utad

Inteligência Artificial

Aprendizagem com Árvores de Decisão

Paulo Moura Oliveira
Departamento de Engenharias
Gabinete F2.15, ECT-1
UTAD
email: oliveira@utad.pt

Como funcionam as árvores de decisão?

- ✓ Considere-se como exemplo a classificação de uma espécie de animais em duas categorias: **mamíferos** e **não mamíferos**.
- ✓ Como determinar se uma nova espécie é um mamífero ou não?
- ✓ Podemos formular questões sobre as características da espécie. Por exemplo:

Q1: A espécie tem sangue-quente ou sangue-frio?

R1: Se tiver sangue-frio não é um mamífero.

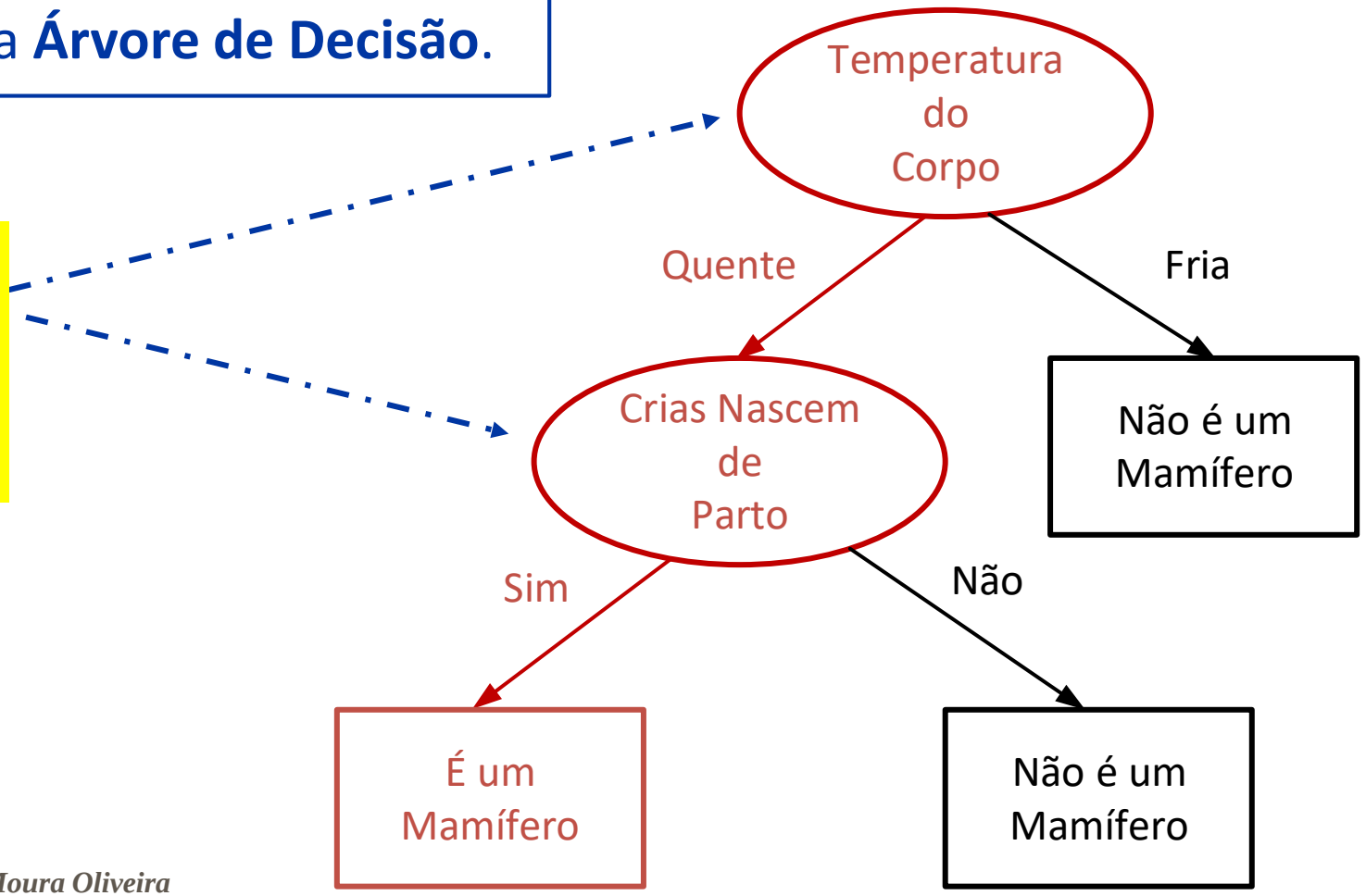
Q2: As crias nascem de parto?

R2: Se nascer de parto é muito provável que seja um mamífero.

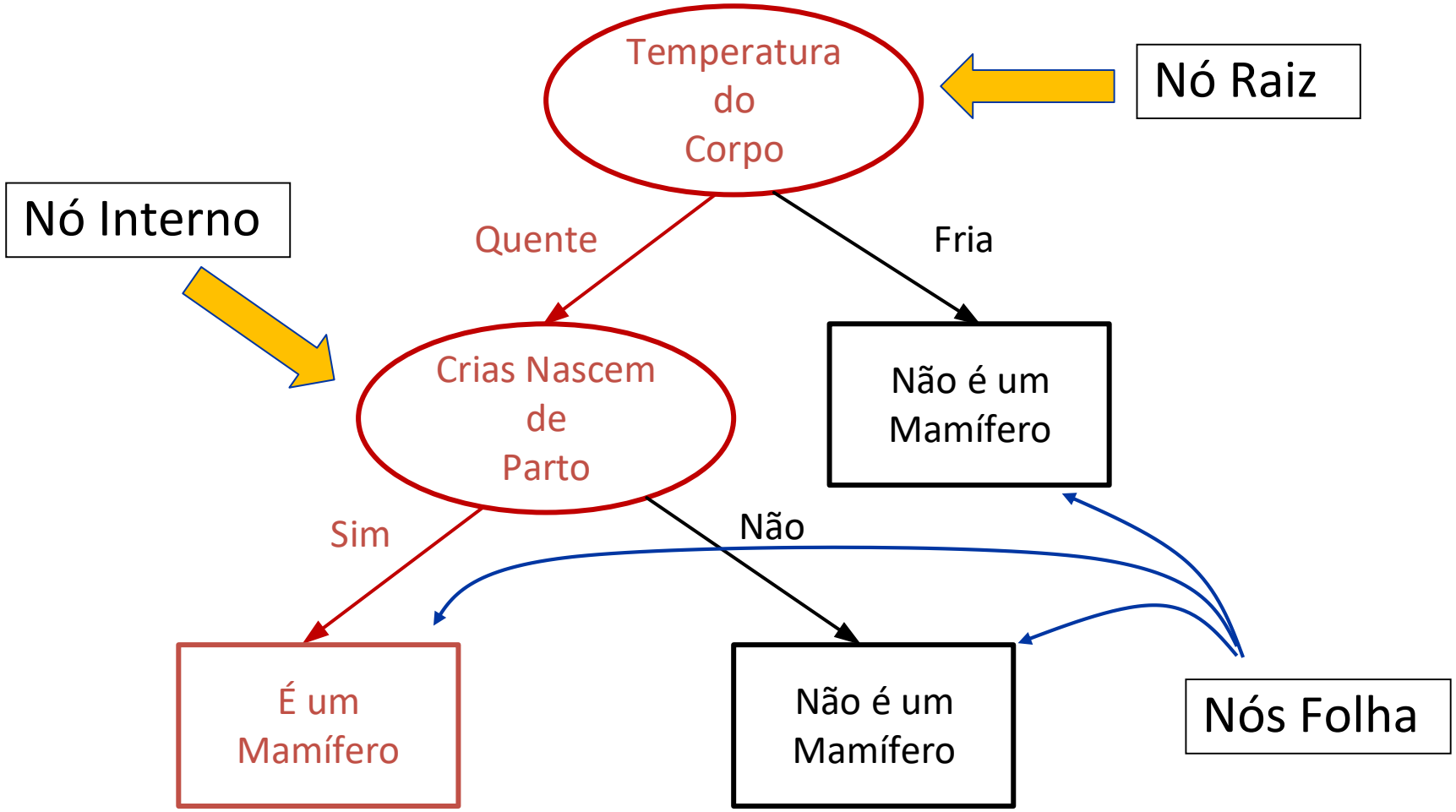
Como funcionam as árvores de decisão?

Pode-se organizar uma **série de questões e respetivas respostas**, utilizadas para efetuar uma **classificação**, de um forma **gráfica** utilizando uma **Árvore de Decisão**.

Atributos utilizados para fazer a separação

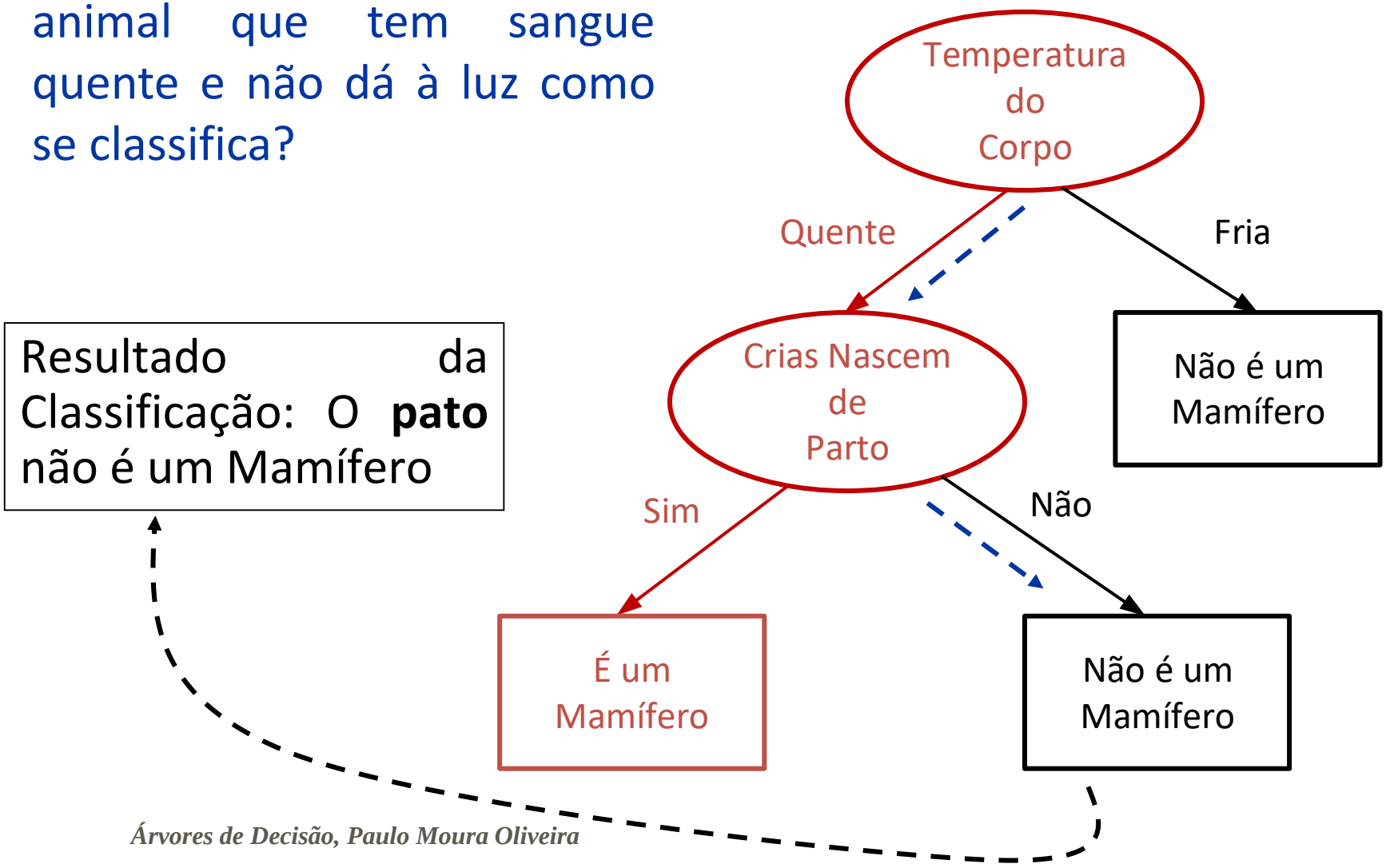


Como funcionam as árvores de decisão?



Como funcionam as árvores de decisão?

Sabendo que um **pato** é um animal que tem sangue quente e não dá à luz como se classifica?



Como construir uma árvore de decisão?

- O número de árvores de decisão que podem ser construídas a partir de um conjunto de atributos pode ser enorme.

- Embora umas árvores sejam mais precisas que outras, determinar a árvore de decisão ótima pode ser um problema complexo devido ao crescimento exponencial do número de soluções.

Existem vários algoritmos que foram desenvolvidos para construir Árvores de Decisão, tais como:

- Algoritmo de Hunt
- ID3
- C4.5
- CART, etc.

Algoritmo de Hunt

Versão Genérica

- ✓ Assumindo que:
  - $\mathbf{X_t}$ representa o conjunto das **amostras de treino** para o nó **t**
  - A designação das classes é representado por $y = \{y_1, y_2, \ldots, y_c\}$

Passo 1

Se todas as amostras de $\mathbf{X_t}$ pertencem à mesma classe $\mathbf{y_t}$, então **t é um nó folha**, designado por $\mathbf{y_t}$

Passo 2

Se $\mathbf{X_t}$ **tem amostras que pertencem a mais do que uma classe**:

- selecionar a **condição teste de atributo** para **particionar** as amostras em subconjuntos mais pequenos.
- criar um **nó folha** para cada subconjunto resultante da primeira condição.
- aplicar o algoritmo **recursivamente** para cada nó **folha**.

Algoritmo de Hunt

Versão Genérica: Exemplo [1]

- ✓ Considere que se pretende **prever** se um candidato a um empréstimo bancário vai ser um devedor: **Cumpridor** ou **Incumpridor** (Classes).

- ✓ A previsão vai ser efetuada considerando um **conjunto de treino** utilizando dados de clientes anteriores.

Amostra #	*Proprietário*	*Estado Civil*	*Rendimento Anual x1000€*	*Devedor Incumpridor*
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

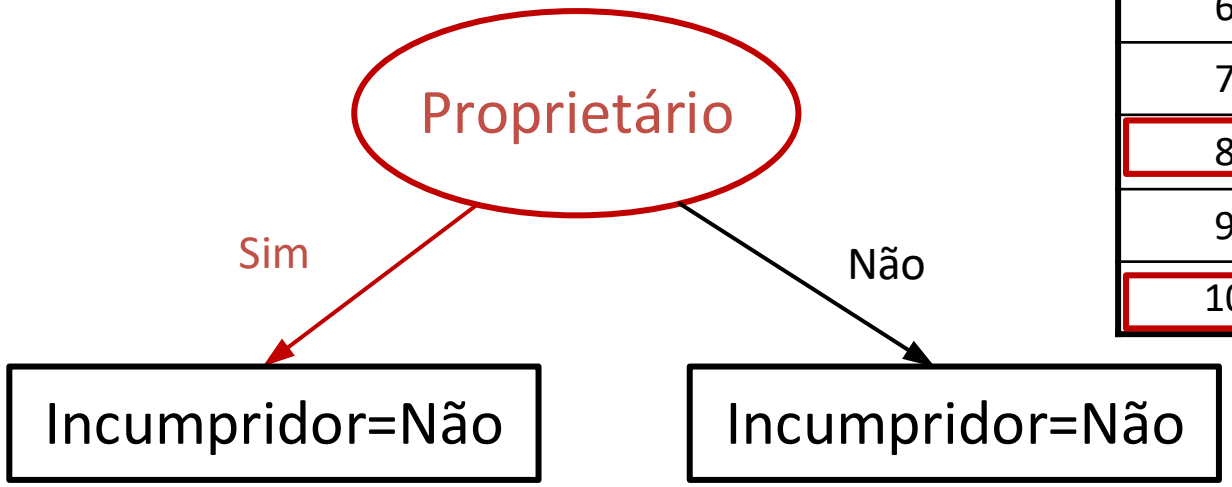
Algoritmo de Hunt

Versão Genérica: Exemplo [1]

- ✓ Nó inicial: a maioria dos casos são **cumpridores**,

Incumpridor=Não

- ✓ Escolhendo o atributo **Proprietário** para dividir os dados obtém-se:



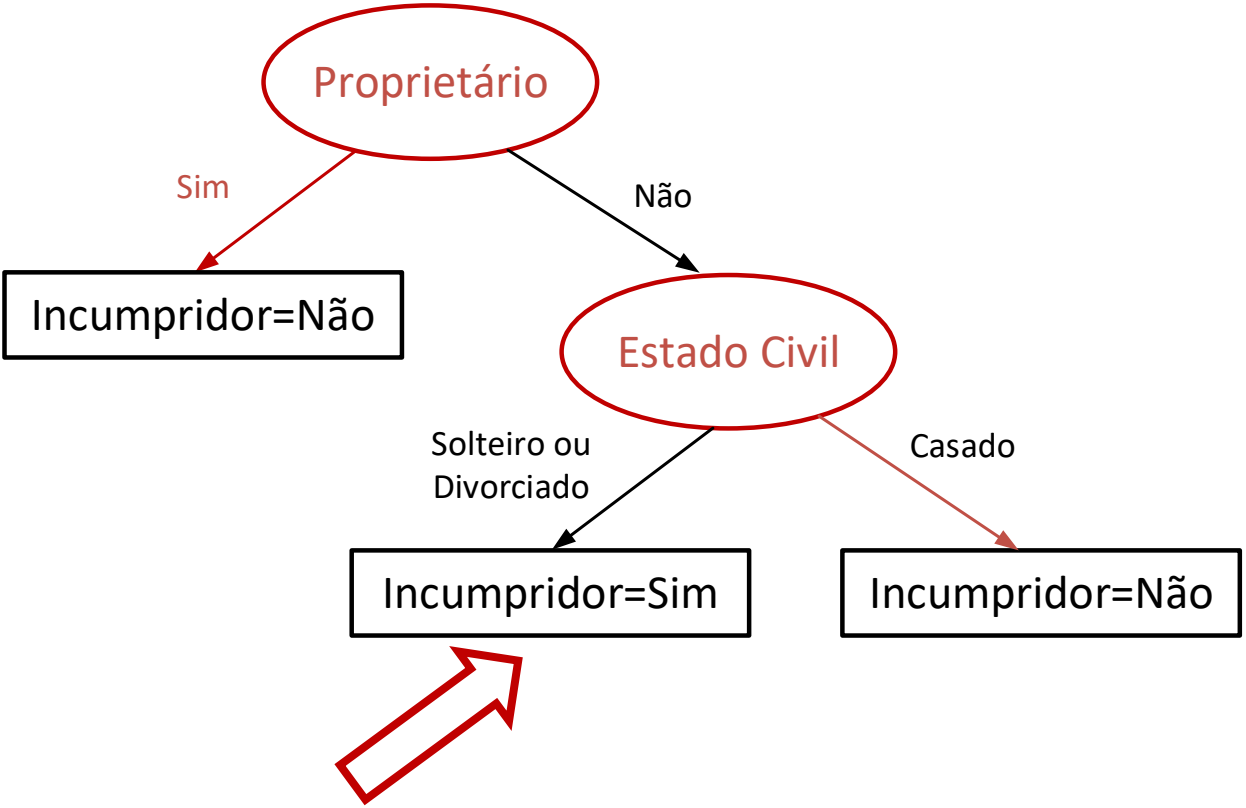
Amostra #	*Proprietário*	*Devedor Incumpridor*
1	Sim	Não
2	Não	Não
3	Não	Não
4	Sim	Não
5	Não	Sim
6	Não	Não
7	Sim	Não
8	Não	Sim
9	Não	Não
10	Não	Sim

- ✓ Continua-se a aplicar o algoritmo recursivamente ao **nó da direita**.

