Preparando terreno no mundo da **Machine Learning**

Um passo de cada vez!

AGENDA

- K-NN
- Árvore de Decisão

K-NN

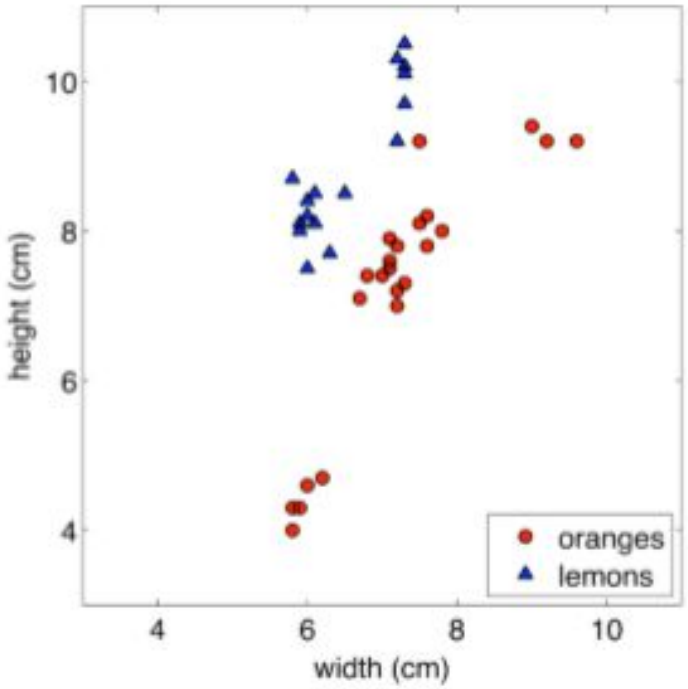
“Diga-me com quem tu andas, que eu digo quem tu és”

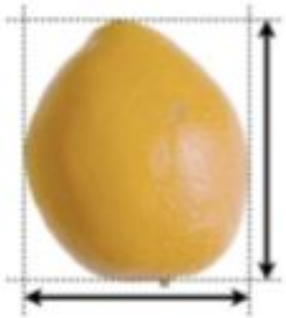
Revisando...

- O que são parâmetros?
  - “Concentram” o que foi aprendido a partir dos dados.
- O que são hiperparâmetros?
  - Definem o “comportamento geral” do modelo.

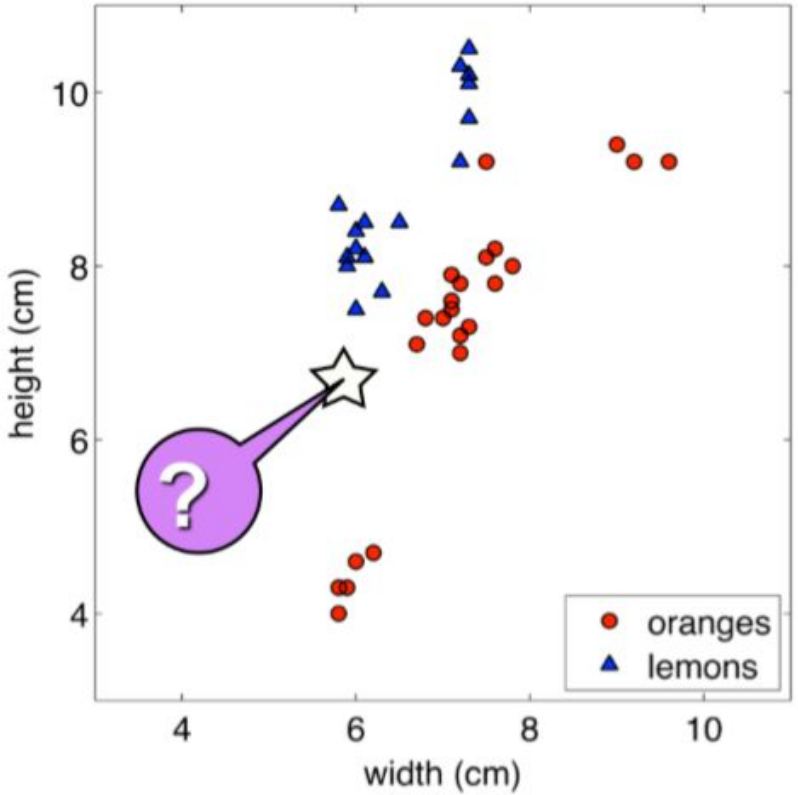
Um modelo pode não ter parâmetros?

