uni1500

CIENTISTA DE DADOS

| Nov 22

CLASSIFICAÇÃO

Modelos de Classificação

- Técnica onde categoriza-se os dados em um determinado número de classes.
- O principal objetivo de um problema de classificação é identificar a categoria/classe na qual um novo dado se enquadrará.
- Os algoritmos de classificação usam dados de treinamento de entrada para prever a probabilidade de que os dados subsequentes caiam em uma das categorias predeterminadas.
- Ex: filtrar e-mails em “spam” ou “não spam”.

CLASSIFICAÇÃO

Terminologias

Classificador

Algoritmo que mapeia os dados de entrada para uma categoria específica.

Modelo de classificação

Tenta tirar conclusões dos valores de entrada fornecidos para treinamento. Irá prever os categorias para os novos dados.

Feature

Propriedade individual mensurável de um fenômeno que está sendo observado.

Classificação Binária

Classificação com dois resultados possíveis.
Ex: Classificação de spam (Spam/Não é spam).

Classificação multiclasse

Mais de duas classes mas apenas um rótulo de destino.
Ex: Classificação de um animal (gato, cachorro ou réptil).

Classificação multi-rótulo

Cada amostra irá para um conjunto de rótulos de destino.
Ex: Artigo de notícias (esportes, pessoa e local).

CLASSIFICAÇÃO

Principais Algoritmos

- Regressão Logística
- Naive Bayes
- K-ésimo vizinho mais próximo
- Árvore de decisão
- Support Vector Machines

CLASSIFICAÇÃO

Regressão Logística

Usado para prever um resultado binário: ou algo acontece ou não.

Ex: Sim/Não; Aprovado/Reprovado; Vivo/Morto.

As probabilidades que descrevem os possíveis resultados são modeladas usando uma função logística.

Vantagens

- Simples e eficiente;
- Baixa variância;
- Fornece score de probabilidade para observações.

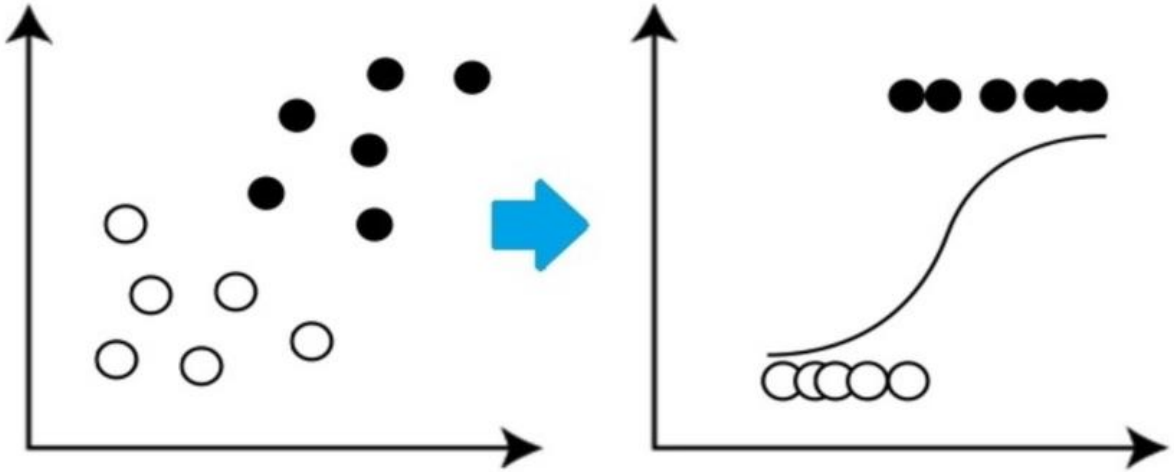
Desvantagens

- Variável target é binária;
- Assume que todos os preditores são independentes uns dos outros;
- Assume que os dados estão livres de valores ausentes.

CLASSIFICAÇÃO

Regressão Logística – Exemplo Visual

Modelo de Regressão Logística que prevê a probabilidade de chuva usando todos os recursos do banco de dados.



CLASSIFICAÇÃO

Naive Bayes

Tem como base o teorema de Bayes que descreve como a probabilidade de um evento é avaliada com base no conhecimento prévio das condições que podem estar relacionadas ao evento.

Vantagens

- Muito rápido;
- Pode ser usado para resolver problemas de previsão multiclasse.

Desvantagens

- Assume que todos os recursos são independentes;
- Conhecido por não ser um bom estimador.

CLASSIFICAÇÃO

Naive Bayes – Exemplo Visual

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table

Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table

Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

CLASSIFICAÇÃO

K-ésimo vizinho mais próximo

Classifica os novos pontos de dados dependendo da classe da maioria dos pontos de dados entre os K vizinhos, onde K é o número de vizinhos a serem considerados.

