Aprendizagem automática

Sessão 12 - T

Modelos baseados em árvores - Parte 1

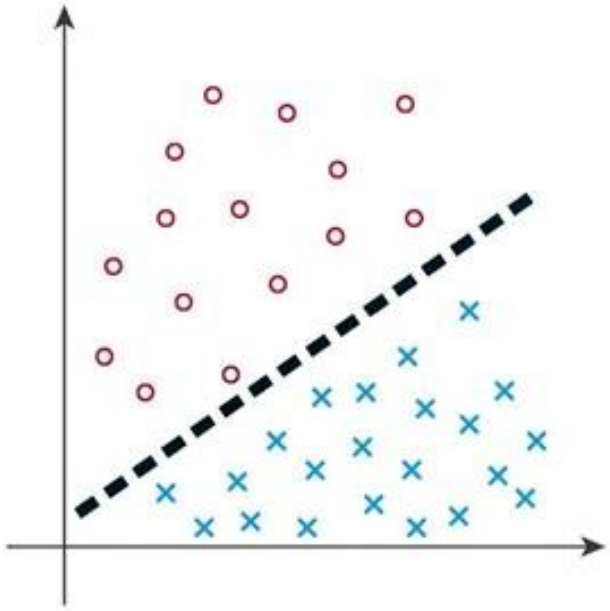
Ciência de Dados Aplicada

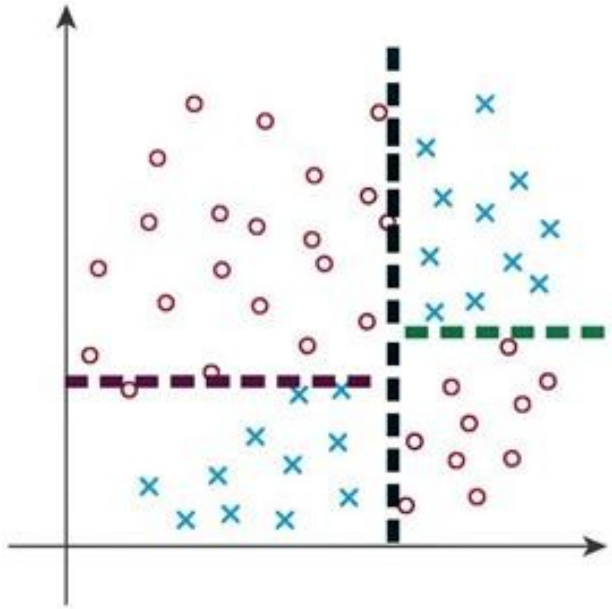
2023/2024

Espaço de características

- **Dados linearmente separáveis** - o espaço de características pode ser bem separado por uma linha ou um hiperplano;

- Dados **linearmente inseparáveis** - o espaço de características não pode ser efetivamente dividido por uma única linha ou hiperplano.





Note-se que as classes ainda estão bem separadas no espaço de características, mas os **limites de decisão não podem ser descritos por equações lineares simples**.

Espaço de características

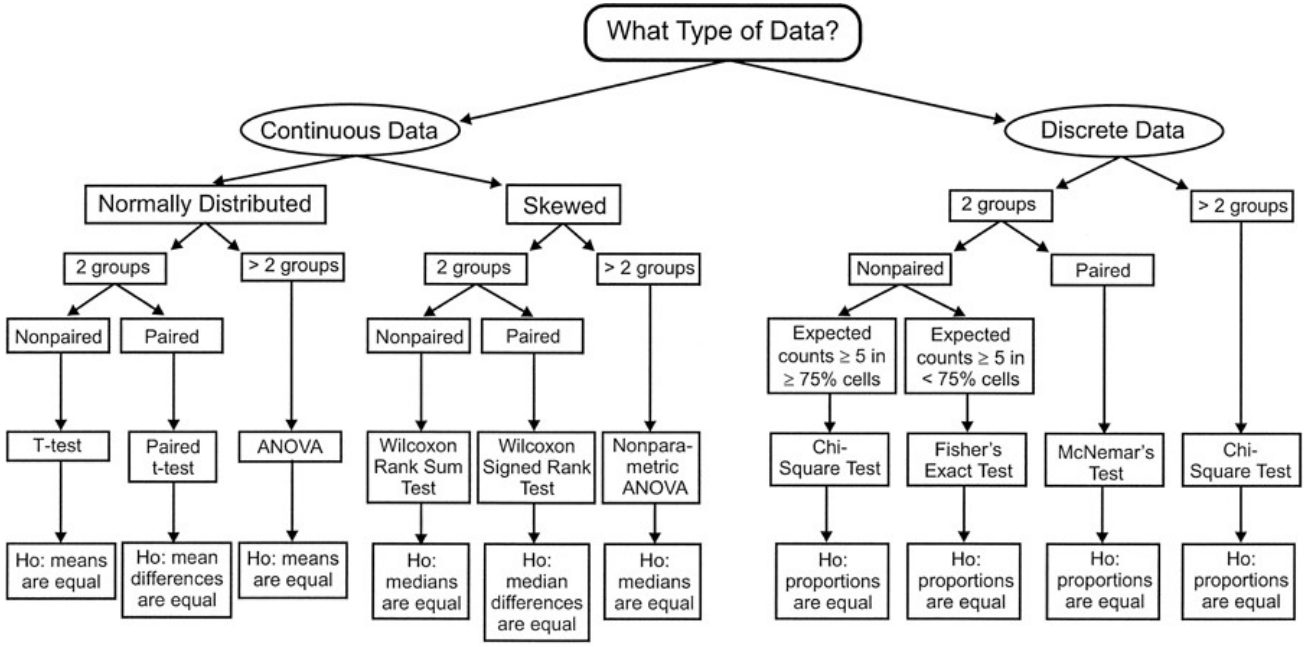
Espaço de características

- Embora os modelos lineares com limites lineares ofereçam uma interpretação intuitiva, a interpretação de limites de decisão não lineares apresenta desafios.

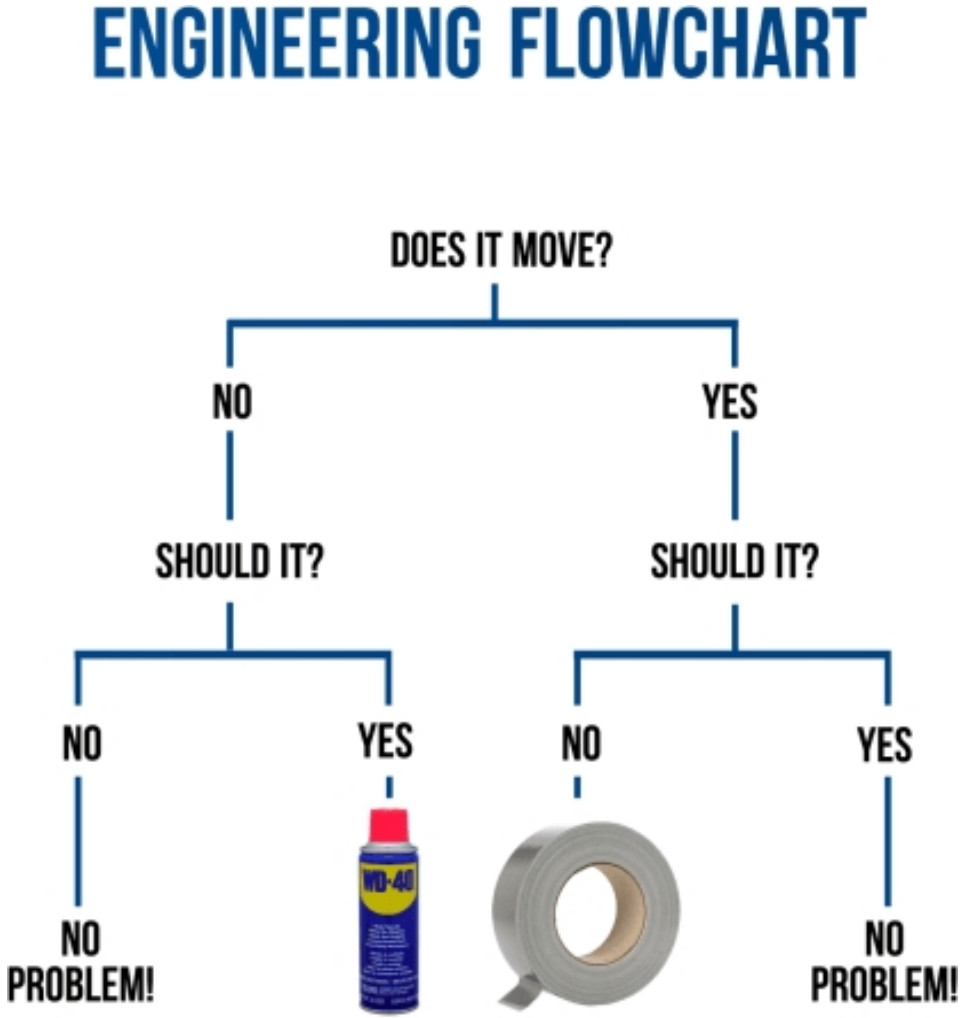
- Por conseguinte, é necessário criar modelos que:
  - permitir **limites de decisão complexos**;
  - são **fáceis de interpretar**.

Modelos interpretáveis

- Historicamente, pessoas de diversas origens têm confiado em **modelos interpretáveis** para distinguir entre várias classes de objectos e fenómenos.



