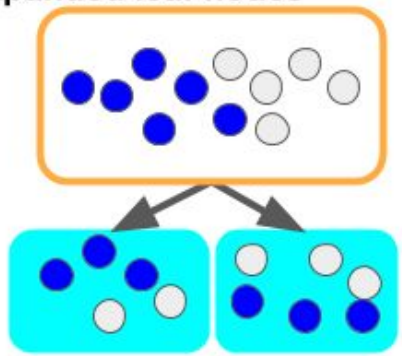
Prof. MSc. Braian Varjão

Entropia

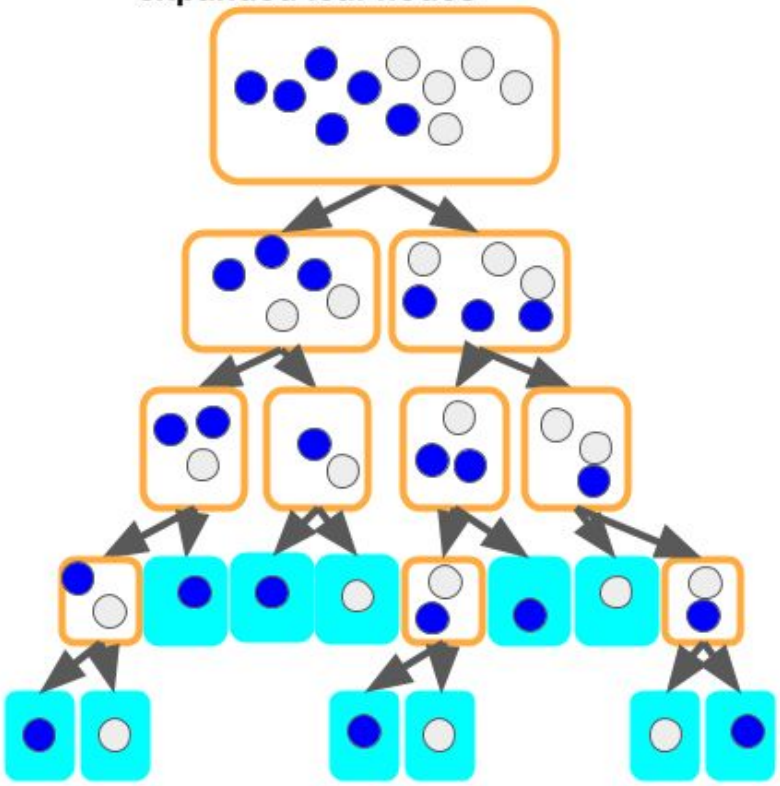
A. root node only



B. decision tree with partially expanded leaf nodes



C. decision tree with fully expanded leaf nodes



A: high in Expected Cross-Entropy

B: Expected Cross-Entropy ?

C: lowest in Expected Cross-Entropy (it is 0)

Prof. MSc. Braian Varjão

Qual a entropia dessa base?

$$Entropia(S) = -p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$$

dia	tempo	temperatura	umidade	vento	jogar_tênis
D1	ensolarado	quente	alta	fraco	**não**
D2	ensolarado	quente	alta	forte	**não**
D3	nublado	quente	alta	fraco	**sim**
D4	chuva	moderada	alta	fraco	**sim**
D5	chuva	fria	normal	fraco	**sim**
D6	chuva	fria	normal	forte	**não**
D7	nublado	fria	normal	forte	**sim**
D8	ensolarado	moderada	alta	fraco	**não**
D9	ensolarado	fria	normal	fraco	**sim**
D10	chuva	moderada	normal	fraco	**sim**
D11	ensolarado	moderada	normal	forte	**sim**
D12	nublado	moderada	alta	forte	**sim**
D13	nublado	quente	normal	fraco	**sim**
D14	chuva	moderada	alta	forte	**não**

Exercício 1: resposta

S possui 14 exemplos, sendo que a classe é constituída por 9 positivos e 5 negativos, ou seja, [9+,5-].

Assim sendo, a entropia de *S* é:

$$\text{Entropia}(S) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0.94$$

Prof. MSc. Braian Varjão

Entropia

Cálculo em problemas multi-classe:

- Seja um conjunto de exemplos S;
- Seja p_i a proporção de instâncias (exemplos) de S pertencendo a classe i;
- Seja c o número total de classes:

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

Prof. MSc. Braian Varjão

Algoritmo ID3

Como definir o nó raiz?

