Introdução à Ciência dos Dados Aula 03

Gabriel Pacheco
Bacharel em Sistemas de Informação (UFMG)
Mestrando em Ciência da Computação (UFMG)
Trabalha em Digicade desde 2016

Curiosidade

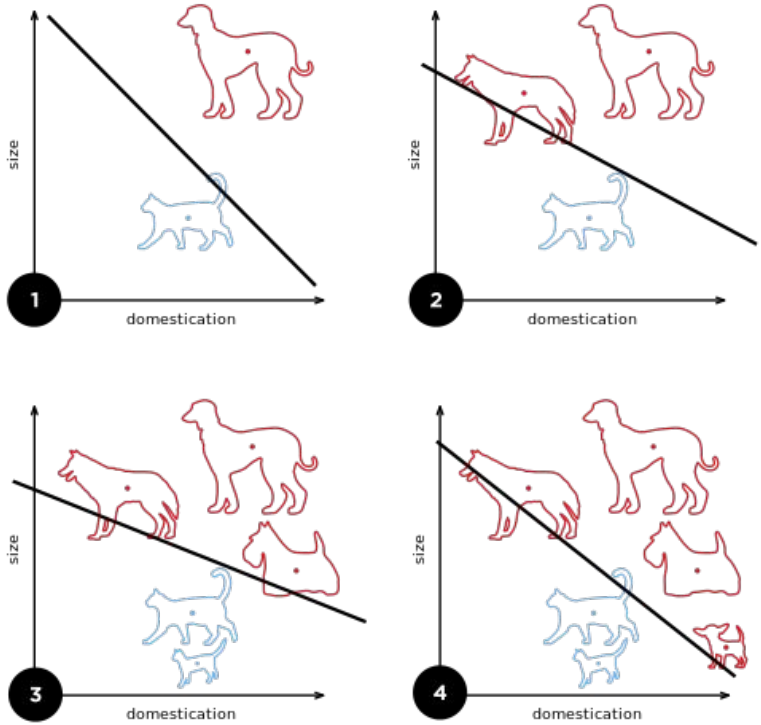
- https://quickdraw.withgoogle.com/
- https://www.tecmundo.com.br/software/151406-novo-profissional-inteligencia-artificial.htm

Cronograma

- Revisão
- Classificação
- Exemplo de análise e visualização dos dados
- Exemplo prático

Classificação

- Detecção de fraudes
- Crédito bancário
- Detecção de Spam
- Análise de Sentimentos
- Reconhecimento de imagem
- Reconhecimento de vídeo
- Análise de ações do mercado financeiro



Técnicas de classificação

- Árvores de decisão (decision trees)
- Florestas aleatórias (random forest)
- Naive Bayes