Source: Waning B, Montagne M: *Pharmacoepidemiology: Principles and Practice*: http://www.accesspharmacy.com



Modelos baseados em árvores

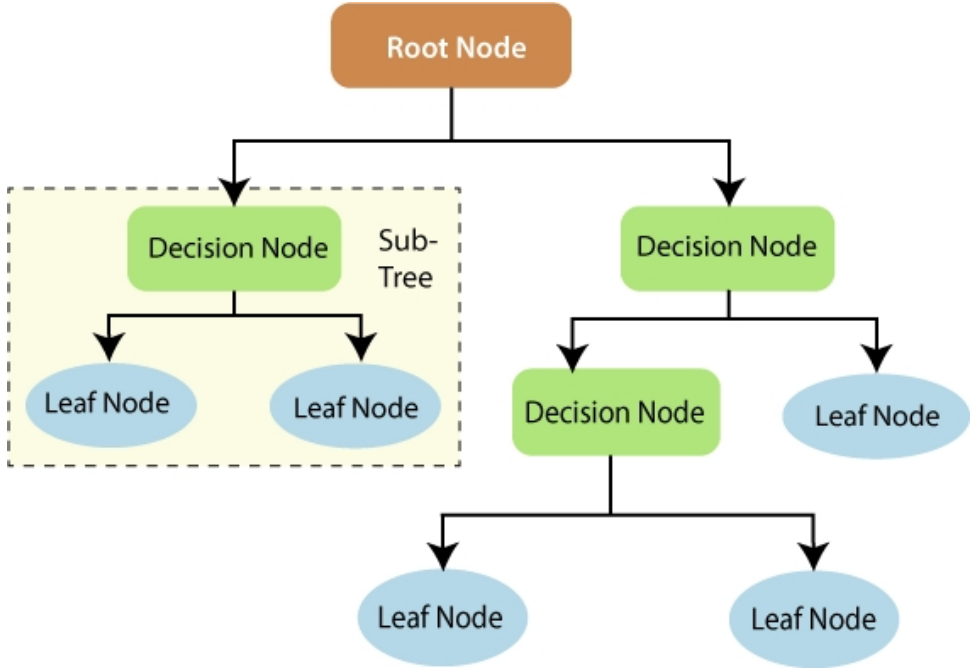
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Os fluxogramas, como nos exemplos anteriores, podem ser formulados como modelos matemáticos (**gráficos**) para classificação e regressão.

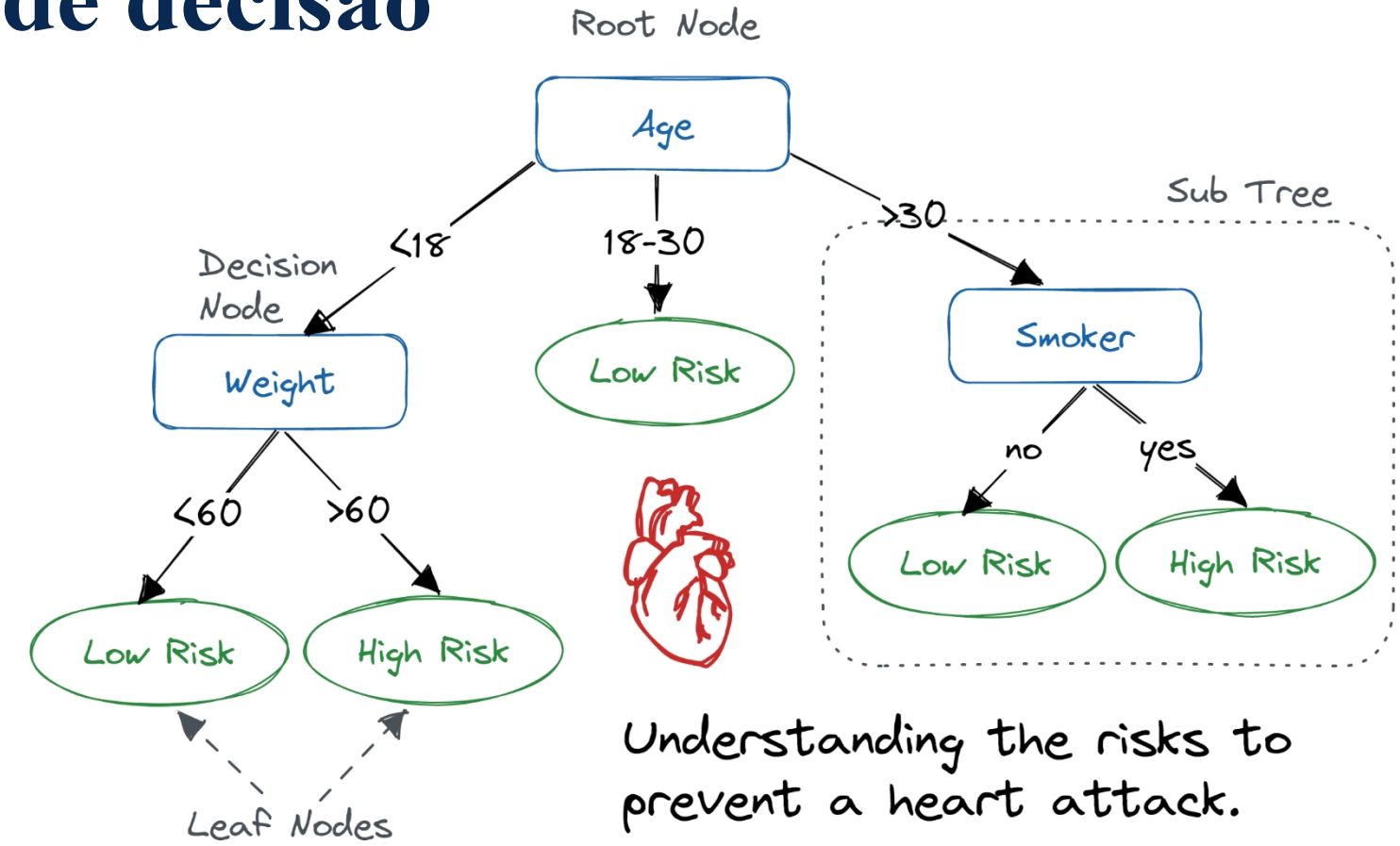
- Estes modelos são:
  - **Interpretável** por humanos;
  - Ter **limites de decisão complexos**;
  - Os limites de decisão são uma **combinação de limites lineares** que são **matematicamente simples de descrever**.

Árvores de decisão

- Matematicamente, uma árvore de decisão pode ser definida como um **gráfico acíclico dirigido**, compreendendo:
  - **Nós:** Representam pontos de decisão ou condições.
  - **Arestas:** Ligam os nós e representam os resultados das decisões.
  - **Nó de raiz:** O ponto de decisão inicial, que representa todo o conjunto de dados.
  - **Nós de decisão:** Pontos de decisão onde é efectuada uma divisão com base numa caraterística ou atributo.
  - **Folha Nós:** Nós terminais nós que representam resultados finais ou previsões.



Árvores de decisão

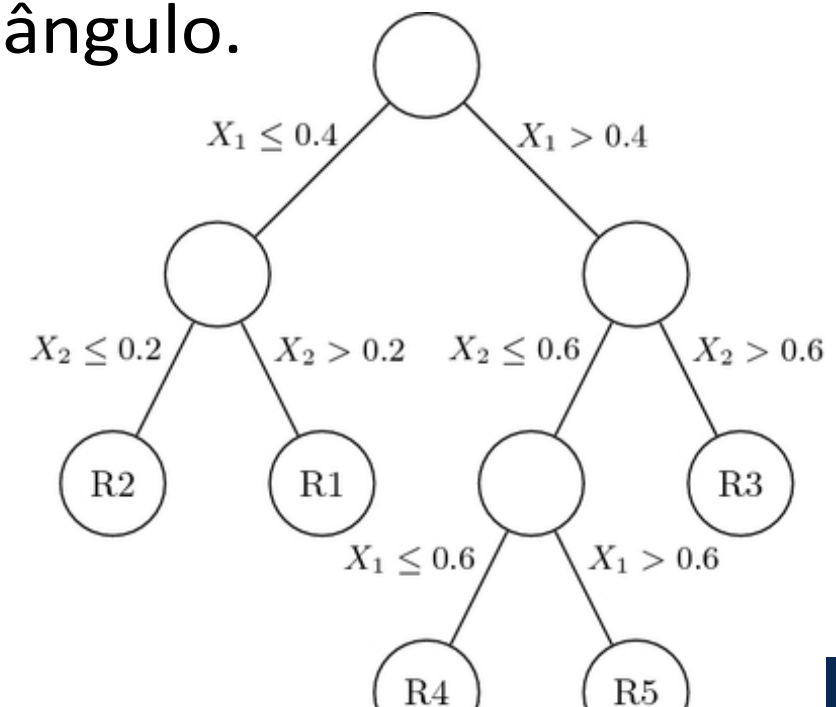


https://www.datacamp.com/tutorial/decision-tree-classification-python

Idade	Peso	Fumador	Previsão
35	80	sim	Risco elevado
25	80	sim	?

Árvores de decisão

- Baseado em árvores baseados métodos trabalho por **particionamento o espaço de características em rectângulos**;
- As previsões são efectuadas através da **média dos valores** ou com base na **classe mais frequente** em cada retângulo.

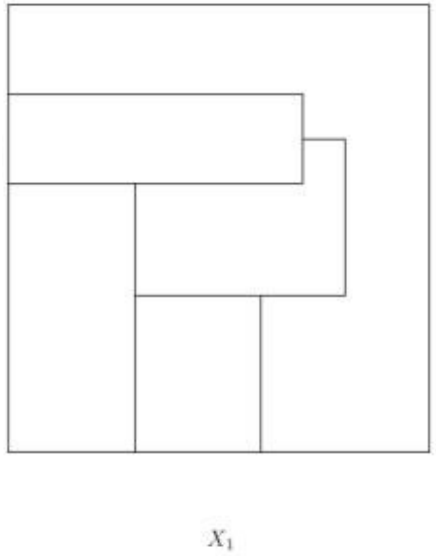


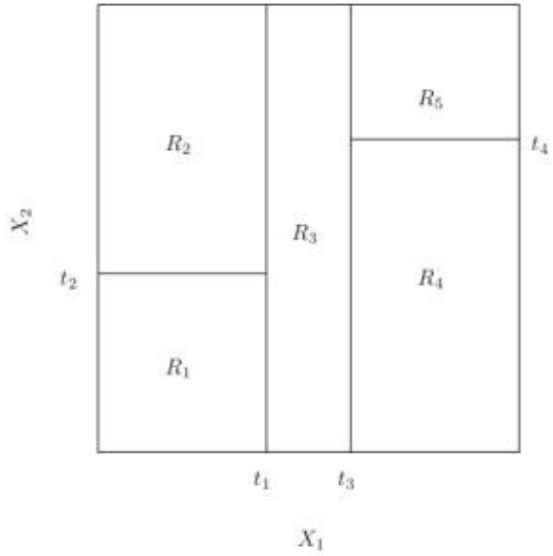
Árvores de decisão

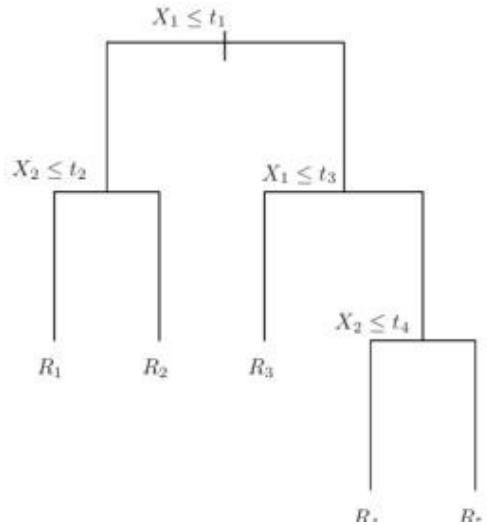
Beaulac, C., C Rosenthal, J. S. (2019). Previsão do sucesso acadêmico e principal dos estudantes universitários usando florestas aleatórias. Em Pesquisa em Ensino Superior (Vol. 60, Edição 7, pp. 1048-1064). Springer Science and Business Media LLC. https://doi.org/10.1007/s11162-019-09546-y