Busca pelo atributo que provê o maior ganho de informação (IG).

O IG permite identificar o atributo cujo particionamento resultará na maior redução da entropia, reduzindo o tamanho das sub árvores “enraizadas” em seus filhos.

Prof. MSc. Braian Varjão

Ganho de informação

Como calcular o IG de um atributo

- Seja um conjunto de exemplo *S*;
- Seja o atributo *A*;
- Seja Valores(*A*) o conjunto de todos os valores possíveis de *A*;
- Seja *Sv* o subconjunto de *S* para o qual o atributo *A* tem valor *v*.

$$Ganho(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|Sv|}{|S|} Entropia(Sv)$$

Escolhendo o nó raiz (1/3

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

$$Ganho(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|Sv|}{|S|} Entropia(Sv)$$

dia	tempo	temperatura	umidade	vento	jogar_tênis
D1	ensolarado	quente	alta	fraco	**não**
D2	ensolarado	quente	alta	forte	**não**
D3	nublado	quente	alta	fraco	**sim**
D4	chuva	moderada	alta	fraco	**sim**
D5	chuva	fria	normal	fraco	**sim**
D6	chuva	fria	normal	forte	**não**
D7	nublado	fria	normal	forte	**sim**
D8	ensolarado	moderada	alta	fraco	**não**
D9	ensolarado	fria	normal	fraco	**sim**
D10	chuva	moderada	normal	fraco	**sim**
D11	ensolarado	moderada	normal	forte	**sim**
D12	nublado	moderada	alta	forte	**sim**
D13	nublado	quente	normal	fraco	**sim**
D14	chuva	moderada	alta	forte	**não**

Prof. MSc. Braian Varjão

Escolhendo o nó raiz (2/3)

Entropia(S) = 0,94

Vento.forte = [6+,2-] Vento.fraco = [3+,3-]

$$Ganho(S,Vento) \equiv Entropia(S) - \sum_{v \in \{fraco, forte\}} \frac{|Sv|}{|S|} Entropia(Sv)$$

$$Ganho(S,Vento) \equiv Entropia(S) - \left(\frac{8}{14}\right) Entropia(Sfraco) - \left(\frac{6}{14}\right) Entropia(Sforte)$$

$$Ganho(S,Vento) \equiv 0.940 - \left(\frac{8}{14}\right) 0.811 - \left(\frac{6}{14}\right) 1.00$$

$$Ganho(S,Vento) \equiv 0.048$$

$-6/8 \log_2 6/8 - 2/8 \log_2 2/8$

$-3/6 \log_2 3/6 - 3/6 \log_2 3/6$

Prof. MSc. Braian Varjão

Escolhendo o nó raiz (3/3)

Definição do nó raiz:

Ganho(S, tempo) = 0.246

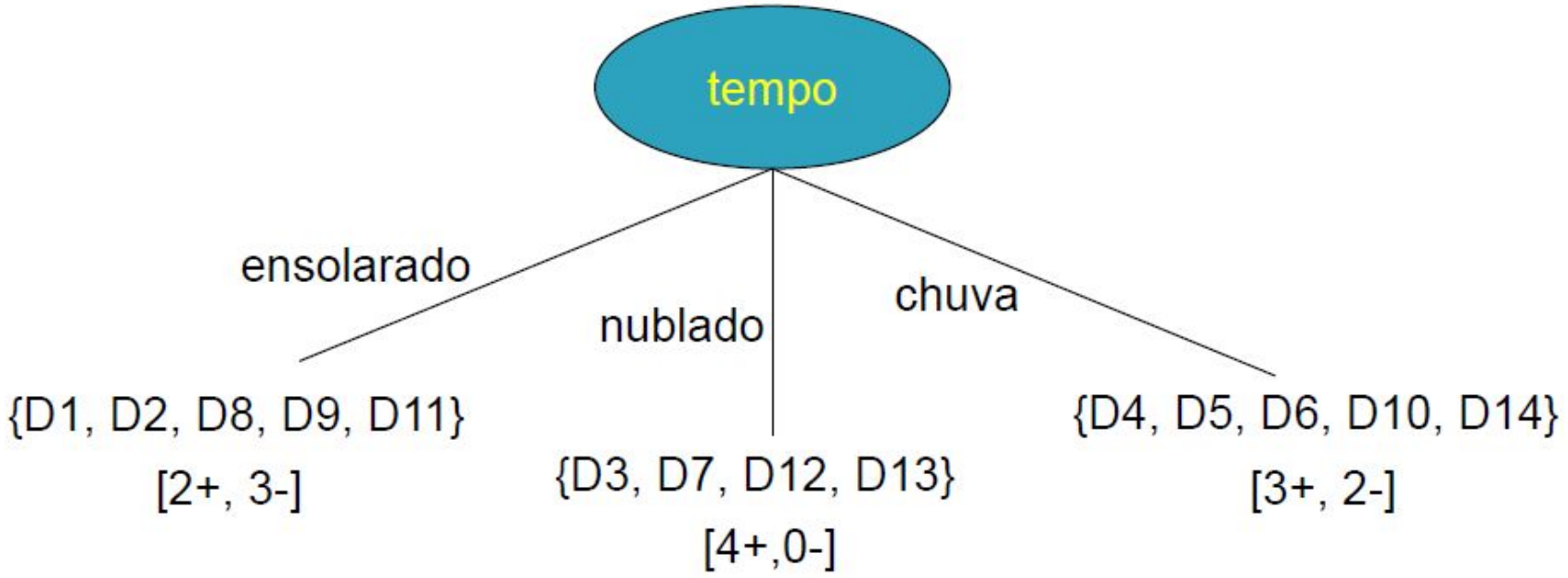
Ganho(S, umidade) = 0.151

Ganho(S, vento) = 0.048

Ganho(S, temperatura) = 0.029

Prof. MSc. Braian Varjão

Algoritmo ID3



Algoritmo ID3

Para cada ramo da árvore o processo continua até que:

1. Todos os atributos já foram incluídos neste caminho da árvore;
2. Os exemplos de treinamento associados com este nó folha sejam todos da mesma classe (entropia zero).

Prof. MSc. Braian Varjão

Algoritmo ID3

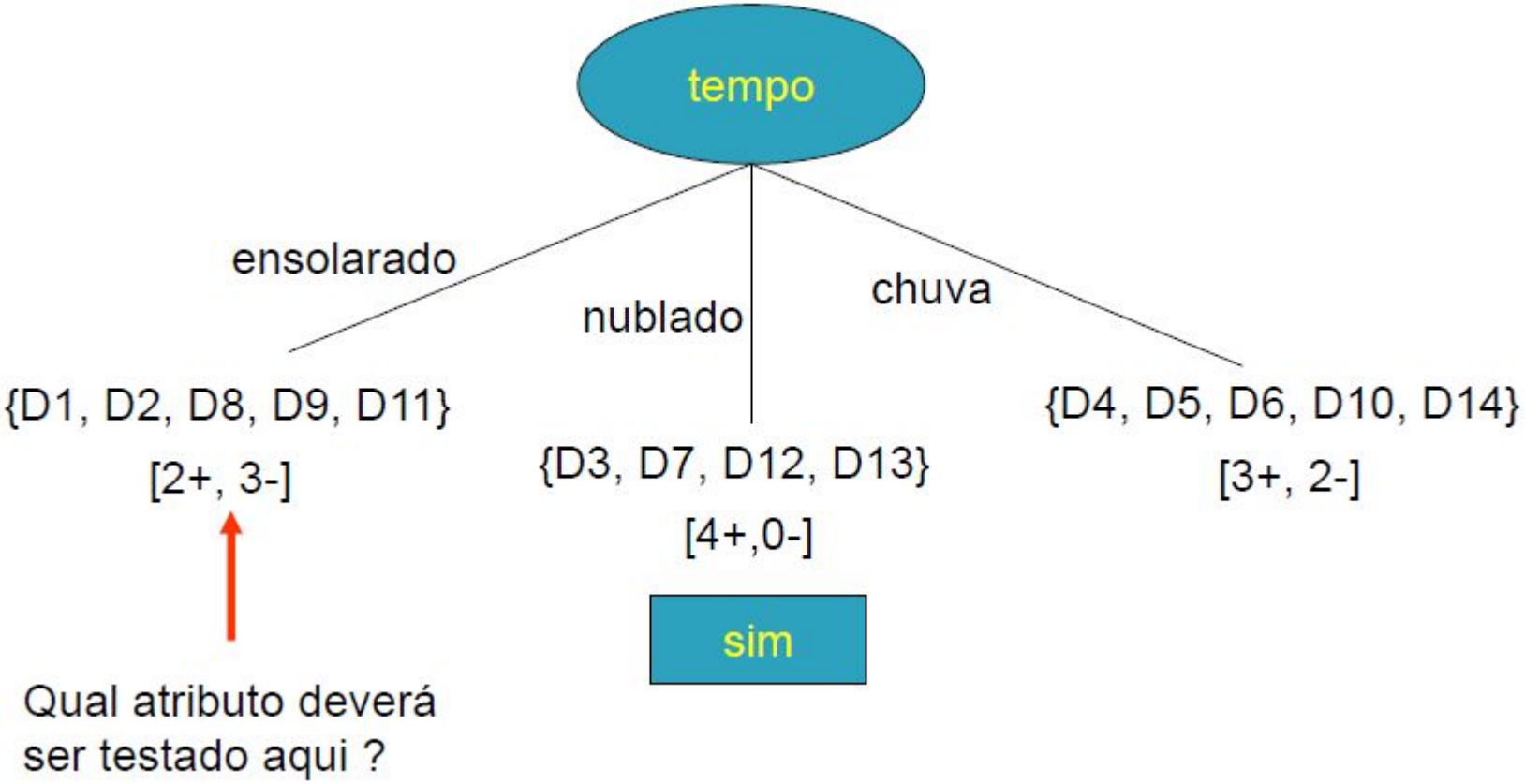
Para cada ramo da árvore o processo continua até que:

1. Todos os atributos já foram incluídos neste caminho da árvore;
2. Os exemplos de treina
nó folha sejam todos da
zero).

Isso porque o ID3 lida apenas com atributos categóricos. Algoritmos como o C4.5, que lidam com atributos contínuos, permitem que um mesmo atributo apareça mais de uma vez num mesmo ramo.

Prof. MSc. Braian Varjão

Algoritmo ID3



Algoritmo ID3

dia	tempo	temperatura	umidade	vento	jogar_tênis
D1	ensolarado	quente	alta	fraco	**não**
D2	ensolarado	quente	alta	forte	**não**
D3	nublado	quente	alta	fraco	**sim**
D4	chuva	moderada	alta	fraco	**sim**
D5	chuva	fria	normal	fraco	**sim**
D6	chuva	fria	normal	forte	**não**
D7	nublado	fria	normal	forte	**sim**
D8	ensolarado	moderada	alta	fraco	**não**
D9	ensolarado	fria	normal	fraco	**sim**
D10	chuva	moderada	normal	fraco	**sim**
D11	ensolarado	moderada	normal	forte	**sim**
D12	nublado	moderada	alta	forte	**sim**
D13	nublado	quente	normal	fraco	**sim**
D14	chuva	moderada	alta	forte	**não**

