
INSIGHT

ÁRVORES DE DECISÃO

Hinessa Caminha - Abril 2019

INSIGHT

DEPARTAMENTO DE COMPUTAÇÃO

UNIVERSIDADE FEDERAL DO CEARÁ

AGENDA

1. Introdução
2. Definição
3. Selecionando o *Best Split*
4. Conclusão
5. Referências

c

1. INTRODUÇÃO

Uma breve motivação

INTRODUÇÃO

- Refrescando a memória...

> **Classificação:** tarefa de aprendizagem de uma função **f** que mapeie cada atributo **x** a uma classe pré-definida **y**.

- Os modelos de classificação podem ser úteis para descrição e sumarização de *datasets*, bem como classificar novas instâncias.

INTRODUÇÃO

- Considere o seguinte dataset:

Nome	Temperatura corporal	Cobertura da pele	Dá à luz	Criatura Aquática	Criatura aérea	Possui pernas	Hiberna	Classe
humano	sangue quente	pêlos	sim	não	não	sim	não	mamífero
cobra	sangue frio	escamas	não	não	não	não	sim	réptil
salmão	sangue frio	escamas	não	sim	não	não	não	peixe
baleia	sangue quente	pêlos	sim	sim	não	não	não	mamífero
pombo	sangue quente	penas	não	não	sim	sim	não	pássaro

INTRODUÇÃO

- Considere o seguinte dataset:

Nome	Temperatura corporal	Cobertura da pele	Dá à luz	Criatura Aquática	Criatura aérea	Possui pernas	Hiberna	Classe
humano	sangue quente	pêlos	sim	não	não	sim	não	mamífero
cobra	sangue frio	escamas	não	não	não	não	sim	réptil
salmão	sangue frio	escamas	não	sim	não	não	não	peixe
baleia	sangue quente	pêlos	sim	sim	não	não	não	mamífero
pombo	sangue quente	penas	não	não	sim	sim	não	pássaro

c

2. DEFINIÇÃO

O que são as árvores de decisão?

DEFINIÇÃO

- As árvores de decisão são um método de aprendizagem de máquina **supervisionado** e **não-paramétrico**.

- As principal vantagem da árvore de decisão é a sua **simplicidade**.

DEFINIÇÃO

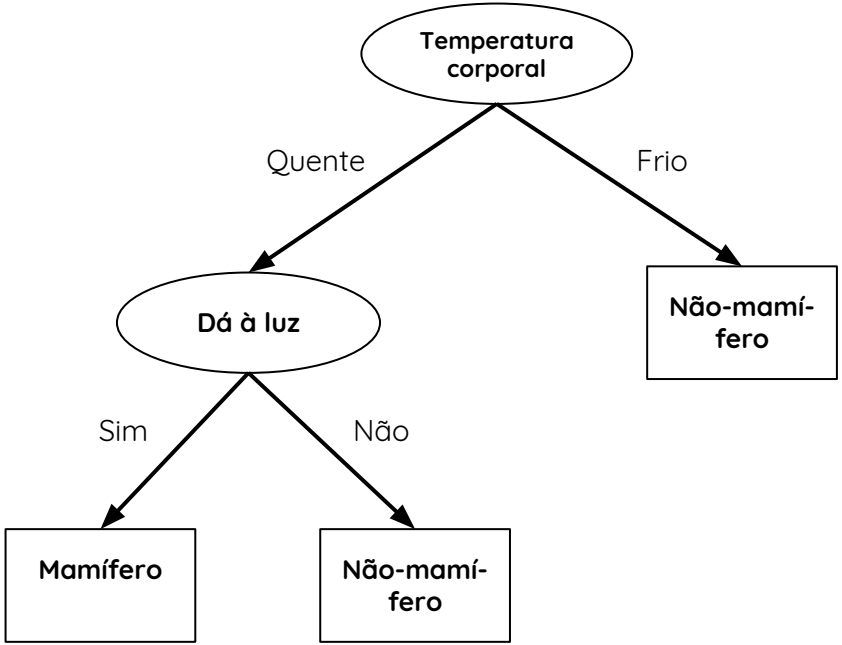

Figura 1 - Árvore de decisão do dataset de animais.

DEFINIÇÃO

- Uma árvore pode pode possuir três tipos de nós distintos:
  - **Raíz:** não possui aresta de entrada e tem zero ou mais arestas de saída;
  - **Interno:** possui exatamente uma aresta de entrada e uma ou duas arestas de saída;
  - **Folha:** possui exatamente uma aresta de entrada e nenhuma aresta de saída.