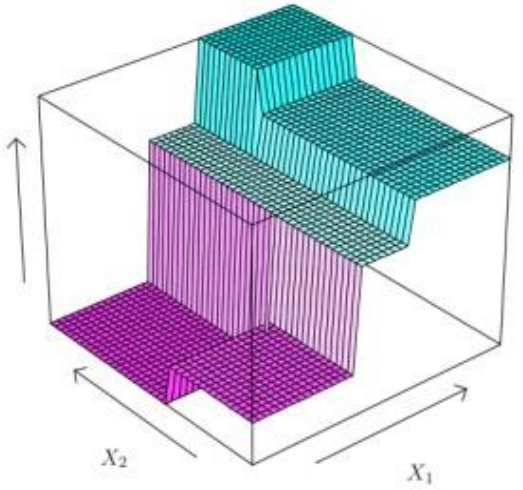
Árvores de decisão

Nunca teremos uma divisão como esta!



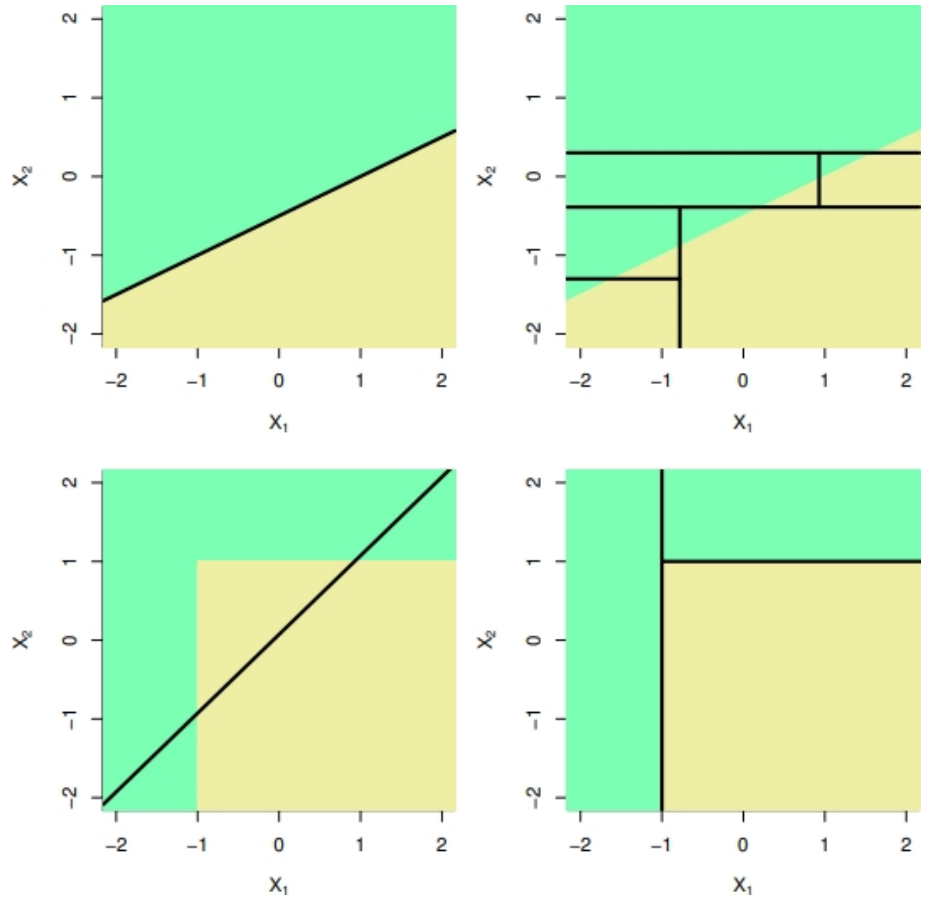






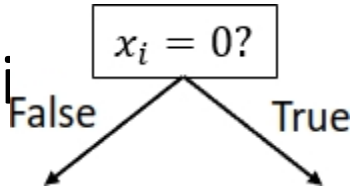
Árvores de decisão

- Modelos lineares vs. árvores de decisão

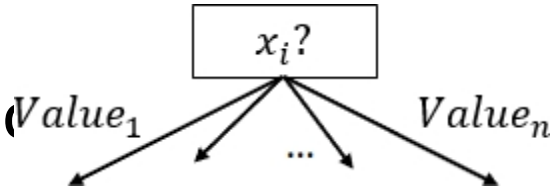


Árvores de decisão: Nós de decisão

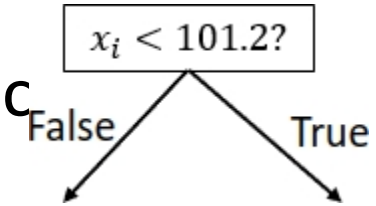
- Característica binári



- Característica categóri

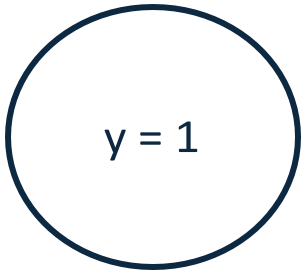


- Característica numéric



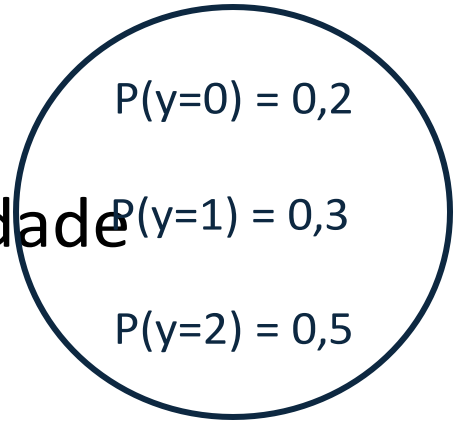
Árvores de decisão: Tipos de folhas

- Classificação



- Regressão

- Estimativa de probabilidade



Árvores de decisão: Algoritmo

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- As árvores são construídas utilizando a **guloso** guloso: **Recursivo** **particionamento binário**

Ganancioso significa que cada divisão é feita de forma a minimizar uma perda **sem olhar** para as divisões futuras!

- Isto implica os seguintes passos:
  - A definição de um **critério de splititing**;
  - A definição de uma **regra de paragem**;
  - **Poda de** árvores.

Árvores de decisão: Critérios de divisão

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- Em cada passo, é selecionada uma nova divisão, **encontrando a caraterística x_j e o ponto de divisão s**
  que **melhor divide os dados em dois meios-espaços**.

$$\{\mathbf{x} : x_j < s\} \quad \{\mathbf{x} : x_j \geq s\}.$$

  - Para a **regressão**, queremos a divisão que **minimiza a soma residual dos quadrados** (RSS)

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2,$$

em que $\hat{y}_{R_j}$ são os valores médios dos dados de treino dentro da j-ésima caixa.

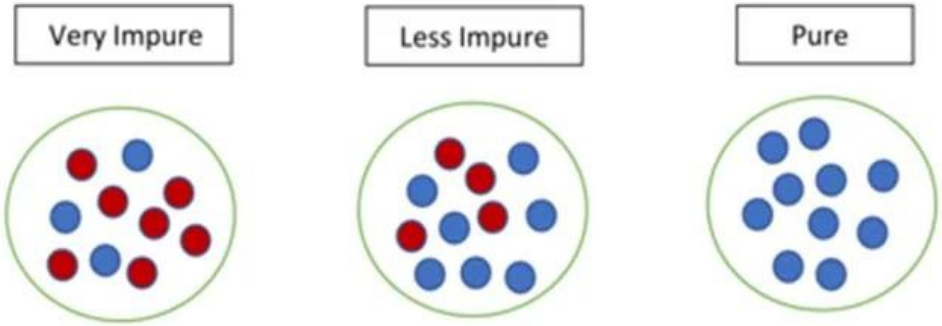
  - Para a **classificação**, podemos utilizar:
    - Entropia e ganho de informação
    - Índice de Gini

Árvores de decisão: Entropia e ganho de informação

- "*Na teoria da informação, a entropia de uma variável aleatória é o nível médio de "informação", "incerteza" ou "surpresa" inerente aos resultados possíveis da variável.*"

- No contexto das árvores de decisão, a entropia mede a **desordem ou impureza de um nó**.

$$E = -\sum_{i=1}^{n} p_i log_2(p_i)$$



p_i é a probabilidade de escolher aleatoriamente um exemplo da classe i.