Algoritmo ID3

S.ensolarado = {D1, D2, D8, D9, D11}

Ganho(S.ensolarado, umidade) =

0.971-(3/5)0.0 – (2/5)0.0 = 0.971

Ganho(S.ensolarado, temperatura) =

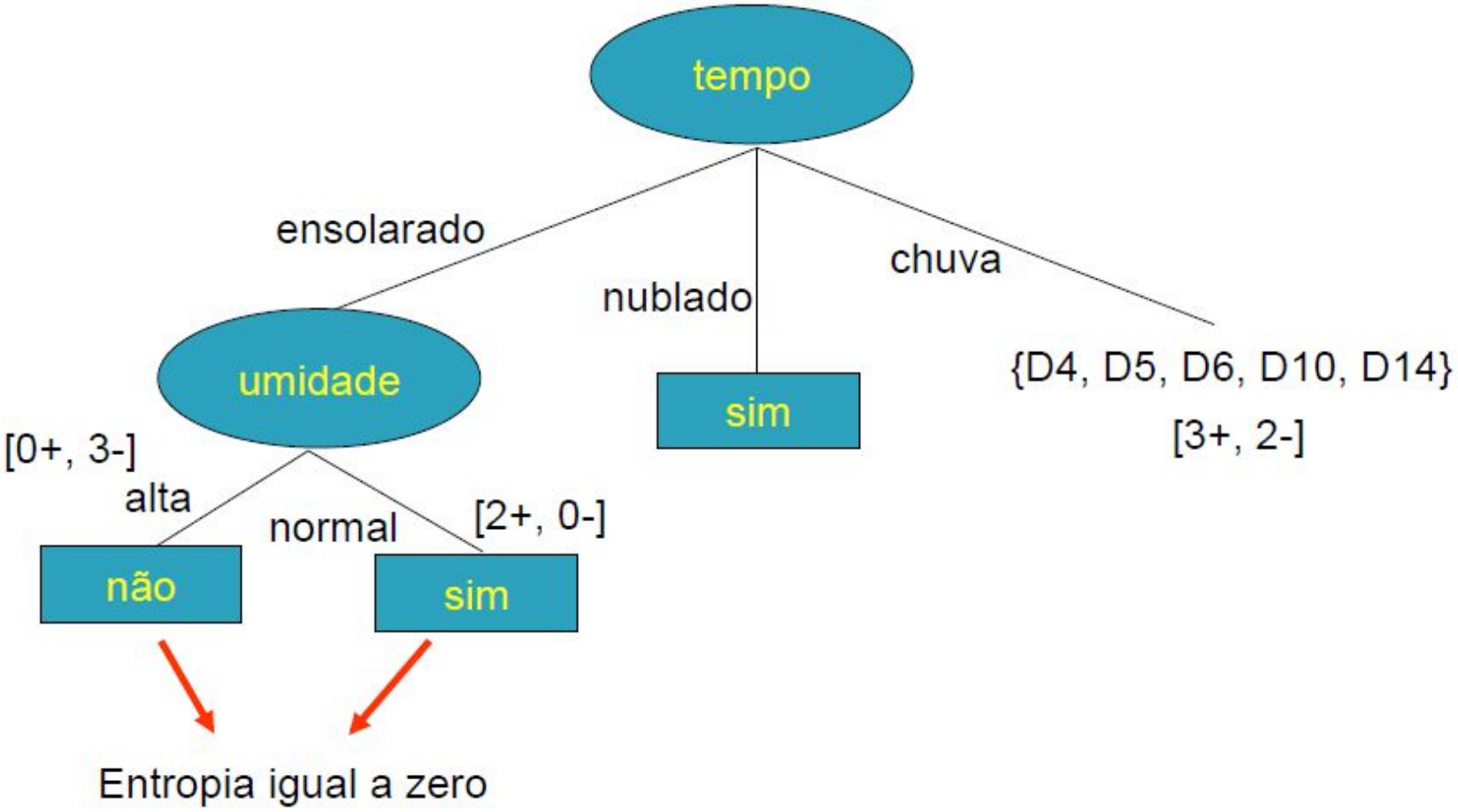
0.971-(2/5)1.0 – (2/5)0.0 – (1/5)0.0 = 0.571