Algoritmo de Hunt

Versão Genérica: Exemplo [1]

- ✓ Resultado:

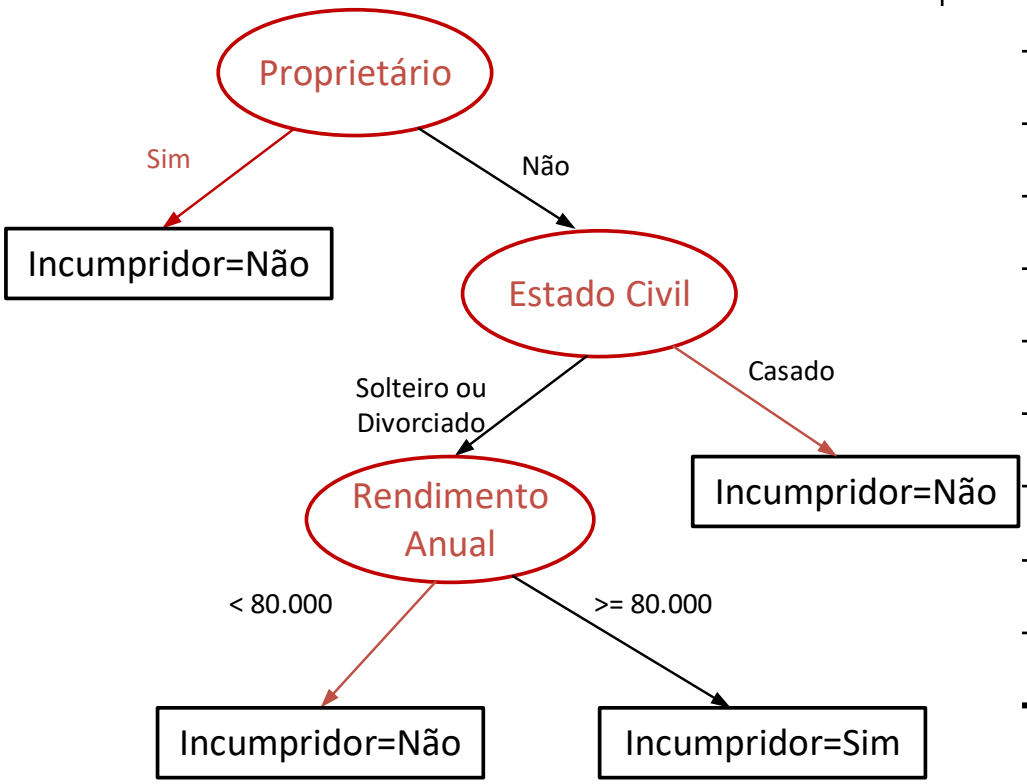


Proprietário	*Estado Civil*	*Devedor Incumpridor*
Sim	Solteiro	Não
Não	Casado	Não
Não	Solteiro	Não
Sim	Casado	Não
Não	Divorciado	Sim
Não	Casado	Não
Sim	Divorciado	Não
Não	Solteiro	Sim
Não	Casado	Não
Não	Solteiro	Sim

- ✓ Continua-se a aplicar o algoritmo recursivamente ao **nó da esquerda**.

Algoritmo de Hunt

Versão Genérica: Exemplo [1]



Proprietário	*Estado Civil*	*Rendimento Anual x1000€*	*Devedor Incumpridor*
Sim	Solteiro	125K	Não
Não	Casado	100K	Não
Não	Solteiro	70K	Não
Sim	Casado	120K	Não
Não	Divorciado	95K	Sim
Não	Casado	60K	Não
Sim	Divorciado	220K	Não
Não	Solteiro	85K	Sim
Não	Casado	75K	Não
Não	Solteiro	90K	Sim

Projeto do Processo de Indução

- ✓ Duas questões fundamentais do **projeto para processo de indução nas árvores de decisão** são:

1. **Como é que as amostras de treino devem ser separadas?**
2. **Quando se deve parar o processo de separação?**

1. Como é que as amostras de treino devem ser separadas?

- ✓ Cada passo recursivo do processo de crescimento da árvore precisa de **escolher uma condição de teste de atributos** para dividir as amostras nos subconjuntos.

- ✓ O algoritmo tem de especificar/fornecer:
  - um **método** para especificar a **condição de teste** para vários tipos de atributo.
  - uma **medida** objetiva para medir o **mérito de cada condição**.

Projeto do Processo de Indução

2. Quando se deve parar o processo de separação?

- ✓ O processo de paragem pode ser:
  - a expansão de um nó até quando **todas as amostras pertencerem à mesma classe**.

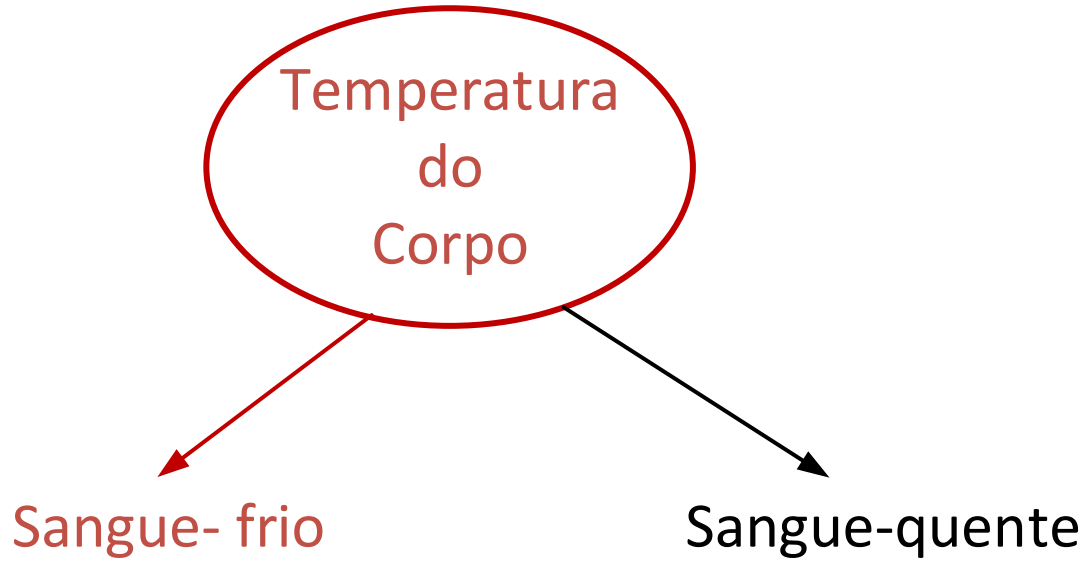
  ou

  - todas as amostras tiverem os **mesmos valores dos atributos**.

Métodos para Exprimir Condições de Teste de Atributos

Atributos Binários

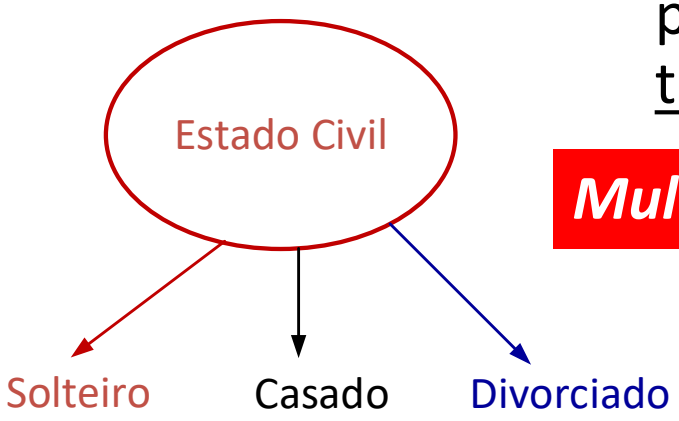
- A **condição de teste** para um atributo binário gera dois resultados possíveis.



Métodos para Exprimir Condições de Teste de Atributos

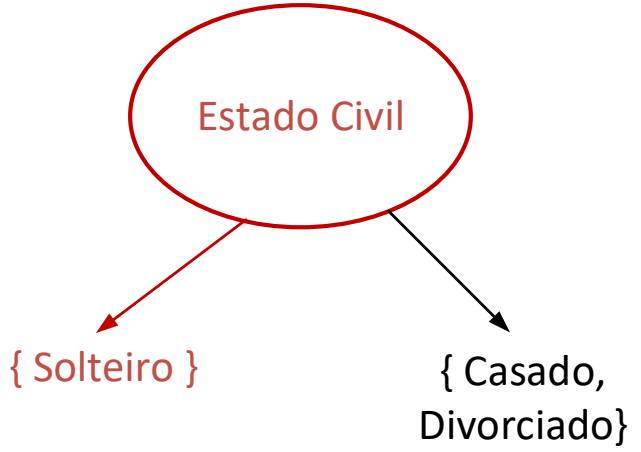
Atributos Nominais

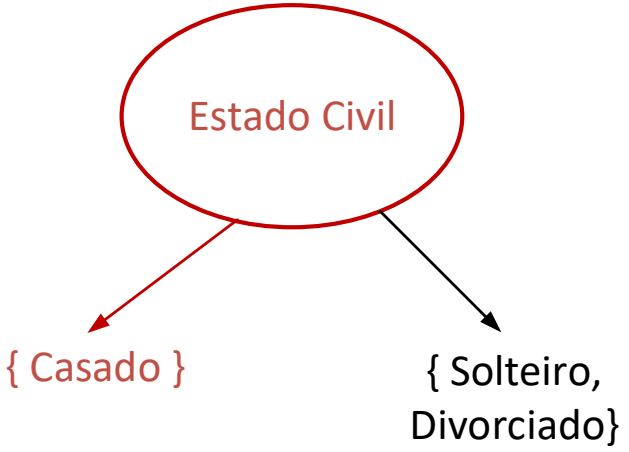
- ✓ A **condição de teste** para um atributo nominal pode ser expressão de várias formas incluindo todos os seus valores.

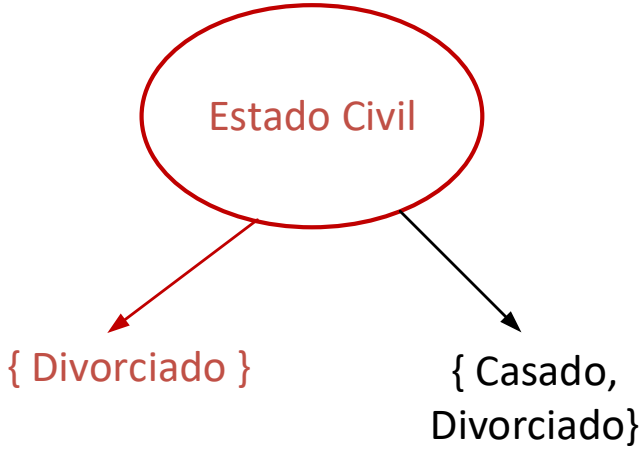


Multiway

Binária





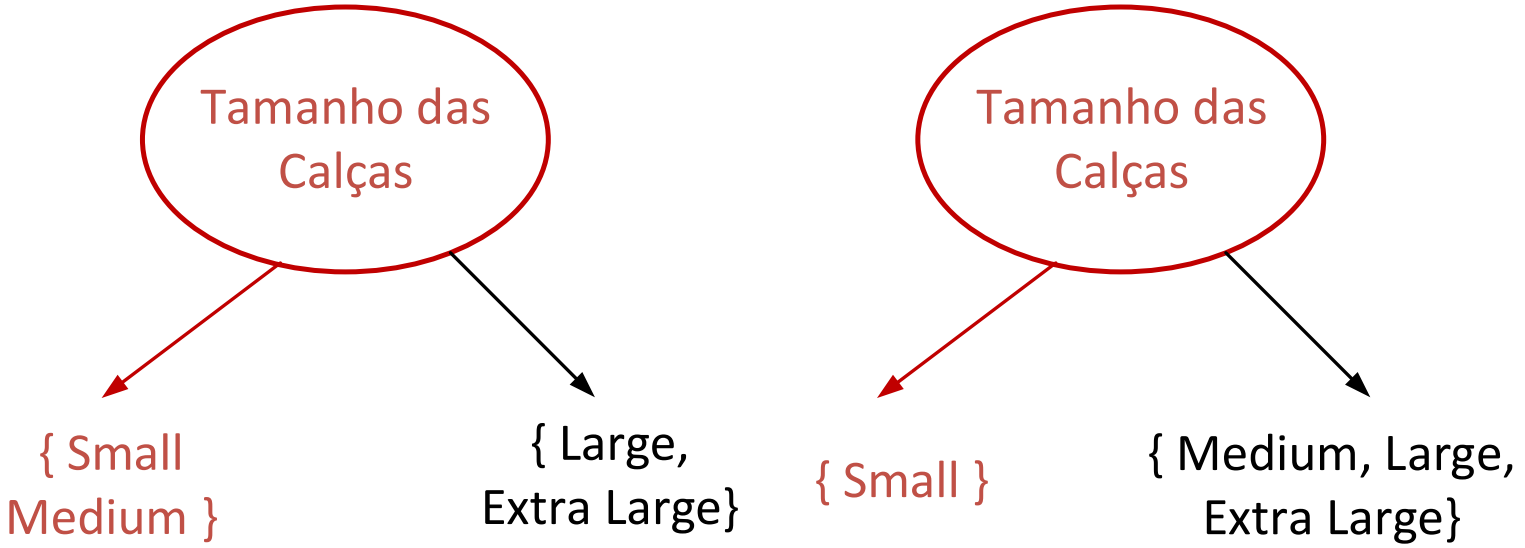


Métodos para Exprimir Condições de Teste de Atributos

Atributos Ordinais

- ✓ Este tipo de atributos podem ser agrupados de forma **binária** ou **multiway**.
- ✓ Os agrupamentos permitidos implicam **cumprir** a **propriedade da ordem** dos **valores do atributo**.