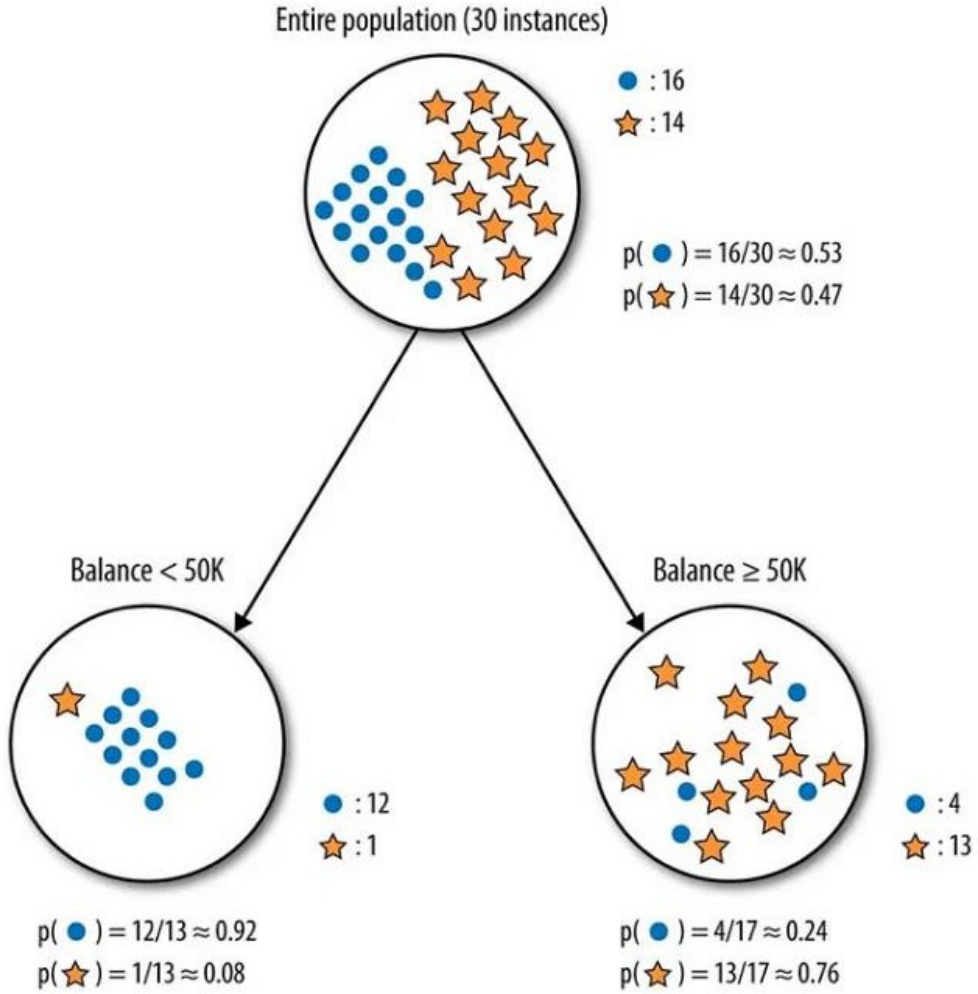
Árvores de decisão: Entropia e ganho de informação

$$InformationGain = Entropy_{parent} - Entropy_{children}$$

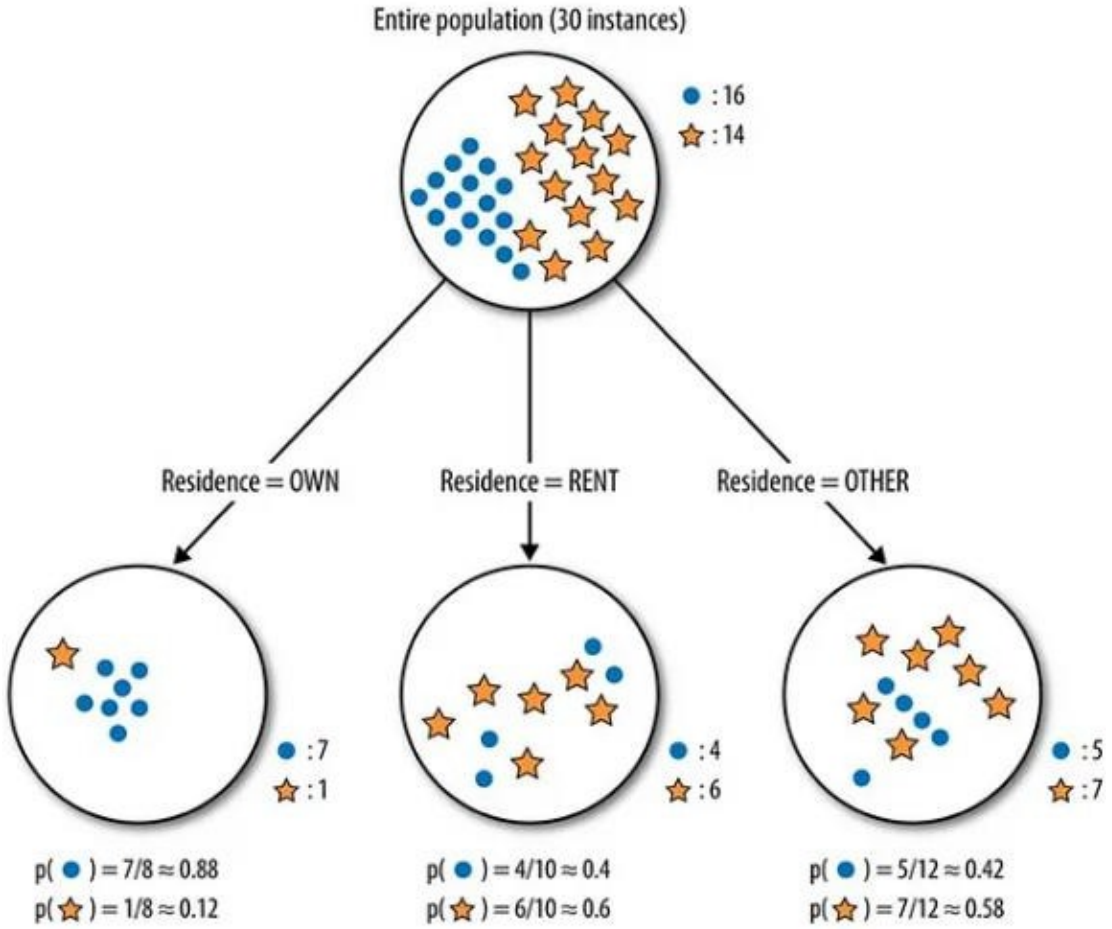
$$InformationGain = Entropy_{parent} - WeightedAvgEntropy_{children}$$

$$\text{Average Entropy} = \frac{n_{subnode_1}}{n_{parent}} E\_\text{subnode}_1 + \frac{n_{subnode_2}}{n_{parent}} E\_subnode_2 + \ldots + \frac{n_{subnode_n}}{n_{parent}} E\_subnode_n$$

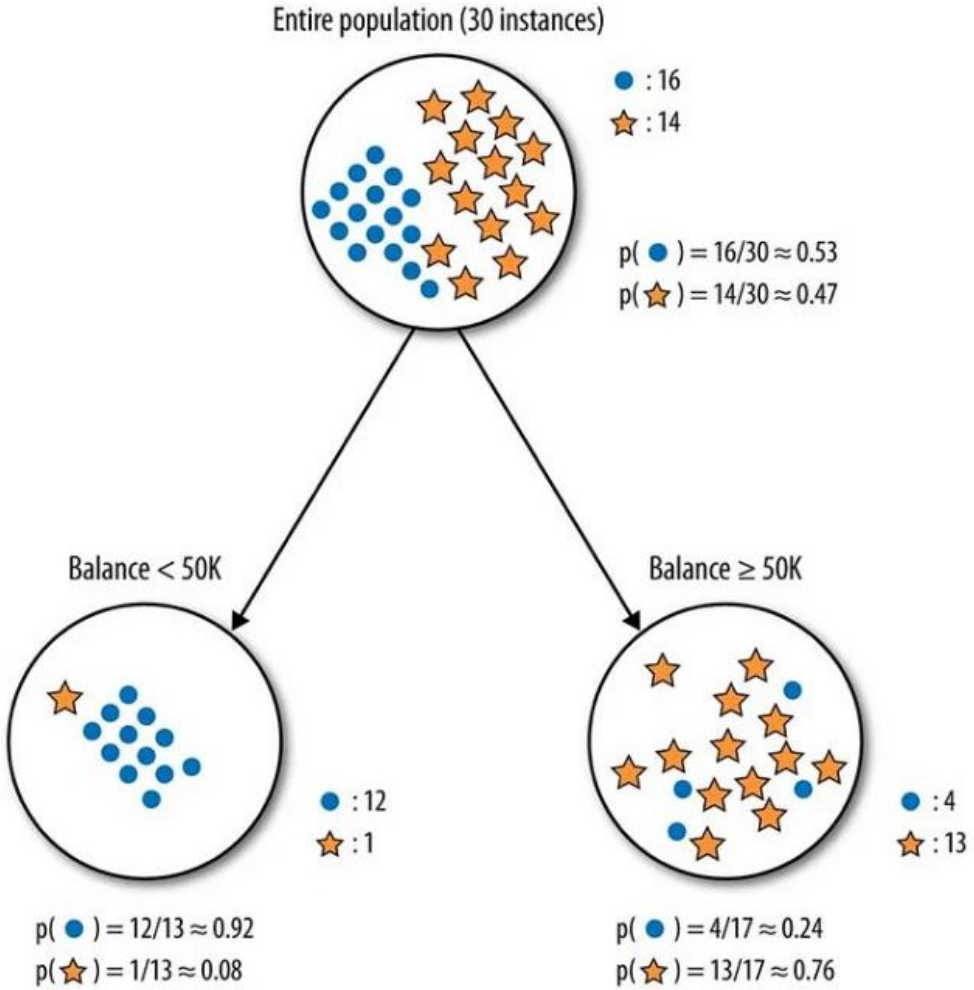
Árvores de decisão: Entropia e ganho de informação



VS



Árvores de decisão: Entropia e ganho de informacão



$$E(Parent) = -\frac{16}{30}\log_2\left(\frac{16}{30}\right) - \frac{14}{30}\log_2\left(\frac{14}{30}\right) \approx 0.99$$

$$E(Balance < 50K) = -\frac{12}{13}\log_2\left(\frac{12}{13}\right) - \frac{1}{13}\log_2\left(\frac{1}{13}\right) \approx 0.39$$

$$E(Balance > 50K) = -\frac{4}{17}\log_2\left(\frac{4}{17}\right) - \frac{13}{17}\log_2\left(\frac{13}{17}\right) \approx 0.79$$