KNN captura a ideia de similaridade (distância ou proximidade) com fórmulas matemáticas de distância (euclidiana, distância de Manhattan, etc).

Vantagens

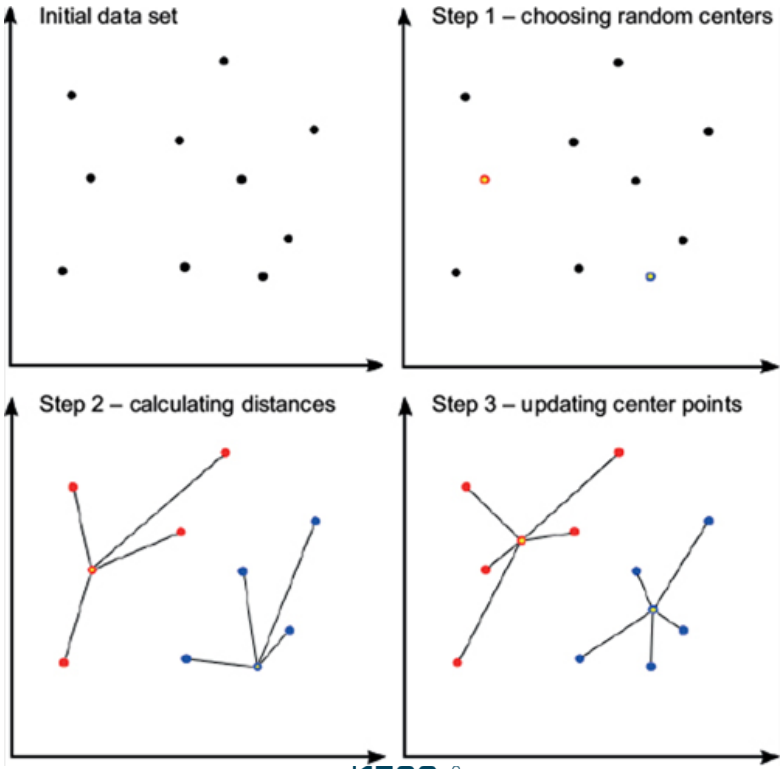
- Simples e fácil de implementar;
- Não há necessidade de construir um modelo;
- Pode ser usado para classificação, regressão e pesquisa -> flexível.

Desvantagens

- Determinar o valor de K;
- Custo computacional alto;
- Influenciado por outliers.

CLASSIFICAÇÃO

K-ésimo vizinho mais próximo – Exemplo Visual



CLASSIFICAÇÃO

Árvore de decisão

Algoritmo de aprendizado supervisionado.

Dividi a população em dois ou mais conjuntos homogêneos com base nos atributos mais significativos, tornando os grupos tão distintos quanto possível.

Vantagens

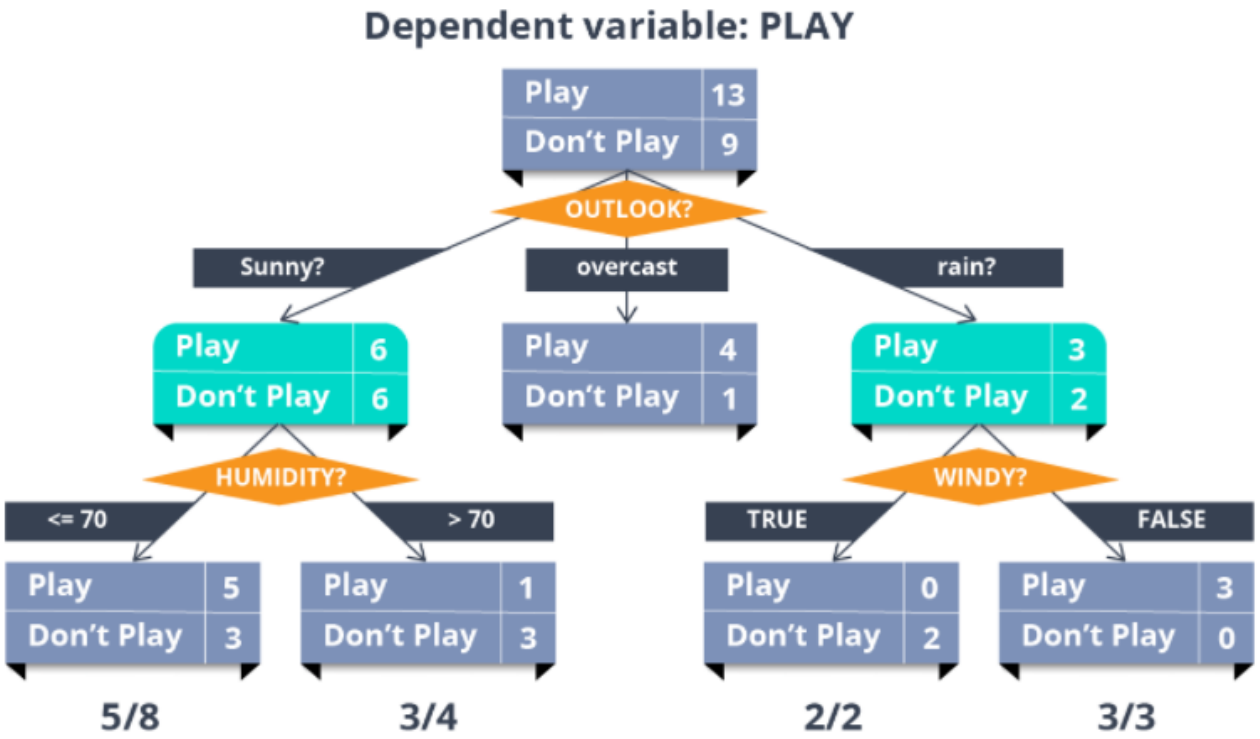
- Simples de entender e visualizar;
- Requer pouca preparação de dados;
- Pode lidar com dados numéricos e categóricos.