Aprendizado baseado em instâncias





Aprendizado baseado em instâncias

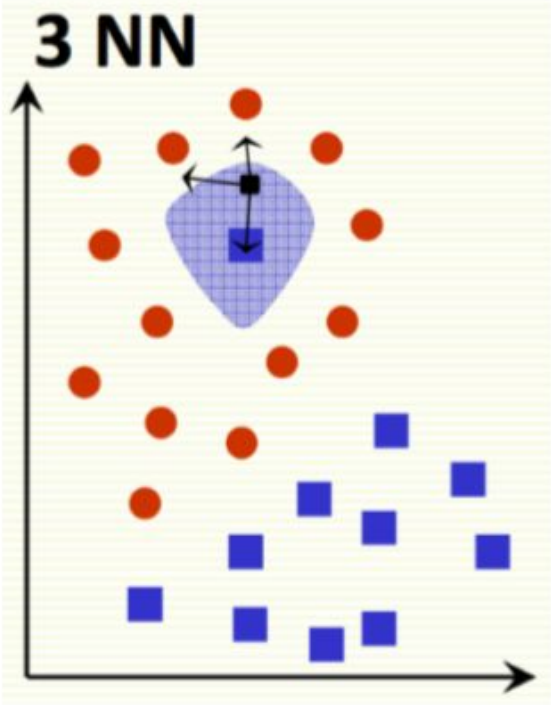


Aprendizado baseado em instâncias

- ➔ Modelos não-paramétricos
- ➔ Não possuem uma etapa de treinamento
- ➔ Predições são baseadas nas instâncias de treinamento mais próximas do padrão de teste.
- ➔ Precisam armazenar os dados de treinamento para realizar predições.

Nearest Neighbors - Vizinhos mais próximos

➔ Ideia: Usar os vizinhos mais próximo



Nearest Neighbors - Vizinhos mais próximos

K Nearest Neighbors (KNN) para classificação

1. Encontre os K padrões $\boldsymbol{x}_k, k \in \{1, \cdots, K\}$ mais próximo do padrão de teste $\boldsymbol{x}_*$:

$$\boldsymbol{x}_{\text{NN}} = \arg \min_{\boldsymbol{x}_i \in \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}} d(\boldsymbol{x}_i, \boldsymbol{x}_*).$$

2. Retorne a classe mais comum entre os K padrões encontrados.

KNN - Observações

- ➔ Valores **muito altos de K** podem incluir informação de dados muito distantes e simplificam a região de decisão.
- ➔ Valores **muito baixos de K** podem ser sensíveis a ruído e tornam a região de decisão mais complexa.

KNN - Observações

→ Distância Euclidiana:

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{d=1}^{D}(x_{id} - x_{jd})^2}.$$

→ Distância de Manhattan:

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_1 = \sum_{d=1}^{D}|x_{id} - x_{jd}|.$$

→ Distância de Mahalanobis:

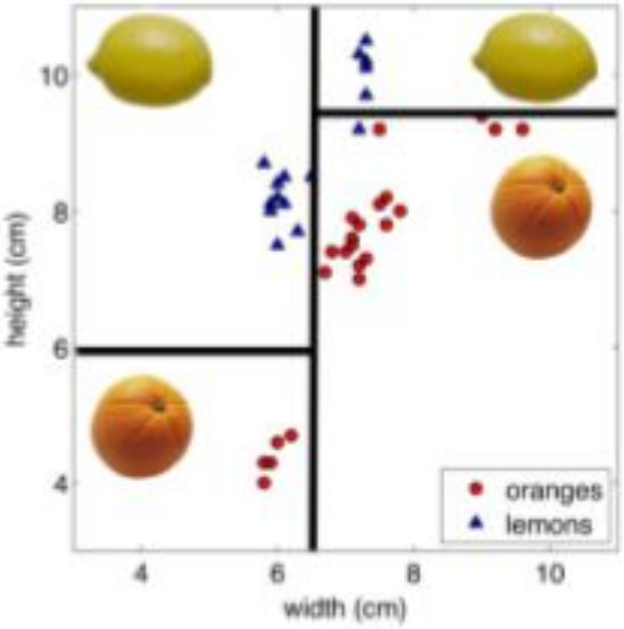
$$d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)},$$