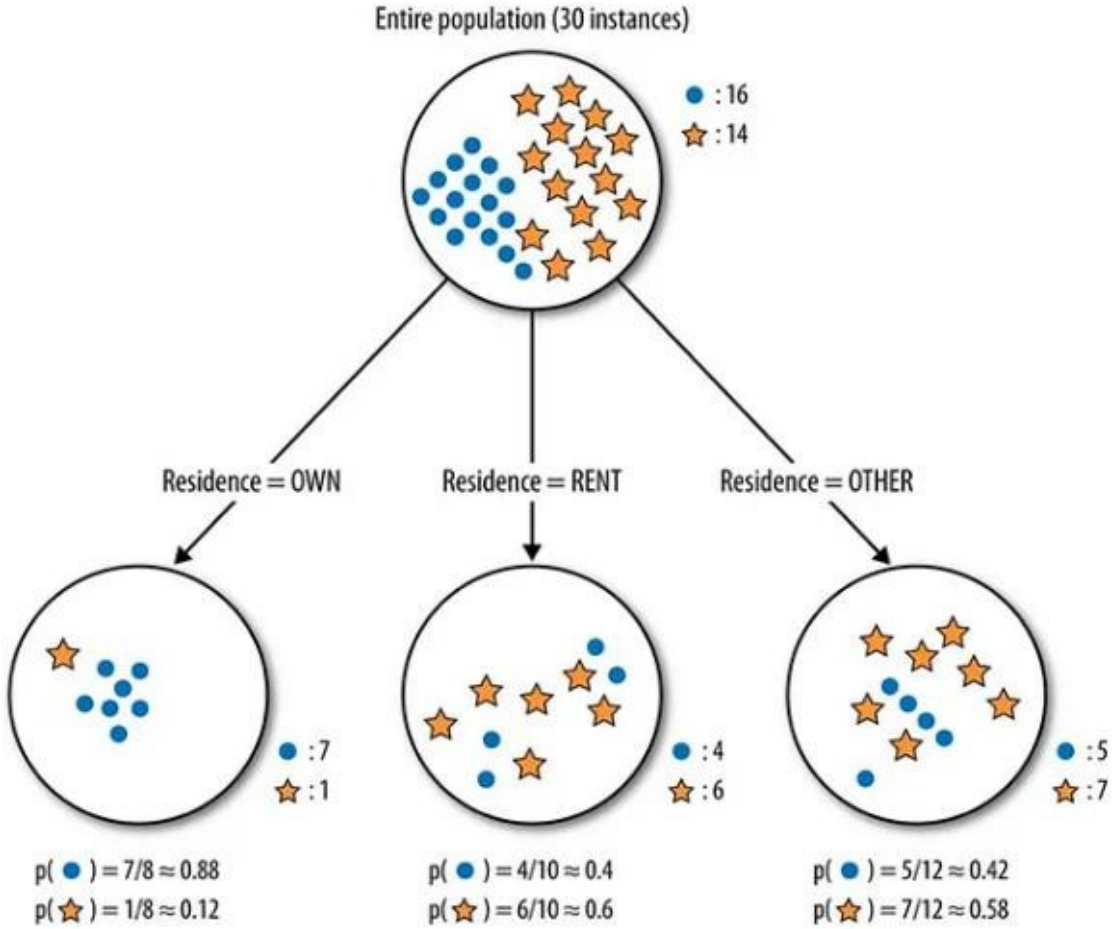
Weighted Average of entropy for each node:

$$E(Balance) = \frac{13}{30} \times 0.39 + \frac{17}{30} \times 0.79$$
$$= 0.62$$

Information Gain:

$$IG(Parent, Balance) = E(Parent) - E(Balance)$$
$$= 0.99 - 0.62$$
$$= 0.37$$

Árvores de decisão: Entropia e ganho de informação



$$E(Parent) = -\frac{16}{30}\log_2\left(\frac{16}{30}\right) - \frac{14}{30}\log_2\left(\frac{14}{30}\right) \approx 0.99$$

$$E(Residence = OWN) = -\frac{7}{8}\log_2\left(\frac{7}{8}\right) - \frac{1}{8}\log_2\left(\frac{1}{8}\right) \approx 0.54$$

$$E(Residence = RENT) = -\frac{4}{10}\log_2\left(\frac{4}{10}\right) - \frac{6}{10}\log_2\left(\frac{6}{10}\right) \approx 0.97$$

$$E(Residence = OTHER) = -\frac{5}{12}\log_2\left(\frac{5}{12}\right) - \frac{7}{12}\log_2\left(\frac{7}{12}\right) \approx 0.98$$

Weighted Average of entropies for each node:

$$E(Residence) = \frac{8}{30} \times 0.54 + \frac{10}{30} \times 0.97 + \frac{12}{30} \times 0.98 = 0.86$$

Information Gain:

$$\begin{aligned} IG(Parent, Residence) &= E(Parent) - E(Residence) \\ &= 0.99 - 0.86 \\ &= 0.13 \end{aligned}$$

Árvores de decisão: Índice de Gini

- O Gini índice mede a probabilidade de classificação incorrecta um elemento escolhido aleatoriamente com base na distribuição de etiquetas;

- Valores mais baixos valores indicam maior pureza e melhor separação das classes num nó da árvore de decisão.

$$Gini = 1 - \sum_{i=1}^{j} P(i)^2 \quad \text{ou} \quad Gini = 1 - \sum_{i=1}^{j} P(i)(1 - P(i))$$

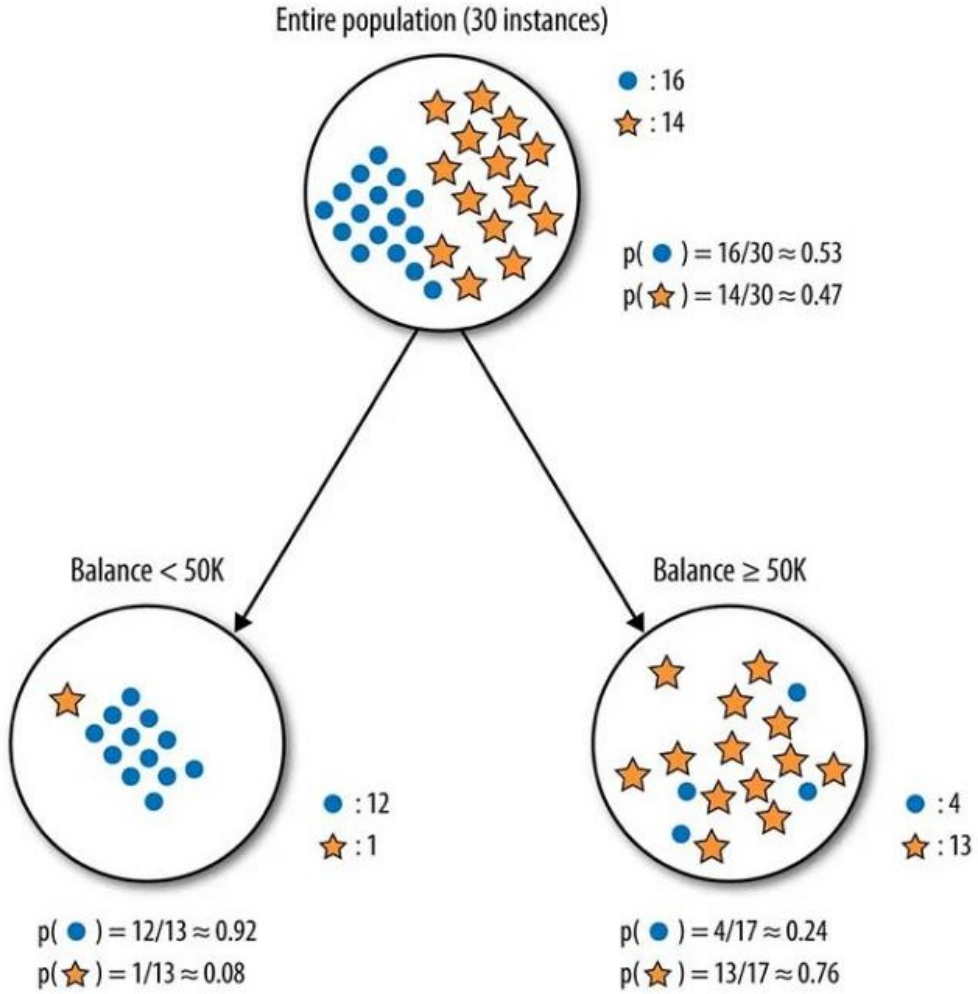
em que j representa o número de classes na variável-alvo

Árvores de decisão: Índice de Gini

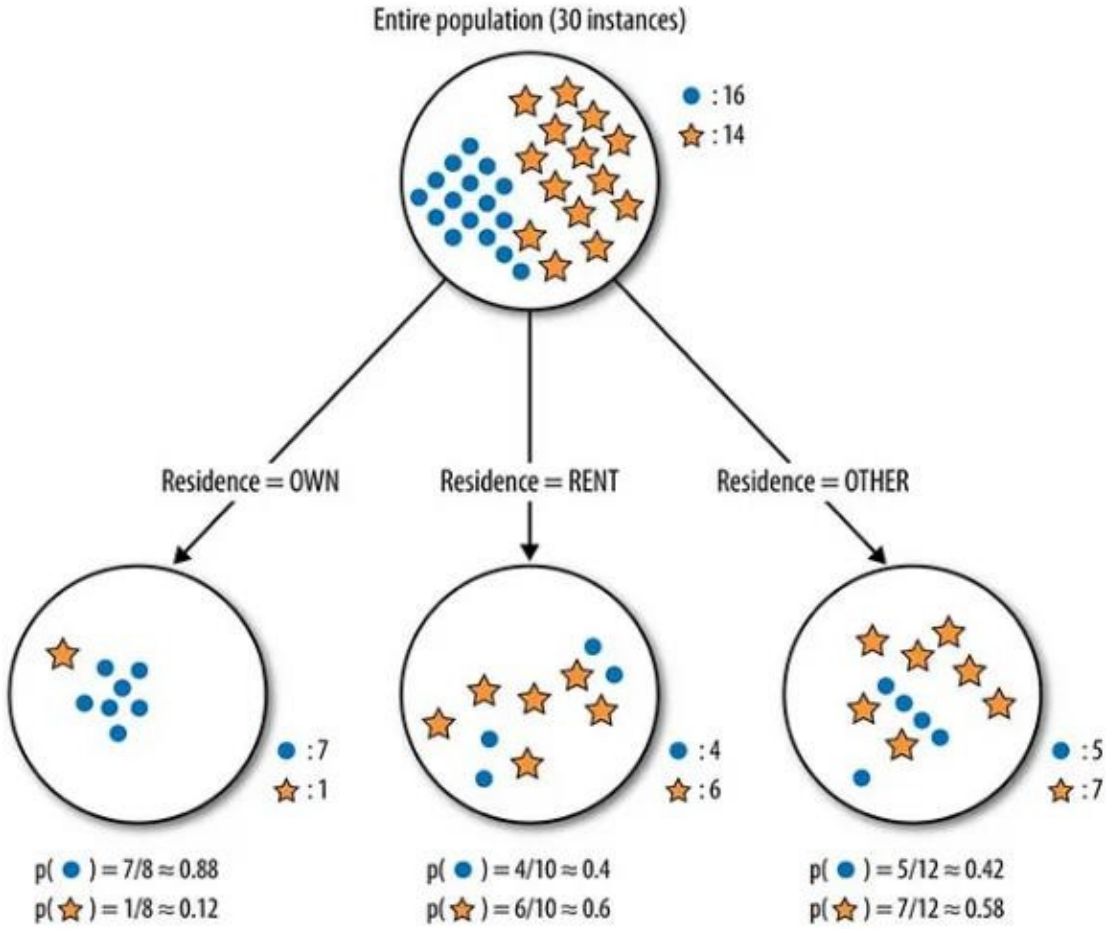
$$Gini_{split} = WeightedAvgGini_{nodes}$$

$$\text{WeightedAvgGini} = \frac{n_{subnode_1}}{n_{parent}} Gini_{subnode_1} + \frac{n_{subnode_2}}{n_{parent}} Gini_{subnode_2} + \ldots + \frac{n_{subnode_n}}{n_{parent}} Gini_{subnode_n}$$