Desvantagens

- Pode criar árvores complexas;
- Árvores podem ser instáveis (pequenas variações nos dados podem resultar árvore diferente).

CLASSIFICAÇÃO

Árvore de decisão – Exemplo Visual



CLASSIFICAÇÃO

Support Vector Machines

Representa os dados de treinamento como pontos no espaço separados em categorias por uma lacuna clara que é a mais ampla possível. Novos dados são então mapeados nesse mesmo espaço e previstos para pertencer a uma categoria com base em qual lado da lacuna eles se enquadram.

Vantagens

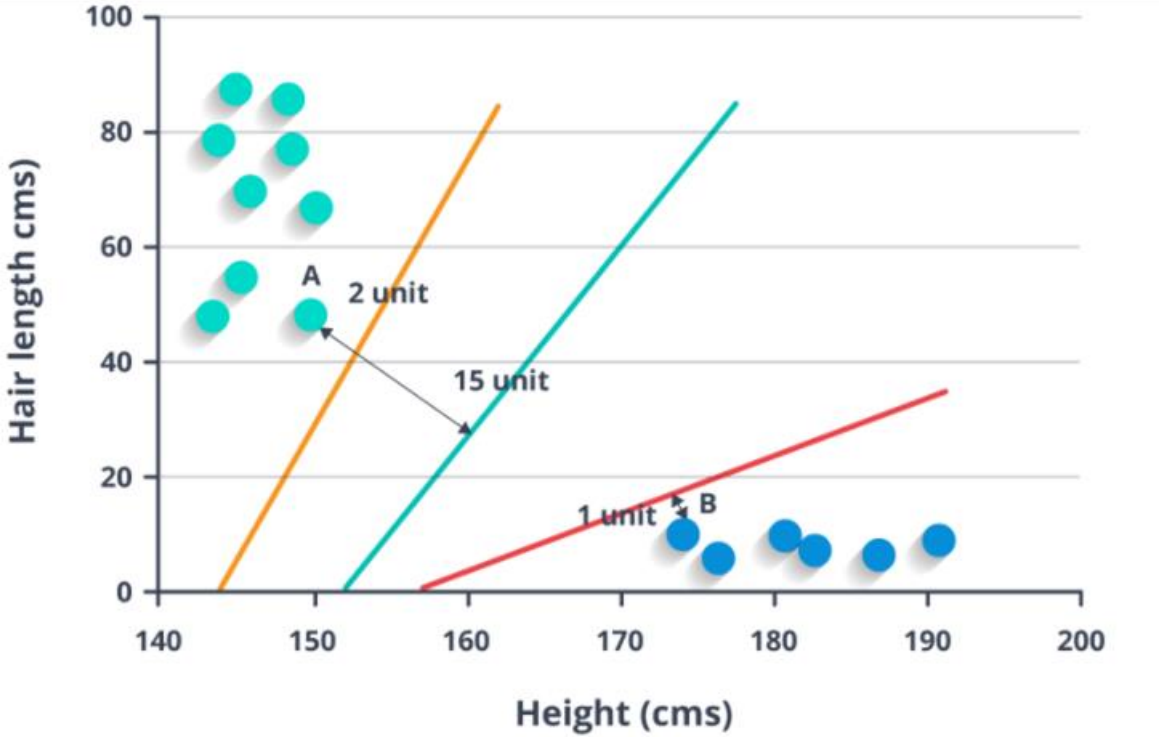
- Funciona relativamente bem quando há uma clara margem de separação entre as classes;
- Mais eficaz em espaços de alta dimensão -> memória.