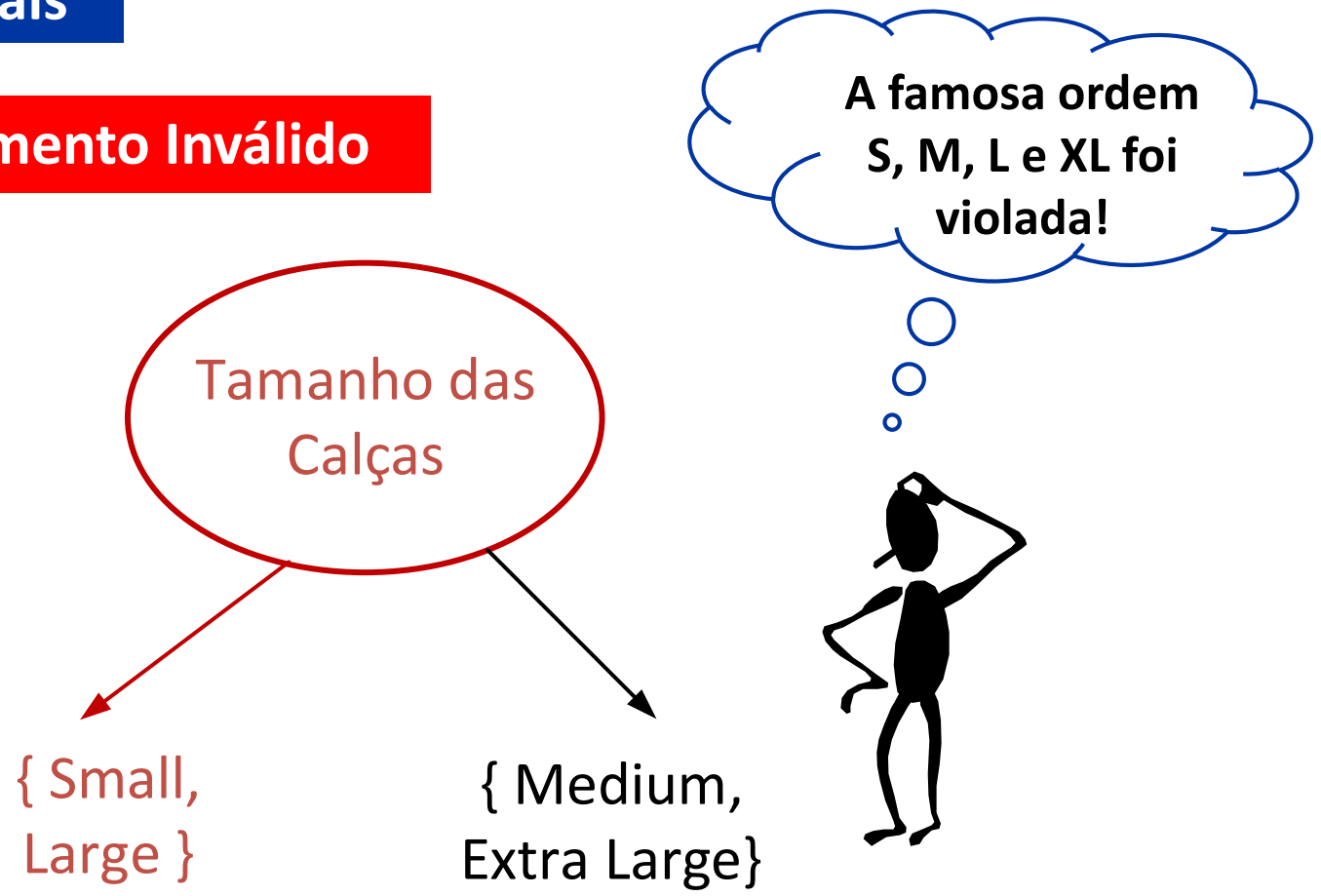
Agrupamentos Válidos



Métodos para Exprimir Condições de Teste de Atributos

Atributos Ordinais

Agrupamento Inválido

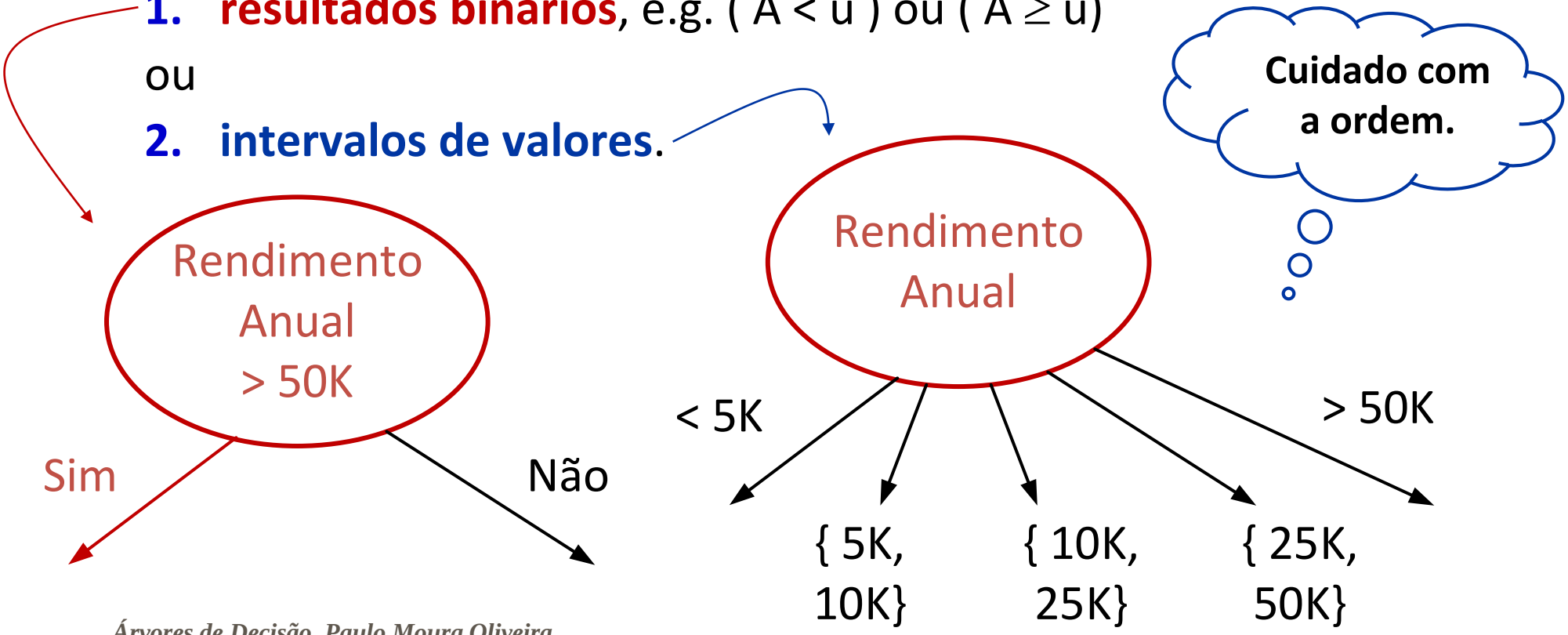


Métodos para Exprimir Condições de Teste de Atributos

Atributos Contínuos

- ✓ Neste caso a **condição de teste** pode ser expressa por **comparações** com:
  1. **resultados binários**, e.g. (A < u) ou (A ≥ u)

  ou
  2. **intervalos de valores**.



Critérios de Separação

- Os métodos desenvolvidos para construir árvores de decisão **variam entre si fundamentalmente no critério de seleção da separação utilizado**.

Qual é o atributo mais útil para ser testado em cada nó ?

Critérios de Seleção da Separação ("Split Criteria")

- Há **muitos critérios** que podem ser utilizados para determinar a **melhor forma de separação dos dados**, tais como os seguintes:
  - Entropia
  - Gini
  - Erro de Classificação
- Estes critérios, ou medidas, são definidos em termos da distribuição das amostras nas classes, antes e depois da classificação.

Método ID3

- ✓ Um dos métodos mais conhecidos é **ID3 (*Iterative Dichotomizer*)** (desenvolvido por Quinlan) para **atributos categóricos (ou qualitativos)**. Posteriormente o ID3 foi adaptado para atributos **contínuos** (C4.5)

S – Conjunto de Treino

Algoritmo ID3 (Básico)

1. Criar um nó raiz
2. **If** todos os exemplos do conjunto S pertencem à mesma classe **Cj**
3. **then** nomear a raiz como **Cj**
4. **else**
   - Selecionar o **atributo**, **A**, mais informativo com **valores v1,v2,…vn**
   - Dividir o conjunto **S** em **S1, S2, …, Sn** de acordo com **v1,v2,…vn**
5. Construir as subárvores **T1, T2,…,Tn** para **S1, S2, …, Sn**

Método ID3

Como parar a separação?

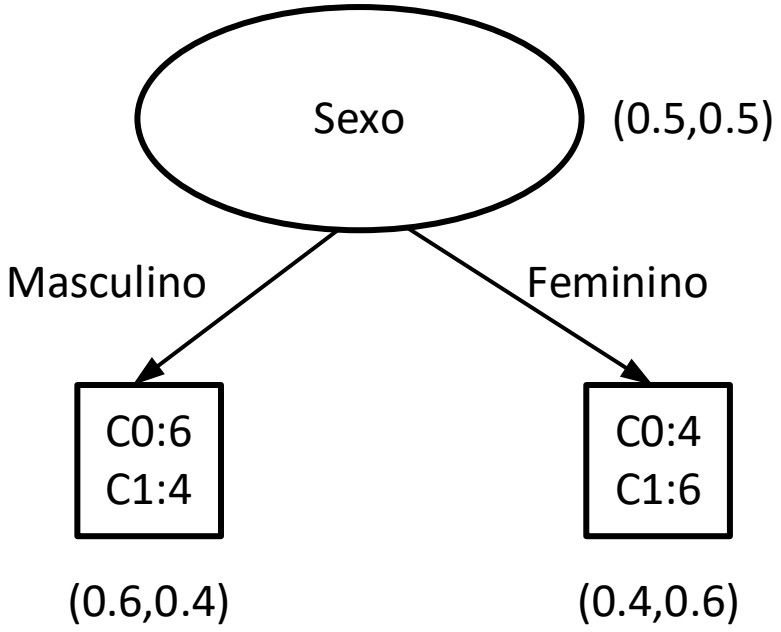
- ✓ Há vários métodos, com vantagens/desvantagens, como por exemplo:
  - Parar quando todas as amostras para um dado nó pertencem à mesma classe;
  - Não há mais atributos para mais partições.

Critérios de Seleção da Separação

- ✓ Considere-se que: $\mathrm{p}(i|t) = p_i$

  representa a ***fração de amostras pertencentes à classe i no nó t.***

- ✓ Num problema com duas classes a distribuição em cada nó pode ser representada como: (p_0,p_1) em que p_1=1- p_0



As medidas desenvolvidas para avaliar a melhor separação baseiam-se no grau de impureza dos nós folha.

- ❑ Exemplos:
  - nó com (0,1) tem uma **impureza zero**,
  - nó com (0.5,0.5) tem **impureza máxima**.

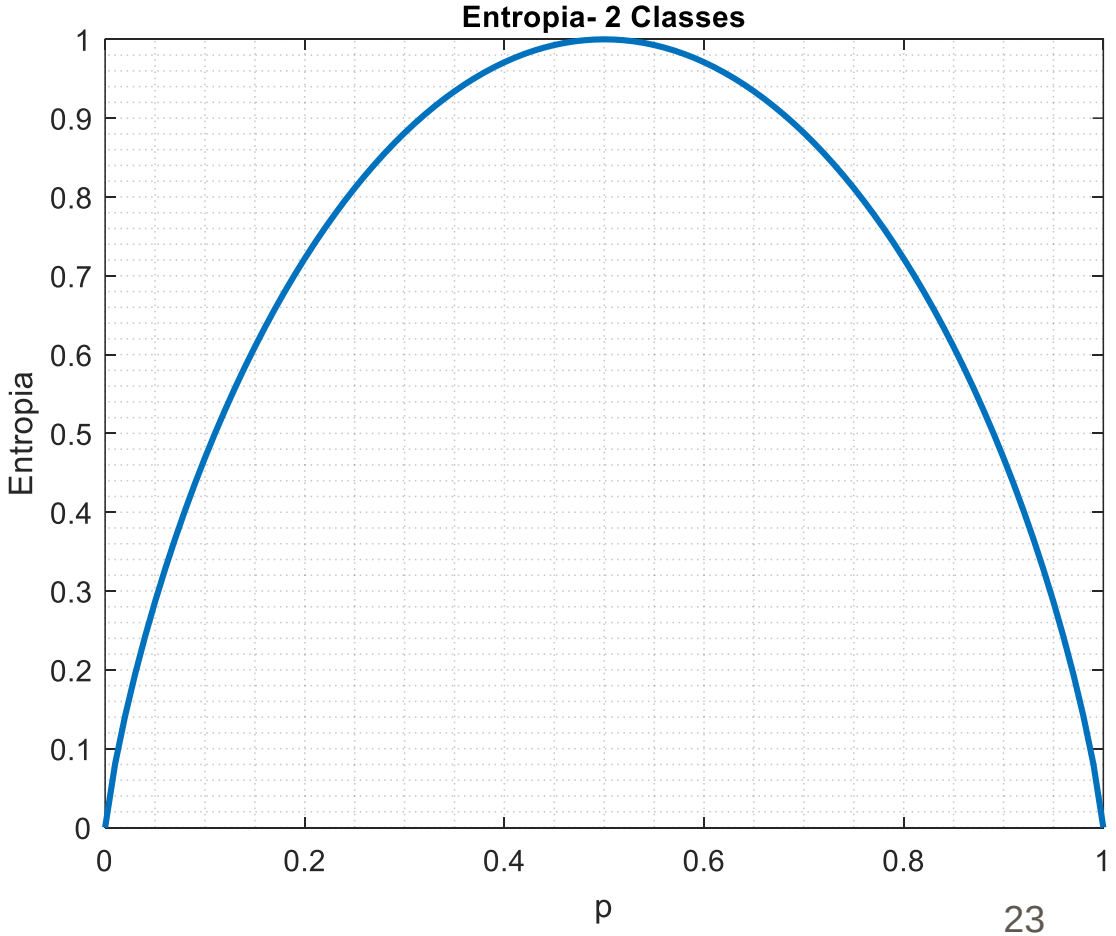
Critérios de Seleção da Separação

Entropia no nó t

$$Entropia(t) = -\sum_{i=0}^{c-1} p_i(i|t) \log_2 p_i(i|t)$$

- ✓ c – número de classes
- ✓ $0 \log_2 0 = 0$ nos cálculos

$$Entropia = -\sum_{i=0}^{c-1} p_i \log_2 p_i$$



Critérios de Seleção da Separação

Entropia no nó t

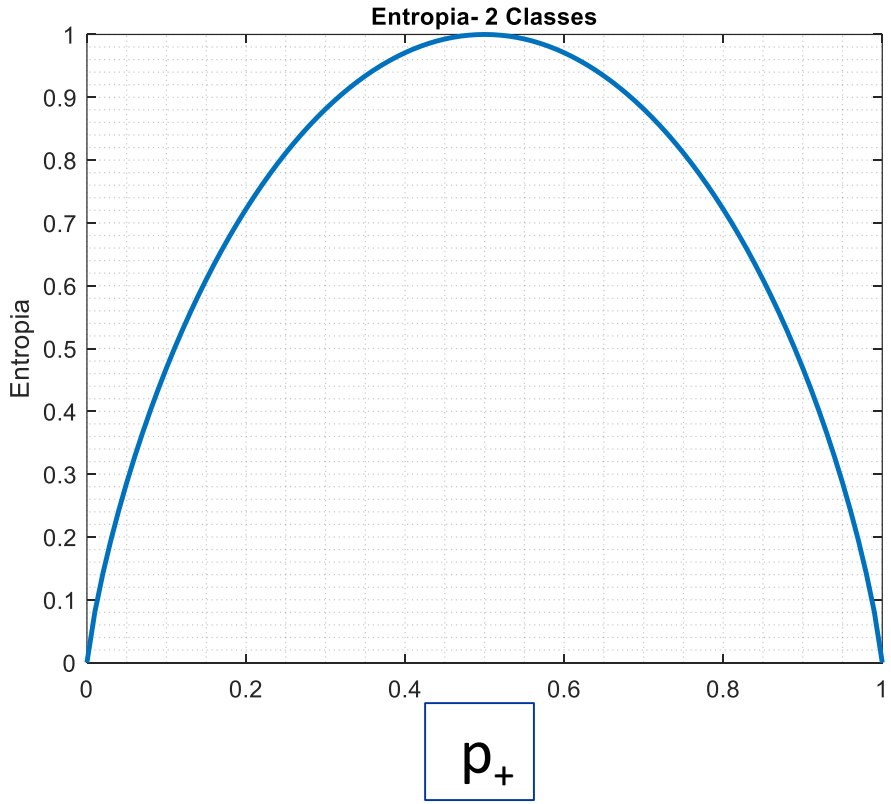
✓ Frequentemente é utilizada a notação seguinte, considerando duas classes:

$$E(S) = -\ p_+ \log_2 p_+ - \ p_- \log_2 p_-$$

Com:

- S conjunto de amostras
- **p_+ fração de amostras na classe +**
- **p_- fração de amostras na classe –**

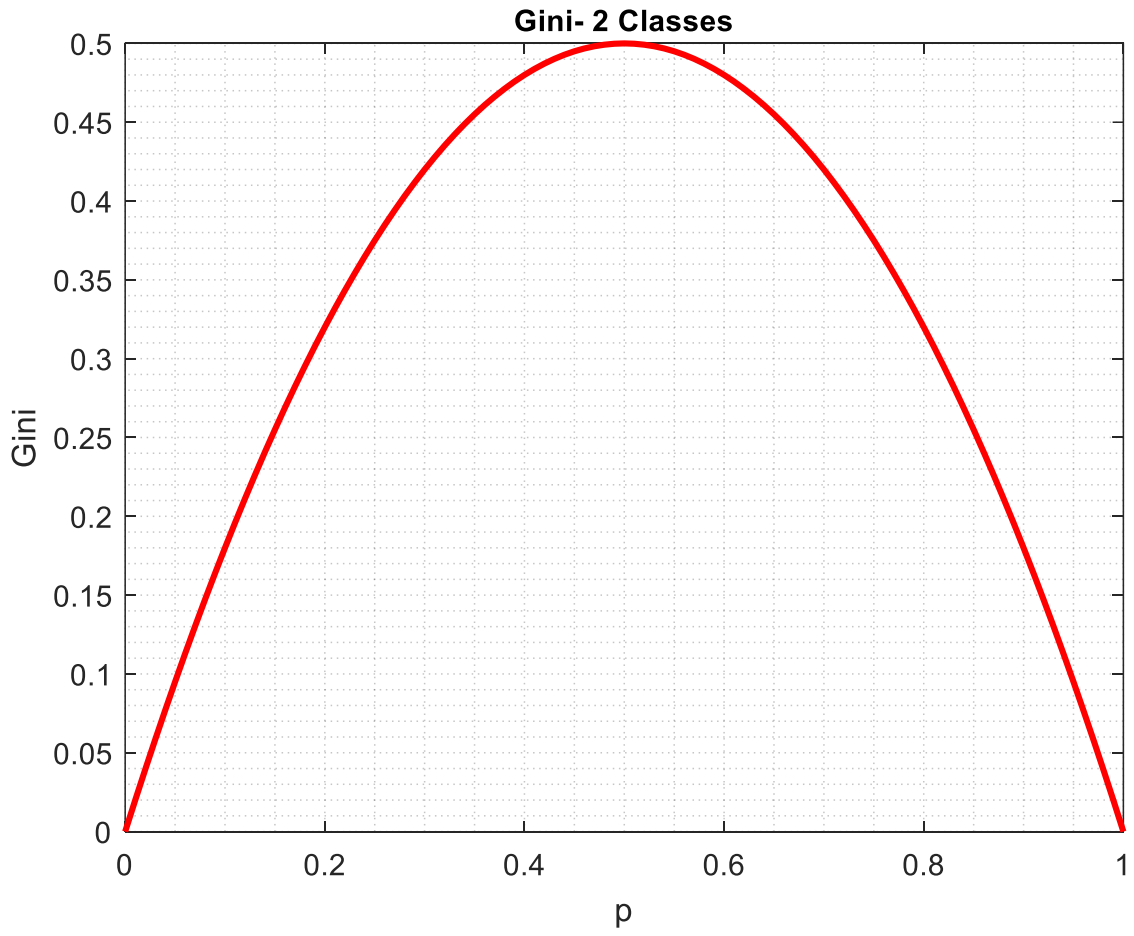
$$p_- = 1 - p_+$$



Critérios de Seleção da Separação

Gini no nó t

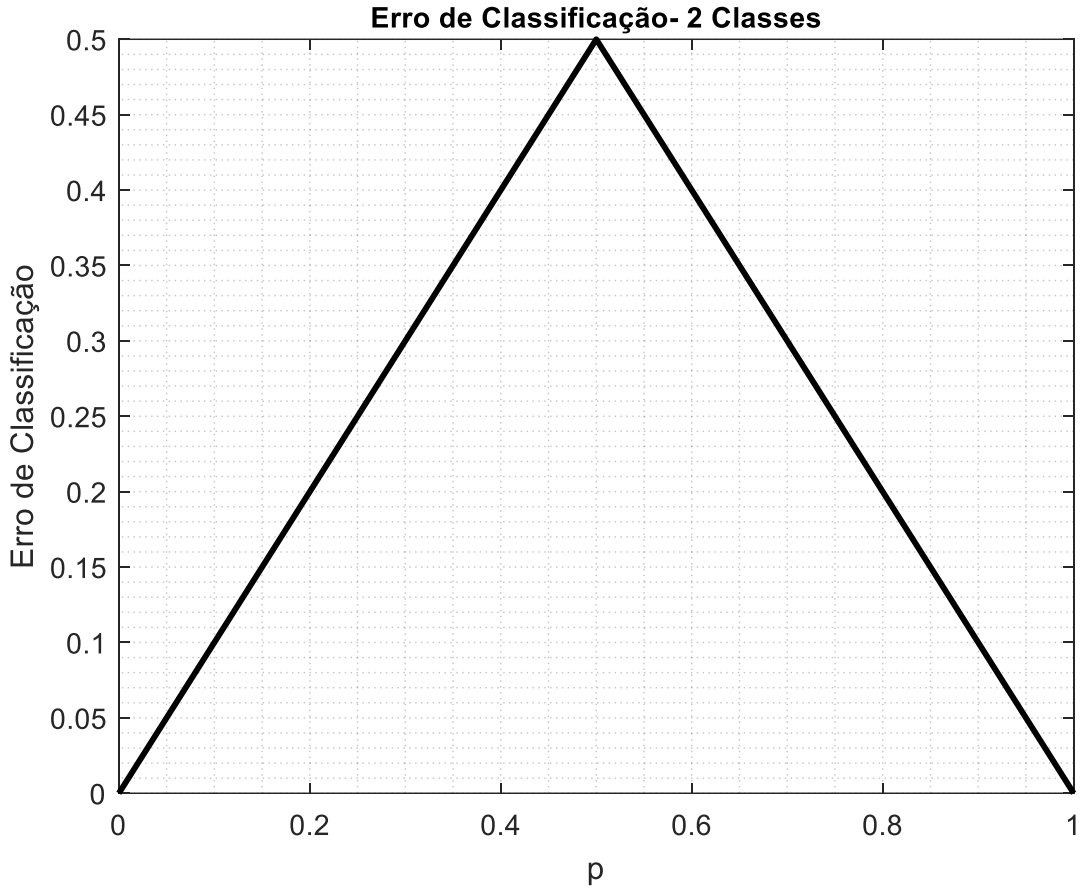
$$Gini\ (t) = 1 - \sum_{i=0}^{c-1} (p_i)^2$$