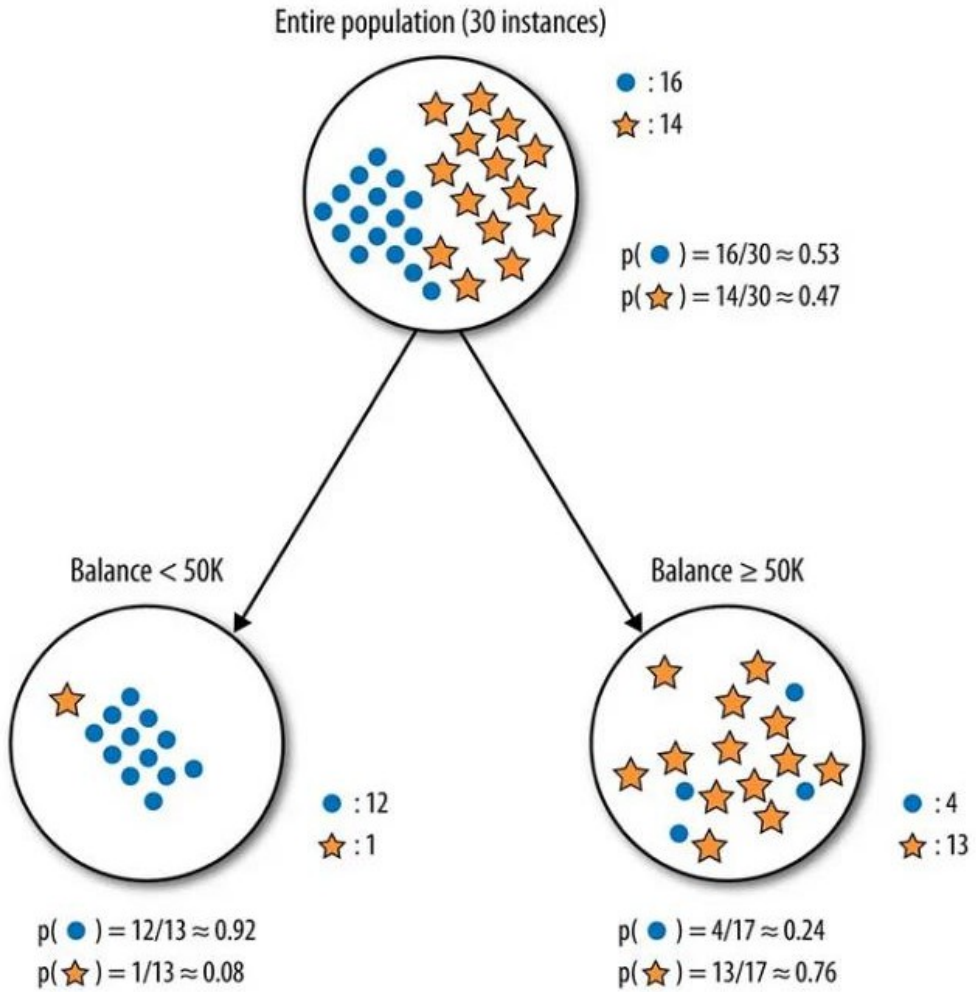
Árvores de decisão: Índice de Gini



VS



Árvores de decisão: Índice de Gini

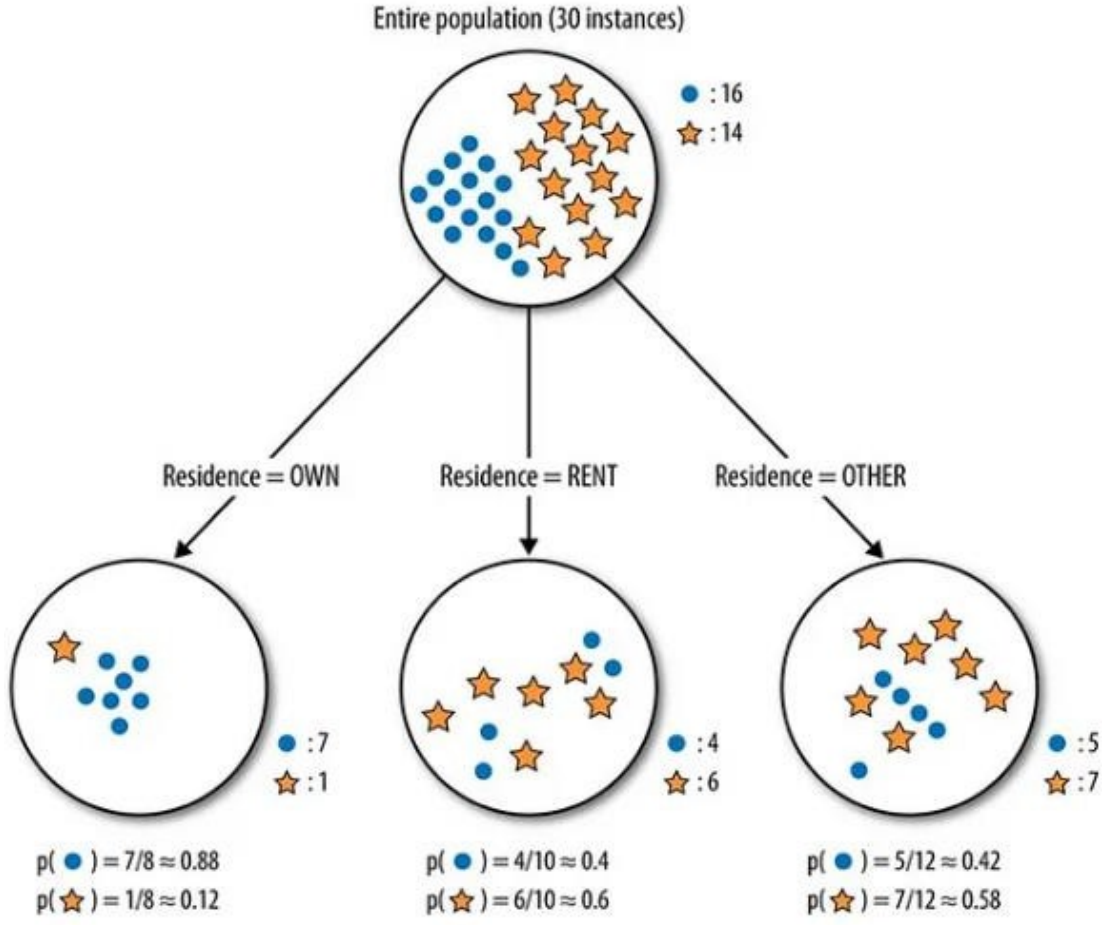


$$Gini_{(Balance<50)} = 1 - (\frac{12}{13})^2 - (\frac{1}{13})^2 = 0.142$$

$$Gini_{(Balance\geq 50)} = 1 - (\frac{4}{17})^2 - (\frac{13}{17})^2 = 0.360$$

$$Gini = \frac{13}{30} * 0.142 + \frac{17}{30} * 0.360 = 0.266$$

Árvores de decisão: Índice de Gini



$$Gini_{(OWN)} = 1 - (\frac{7}{8})^2 - (\frac{1}{8})^2 = 0.219$$

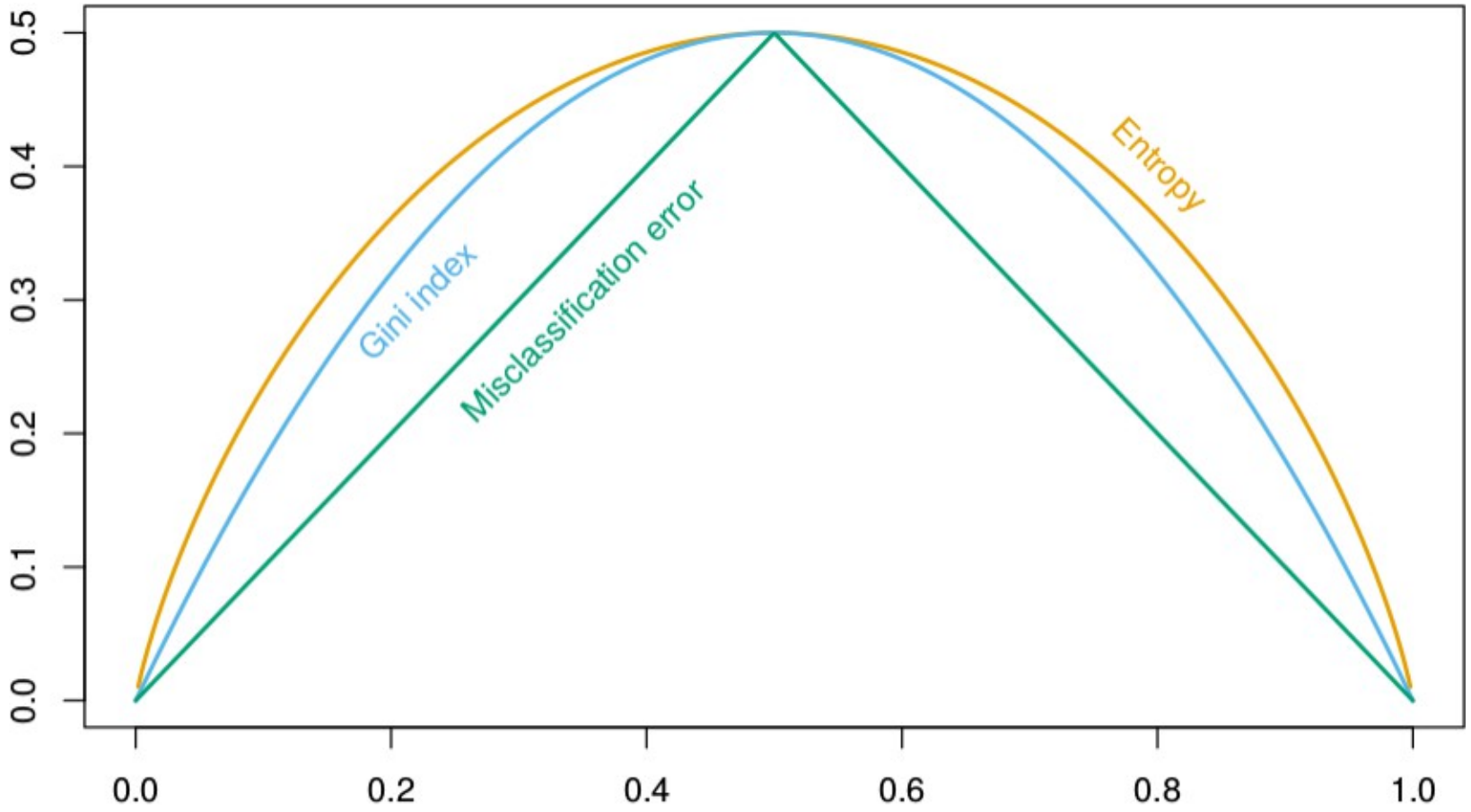
$$Gini_{(RENT)} = 1 - (\frac{4}{10})^2 - (\frac{6}{10})^2 = 0.48$$