Ganho(S.ensolarado, vento) =

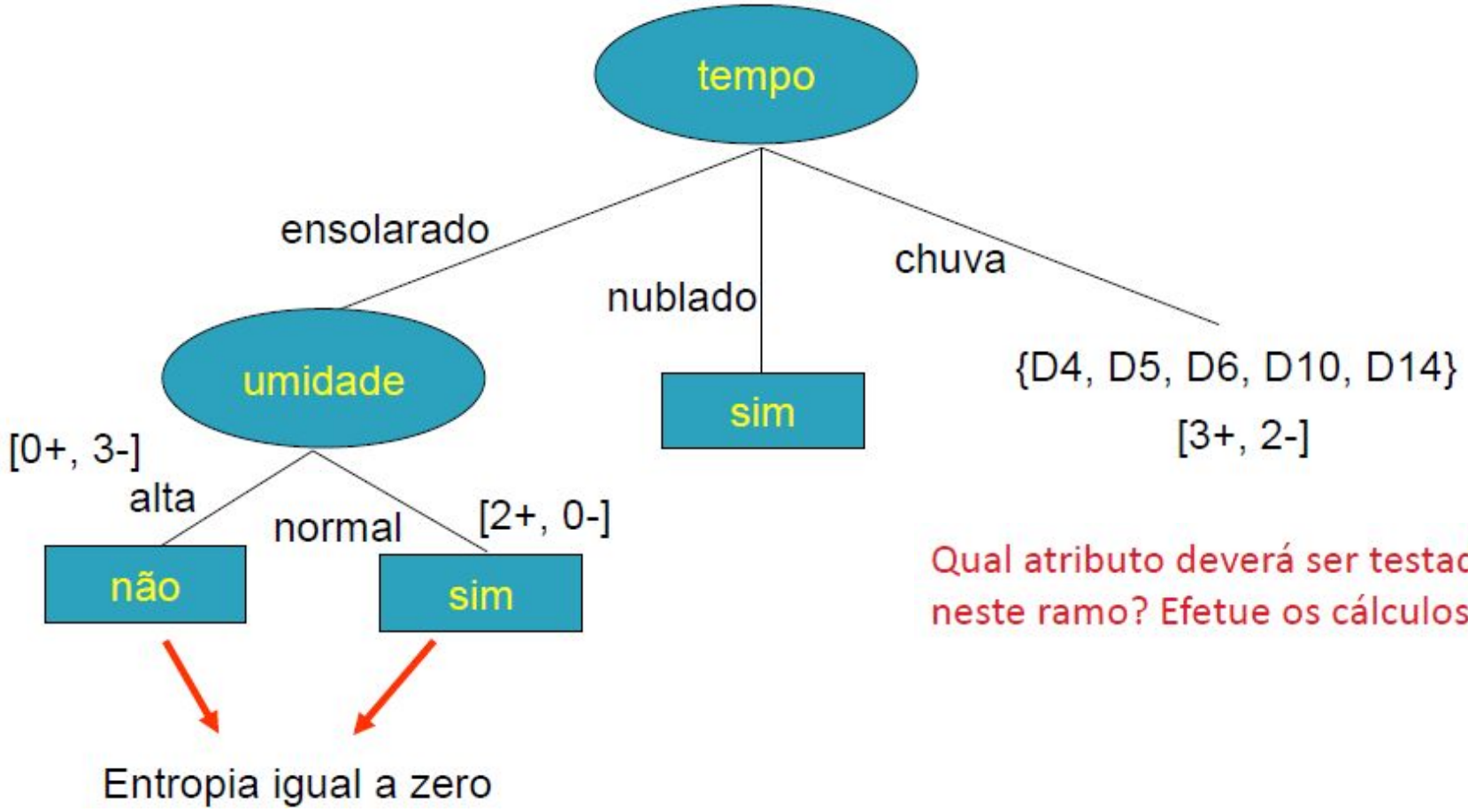
0.970-(2/5)1.0 – (3/5)0.918 = 0.020

Prof. MSc. Braian Varjão

Algoritmo ID3



Exercício 3



Qual atributo deverá ser testado neste ramo? Efetue os cálculos!

Prof. MSc. Braian Varjão

Questões importantes

Atributos contínuos

Dados faltantes

Overffiting

Poda

Prof. MSc. Braian Varjão

Árvores de decisão: atributos contínuos

dia	hora	tempo	temperatura	umidade	vento	jogar tênis
D1	6	ensolarado	quente	alta	fraco	*não*
D2	9	ensolarado	quente	alta	forte	*não*
D3	10	nublado	quente	alta	fraco	*sim*
D4	15	chuva	moderada	alta	fraco	*sim*
D5	7	chuva	fria	normal	fraco	*sim*
D6	8	chuva	fria	normal	forte	*não*
D7	16	nublado	fria	normal	forte	*sim*
D8	11	ensolarado	moderada	alta	fraco	*não*
D9	20	ensolarado	fria	normal	fraco	*sim*
D10	21	chuva	moderada	normal	fraco	*sim*
D11	13	ensolarado	moderada	normal	forte	*sim*
D12	12	nublado	moderada	alta	forte	*sim*
D13	19	nublado	quente	normal	fraco	*sim*
D14	18	chuva	moderada	alta	forte	*não*

Prof. MSc. Braian Varjão

Dados ausentes

Na construção da árvore:

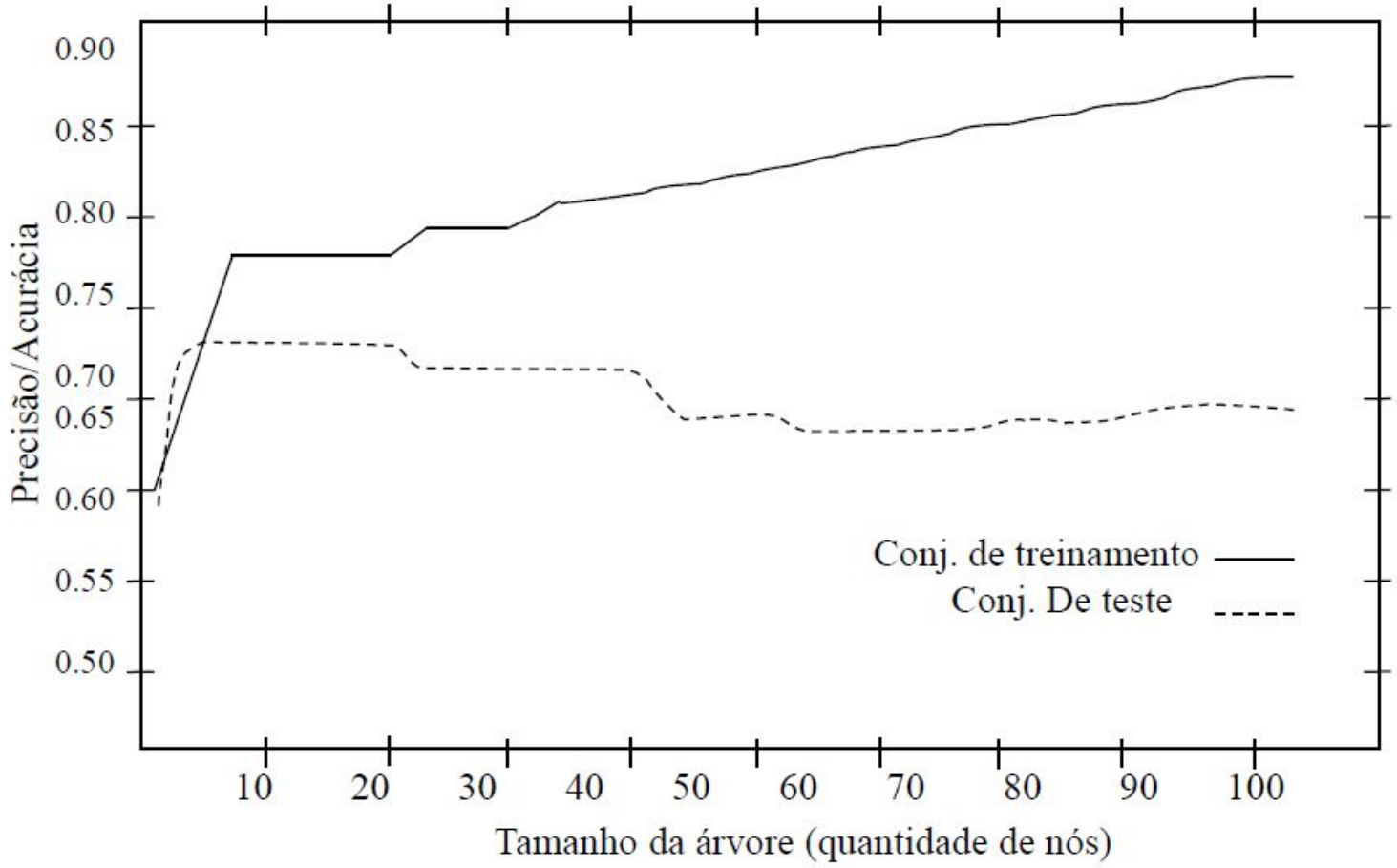
Ignora as instâncias com dados ausentes.

Na classificação:

Diante de um valor ausente, segue-se pelo ramo com mais nós (moda).

Prof. MSc. Braian Varjão

Overfitting



Poda

Pré-poda

Realizado durante o processo de construção da árvore.

Estratégias:

- IG mínimo;
- Profundidade máxima;
- Número mínimo de exemplos.

Poda

Pós poda

Remover ramos completos após a construção da árvore.

Testa a árvore com e sem o ramo com um conjunto de teste. Se o resultado da classificação for similar ou melhor, corta-se o ramo.

Prof. MSc. Braian Varjão

Árvores de decisão

Vantagens

- Fáceis de interpretar e visualizar;
- Lida facilmente padrões não lineares;
- Não há necessidade de tarefas de pré-processamento como normalização;
- Robusto a dados ausentes;
- Não tem exigências relativas à distribuição dos dados.

Prof. MSc. Braian Varjão

Árvores de decisão

Desvantagens

- Tendência a overfitting;
- Sensíveis ao desbalanceamento entre classes;
- Instáveis, ou seja, uma pequena variação nos dados pode gerar uma árvore completamente diferente.

