
INSIGHT

INSIGHT

ALGORITMO kNN E UMA INTRODUÇÃO AO SCIKIT-LEARN

Disciplina de Garimpagem de Dados [11/10/2017]

DEPARTAMENTO DE COMPUTAÇÃO

UNIVERSIDADE FEDERAL DO CEARÁ

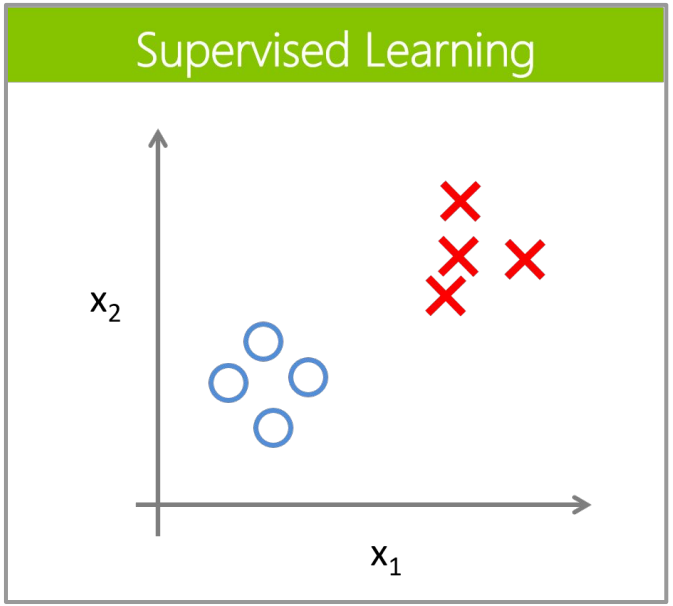
AGENDA

1. Supervisionada vs Não Supervisionada
2. Algoritmos de Classificação
3. k-Nearest Neighbors
4. SCIKIT-LEARN

1. APRENDIZAGEM SUPERVISIONADA E NÃO SUPERVISIONADA

APRENDIZAGEM SUPERVISIONADA

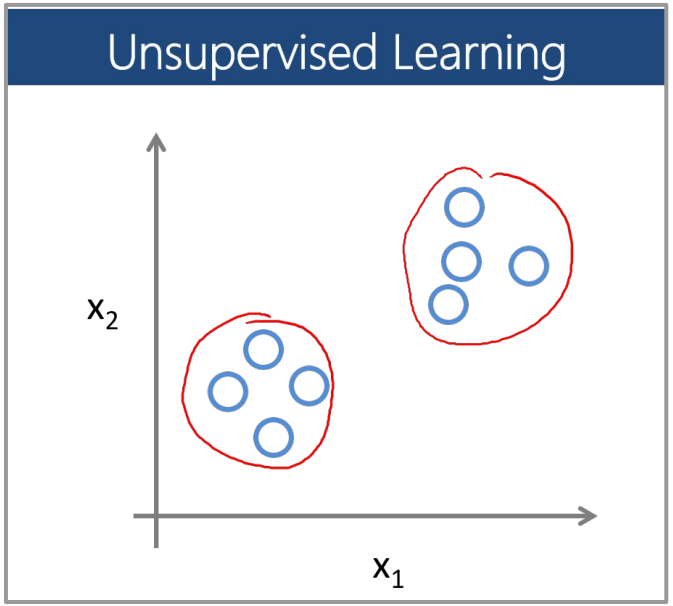
- É necessário um **dataset rotulado** (*labeled*) para **treinamento**
- A partir de uma **análise** desse dataset uma **função** que pode ser usada para **mapear** novos exemplos
- Em um cenário ótimo isso permite que o algoritmo determine corretamente a **classe**



scikit-learn. **Supervised learning**, 2017. [http://scikit-learn.org/stable/tutorial/statistical_inference/supervised_learning.html]

APRENDIZAGEM NÃO SUPERVISIONADA

- É quando uma algoritmo pode **automaticamente** encontrar **padrões** e **relações**
- Baseada na **observação** e **descoberta**
- Não são definidas classes, o algoritmo necessita analisar os dados e reconhecer os padrões por si próprio



scikit-learn. **Unsupervised learning**, 2017. [http://scikit-learn.org/stable/tutorial/statistical_inference/unsupervised_learning.html]

1.
ALGORITMOS DE CLASSIFICAÇÃO

ALGORITMOS DE CLASSIFICAÇÃO

- É uma técnica para prever a associação de grupos para instâncias de dados
- A ideia é **predizer** uma **classe alvo** através da análise de um **dataset de treino**
- Isso pode ser feito quando conseguimos definir as **fronteiras** de cada classe
- As classes são **mutuamente exclusivas**
  - O email é um spam?
  - A transação do cartão de crédito é fraudul a?
  - A fruta é banana, maçã ou uva?
- Classificação **binária** ou **multiclasse**