$$Gini_{(OTHER)} = 1 - (\frac{5}{12})^2 - (\frac{7}{12})^2 = 0.486$$

$$Gini = \frac{8}{30} * 0.219 + \frac{10}{30} * 0.48 + \frac{12}{30} * 0.486 = 0.4128$$

Árvores de decisão: Critérios de divisão

- Porque não minimizar o erro de classificação incorrecta?



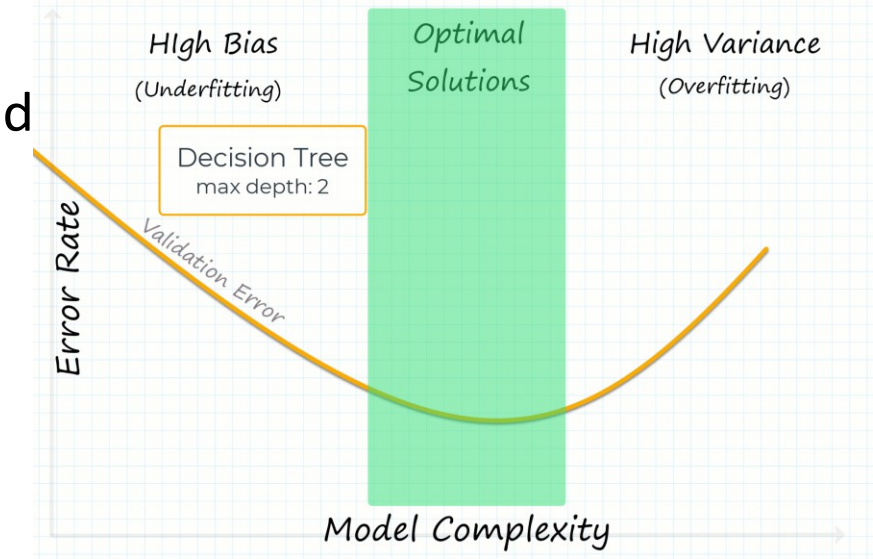
Árvores de decisão: Regras de paragem

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Profundidade máxima:** limita a profundidade da árvore;
- **Mínimo de amostras por folha:** limita o número mínimo de amostras que um nó folha pode ter;
- **Mínimo de amostras por divisão:** limita o número mínimo de amostras necessárias para efetuar uma divisão;
- **Número máximo de nós de folha:** limita o número total de nós principais numa árvore;
- **Limiar de impureza:** uma divisão só é efectuada se reduzir a impureza numa determinada quantidade;

Árvores de decisão

- A flexibilidade/complexidade das árvores de decisão é determinada principalmente pela **profundidade** da **árvore**:
  - Para obter uma **pequena polarização**, precisamos d
  - No entanto, isto resulta numa **grande variação**!



- Para melhorar o desempenho:
  - **Poda:** cultivar árvores profundas (pequena tendência, alta variância) que depois são podadas em árvores mais pequenas (reduzir a variância);
  - **Métodos de conjunto (próxima sessão):** combinar várias árvores simples.
    - Ensacamento e florestas aleatórias
    - Árvores reforçadas

Árvores de decisão: Poda de árvores

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **As árvores profundas ajustam-se frequentemente em excesso** aos dados de treino, o que resulta num **fraco desempenho nos testes**;

- Poderíamos parar de dividir assim que o ganho de informação não melhorasse pelo menos um valor pré-especificado;

- No entanto, **os desdobramentos "fracos" no início podem, por vezes, conduzir a um desdobramento realmente bom mais tarde**;

- **Solução:** Cultivar uma árvore profunda e depois **podá-la**.

Árvores de decisão: Poda por complexidade de custos

- **Poda de complexidade de custos**, também conhecida como **poda do elo mais fraco**:

- Matematicamente, a medida de complexidade do custo de uma árvore T é dada por:

$$\mathrm{R}_{\alpha}\left(\mathrm{T}\right) = \mathrm{R}(\mathrm{T}) + \alpha|\mathrm{T}|$$

Onde:

* $\mathrm{R}(\mathrm{T})$ é o risco da árvore T (RSS global, Gini/Entropia/etc)
* $|\mathrm{T}|$ é o número de nós folha na árvore T

Árvores de decisão: Poda por complexidade de custos

* α é o parâmetro de penalização/regularização

Árvores de decisão: Poda por complexidade de custos

$$R(T) = \sum_{t \in leaf\ \text{nodes}} (p_t \cdot M_t) = \sum_{t \in leaf\ \text{nodes}} R(t)$$

Risco de árvore T

Quanto mais pequeno, melhor

Proporção de amostras no nó de folha t

Métrica do nó folha t

Risco do nó folha t

- **Objetivo**: minimizar $\mathrm{R}_\alpha(\mathrm{T}) = \mathrm{R}(\mathrm{T}) + \alpha|\mathrm{T}|$

Dá-nos custos

Dá-nos complexidade

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

Árvores de decisão: Poda por complexidade de custos

Árvores de decisão: Poda por complexidade de custos

- **Regra de poda:**
  - podar todos os nós filhos de t se:

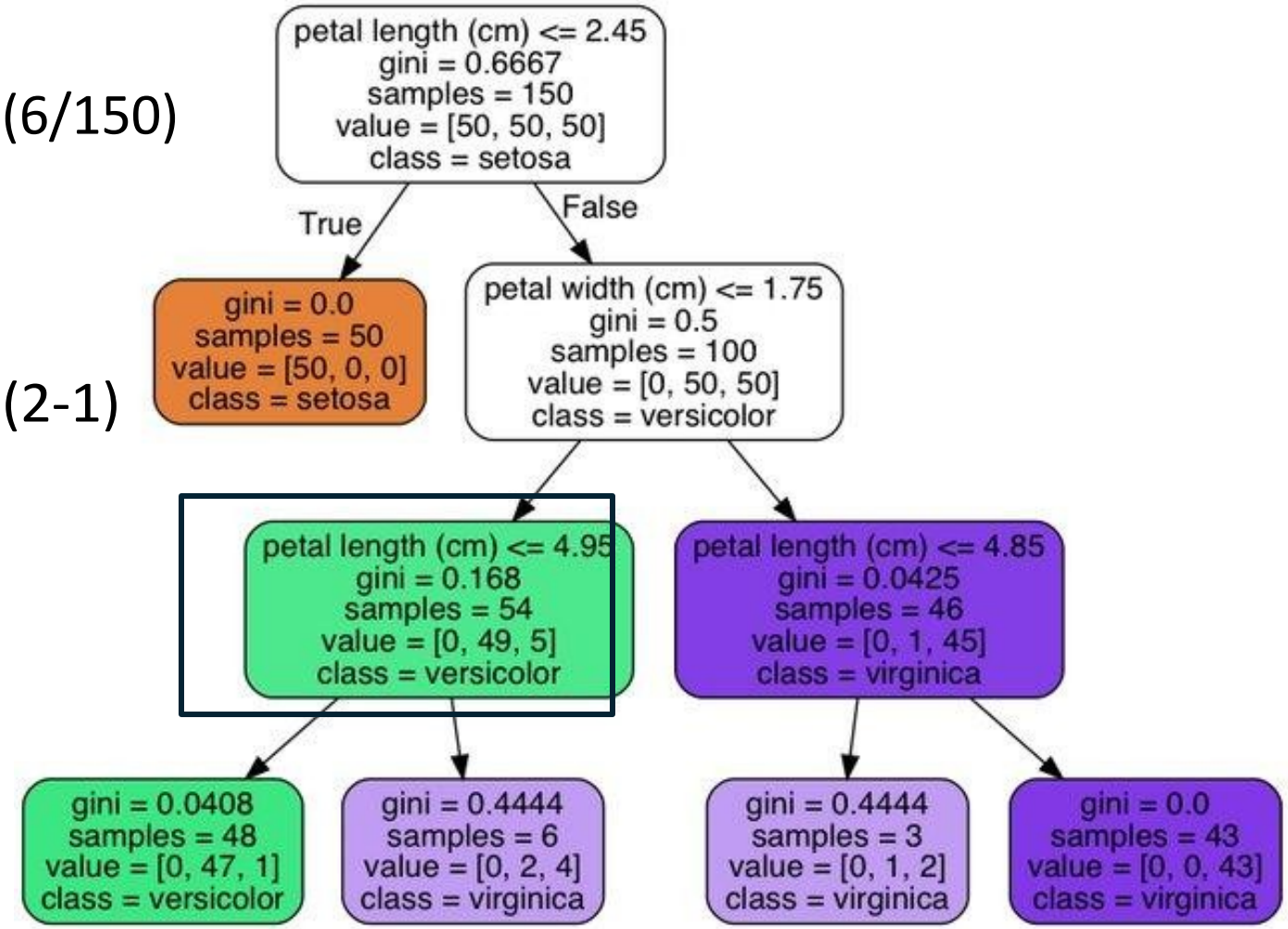
$$\underbrace{(|T_t| - 1)\alpha}_{\text{Penalização}} > \underbrace{R(t) - R(T_t)}_{\text{Prémio}} \Rightarrow \alpha > \frac{R(t) - R(T_t)}{|T_t| - 1}$$