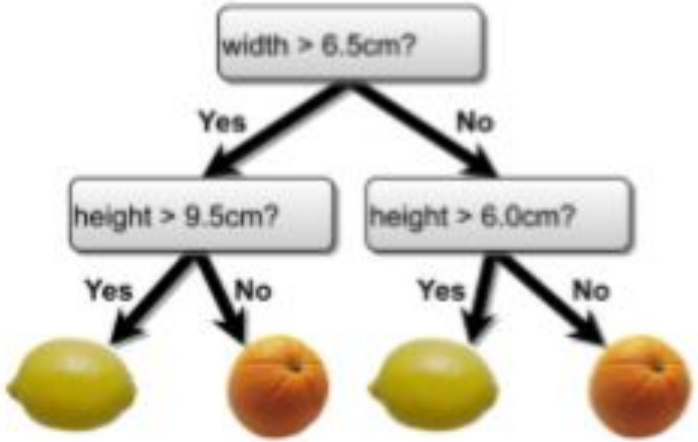
em que $\boldsymbol{\Sigma}$ é matriz de covariância dos dados de treinamento.

Árvore de Decisão

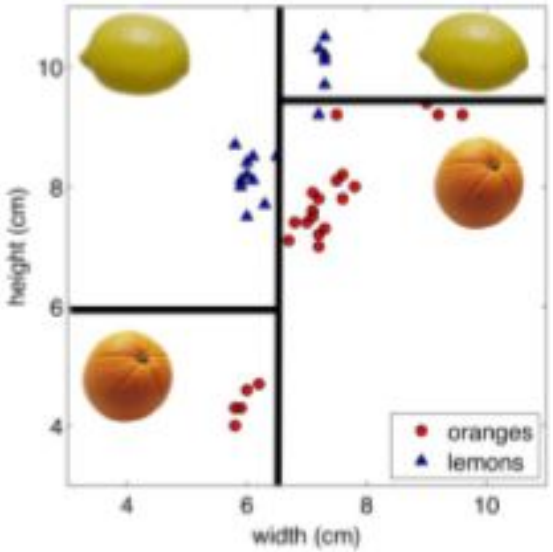
Árvores de Decisão

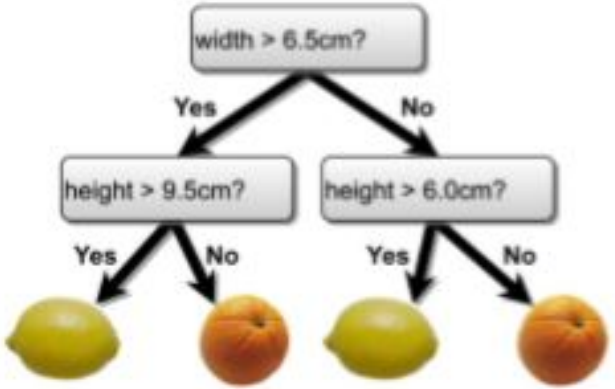
➔ Ideia: Usamos regras lógicas (se-então) para separar as frutas





Árvores de Decisão





- ➔ Nós internos verificam os valores dos atributos
- ➔ Ramificação é feita de acordo com o limiar (threshold) escolhido.
- ➔ Nós terminais (folhas) estão associados a uma classe específica.