Prof. MSc. Braian Varjão

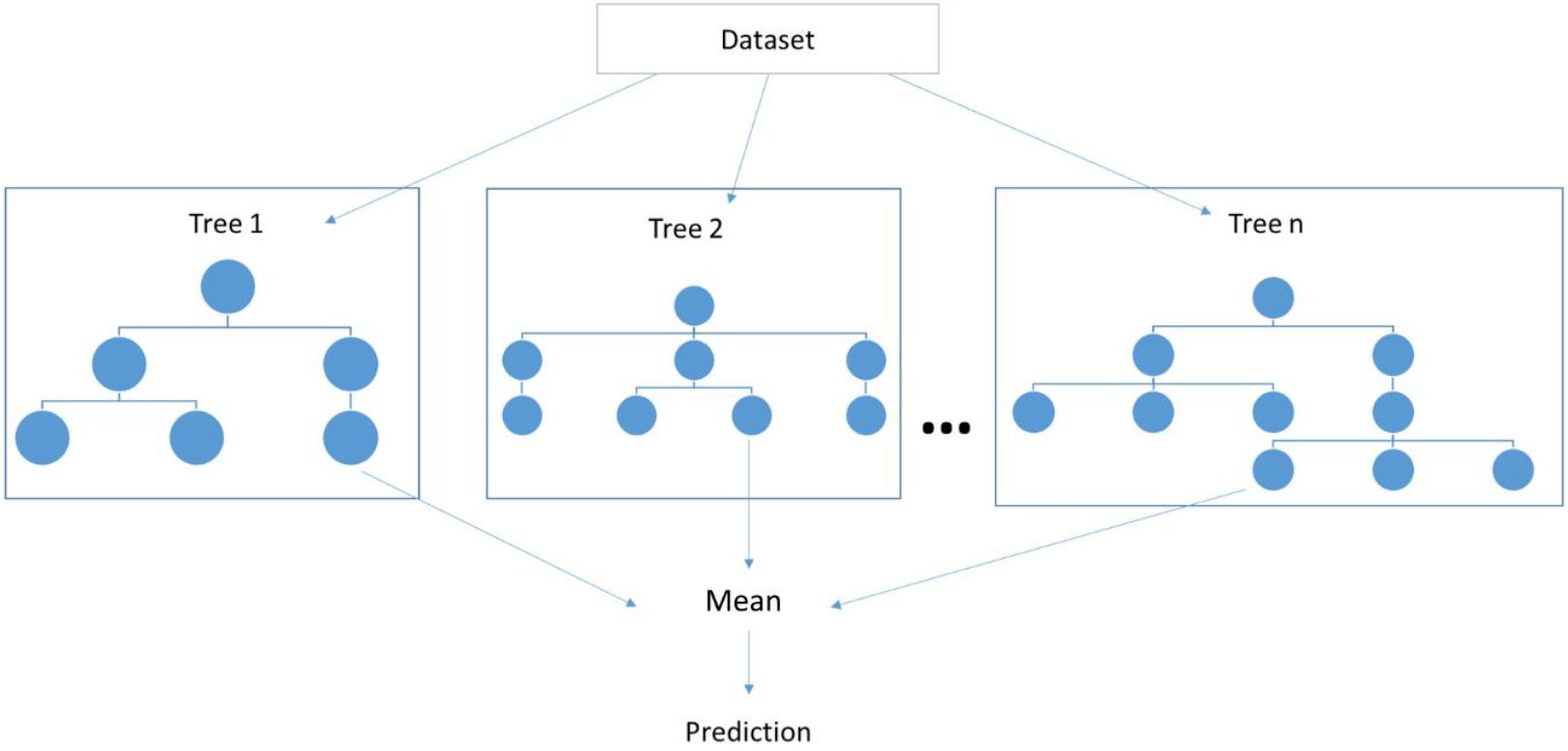
Árvores de decisão

Desvantagens

- Tendência a overfitting;
- Sensíveis ao desbalanceamento entre classes;
- Instáveis, ou seja, uma pequena variação nos dados pode gerar uma árvore completamente diferente.
  - **Esse problema pode ser minimizado com o uso de um ensemble, como o Randon Forest.**

Prof. MSc. Braian Varjão

Random Forest



Prof. MSc. Braian Varjão

E como faz no python?

```
from sklearn.tree import DecisionTreeClassifier

model= DecisionTreeClassifier()
model.fit(Features, classes)
model.predict(new_object)
```

```
from sklearn.ensemble import RandomForestClassifier

model=RandomForestClassifier(n_estimators=25)
model.fit(Features, classes)
model.predict(new_object)
```

Prof. MSc. Braian Varjão