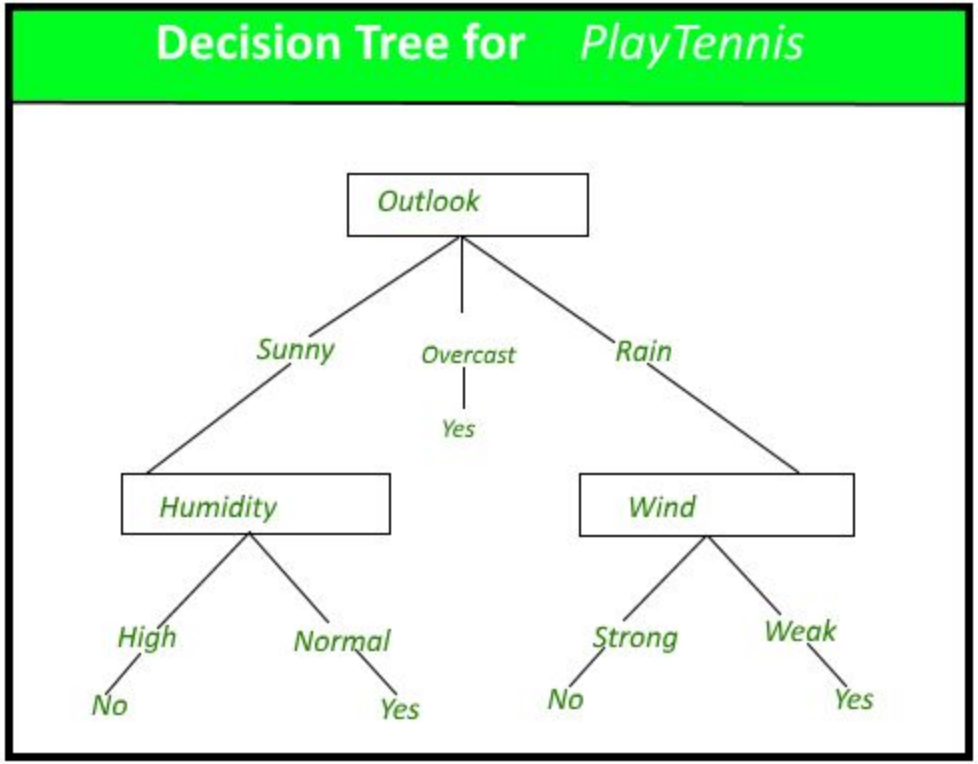
	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Árvores de decisão

- Fácil de criar
- Fácil de utilizar
- Fácil de interpretar

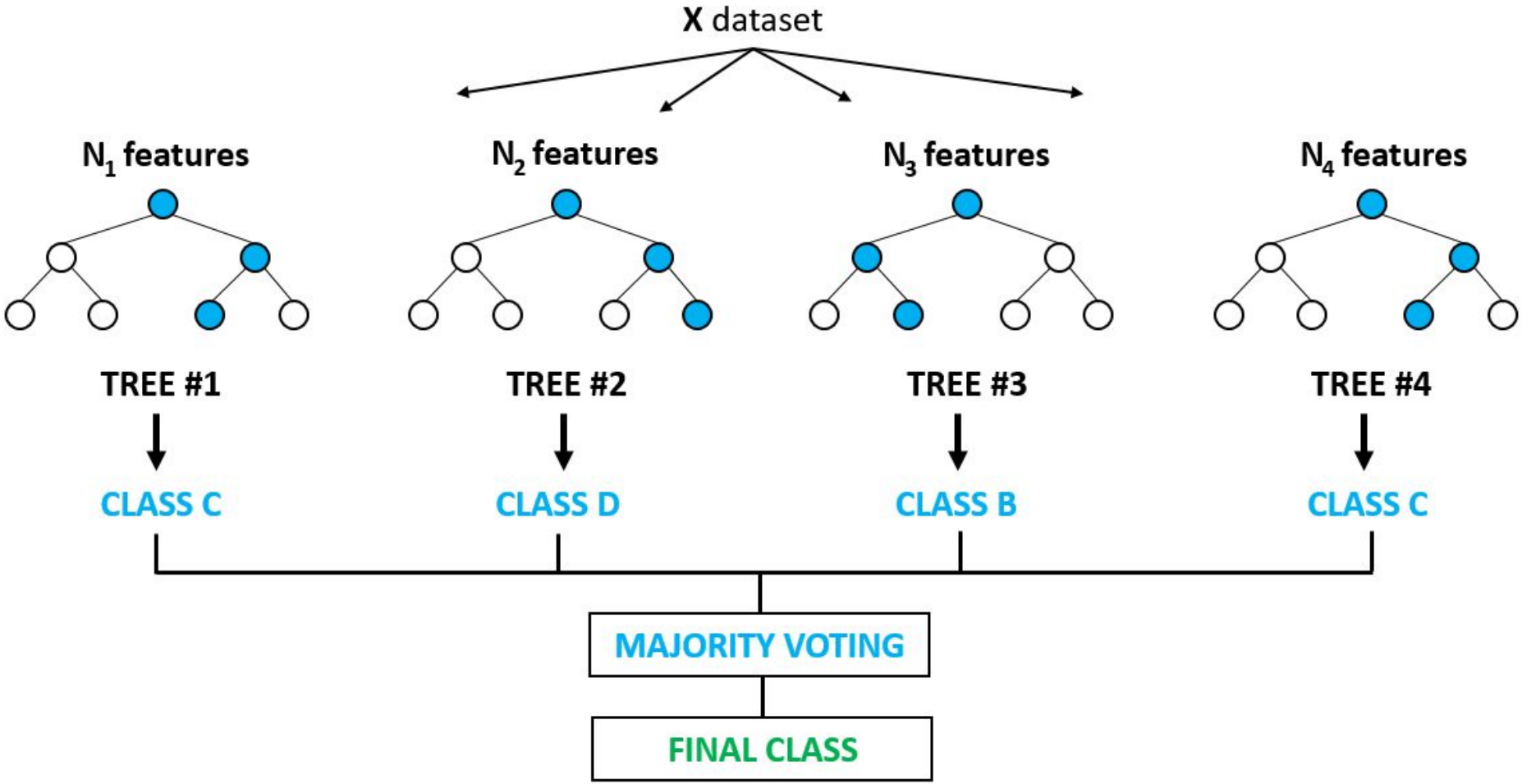
Porém...