Desvantagens

- Não é adequado para grandes conjuntos de dados;
- Não funciona muito bem com conjunto de dados com mais ruído (overlapping).

CLASSIFICAÇÃO

Support Vector Machines – Exemplo Visual



uni1500

CLASSIFICAÇÃO

Qual modelo escolher?

- **Identificação do problema:** entender completamente a tarefa em mãos.

Classificação supervisionada -> Regressão Logística, Árvore de Decisão;

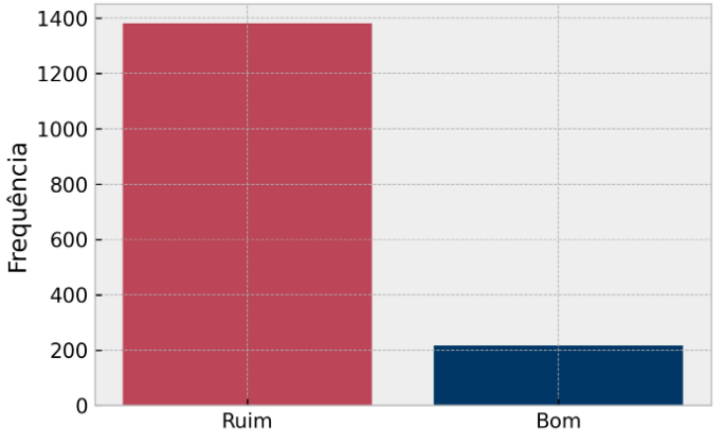
Classificação não supervisionada -> algoritmos de agrupamento.

- **Tamanho do conjunto de dados:** pequeno -> algoritmos de baixa viés/alta variância: como Naive Bayes. Grande ou nº de recursos for alto -> KNN, árvores de decisão.
- **Tempo de treinamento:** SVM e Random Forests podem levar muito tempo para computação. Além disso, maior precisão e grandes conjuntos de dados exigem mais tempo para aprender o padrão. Regressão Logística são mais fáceis de implementar e economizam tempo.
- **Linearidade do Dataset:** Verificar sempre existe uma relação linear entre as variáveis de entrada e as variáveis de destino, pois alguns deles são restritos a conjuntos de dados lineares.

CLASSIFICAÇÃO

Desbalanceamento

- Dados desbalanceados podem ser definidos pela pequena incidência de uma categoria dentro de um dataset (classe minoritária) em comparação com as demais categorias (classes majoritárias).
- Na maioria dos casos, isso faz com que tenhamos muitas informações a respeito das categorias mais incidentes, e menos das minoritárias, o que pode, em muito casos, interferir no workflow padrão de um Cientista de Dados.

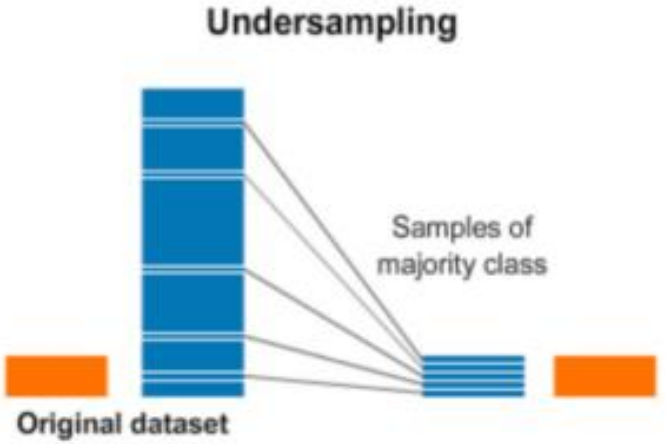


Exemplo de Dados Desbalanceados: Vinho Bom e Ruim

CLASSIFICAÇÃO

Desbalanceamento: Undersampling

- Consiste em reduzir o número de observações da classe majoritária para diminuir a diferença entre as categorias.
- Há duas formas de realizar o Undersampling:
  - Random Undersampling: consiste na retirada aleatória de dados da classe majoritária (o que acarreta em uma perda grave de informação);
  - Utilizar métodos para unir duas ou mais observações de classes majoritárias em apenas uma, o que acarreta em uma menor perda de informação.



CLASSIFICAÇÃO

Desbalanceamento: Oversampling

- Consiste em criar sinteticamente novas observações da classe minoritária, com o objetivo de igualar a proporção das categorias.
- A maneira mais primitiva de fazer um Oversampling é por meio de cópias de dados já existentes na classe minoritária.