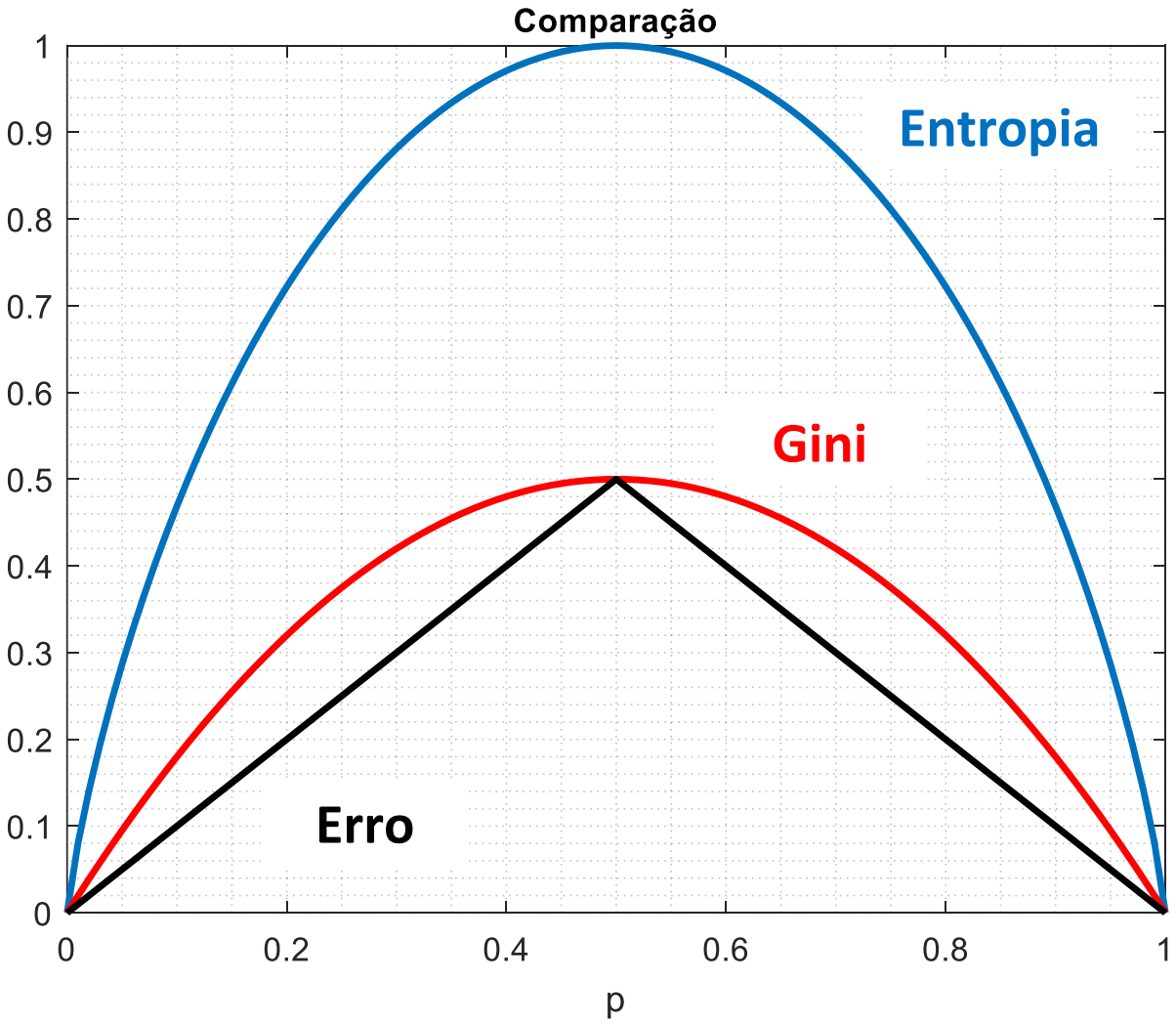
Critérios de Seleção da Separação

Erro de Classificação (*Classification Error*) no nó t

$$Erro\ de\ Classificação\ (t) = 1 - \max_i p_i$$



Critérios de Seleção da Separação



Critérios de Seleção da Separação

- Exemplos [1]:

$$Entropia = -\sum_{i=0}^{c-1} p_i \log_2 p_i$$

$$Gini\ (t) = 1 - \sum_{i=0}^{c-1} (p_i)^2$$

Nó N1	#
C0	0
C1	6

- Entropia = - (0/6)*log2(0/6) - (6/6)*log2(6/6) = 0
- Gini = 1 - (0/6)^2 - (6/6)^2 = 0
- Erro = 1 - max [(0/6), (6/6)] = 0

Nó N2	#
C0	1
C1	5

- Entropia = - (1/6)*log2(1/6) - (5/6)*log2(5/6) = 0.650
- Gini = 1 - (1/6)^2 - (5/6)^2 = 0.278
- Erro = 1 - max [(1/6), (5/6)] = 0.167

Nó N3	#
C0	3
C1	3

- Entropia = - (3/6)*log2(3/6) - (3/6)*log2(3/6) = 1
- Gini = 1 - (3/6)^2 - (3/6)^2 = 0.5
- Erro = 1 - max [(3/6), (3/6)] = 0.167 = 0.5

Grau de Impureza Aumenta

Critérios de Seleção da Separação

Para avaliar o **desempenho duma condição de teste** tem de se comparar:

- o **grau de impureza do nó pai (antes da separação)** com
- o **grau de impureza do nós filhos (depois da separação).**

Quanto maior a diferença, melhor é a condição de teste!

Ganho - Δ

$$\Delta = G = I(pai) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

✓ Em que:

- I(.) - representa o grau de impureza de um dado nó,
- N – número total de amostras no nó pai
- k – número de atributos
- $N(v_j)$ – número de amostras associadas com o nó filho v_j

Critérios de Seleção da Separação

Os algoritmos de indução na árvores de decisão selecionam condições de teste que Maximizam o Ganho Δ (ou G) .

$$\Delta = I(pai) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

Média Ponderada das medidas de impureza dos nós filhos

Se a medida de impureza utilizada, I, for a Entropia então este ganho é designado como Ganho de Informação, Δ_{info}

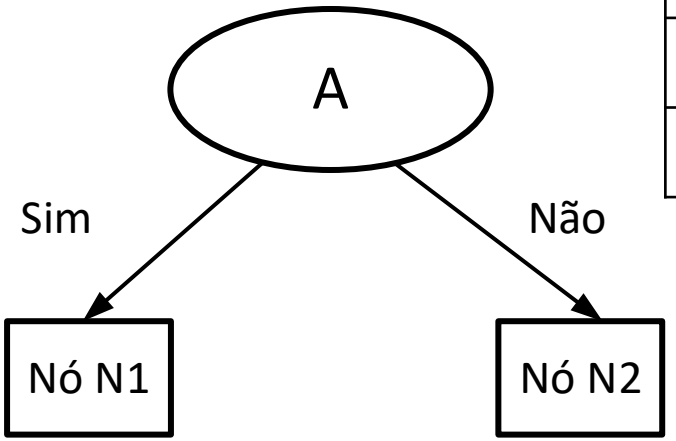
Separação de Atributos Binários

- Exemplos [1]:

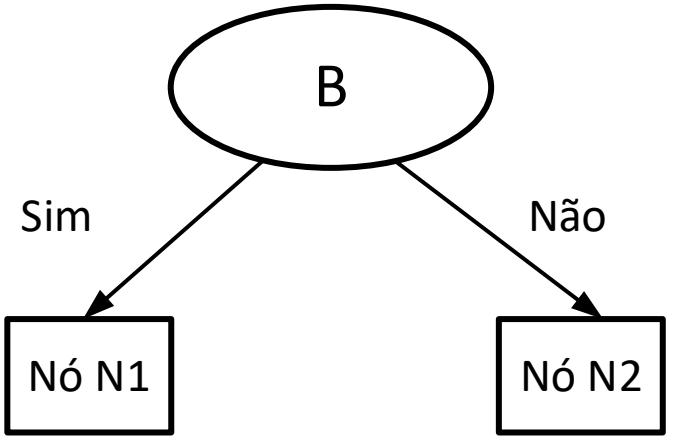
A e B: Atributos

$$Gini\ (t) = 1 - \sum_{i=0}^{c-1} (p_i)^2$$

	Pai
C0	6
C1	6
Gini=0.5	



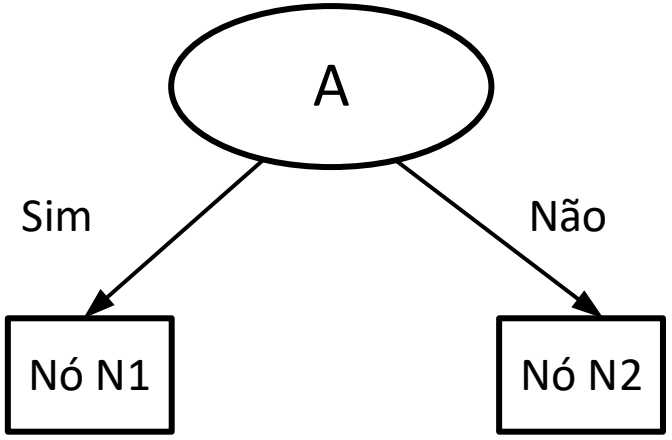
	N1	N2
C0	4	2
C1	3	3
Gini	?	?
Gini=?		



	N1	N2
C0	1	5
C1	4	2
Gini	?	?
Gini=?		

Separação de Atributos Binários

- Exemplos [1]:



	Pai
C0	6
C1	6
Gini=0.5	

$$Gini\ (t) = 1 - \sum_{i=0}^{c-1} (p_i)^2$$

	N1	N2
C0	4	2
C1	3	3
Gini	0.4898	0.480
	Gini=0.486	

- Gini (N1) = 1 - (4/7)^2 - (3/7)^2 = 0.4898
- Gini (N2) = 1 - (2/5)^2 - (3/5)^2 = 0.4800

Gini Média Ponderada:

- Gini = (7/12)*0.4898+ (5/12)*0.4800=

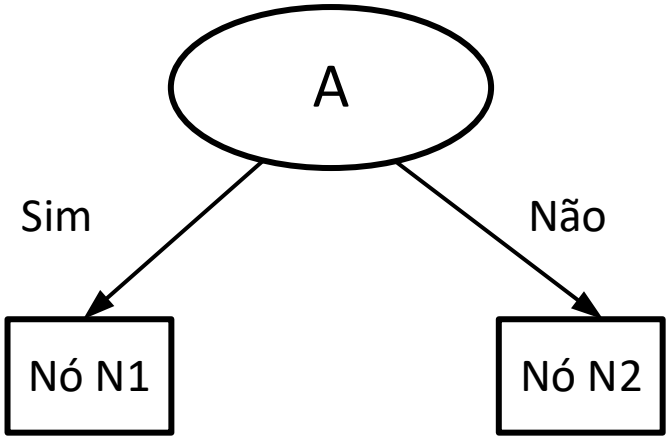
0.486

$$\Delta = I(pai) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

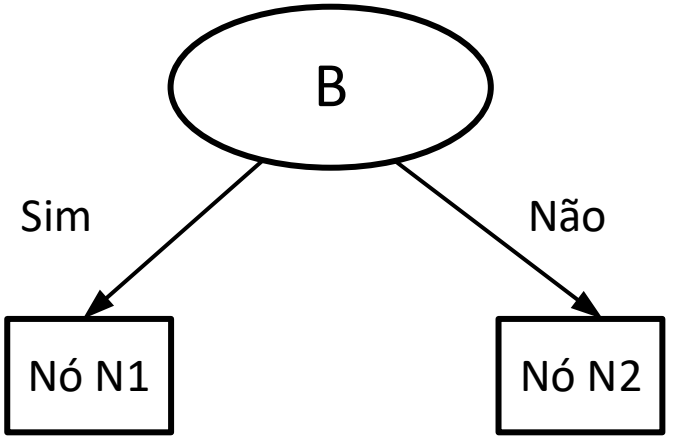
Separação de Atributos Binários

- Exemplos [1]:

	Pai
C0	6
C1	6
Gini=0.5	



	N1	N2
C0	4	2
C1	3	3
Gini	0.4898	0.480
Gini=0.486		

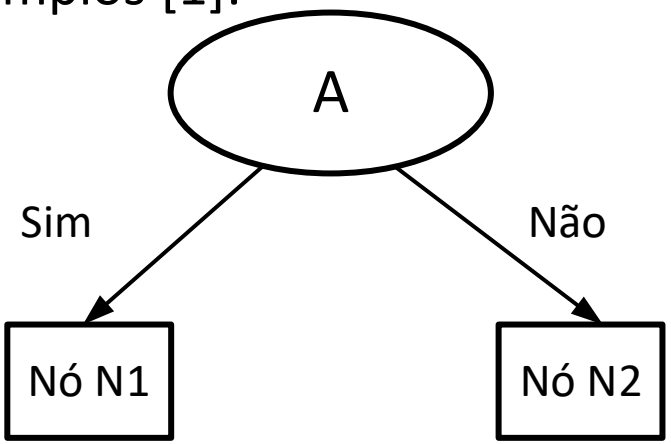


	N1	N2
C0	1	5
C1	4	2
Gini	0.320	0.482
Gini=0.371		

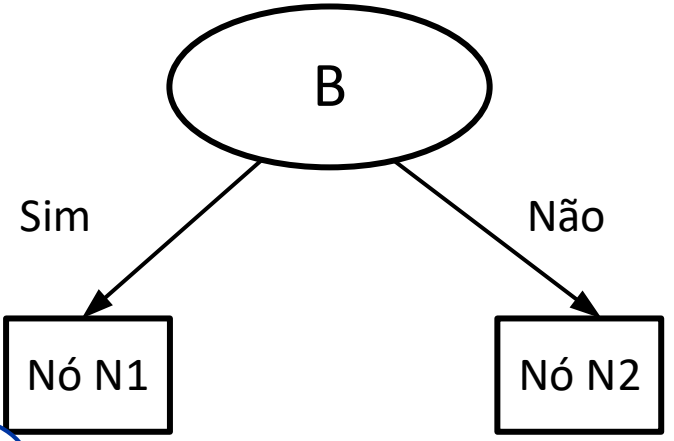
Como os valores de Gini para a separação com base no atributo B é inferior ao do A, o B é o escolhido.

Separação de Atributos Binários

- Exemplos [1]:



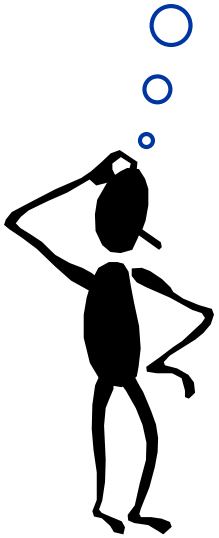
	Pai
C0	6
C1	6
Gini=0.5	



	N1	N2
C0	4	2
C1	3	3
Gini	0.4898	0.480
Gini=0.486		

- Δ_A= 0.5 – 0.486 = 0.014

Menor Gini, maior Ganho!