Árvores de Decisão

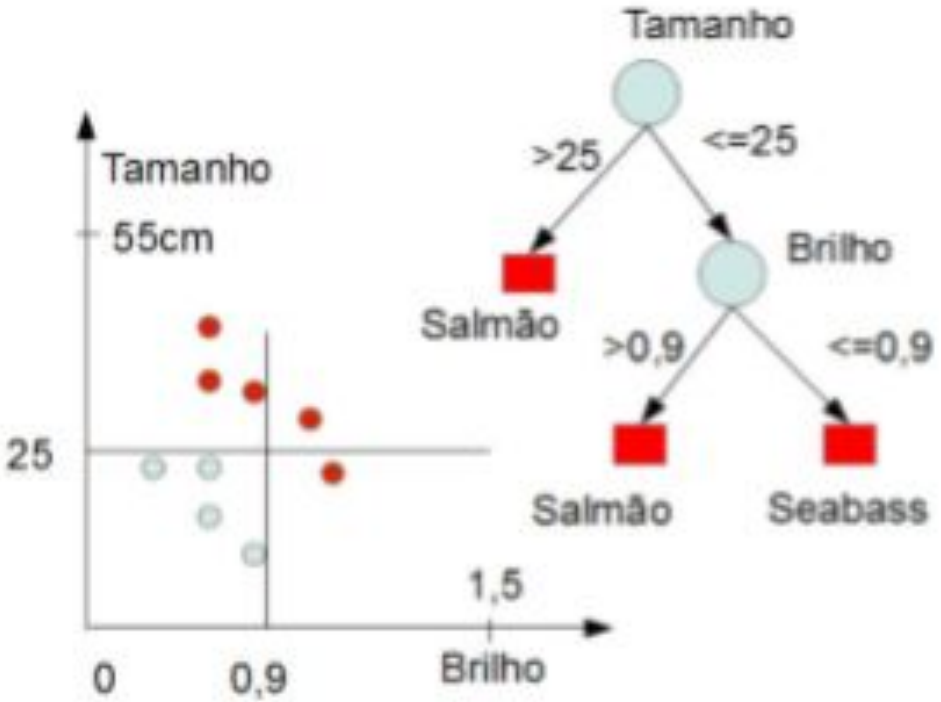
Predições usando árvores de decisão

Dada uma árvore de decisão já existente e um padrão de teste:

1. Inicie no nó mais superior (raiz da árvore).
2. Considere o atributo do nó em questão.
3. Verifique o limiar do nó atual e siga um dos ramos existentes.
4. Caso chegue em um nó terminal (folha), retorne a saída associada. Caso contrário, desça para o próximo nó interno e continue.

Árvores de Decisão

Brilho	Tamanho	Classe
1.2	23	Salmão
1.1	30	Salmão
0.9	36	Salmão
0.8	45	Salmão
0.8	38	Salmão
0.9	15	Seabass
0.8	20	Seabass
0.8	25	Seabass
0.7	25	Seabass



Árvores de Decisão

- **Problema:** Como obter a árvore de decisão automaticamente a partir dos dados de treinamento?

- **Problema:** Construir a menor árvore (mais concisa) é um problema NP completo.

- **Ideia:** Seguir uma abordagem heurística gulosa (greedy):
  - Comece de uma árvore vazia;
  - Encontre o melhor atributo para realizar uma divisão;
  - Repita recursivamente o passo anterior para o próximo nó até encontrar uma folha.

Árvores de Decisão

- **Problema:** Como encontrar o melhor atributo para realizar a divisão?
- **Ideia:** Usar índices de pureza.
  - **Pureza máxima:** Somente exemplos de uma mesma classe em uma folha.
  - **Pureza mínima:** Quantidades iguais de todas as classes em uma folha.
  - Distribuições intermediárias devem ser quantificadas por um índice.

Árvores de Decisão

Entropia (teoria da informação)

- Taxa de informação gerada por uma fonte de dados.
- Dados improváveis fornecem mais informação (mais "surpresa").
- Maior a pureza, menor a entropia, sendo quantificada por:

$$H = -\sum_{k} P(C_k) \log_2 P(C_k)$$

Árvores de Decisão

Índice (ou impureza de) Gini

- Frequência em que um exemplo aleatório seja incorretamente classificado.
- Pode ser quantificada por:

$$G = \sum_k P(C_k)(1 - P(C_k)) = 1 - \sum_k P(C_k)^2$$

Árvores de Decisão

Treinamento guloso (*greedy*) de árvores de decisão

1. Calcule o índice de pureza/impureza do nó atual (nó pai);
2. Crie ramificações a partir de um atributo e um limiar candidatos;
3. Escolha a ramificação com maior queda de impureza (maior pureza) em relação ao nó pai;
4. Para cada nó criado pela ramificação escolhida:
   - Se não houver exemplos de treinamento, retorne a classe mais comum no nó pai.
   - Se todos os exemplos são de uma mesma classe, retorne-a.
   - Caso contrário, retorne ao primeiro passo.

HANDS-ON!