- Imprecisa, ou seja, não funciona bem para classificar novas amostras


Decision Tree for PlayTennis
Outlook
Sunny
Overcast
Rain
Yes
Humidity
Wind
High
Normal
Strong
Weak
No
Yes
No
Yes

Florestas aleatórias

- Técnica baseada em árvores de decisão
- Adiciona conceitos de aleatoriedade para aumentar a precisão


X dataset
N1 features
N2 features
N3 features
N4 features
TREE #1
TREE #2
TREE #3
TREE #4
CLASS C
CLASS D
CLASS B
CLASS C
MAJORITY VOTING
FINAL CLASS

Naive Bayes

- Simples de entender
- Classificador probabilístico, assume que as variáveis são independentes
- Categorização de texto, detecção de spam (frequência de palavras como atributos)
- Diagnóstico médico automático

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

Naive Bayes

- Hoje = (Sunny, Hot, Normal, False)

$$P(Yes|today) \propto \frac{2}{9}.\frac{2}{9}.\frac{6}{9}.\frac{6}{9}.\frac{9}{14} \approx 0.0141$$

$$P(No|today) \propto \frac{3}{5}.\frac{2}{5}.\frac{1}{5}.\frac{2}{5}.\frac{5}{14} \approx 0.0068$$

$$P(Yes|today) = \frac{0.0141}{0.0141+0.0068} = 0.67$$

$$P(No|today) = \frac{0.0068}{0.0141+0.0068} = 0.33$$

Análise e visualização de dados: What is the secret of academic success?

- https://www.kaggle.com/datasets
- https://www.kaggle.com/uciml/student-alcohol-consumption
- https://www.kaggle.com/hely333/what-is-the-secret-of-academic-success

Iris Species

- Toy example
- Serve para aplicarmos técnicas de regressão, classificação e agrupamento
- https://www.kaggle.com/uciml/iris

Exemplo prático: Diagnosis of COVID-19 and its clinical spectrum

- Desafio!
- Grande quantidade de atributos
- Transformações necessárias de acordo com o modelo escolhido
- https://www.kaggle.com/einsteindata4u/covid19
- https://www.kaggle.com/caesarlupum/brazil-against-the-advance-of-covid-19

Tutoriais

- https://www.geeksforgeeks.org/naive-bayes-classifiers/

Encerramento

- Materiais do curso: https://github.com/gapacheco/ICD.git

Contato

- Email: gacampacheco@gmail.com
- Linkedin: www.linkedin.com/in/gabriel-oc-pacheco