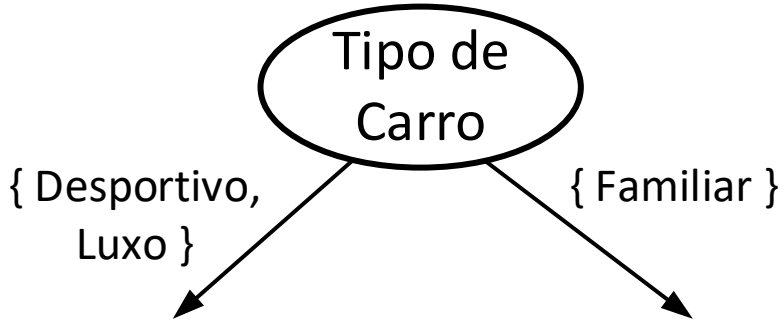


	N1	N2
C0	1	5
C1	4	2
Gini	0.320	0.482
Gini=0.371		

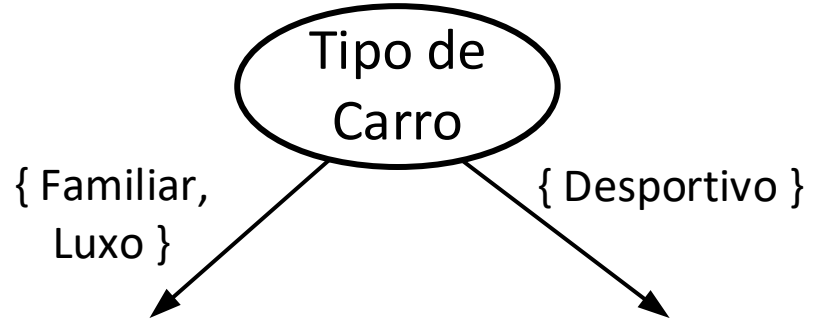
- Δ_A= 0.5 – 0.371 = 0.129

Separação de Atributos Nominais

- Exemplos [1]: **Binária**



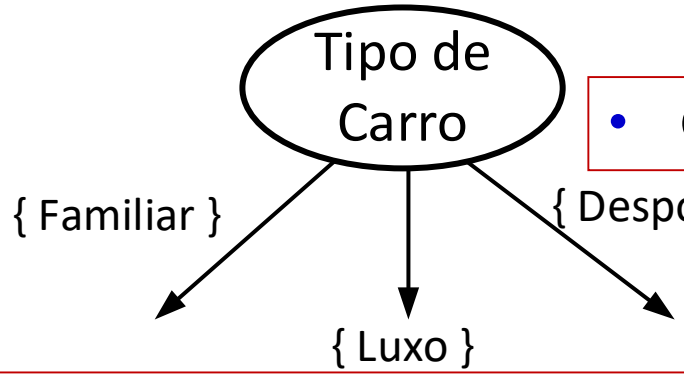
	Tipo de Carro	
	{ Desportivo, Luxo }	{ Familiar }
C0	9	1
C1	7	3
Gini	0.492	0.375
Gini=0.468		



	Tipo de Carro	
	{ Familiar, Luxo }	{Desportivo}
C0	8	2
C1	0	10
Gini	0	0.277
Gini=0.167		

Separação de Atributos Nominais

- Exemplos [1]: **Multiway**



- Gini ({Desportivo}) = 1 - (1/4)^2 - (3/4)^2 = 0.375
- Gini ({Luxo}) = 1 - (8/8)^2 - (0/8)^2 = 0
- Gini ({Familiar}) = 1 - (1/8)^2 - (7/8)^2 = 0.219

	Tipo de Carro		
	{ Desportivo}	{ Luxo }	{ Familiar }
C0	1	8	1
C1	3	0	7
Gini	0.375	0	0.219
Gini=0.163			

- Gini = (4/20)*0.375 + (8/20)*0.219 = 0.163

Exemplo

- *Determine o melhor atributo para efetuar a separação numa árvore de decisão para classificar se um animal é um mamífero i) com base no critério Gini.*

Nome	Temperatura Corpo	Dá a luz	Quadrupede	Hiberna	Classe (Mamífero?)
Porco Espinho	Quente	Sim	Sim	Sim	Sim
Gato	Quente	Sim	Sim	Não	Sim
Morcego	Quente	Sim	Não	Sim	Não
Baleia	Quente	Sim	Não	Não	Não
Salamandra	Frio	Não	Sim	Sim	Não
Dragão de Comodo	Frio	Não	Sim	Não	Não
Pitão	Frio	Não	Não	Sim	Não
Salmão	Frio	Não	Não	Não	Não
Águia	Quente	Não	Não	Não	Não
Peixe Guppy	Frio	Sim	Não	Não	Não

Exemplo

Nome	Temperatura Corpo	Dá a luz	Quadrupede	Hiberna	Classe (Mamífero?)
Porco Espinho	Quente	Sim	Sim	Sim	Sim
Gato	Quente	Sim	Sim	Não	Sim
Morcego	Quente	Sim	Não	Sim	Não
Baleia	Quente	Sim	Não	Não	Não
Salamandra	Frio	Não	Sim	Sim	Não
Dragão de Comodo	Frio	Não	Sim	Não	Não
Pitão	Frio	Não	Não	Sim	Não
Salmão	Frio	Não	Não	Não	Não
Águia	Quente	Não	Não	Não	Não
Peixe Guppy	Frio	Sim	Não	Não	Não

$$Gini\ (t) = 1 - \sum_{i=0}^{c-1} (p_i)^2$$

Sim (+): 2 amostras

Não (-): 8 amostras

✓ S= [2+,8-]

- $Gini_{Inicial}$ = 1 - (2/10)^2 - (8/10)^2 = 0.32

Exemplo

Nome	Temperatura Corpo	Classe (Mamífero?)
Porco Espinho	Quente	Sim
Gato	Quente	Sim
Morcego	Quente	Não
Baleia	Quente	Não
Salamandra	Frio	Não
Dragão de Comodo	Frio	Não
Pitão	Frio	Não
Salmão	Frio	Não
Águia	Quente	Não
Peixe Guppy	Frio	Não

Quente – 5 amostras

Frio - 5 amostras

Total: 10 amostras

Separar com base no atributo "Temperatura do Corpo"?

$$Gini\ (t) = 1 - \sum_{i=0}^{c-1} (p_i)^2$$

- $S_{Temp=Quente}$= [2+,3-]
- $S_{Temp=Frio}$= [0+,5-]

- $Gini_{Temp=Quente}$ = 1 - (2/5)^2 - (3/5)^2 = 0.48
- $Gini_{Temp=Frio}$ = 1 - (0/5)^2 - (5/5)^2 = 0
- Média ponderada = (5/10)*0.48 + (5/10)*0 = 0.24

- $Gini_{Temp}$= 0.32 – 0.24 = 0.08

Exemplo

Nome	Dá à luz	Classe (Mamífero?)
Porco Espinho	Sim	Sim
Gato	Sim	Sim
Morcego	Sim	Não
Baleia	Sim	Não
Salamandra	Não	Não
Dragão de Comodo	Não	Não
Pitão	Não	Não
Salmão	Não	Não
Águia	Não	Não
Peixe Guppy	Sim	Não

Sim – 5 amostras

Não - 5 amostras

Separar com base no atributo "Dá à luz"?

$$Gini\ (t) = 1 - \sum_{i=0}^{c-1} (p_i)^2$$

- $S_{Dá\_à\_luz=Sim}$= [2+,3-]
- $S_{Dá\_à\_luz=Não}$= [0+,5-]

- $Gini_{Dá\_à\_luz=Sim}$ = 1 - (2/5)^2 - (3/5)^2 = 0.48
- $Gini_{Dá\_à\_luz=Não}$ = 1 - (0/5)^2 - (5/5)^2 = 0
- Média ponderada= (5/10)*0.48 + (5/10)*0 = 0.24

- $Gini_{Dá\_à\_luz}$= 0.32 – 0.24 = 0.08

Exemplo

Nome	Quadrupede	Classe (Mamífero?)
Porco Espinho	Sim	Sim
Gato	Sim	Sim
Morcego	Não	Não
Baleia	Não	Não
Salamandra	Sim	Não
Dragão de Comodo	Sim	Não
Pitão	Não	Não
Salmão	Não	Não
Águia	Não	Não
Peixe Guppy	Não	Não

Sim – 4 amostras

Não - 6 amostras

Separar com base no atributo "Quadrupede"?

$$Gini\ (t) = 1 - \sum_{i=0}^{c-1} (p_i)^2$$

- $S_{Quadrupede=Sim}$= [2+,2-]
- $S_{Quadrúpede=Não}$= [0+,6-]

- $Gini_{Quad=Sim}$ = 1 - (2/4)^2 - (2/4)^2 = 0.5
- $Gini_{Quad=Não}$ = 1 - (0/6)^2 - (6/6)^2 = 0
- Média ponderada= (4/10)*0.5 + (6/10)*0 = 0.2
- $Gini_{Quad}$= 0.32 – 0.2 = 0.12

Exemplo

Nome	Hiberna	Classe (Mamífero?)
Porco Espinho	Sim	Sim
Gato	Não	Sim
Morcego	Sim	Não
Baleia	Não	Não
Salamandra	Sim	Não
Dragão de Comodo	Não	Não
Pitão	Sim	Não
Salmão	Não	Não
Águia	Não	Não
Peixe Guppy	Não	Não

Sim – 4 amostras

Não - 6 amostras

Separar com base no atributo "Hiberna"?

$$Gini\ (t) = 1 - \sum_{i=0}^{c-1} (p_i)^2$$

- $S_{Hiberna=Sim}$= [1+,3-]
- $S_{Hiberna=Não}$= [1+,5-]

- $Gini_{Hiberna=Sim}$ = 1 - (1/4)^2 - (3/4)^2 = 0.375
- $Gini_{Hiberna=Não}$ = 1 - (1/6)^2 - (5/6)^2 = 0.278
- Média ponderada =

(4/10)*0.375 + (6/10)*0.278 = 0.3168

- $Gini_{Hiberna}$= 0.32 – 0.3168 = 0.032

Exemplo

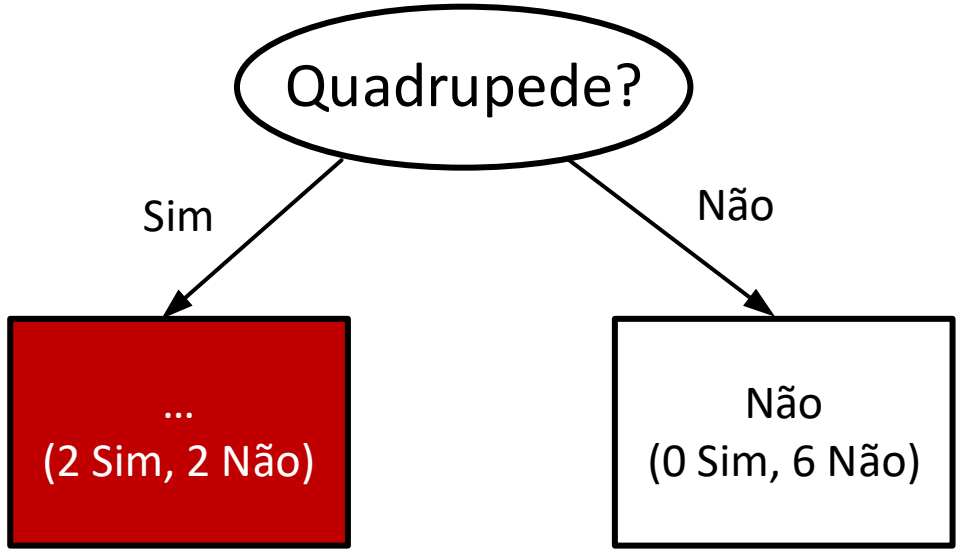
➢ *Resumindo:*

- $Gini_{Temp} = 0.32 - 0.24 = 0.08$
- $Gini_{dá\_à\_luz} = 0.32 - 0.24 = 0.08$
- $Gini_{Quad} = 0.32 - 0.2 = 0.12$
- $Gini_{Hiberna} = 0.32 - 0.3168 = 0.032$

✓ *Como o atributo Quadrupede tem o* ***maior Ganho*** *é o selecionado para efetuar a separação.*

Exemplo

Nome	*Quadrupede*	*Classe (Mamífero?)*
Porco Espinho	Sim	Sim
Gato	Sim	Sim
Morcego	Não	Não
Baleia	Não	Não
Salamandra	Sim	Não
Dragão de Comodo	Sim	Não
Pitão	Não	Não
Salmão	Não	Não
Águia	Não	Não
Peixe Guppy	Não	Não



- Gini = 1 - (2/4)^2 - (2/4)^2 = 0.5

✓ *Que atributo utilizar para continuar a separação?*

Exemplo

Nome	Dá a luz	Quadrupede	Classe (Mamífero?)
Porco Espinho	Sim	Sim	Sim
Gato	Sim	Sim	Sim
Morcego	Sim	Não	Não
Baleia	Sim	Não	Não
Salamandra	Não	Sim	Não
Dragão de Comodo	Não	Sim	Não
Pitão	Não	Não	Não
Salmão	Não	Não	Não
Águia	Não	Não	Não
Peixe Guppy	Sim	Não	Não

Sim – 2 amostras

Não - 2 amostras

Nova Separação com base no atributo "Dá à luz"?

- $S_{Dá\_à\_luz=Sim}$= [2+,0-]
- $S_{Dá\_à\_luz=Não}$= [0+,2-]

- $Gini_{dá\_à\_luz=Sim}$ =

1 - (2/2)^2 - (0/2)^2 = 0

- $Gini_{dá\_à\_luz=Não}$ = 0
- Média ponderada=0

- $Gini_{dá\_à\_luz}$= 0.5 – 0 = 0.5

Exemplo

Nome	Quadrupede	Hiberna	Classe (Mamífero?)
Porco Espinho	Sim	Sim	Sim
Gato	Sim	Não	Sim
Morcego	Não	Sim	Não
Baleia	Não	Não	Não
Salamandra	Sim	Sim	Não
Dragão de Comodo	Sim	Não	Não
Pitão	Não	Sim	Não
Salmão	Não	Não	Não
Águia	Não	Não	Não
Peixe Guppy	Não	Não	Não

Sim – 2 amostras

Não - 2 amostras

Nova Separação com base no atributo "Hiberna"?

- $S_{Hiberna=Sim}$= [1+,1-]
- $S_{Hiberna=Não}$= [1+,1-]

- $Gini_{Hiberna=Sim}$ =

1 - (1/2)^2 - (1/2)^2 = 0.5

- $Gini_{Hiberna=Não}$ =

1 - (1/2)^2 - (1/2)^2 = 0.5

- Média ponderada =

(2/4)*0.5 + (2/4)*0.5 = 0.5

- $Gini_{Hiberna}$= 0.5 – 0.5 = 0

Exemplo

Nome	Temperatura Corpo	Quadrupede	Classe (Mamífero?)
Porco Espinho	Quente	Sim	Sim
Gato	Quente	Sim	Sim
Morcego	Quente	Não	Não
Baleia	Quente	Não	Não
Salamandra	Frio	Sim	Não
Dragão de Comodo	Frio	Sim	Não
Pitão	Frio	Não	Não
Salmão	Frio	Não	Não
Águia	Quente	Não	Não
Peixe Guppy	Frio	Não	Não

Quente – 2 amostras

Frio - 2 amostras

Nova Separação com base no atributo "Temperatura do Corpo"?

- $S_{Temp=Quente}$= [2+,0-]
- $S_{Temp=Frio}$= [0+,2-]

- $Gini_{Temp=Quente}$ =

1 - (2/2)^2 - (0/2)^2 = 0

- $Gini_{Temp=Frio}$ =

1 - (0/2)^2 - (2/2)^2 = 0

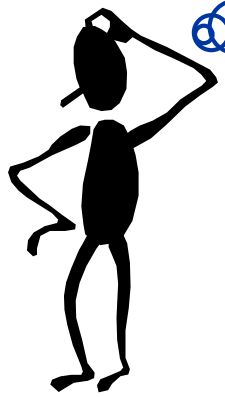
- Média ponderada = 0

- $Gini_{Temp}$= 0.5 – 0 = 0.5

Exemplo

- *Resumindo:*

- $Gini_{dá\_à\_luz} = 0.5 - 0 = 0.5$
- $Gini_{Hiberna} = 0.5 - 0.5 = 0$
- $Gini_{Temp} = 0.5 - 0 = 0.5$



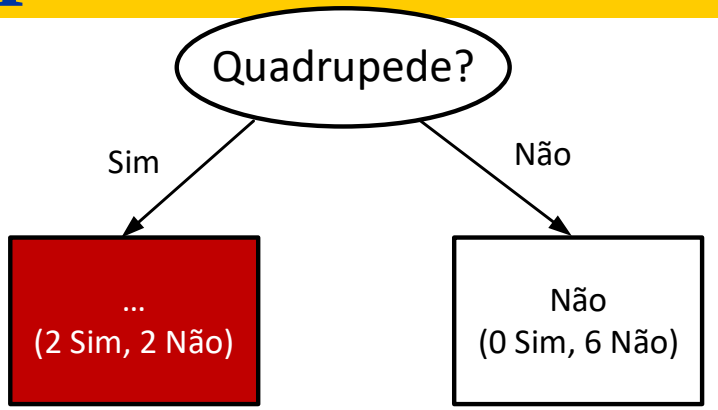
Qual o atributo que tem melhor ganho Gini?

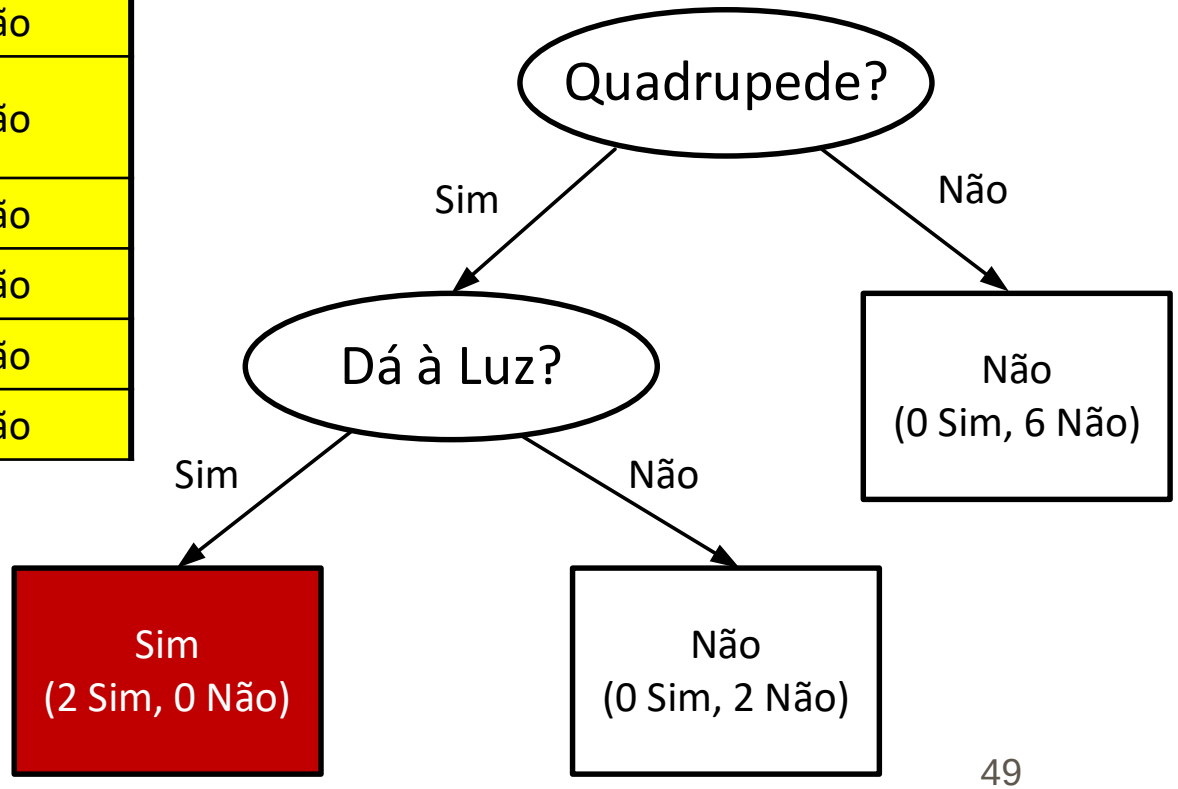
✓ *Como tanto o atributo "dá à luz" como o atributo "Temperatura" têm ambos o maior Ganho (máximo) ambas resultam numa separação ideal.*

✓ *Vamos aplicar a separação usando o atributo "dá à luz".*

Exemplo

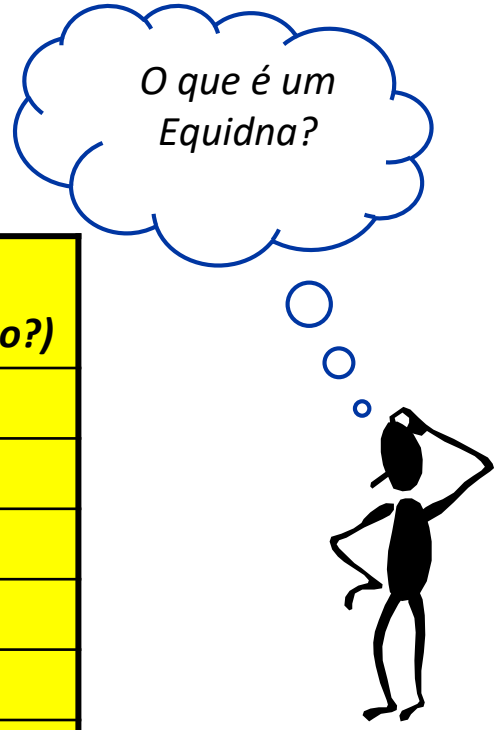
Nome	*Dá a luz*	*Quadrupede*	*Classe (Mamífero?)*
Porco Espinho	Sim	Sim	Sim
Gato	Sim	Sim	Sim
Morcego	Sim	Não	Não
Baleia	Sim	Não	Não
Salamandra	Não	Sim	Não
Dragão de Comodo	Não	Sim	Não
Pitão	Não	Não	Não
Salmão	Não	Não	Não
Águia	Não	Não	Não
Peixe Guppy	Sim	Não	Não