DEFINIÇÃO

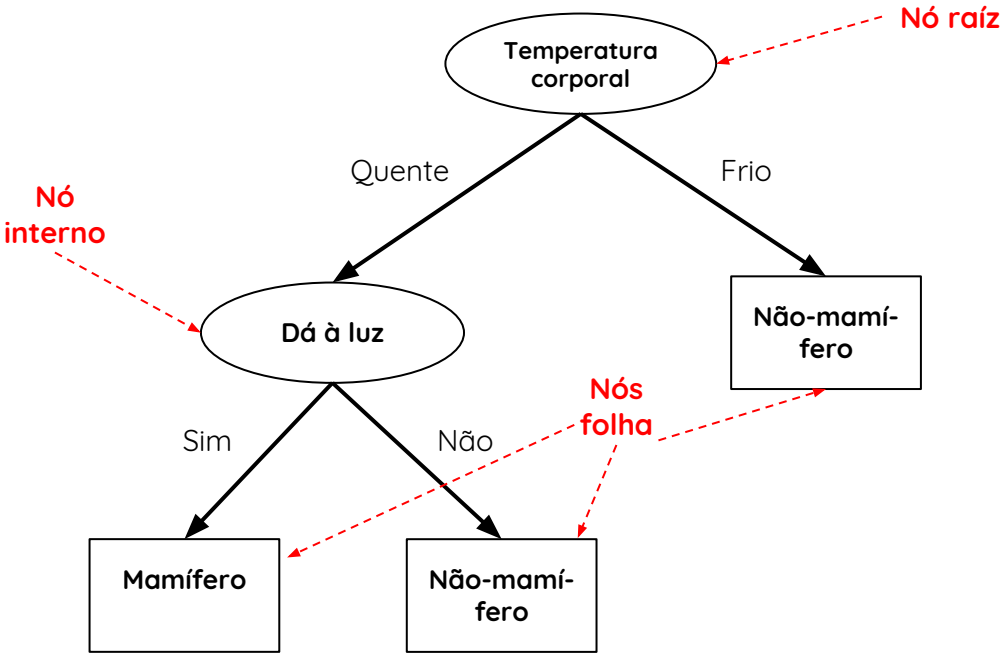

Figura 1 - Árvore de decisão do dataset de animais.

c

3. SELECIONANDO O BEST SPLIT

Como escolher o best split?

SELECIONANDO O BEST SPLIT

- Existem diversas métricas que podem ser utilizadas para definir a melhor forma de dividir os dados;
- Essas medidas são definidas em termos da distribuição da classe dos dados antes e depois do *split*.
- Tais medidas calculam o grau de impureza do nós filhos:
  - Quanto menor o grau de impureza, mais enviesados estão os dados.

SELECIONANDO O BEST SPLIT

- Medidas de impureza mais comuns:
  - Entropia e Índice Gini.

Entropia

$$E(t) = \Sigma\ -pi\ \log_2 pi$$

Índice Gini

$$G(t) = 1 - \Sigma\ [pi]^2$$

SELECIONANDO O BEST SPLIT

- Medidas de impureza mais comuns:
  - Entropia e Índice Gini.

Entropia

$$E(t) = \Sigma\ -pi\ \log_2 pi$$

Nó t

Probabilidade de ocorrência da classe

Atingem o valor máximo quando as probabilidades de todas as classes são iguais!

Índice Gini

$$G(t) = 1 - \Sigma\ [pi]^2$$

SELECIONANDO O BEST SPLIT

Mamíferos: 2 registros
Não-mamíferos: 3 registros

- Calculando a entropia do *dataset* de animais:

Nome	Temperatura corporal	Cobertura da pele	Dá à luz	Criatura Aquática	Criatura aérea	Possui pernas	Hiberna	Classe
humano	sangue quente	pêlos	sim	não	não	sim	não	**mamífero**
cobra	sangue frio	escamas	não	não	não	não	sim	réptil
salmão	sangue frio	escamas	não	sim	não	não	não	peixe
baleia	sangue quente	pêlos	sim	sim	não	não	não	**mamífero**
pombo	sangue quente	penas	não	não	sim	sim	não	pássaro

SELECIONANDO O BEST SPLIT

Mamíferos: 2 registros
Não-mamíferos: 3 registros

- Calculando a entropia do *dataset* de animais:

  - $E(t) = \Sigma\ -pi\ \log_2 pi$
  - $E(2,3) = -(\frac{2}{5})\ \log_2 (\frac{2}{5}) - (\frac{3}{5})\ \log_2\ (\frac{3}{5})$
  - $E(2,3) = 0.4\ \log_2 0.4 - 0.6\ \log_2 0.6$
  - $E(2,3) = 0.971$

SELECIONANDO O BEST SPLIT

Mamíferos: 2 registros
Não-mamíferos: 3 registros

- Calculando a entropia do *dataset* de animais:
  - **$E(t) = \Sigma -pi \log_2 pi$**
  - **$E(2,3) = -(\frac{2}{5}) \log_2 (\frac{2}{5}) - (\frac{3}{5}) \log_2 (\frac{3}{5})$**
  - **$E(2,3) = 0.4 \log_2 0.4 - 0.6 \log_2 0.6$**
  - **$E(2,3) = 0.971$**

- O alto valor de entropia nos indica que os dados não estão enviesados em uma única classe, logo os atributos escolhidos para split são bons!

c

4. CONCLUSÃO

Juntando tudo...

CONCLUSÃO

- Método **não-paramétrico**;
- Encontrar uma árvore ótima é **NP-Difícil**;
- São relativamente **fáceis** de interpretar;
- Uma vez construída, a árvore classifica novas instâncias em tempo **O(w)**, com w sendo a profundidade da árvore;
- Atributos altamente correlacionados não influenciam na acurácia da árvore, entretanto, atributos com pouca relevância podem expandir a árvore mais do que o necessário!
- As medidas de impureza causam **pouco impacto** no desempenho das árvores de decisão.

REFERÊNCIAS

- **TAN, Pang-Ning**. Introduction to data mining. Pearson Education India, 2018.

OBRIGADA!

Dúvidas?

Você pode me encontrar em

- hinessa@insightlab.ufc.br