Regra de poda

Árvores de decisão: Poda por complexidade de custos

- R(t) = 0,168 * (54/150) = 0,06048
- $R(T_t)$ = 0,0408 * (58/150) + 0,4444 * (6/150)
  = 0.033552
- |T| = 2
- $\frac{R(t) - R(T_t)}{|T| - 1}$ = (0.06048 - 0.033552) / (2-1)
  = 0.026928
- Então, se:
  - α = 0,02 não podamos
  - α = 0,03 podamos

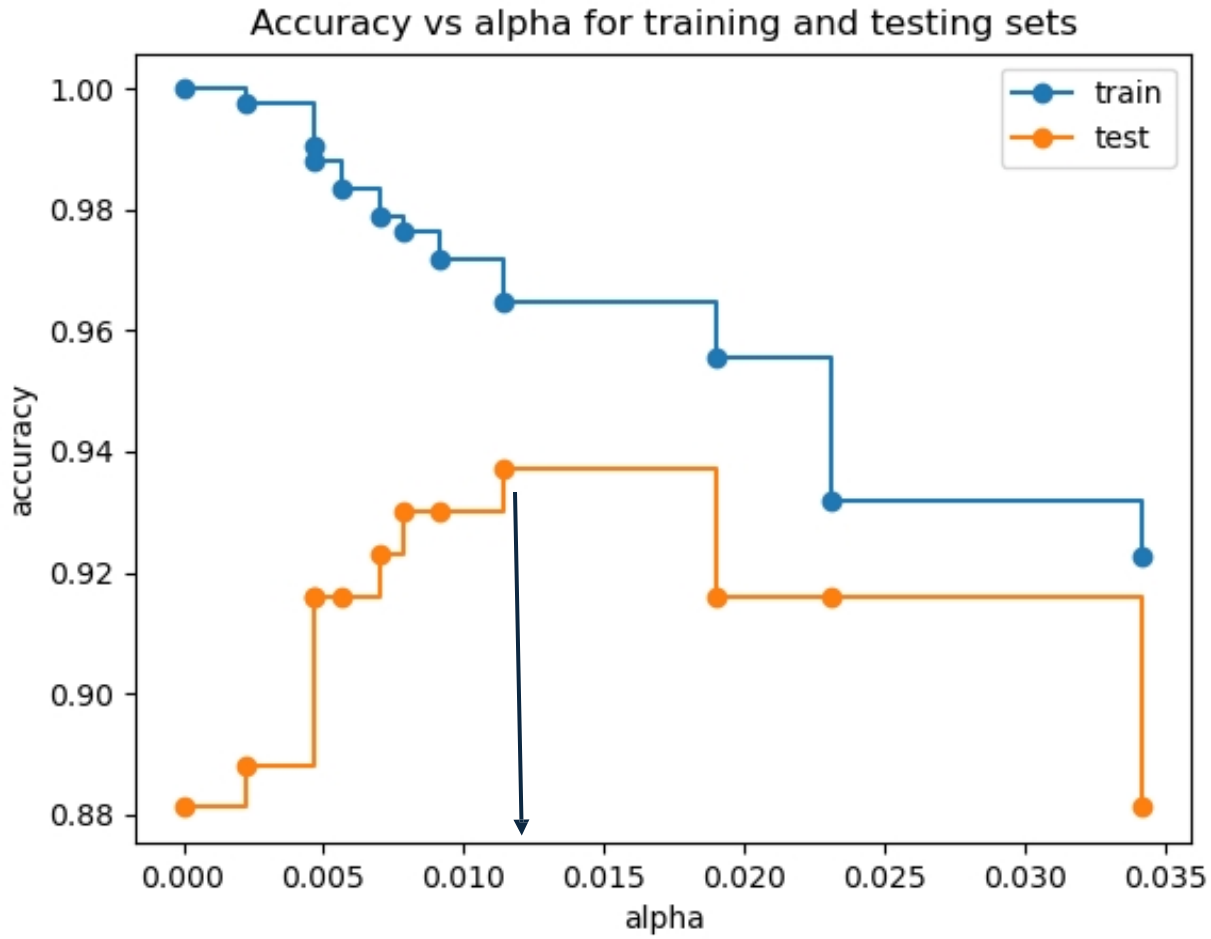
- **Pergunta:**
  - **Como escolher o valor de α ?**



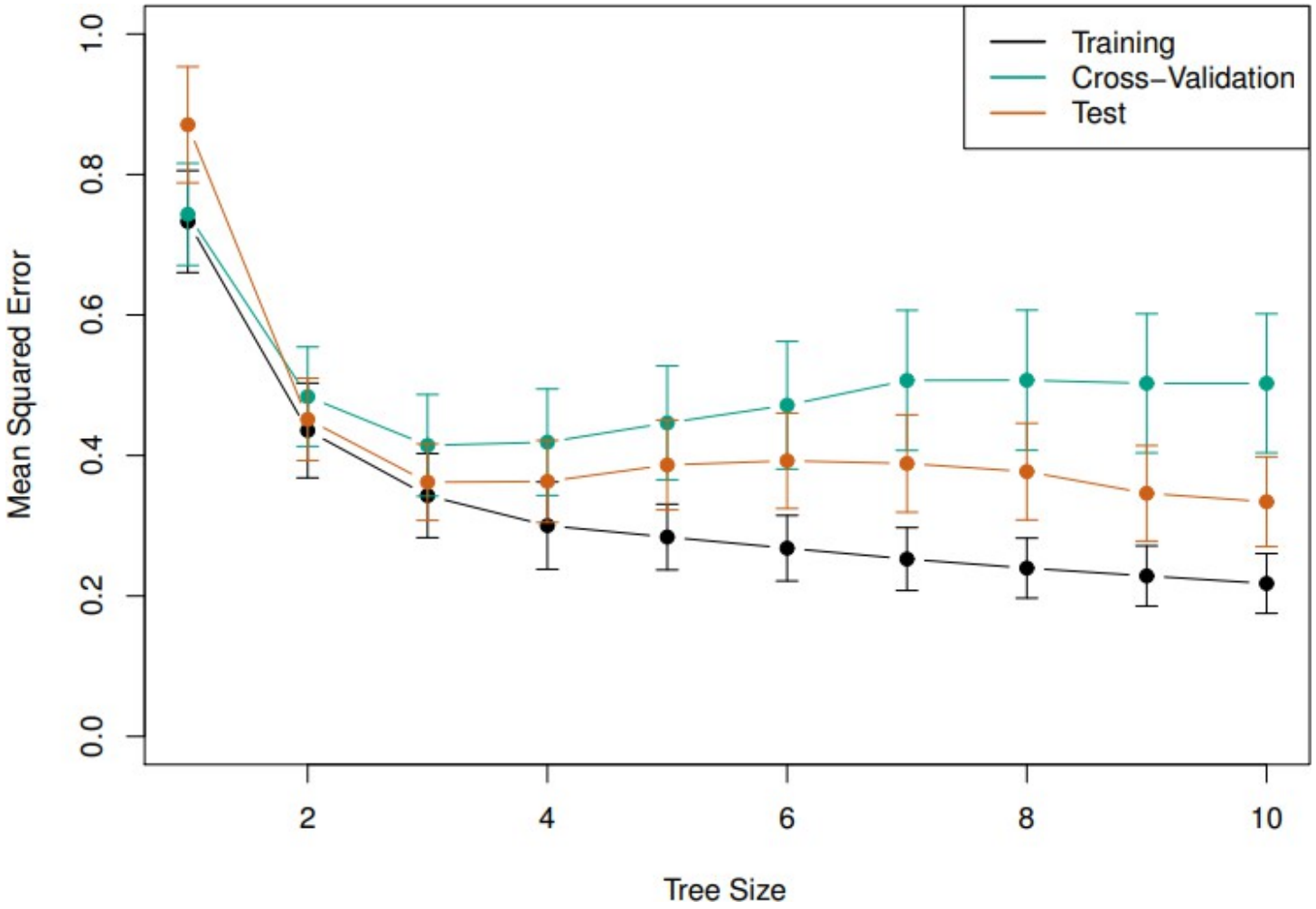
Árvores de decisão: Poda por complexidade de custos

- **Pergunta:**
  - **Como escolher o valor de α?**

- Utilizar a validação cruzada!



Árvores de decisão: Profundidade vs Erro



- Parece que uma pequena árvore de 3 folhas tem o erro CV mais baixo!

Árvores de decisão: Vantagens

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Interpretabilidade:** fáceis de compreender e interpretar, o que as torna adequadas para explicar o raciocínio subjacente às decisões a não especialistas.
- **Sem pré-processamento de dados:** pode tratar dados numéricos e categóricos sem necessidade de pré-processamento extensivo, como normalização ou escalonamento.
- **Lida com relações não lineares:** pode captar relações não lineares entre características e a variável-alvo sem as modelar explicitamente.
- **Lida com valores em falta:** pode lidar com valores em falta excluindo-os simplesmente do processo de divisão, tornando-os robustos para dados em falta.
- **Importância das características:** fornece uma medida da importância das características, que pode ajudar a identificar as características mais influentes no conjunto de dados.
- **Eficiência:** têm um tempo de formação relativamente rápido, especialmente para conjuntos de dados mais pequenos, em comparação com algoritmos mais complexos.

Árvores de decisão: Limitações

UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Sobreajuste:** são propensos ao sobreajuste, especialmente quando crescem demasiado em profundidade ou não são podados corretamente, captando ruído ou padrões específicos nos dados de treino que não generalizam bem.

- **Instabilidade:** pequenas variações nos dados podem levar a estruturas de árvore diferentes, tornando as árvores de decisão instáveis e sensíveis a alterações nos dados de formação.

- **Enviesamento para as classes dominantes:** em tarefas de classificação com classes desequilibradas, as árvores de decisão podem apresentar um enviesamento para as classes dominantes, levando a um fraco desempenho nas classes minoritárias.

- **Natureza gulosa:** utilizar uma abordagem gulosa, de cima para baixo, para particionar recursivamente o espaço de características, o que pode nem sempre conduzir à estrutura de árvore globalmente óptima.

Recursos

- Koning, M., C Smith, C. (2017). Árvores de decisão e florestas aleatórias. Publicado de forma independente.

- https://www.youtube.com/watch?v=_L39rN6gz7Y

- https://www.youtube.com/watch?v=_L39rN6gz7Y

- https://www.youtube.com/watch?v=wpNl-JwwplA

- https://www.youtube.com/watch?v=D0efHEJsfHo