Treino-Validação-Teste

✓ *Consideremos agora o seguinte conjunto de amostras novas, não consideradas no treino, como conjunto de teste.*



Nome	*Temperatura Corpo*	*Dá a luz*	*Quadrupede*	*Hiberna*	*Classe (Mamífero?)*
Humano	Quente	Sim	Não	Não	Sim
Pombo	Quente	Não	Não	Não	Não
Elefante	Quente	Sim	Sim	Não	Sim
Tubarão	Fria	Sim	Não	Não	Não
Tartaruga	Fria	Não	Sim	Não	Não
Pinguim	Fria	Não	Não	Não	Não
Enguia	Fria	Não	Não	Não	Não
Golfinho	Quente	Sim	Não	Não	Sim
Equidna	Quente	Não	Não	Não	Sim
Lagarto	Frio	Não	Sim	Sim	Não

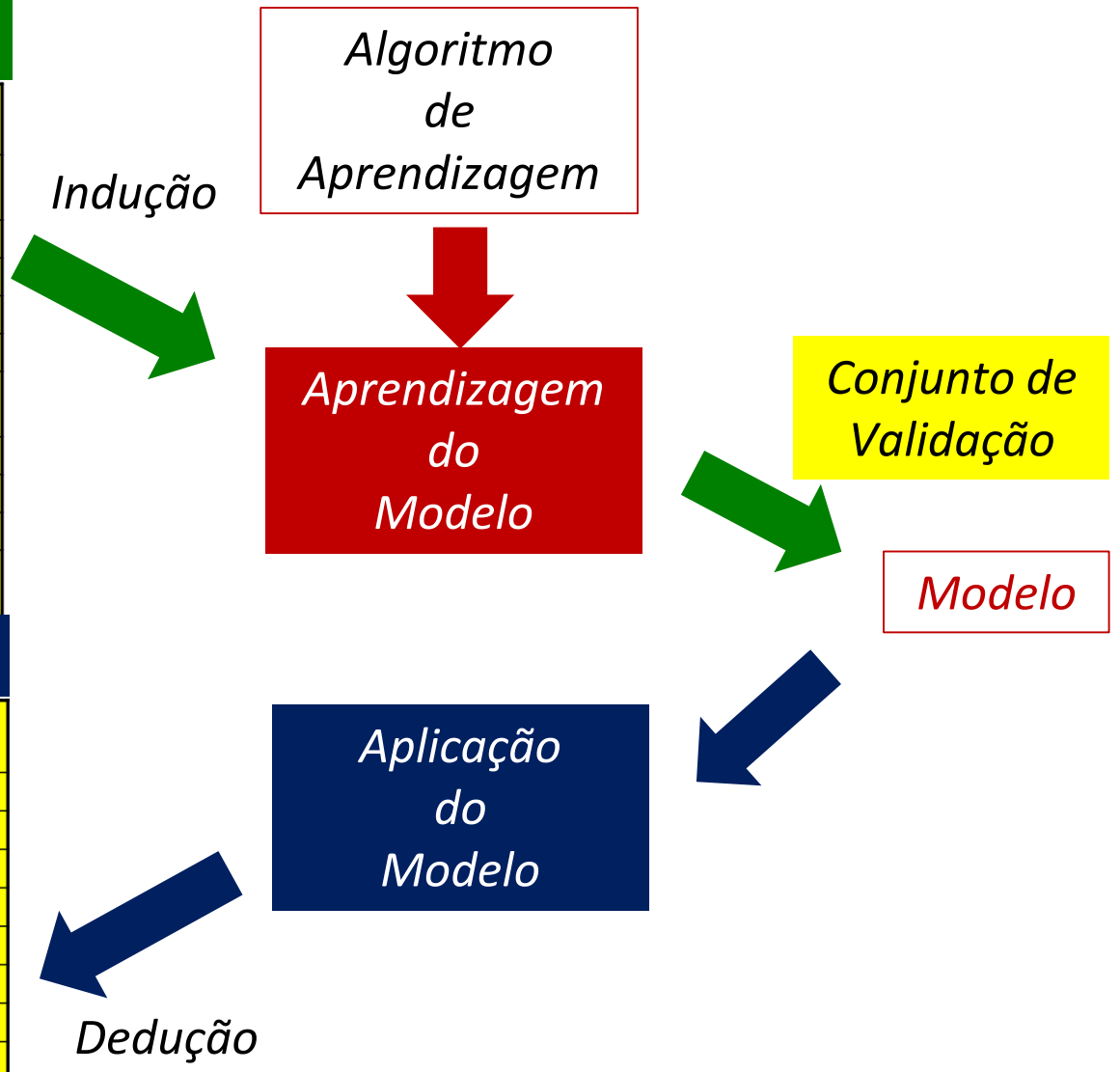
Treino-Validação-Teste

Conjunto de Treino

Nome	Temperatura Corpo	Dá a luz	Quadrupede	Hiberna	Classe (Mamífero?)
Porco Espinho	Quente	Sim	Sim	Sim	Sim
Gato	Quente	Sim	Sim	Não	Sim
Morcego	Quente	Sim	Não	Sim	Não
Baleia	Quente	Sim	Não	Não	Não
Salamandra	Frio	Não	Sim	Sim	Não
Dragão de Comodo	Frio	Não	Sim	Não	Não
Pitão	Frio	Não	Não	Sim	Não
Salmão	Frio	Não	Não	Não	Não
Águia	Quente	Não	Não	Não	Não
Peixe Guppy	Frio	Sim	Não	Não	Não

Conjunto de Teste

Nome	Temperatura Corpo	Dá a luz	Quadrupede	Hiberna	Classe (Mamífero?)
Humano	Quente	Sim	Não	Não	Sim
Pombo	Quente	Não	Não	Não	Não
Elefante	Quente	Sim	Sim	Não	Sim
Tubarão	Fria	Sim	Não	Não	Não
Tartaruga	Fria	Não	Sim	Não	Não
Pinguim	Fria	Não	Não	Não	Não
Enguia	Fria	Não	Não	Não	Não
Golfinho	Quente	Sim	Não	Não	Sim
Equidna	Quente	Não	Não	Não	Sim
Lagarto	Frio	Não	Sim	Sim	Não



Critérios de Avaliação do Modelo

- ✓ Como podemos avaliar a qualidade do modelo testando-o no conjunto de teste?

Taxa de Erro – *Error Rate*

$$= \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- ✓ Em que:
- n – número de amostras do conjunto de teste
- $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$
- $\hat{y}_i$ - classe prevista para a amostra i

$$I(y_i \neq \hat{y}_i) = \begin{cases} 1 \Leftarrow y_i \neq \hat{y}_i \\ 0 \Leftarrow y_i = \hat{y}_i \end{cases}$$

Overfitting do Modelo

- ✓ **Já vimos (RN)** que quando o modelo "memoriza" os dados de treino, temos o fenómeno de *overfitting*, no qual se pode observar:
  - **uma taxa de erro muito baixa no conjunto de treino;**
  - **uma taxa de erro muito elevada no conjunto de teste.**

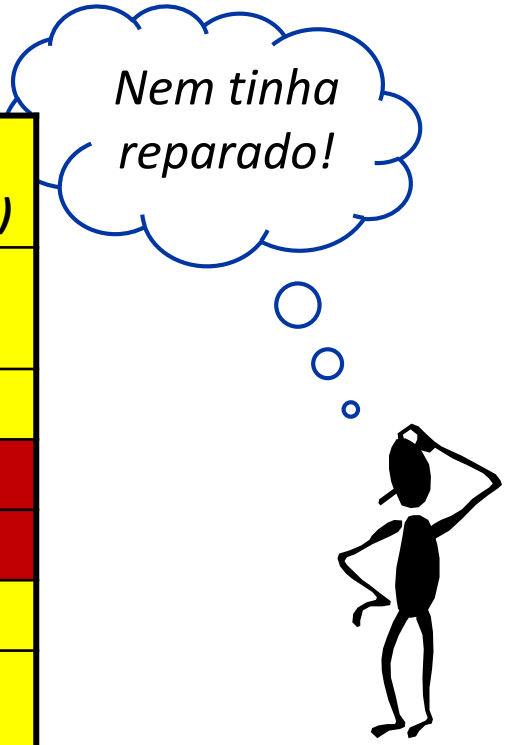
Tem de evitar *overfitting* para que o modelo tenha capacidade de generalização, i.e., consiga classificar bem amostras não utilizadas no conjunto de treino.

- ✓ **Erros de classificação do modelo podem ser provocados por dados errados utilizados no treino (amostras erradas ou com ruído).**

Overfitting do Modelo

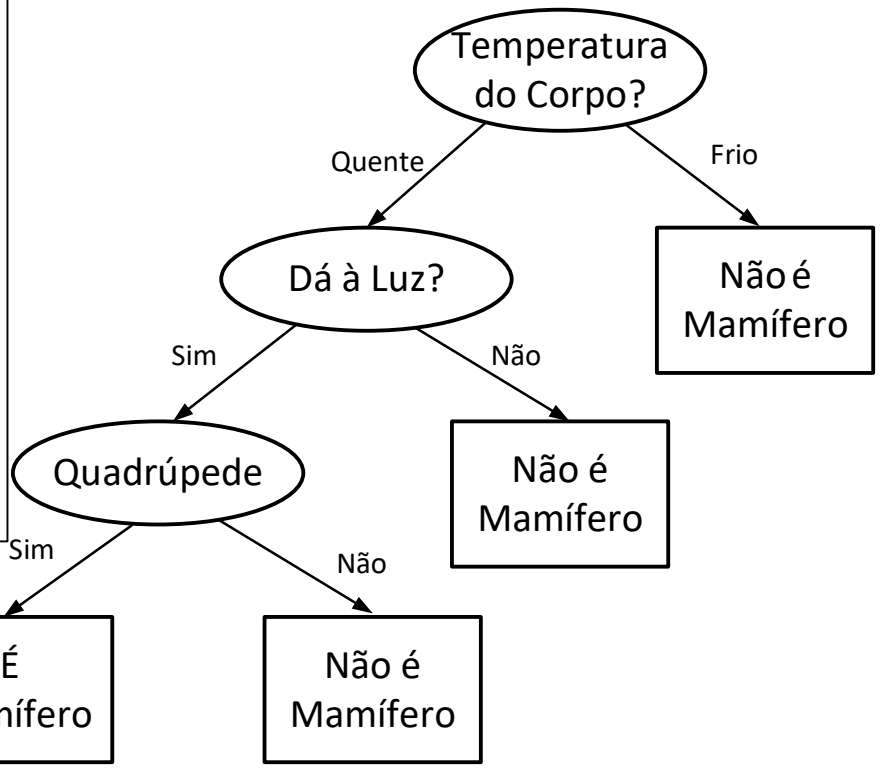
- ✓ No exemplo de classificação de um mamífero existem dois erros nos dados utilizados no treino:

Nome	*Temperatura Corpo*	*Dá a luz*	*Quadrupede*	*Hiberna*	*Classe (Mamífero?)*
Porco Espinho	Quente	Sim	Sim	Sim	Sim
Gato	Quente	Sim	Sim	Não	Sim
Morcego	Quente	Sim	Não	Sim	Não
Baleia	Quente	Sim	Não	Não	Não
Salamandra	Frio	Não	Sim	Sim	Não
Dragão de Comodo	Frio	Não	Sim	Não	Não
Pitão	Frio	Não	Não	Sim	Não
Salmão	Frio	Não	Não	Não	Não
Águia	Quente	Não	Não	Não	Não
Peixe Guppy	Frio	Sim	Não	Não	Não



Overfitting do Modelo

Nome	*Temperatura Corpo*	*Dá a luz*	*Quadrupede*	*Hiberna*	*Classe (Mamífero?)*
Porco Espinho	Quente	Sim	Sim	Sim	Sim
Gato	Quente	Sim	Sim	Não	Sim
Morcego	Quente	Sim	Não	Sim	Não
Baleia	Quente	Sim	Não	Não	Não
Salamandra	Frio	Não	Sim	Sim	Não
Dragão de Comodo	Frio	Não	Sim	Não	Não
Pitão	Frio	Não	Não	Sim	Não
Salmão	Frio	Não	Não	Não	Não
Águia	Quente	Não	Não	Não	Não
Peixe Guppy	Frio	Sim	Não	Não	Não



✓ Esta árvore ajusta-se perfeitamente ao conjunto de dados de treino.

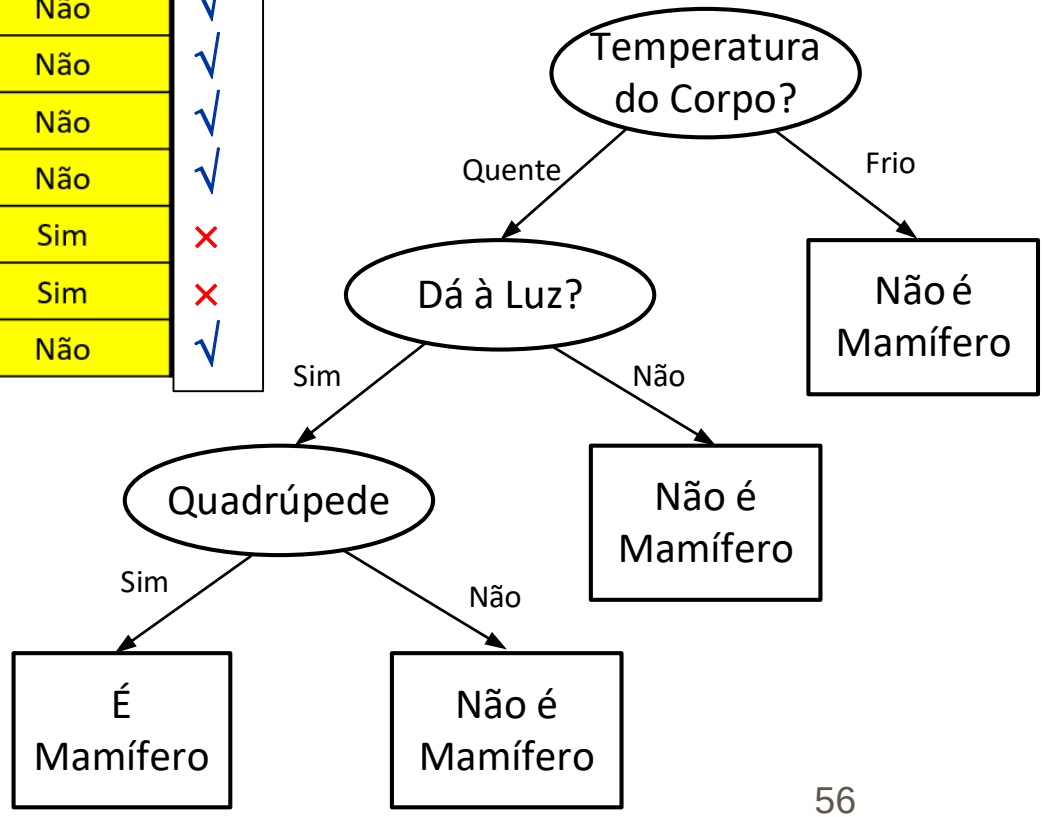
Overfitting do Modelo

✓ Como se comporta esta árvore [1] no conjunto de teste seguinte:

Nome	*Temperatura Corpo*	*Dá a luz*	*Quadrupede*	*Hiberna*	*Classe (Mamífero?)*	
Humano	Quente	Sim	Não	Não	Sim	×
Pombo	Quente	Não	Não	Não	Não	√
Elefante	Quente	Sim	Sim	Não	Sim	√
Tubarão	Fria	Sim	Não	Não	Não	√
Tartaruga	Fria	Não	Sim	Não	Não	√
Pinguim	Fria	Não	Não	Não	Não	√
Enguia	Fria	Não	Não	Não	Não	√
Golfinho	Quente	Sim	Não	Não	Sim	×
Equidna	Quente	Não	Não	Não	Sim	×
Lagarto	Frio	Não	Sim	Sim	Não	√

Taxa de erro no teste=30%

O Equidna é um caso excecional, no sentido em que contraria a maioria dos casos de treino.



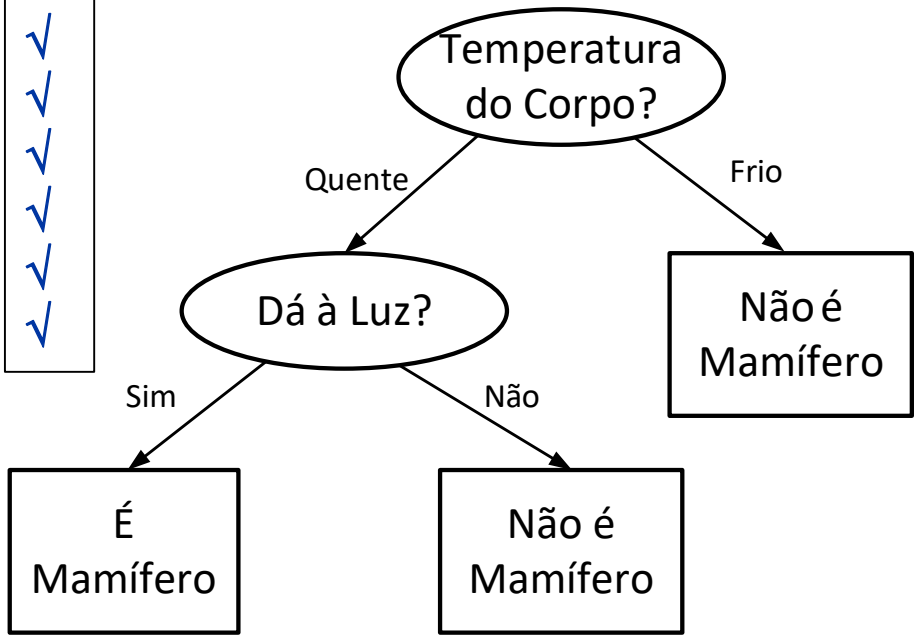
Overfitting do Modelo

✓ Vamos agora considerar outra árvore mais simples [1]:

Nome	*Temperatura Corpo*	*Dá a luz*	*Quadrupede*	*Hiberna*	*Classe (Mamífero?)*
Porco Espinho	Quente	Sim	Sim	Sim	Sim
Gato	Quente	Sim	Sim	Não	Sim
Morcego	Quente	Sim	Não	Sim	Não
Baleia	Quente	Sim	Não	Não	Não
Salamandra	Frio	Não	Sim	Sim	Não
Dragão de Comodo	Frio	Não	Sim	Não	Não
Pitão	Frio	Não	Não	Sim	Não
Salmão	Frio	Não	Não	Não	Não
Águia	Quente	Não	Não	Não	Não
Peixe Guppy	Frio	Sim	Não	Não	Não

√ √ × × √ √ √ √ √ √

Taxa de erro de treino=20%



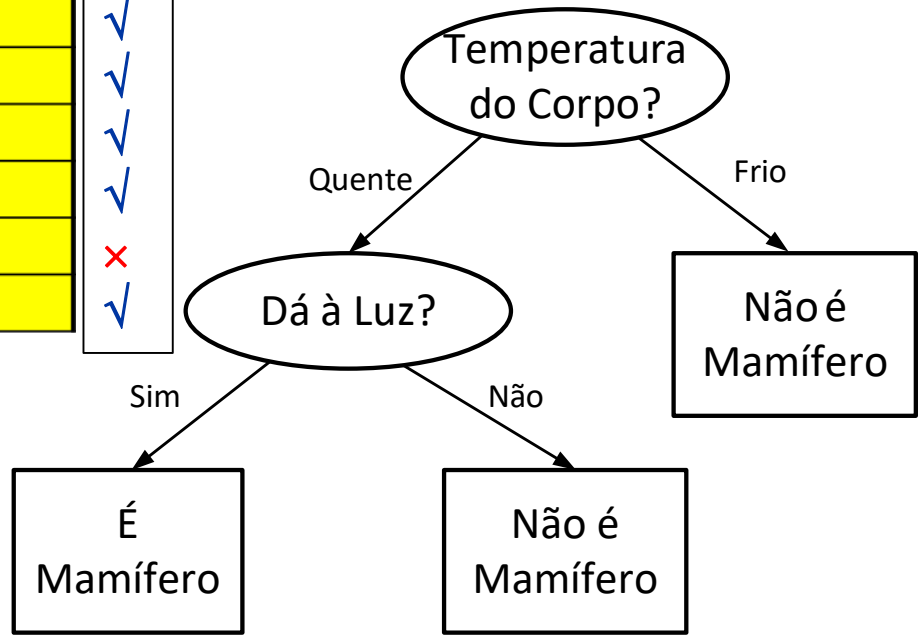
Overfitting do Modelo

✓ Vamos agora considerar outra árvore mais simples [1]:

Nome	*Temperatura Corpo*	*Dá a luz*	*Quadrupede*	*Hiberna*	*Classe (Mamífero?)*
Humano	Quente	Sim	Não	Não	Sim
Pombo	Quente	Não	Não	Não	Não
Elefante	Quente	Sim	Sim	Não	Sim
Tubarão	Fria	Sim	Não	Não	Não
Tartaruga	Fria	Não	Sim	Não	Não
Pinguim	Fria	Não	Não	Não	Não
Enguia	Fria	Não	Não	Não	Não
Golfinho	Quente	Sim	Não	Não	Sim
Equidna	Quente	Não	Não	Não	Sim
Lagarto	Frio	Não	Sim	Sim	Não

√ √ √ √ √ √ √ √ √ × √

Taxa de erro de teste=10%



Notar que mesmo tendo uma taxa erro de treino maior o segundo modelo tem uma taxa de teste menor. Tem menos overfitting.

Exemplo: Ganho de Informação

- ✓ Considere o exemplo seguinte em a questão é " O tempo está para jogar ténis?
- ✓ Pretende-se construir uma árvore de decisão com base na Entropia.

Dia	Previsão	Temperatura	Humidade	Vento	Jogar Ténis ?
1	Sol	Quente	Alta	Leve	Não (-)
2	Sol	Quente	Alta	Forte	Não (-)
3	Nuvens	Quente	Alta	Leve	Sim (+)
4	Chuva	Média	Alta	Leve	Sim (+)
5	Chuva	Frio	Normal	Leve	Sim (+)
6	Chuva	Frio	Normal	Forte	Não (-)
7	Nuvens	Frio	Normal	Forte	Sim (+)
8	Sol	Média	Alta	Leve	Não (-)
9	Sol	Frio	Normal	Leve	Sim (+)
10	Chuva	Média	Normal	Leve	Sim (+)
11	Sol	Média	Normal	Forte	Sim (+)
12	Nuvens	Média	Alta	Forte	Sim (+)
13	Nuvens	Quente	Normal	Leve	Sim (+)
14	Chuva	Média	Alta	Forte	Não (-)

Exemplo: Ganho de Informação (cont.)

Dia	Previsão	Temperatura	Humidade	Vento	Jogar Ténis ?
1	Sol	Quente	Alta	Leve	Não (-)
2	Sol	Quente	Alta	Forte	Não (-)
3	Nuvens	Quente	Alta	Leve	Sim (+)
4	Chuva	Média	Alta	Leve	Sim (+)
5	Chuva	Frio	Normal	Leve	Sim (+)
6	Chuva	Frio	Normal	Forte	Não (-)
7	Nuvens	Frio	Normal	Forte	Sim (+)
8	Sol	Média	Alta	Leve	Não (-)
9	Sol	Frio	Normal	Leve	Sim (+)
10	Chuva	Média	Normal	Leve	Sim (+)
11	Sol	Média	Normal	Forte	Sim (+)
12	Nuvens	Média	Alta	Forte	Sim (+)
13	Nuvens	Quente	Normal	Leve	Sim (+)
14	Chuva	Média	Alta	Forte	Não (-)

$$E(S) = - \ p_+ \log_2 p_+ - \ p_- \log_2 p_-$$

- ✓ S= [9+,5-]
- ✓ Qual o melhor atributo para iniciar a separação?
- ✓ Vamos começar por testar a Previsão.

- E(S) = - (9/14)*log2(9/14) - (5/14)*log2(5/14) = 0.940

Exemplo: Ganho de Informação (cont.)

Dia	Previsão	Jogar Ténis ?
1	Sol	Não (-)
2	Sol	Não (-)
3	Nuvens	Sim (+)
4	Chuva	Sim (+)
5	Chuva	Sim (+)
6	Chuva	Não (-)
7	Nuvens	Sim (+)
8	Sol	Não (-)
9	Sol	Sim (+)
10	Chuva	Sim (+)
11	Sol	Sim (+)
12	Nuvens	Sim (+)
13	Nuvens	Sim (+)
14	Chuva	Não (-)

$$E(S) = - \ p_+ \log_2 p_+ - \ p_- \log_2 p_-$$

✓ **Previsão?**

- S_{Sol}= [2+,3-]
- $E(S_{Sol})$ = - (2/5)*log2(2/5) - (3/4)*log2(3/5) = 0.971
- S_{Nuvens}= [4+,0-]
- $E(S_{Nuvens})$ = 0
- S_{Chuva}= [3+,2-]
- $E(S_{Chuva})$ = - (3/5)*log2(3/5) - (3/5)*log2(2/5) = 0.971
- Média Pond.= (5/14)*0.971 + 0 + (5/14)*0.971 = 0.694

- $GI_{Previsão}$= 0.940-0.694=0.246

Exemplo: Ganho de Informação (cont.)

$$E(S) = -\ p_+ \log_2 p_+ -\ p_- \log_2 p_-$$

Dia	Temperatura	Jogar Ténis ?
1	Quente	Não (-)
2	Quente	Não (-)
3	Quente	Sim (+)
4	Média	Sim (+)
5	Frio	Sim (+)
6	Frio	Não (-)
7	Frio	Sim (+)
8	Média	Não (-)
9	Frio	Sim (+)
10	Média	Sim (+)
11	Média	Sim (+)
12	Média	Sim (+)
13	Quente	Sim (+)
14	Média	Não (-)

✓ **Temperatura?**

- S_{Quente}= [2+,2-]
- $E(S_{Sol})$ = - (2/4)*log2(2/4) - (2/4)*log2(2/4) = 1
- $S_{Média}$= [4+,2-]
- $E(S_{Média})$ = - (4/6)*log2(4/6) - (2/6)*log2(2/6) = 0.918
- S_{Frio}= [3+,1-]
- $E(S_{Frio})$ = - (3/4)*log2(3/4) - (1/4)*log2(1/4) = 0.811
- Média Pond.= (4/14)*1 + (6/14)*0.918 + (4/14)*0.811 = 0.911

- $GI_{Temperatura}$= 0.940-0.911=0.029

Exemplo: Ganho de Informação (cont.)

$$E(S) = -\ p_+ \log_2 p_+ -\ p_- \log_2 p_-$$

Dia	Humidade	Jogar Ténis ?
1	Alta	Não (-)
2	Alta	Não (-)
3	Alta	Sim (+)
4	Alta	Sim (+)
5	Normal	Sim (+)
6	Normal	Não (-)
7	Normal	Sim (+)
8	Alta	Não (-)
9	Normal	Sim (+)
10	Normal	Sim (+)
11	Normal	Sim (+)
12	Alta	Sim (+)
13	Normal	Sim (+)
14	Alta	Não (-)

✓ **Humidade?**

- S_{Alta}= [3+,4-]
- $E(S_{Sol})$ = - (3/7)*log2(3/7) - (4/7)*log2(4/7) =0.985
- S_{Normal}= [6+,1-]
- $E(S_{Normal})$ = - (6/7)*log2(6/7) - (1/7)*log2(1/7) = 0.592
- Média Pond.= (7/14)*0.985 + (7/14)*0.592 = 0.789
- $GI_{Humidade}$= 0.940-0.789=0.151

Exemplo: Ganho de Informação (cont.)

$$E(S) = -\ p_+ \log_2 p_+ -\ p_- \log_2 p_-$$

Dia	Vento	Jogar Ténis ?
1	Leve	Não (-)
2	Forte	Não (-)
3	Leve	Sim (+)
4	Leve	Sim (+)
5	Leve	Sim (+)
6	Forte	Não (-)
7	Forte	Sim (+)
8	Leve	Não (-)
9	Leve	Sim (+)
10	Leve	Sim (+)
11	Forte	Sim (+)
12	Forte	Sim (+)
13	Leve	Sim (+)
14	Forte	Não (-)

✓ **Vento?**

- S_{Leve}= [6+,2-]
- $E(S_{Leve})$ = - (6/8)*log2(6/8) - (2/8)*log2(2/8) = 0.811
- S_{Forte}= [3+,3-]
- $E(S_{Forte})$ = - (3/6)*log2(3/6) - (3/6)*log2(3/6) = 1
- Média Pond.= (8/14)*0.811 + (6/14)*1 = 0.892

- GI_{Vento}= 0.940-0.892=0.048

Exemplo: Ganho de Informação (cont.)

Dia	Previsão	Temperatura	Humidade	Vento	Jogar Ténis ?
1	Sol	Quente	Alta	Leve	Não (-)
2	Sol	Quente	Alta	Forte	Não (-)
3	Nuvens	Quente	Alta	Leve	Sim (+)
4	Chuva	Média	Alta	Leve	Sim (+)
5	Chuva	Frio	Normal	Leve	Sim (+)
6	Chuva	Frio	Normal	Forte	Não (-)
7	Nuvens	Frio	Normal	Forte	Sim (+)
8	Sol	Média	Alta	Leve	Não (-)
9	Sol	Frio	Normal	Leve	Sim (+)
10	Chuva	Média	Normal	Leve	Sim (+)
11	Sol	Média	Normal	Forte	Sim (+)
12	Nuvens	Média	Alta	Forte	Sim (+)
13	Nuvens	Quente	Normal	Leve	Sim (+)
14	Chuva	Média	Alta	Forte	Não (-)

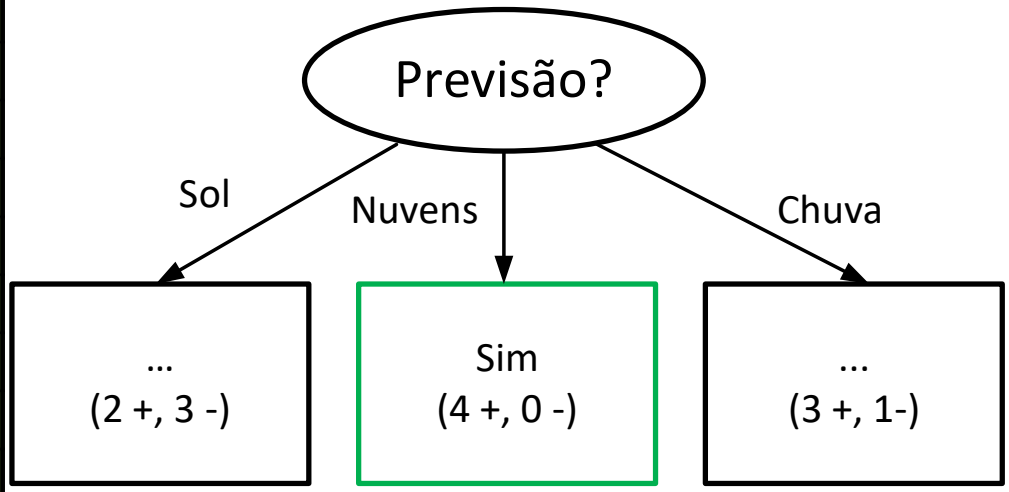
✓ Qual o melhor atributo para iniciar a separação?

- $Ganho_{Previsão} = 0.246$
- $Ganho_{Temperatura} = 0.029$
- $Ganho_{Humidade} = 0.151$
- $Ganho_{Vento} = 0.048$

Exemplo: Ganho de Informação (cont.)

Dia	Previsão	Temperatura	Humidade	Vento	Jogar Ténis ?
1	Sol	Quente	Alta	Leve	Não (-)
2	Sol	Quente	Alta	Forte	Não (-)
3	Nuvens	Quente	Alta	Leve	Sim (+)
4	Chuva	Média	Alta	Leve	Sim (+)
5	Chuva	Frio	Normal	Leve	Sim (+)
6	Chuva	Frio	Normal	Forte	Não (-)
7	Nuvens	Frio	Normal	Forte	Sim (+)
8	Sol	Média	Alta	Leve	Não (-)
9	Sol	Frio	Normal	Leve	Sim (+)
10	Chuva	Média	Normal	Leve	Sim (+)
11	Sol	Média	Normal	Forte	Sim (+)
12	Nuvens	Média	Alta	Forte	Sim (+)
13	Nuvens	Quente	Normal	Leve	Sim (+)
14	Chuva	Média	Alta	Forte	Não (-)

- $Ganho_{Previsão} = 0.246$



✓ Qual o melhor atributo para iniciar a separação a partir de **Previsão=Sol**?

- $E(S_{Sol}) = - (2/5)*\log2(2/5) - (3/5)*\log2(3/5) = 0.971$

Exemplo: Ganho de Informação (cont.)

Dia	Previsão	Temperatura	Jogar Ténis ?
1	Sol	Quente	Não (-)
2	Sol	Quente	Não (-)
3	Nuvens	Quente	Sim (+)
4	Chuva	Média	Sim (+)
5	Chuva	Frio	Sim (+)
6	Chuva	Frio	Não (-)
7	Nuvens	Frio	Sim (+)
8	Sol	Média	Não (-)
9	Sol	Frio	Sim (+)
10	Chuva	Média	Sim (+)
11	Sol	Média	Sim (+)
12	Nuvens	Média	Sim (+)
13	Nuvens	Quente	Sim (+)
14	Chuva	Média	Não (-)

- S_{Sol}= [2+,3-]

✓ Vamos começar pelo atributo **Temperatura**.

- $S_{Temp=Quente}$= [0+,2-]
- $S_{Temp=Média}$= [1+,1-]
- $S_{Temp=Frio}$= [1+,0-]
- $E(S_{Sol},Temp=Quente) = - (0/2)*log2(0/2) - (2/2)*log2(2/2) = 0$
- $E(S_{Sol},Temp=Média) = - (1/2)*log2(1/2) - (1/2)*log2(1/2) = 1$
- $E(S_{Sol},Temp=Frio) = - (1/1)*log2(1/1) - (0/1)*log2(0/1) = 0$
- Média Pond.= (2/5)*0 + (2/5)*1+(1/5)*0 = 0.400
- $GI(S_{Sol},Temp) = 0.971 – 0.400=0.571$

Exemplo: Ganho de Informação (cont.)

Dia	Previsão	Vento	Jogar Ténis ?
1	Sol	Leve	Não (-)
2	Sol	Forte	Não (-)
3	Nuvens	Leve	Sim (+)
4	Chuva	Leve	Sim (+)
5	Chuva	Leve	Sim (+)
6	Chuva	Forte	Não (-)
7	Nuvens	Forte	Sim (+)
8	Sol	Leve	Não (-)
9	Sol	Leve	Sim (+)
10	Chuva	Leve	Sim (+)
11	Sol	Forte	Sim (+)
12	Nuvens	Forte	Sim (+)
13	Nuvens	Leve	Sim (+)
14	Chuva	Forte	Não (-)

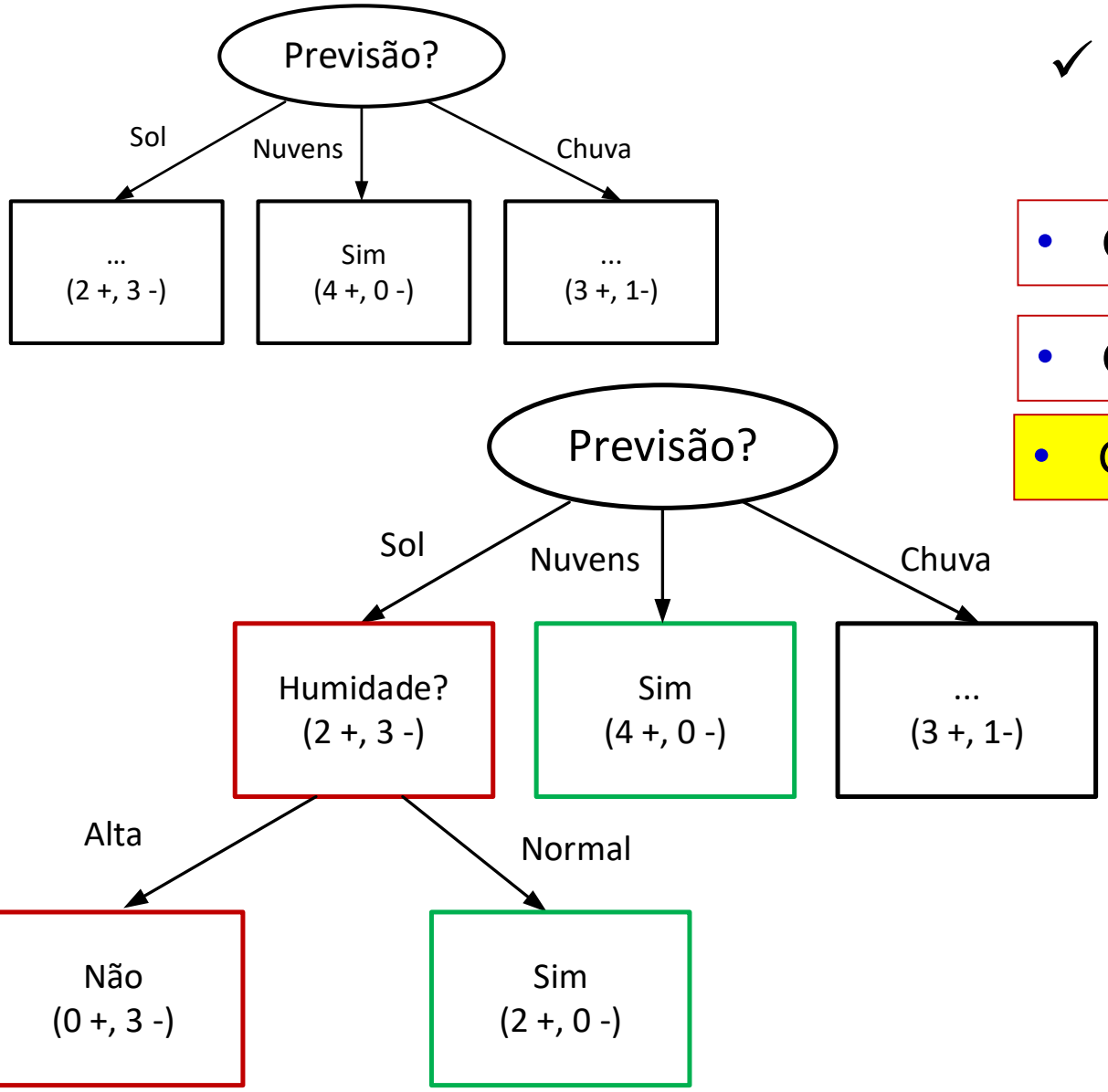
- S_{Sol}= [2+,3-] ✓ Vento?
- $S_{Vento=Leve}$= [1+,2-] • $S_{Vento=Forte}$= [1+,1-]
- $E(S_{Sol},Vento=Leve) = - (1/3)*\log2(1/3) - (2/3)*\log2(2/3) = 0.918$
- $E(S_{Sol},Vento=Forte) = - (1/2)*\log2(1/2) - (1/2)*\log2(1/2) = 1$
- Média Pond.= $(3/5)*0.918 + (2/5)*1 = 0.951$
- $GI(S_{Sol},Vento) = 0.971 – 0.951=0.02$

Exemplo: Ganho de Informação (cont.)

Dia	Previsão	Humidade	Jogar Ténis ?
1	Sol	Alta	Não (-)
2	Sol	Alta	Não (-)
3	Nuvens	Alta	Sim (+)
4	Chuva	Alta	Sim (+)
5	Chuva	Normal	Sim (+)
6	Chuva	Normal	Não (-)
7	Nuvens	Normal	Sim (+)
8	Sol	Alta	Não (-)
9	Sol	Normal	Sim (+)
10	Chuva	Normal	Sim (+)
11	Sol	Normal	Sim (+)
12	Nuvens	Alta	Sim (+)
13	Nuvens	Normal	Sim (+)
14	Chuva	Alta	Não (-)

- S_{Sol}= [2+,3-] ✓ Humidade?
- $S_{Humi=Normal}$= [2+,0-] • $S_{Humi=Alta}$= [0+,3-]
- E(S_{Sol},Humi=Normal) = 0
- E(S_{Sol},Humi=Alta) = 0
- Média Pond.= 0
- GI(S_{Sol},Humidade) = 0.971 – 0=0.971

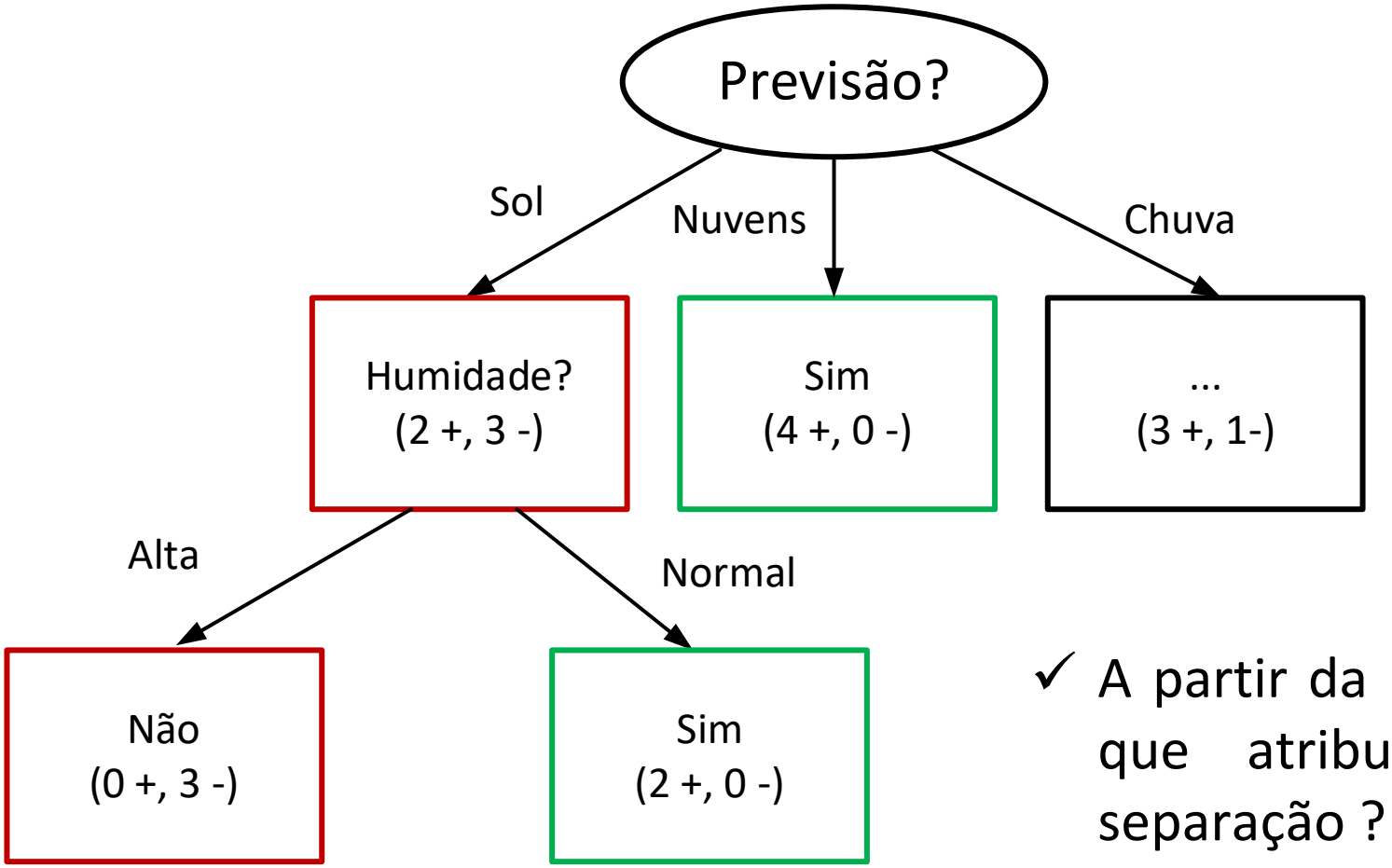
Exemplo: Ganho de Informação (cont.)



- ✓ Como base nos seguintes ganhos:

- $GI(S_{Sol}, Temp) = 0.571$
- $GI(S_{Sol}, Vento) = 0.02$
- $GI(S_{Sol}, Humidade) = 0.971$

Exemplo: Ganho de Informação (cont.)



- ✓ A partir da **Previsão=Chuva** que atributo utilizar na separação ?

Exemplo: Ganho de Informação (cont.)

Dia	Previsão	Temperatura	Humidade	Vento	Jogar Ténis ?
1	Sol	Quente	Alta	Leve	Não (-)
2	Sol	Quente	Alta	Forte	Não (-)
3	Nuvens	Quente	Alta	Leve	Sim (+)
4	Chuva	Média	Alta	Leve	Sim (+)
5	Chuva	Frio	Normal	Leve	Sim (+)
6	Chuva	Frio	Normal	Forte	Não (-)
7	Nuvens	Frio	Normal	Forte	Sim (+)
8	Sol	Média	Alta	Leve	Não (-)
9	Sol	Frio	Normal	Leve	Sim (+)
10	Chuva	Média	Normal	Leve	Sim (+)
11	Sol	Média	Normal	Forte	Sim (+)
12	Nuvens	Média	Alta	Forte	Sim (+)
13	Nuvens	Quente	Normal	Leve	Sim (+)
14	Chuva	Média	Alta	Forte	Não (-)

- ✓ A partir do atributo Chuva que atributo utilizar na separação ?

- S_{Chuva}= [3+,2-]

- $GI(S_{Chuva}) = - (3/5)*\log2(3/5) - (2/5)*\log2(2/5) = 0.971$

Exemplo: Ganho de Informação (cont.)

Dia	Previsão	Temperatura	Jogar Ténis ?
1	Sol	Quente	Não (-)
2	Sol	Quente	Não (-)
3	Nuvens	Quente	Sim (+)
4	Chuva	Média	Sim (+)
5	Chuva	Frio	Sim (+)
6	Chuva	Frio	Não (-)
7	Nuvens	Frio	Sim (+)
8	Sol	Média	Não (-)
9	Sol	Frio	Sim (+)
10	Chuva	Média	Sim (+)
11	Sol	Média	Sim (+)
12	Nuvens	Média	Sim (+)
13	Nuvens	Quente	Sim (+)
14	Chuva	Média	Não (-)

- S_{Chuva}= [3+,2-] ✓ Vamos começar pelo atributo **Temperatura**.
- $S_{Temp=Quente}$= [0+,0-] • $S_{Temp=Média}$= [2+,1-]
- $S_{Temp=Frio}$= [1+,1-]
- $E(S_{Sol},Temp=Quente) = 0$
- $E(S_{Sol},Temp=Média) = - (2/3)*log2(2/3) - (1/3)*log2(1/3) = 0.528$
- $E(S_{Sol},Temp=Frio) = - (1/2)*log2(1/2) - (1/2)*log2(1/2) = 0.5$
- Média Pond.= 0 + (3/5)*0.528+(2/5)*0.5 = 0.517
- $GI(S_{Sol},Temp) = 0.971 – 0.400=0.571$

Exemplo: Ganho de Informação (cont.)

Dia	Previsão	Humidade	Jogar Ténis ?
1	Sol	Alta	Não (-)
2	Sol	Alta	Não (-)
3	Nuvens	Alta	Sim (+)
4	Chuva	Alta	Sim (+)
5	Chuva	Normal	Sim (+)
6	Chuva	Normal	Não (-)
7	Nuvens	Normal	Sim (+)
8	Sol	Alta	Não (-)
9	Sol	Normal	Sim (+)
10	Chuva	Normal	Sim (+)
11	Sol	Normal	Sim (+)
12	Nuvens	Alta	Sim (+)
13	Nuvens	Normal	Sim (+)
14	Chuva	Alta	Não (-)

- S_{Chuva}= [3+,2-] ✓ **Humidade.**
- $S_{Humi=Alta}$= [1+,1-]
- $S_{Humi=Normal}$= [2+,1-]
- $E(S_{Chuva}, Humi=Alta) = - (1/2)*\log2(1/2) - (1/2)*\log2(1/2) = 0$
- $E(S_{Chuva}, Humi=Normal) = - (2/3)*\log2(2/3) - (1/3)*\log2(1/3) = 0.528$
- Média Pond.= 0 + (3/5)*0.528 = 0.317
- $GI(S_{Chuva}, Humi) = 0.971 - 0.317 = 0.654$

Exemplo: Ganho de Informação (cont.)

Dia	Previsão	Vento	Jogar Ténis ?
1	Sol	Leve	Não (-)
2	Sol	Forte	Não (-)
3	Nuvens	Leve	Sim (+)
4	Chuva	Leve	Sim (+)
5	Chuva	Leve	Sim (+)
6	Chuva	Forte	Não (-)
7	Nuvens	Forte	Sim (+)
8	Sol	Leve	Não (-)
9	Sol	Leve	Sim (+)
10	Chuva	Leve	Sim (+)
11	Sol	Forte	Sim (+)
12	Nuvens	Forte	Sim (+)
13	Nuvens	Leve	Sim (+)
14	Chuva	Forte	Não (-)

- S_{Chuva}= [3+,2-]
- $S_{Vento=Leve}$= [3+,0-]
- $S_{Vento=Forte}$= [0+,2-]

- E(S_{Chuva},Vento=Leve) = 0
- E(S_{Chuva},Vento=Forte)= 0

- Média Pond.= 0

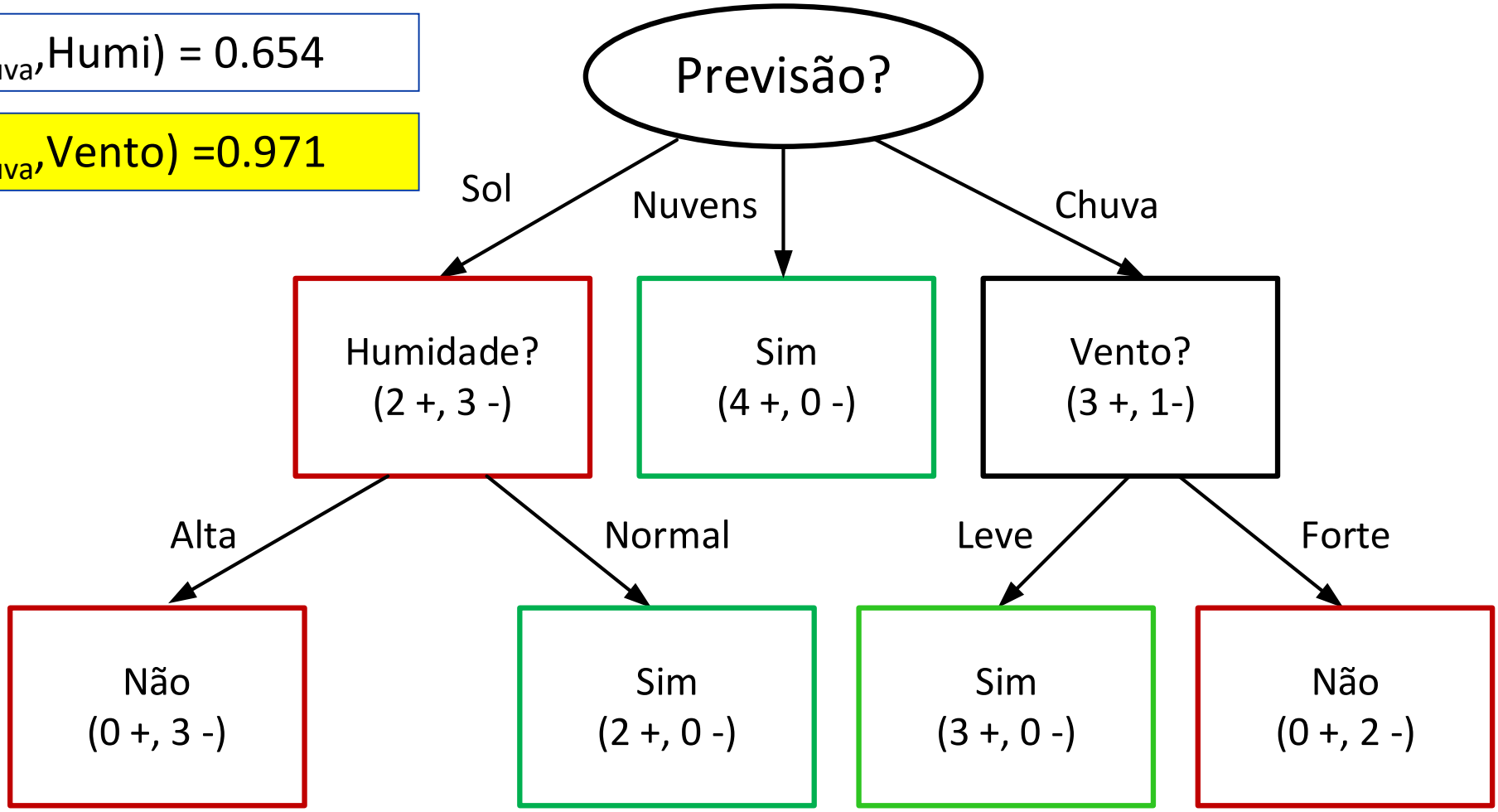
- GI(S_{Chuva},Vento) = 0.971 –0=0.971

✓ **Vento**.

Exemplo: Ganho de Informação (cont.)

✓ Como base nos seguintes ganhos:

- $E(S_{Sol}, Temp) = 0.571$
- $E(S_{Chuva}, Humi) = 0.654$
- $E(S_{Chuva}, Vento) = 0.971$



Rácio de Ganho (*Gain Ratio*)

- ✓ As medidas de impureza como a Entropia e Gini tendem a favorecer atributos que possuam um elevado número de valores distintos.

- ✓ Duas estratégias para evitar este problema são:

1. Restringir as condições de teste a separações binárias (usada no algoritmo CART – que é um *Decision Tree Algorithm*);
2. Modificar o critério de separação de forma a considerar o número de resultados obtidos pela condição de teste:

$$Gain\ Ratio\ (S,A) = \frac{G_{info}(S,A)}{Split_{info}(S,A)}$$

$$Split_{info}(S,A) = -\sum \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}$$

Rácio de Ganho (*Gain Ratio*)

✓ Para o exemplo da classificação de jogar Ténis:

- S_{Sol}= [2+,3-] $\Rightarrow |S_{sol}| = 5$
- S_{Nuvens}= [4+,0-] $\Rightarrow |S_{Nuvens}| = 4$
- S_{Chuva}= [3+,2-] $\Rightarrow |S_{Chuva}| = 5$

$$Split_{info}(S, Previsão) = -\sum \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}$$

$$Split_{info}(S, Previsão) = -\left(\frac{5}{14}\right) \log \left(\frac{5}{14}\right) - \left(\frac{4}{14}\right) \log \left(\frac{4}{14}\right) - \left(\frac{5}{14}\right) \log \left(\frac{5}{14}\right) = 1.577$$

$$Gain\ Ratio\ (S, A) = \frac{G_{info}(S,A)}{Split_{info}(S,A)} = \frac{0.246}{1.577} = 0.156$$

Bibliografia

[1] Tan P. N., Steibach M. e Kumar V, (2016), **Introduction to Data Mining**, Pearson,Addison-Wesley,