ALGORITMOS DE CLASSIFICAÇÃO

- As observações **individuais** são analisadas em um conjunto de propriedades **quantificáveis**, conhecidas como **variáveis explicativas** ou ***features***
  - Categóricas (i.e.: A, B, AB ou O)
  - Ordinais (i.e.: grande, médio ou pequeno)
  - Integer-valued (i.e.: número de ocorrências de uma determinada palavra)
  - Real-valued (i.e.: medida da temperatura corporal)
- Resumindo, uma **função matemática** que mapeia dados de entrada para uma categoria/classe/label

2.
k-NEAREST NEIGHBORS

k-NEAREST NEIGHBORS

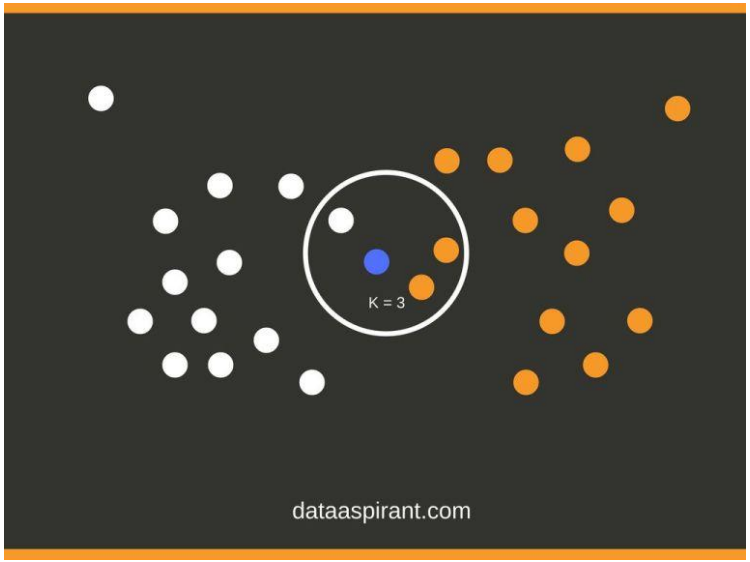
- É um algoritmo de classificação clássico proposto em 1951 e conhecido como **kNN**
- Prever uma **classe alvo** ao encontrar a **classe vizinha mais próxima**
- A classe mais próxima é identificada usando medidas de distância no **espaço de características**, como a **euclidiana**

Duas classes: branco e laranja

Dataset de treino: 26

k: 3

Qual a classe da bola azul?



(1) Calcular a distância entre o exemplo desconhecido e o outros exemplos do conjunto de treinamento.

(2) Identificar os K vizinhos mais próximos.

(3) Utilizar o rótulo da classe dos vizinhos mais próximos para determinar o rótulo de classe do exemplo desconhecido (votação majoritária)

k-NEAREST NEIGHBORS

- Como escolher o valor de **k**?
  - Um **pequeno** valor de k significa que um **ruído** terá uma maior influência sobre o resultado (*overfitting*)
  - Um **grande** valor de k torna o **processamento** muito caro e **derruba** a ideia básica do kNN (pontos próximos podem ter classes semelhantes)
  - É necessário sempre escolher um **valor ímpar** para k, assim evitamos **empates** na votação

k-NEAREST NEIGHBORS

- A **precisão** da classificação utilizando o algoritmo kNN depende fortemente do modelo de dados

- Na maioria das vezes os atributos precisam ser **normalizados** para evitar que as medidas de distância sejam dominadas por um único atributo. Exemplos:
  - Altura de uma pessoa pode variar de 1,20m a 2,10m
  - Peso de uma pessoa pode variar de 40kg a 150kg
  - O salário de uma pessoa podem variar de R$ 800 a R$ 20.000

k-NEAREST NEIGHBORS

- **Vantagens**
  - Técnica simples e facilmente implementada
  - Bastante flexível
  - Em alguns casos apresenta ótimos resultados
  - Não é necessária nenhum novo treino quando um novo dado é adicionado

- **Desvantagens**
  - A precisão pode ser severamente degradada pela presença de ruídos
  - Para cada novo dado, a distância deverá ser calculada entre o dado e todo o dataset de treino

2.
SCIKIT-LEARN

SCIKIT-LEARN

- **Toolbox** de propósito geral para ***machine learning*** em Python
- Prover uma variedade de técnicas supervisionadas e não supervisionadas de machine learning
- Prover também utilitários comuns como ***model selection***, ***feature extraction*** e ***feature selection***
- Scikit-learn fornece uma interface orientada a objetos centrada em torno do conceito de ***Estimator***
  - *def **fit**(train_data)*
  - *def **predict**(test_data)*
- Fornecer uma variedade de ***datasets*** padrões

Scikit-Learn, 2017. [http://scikit-learn.org]

INSTALAÇÃO

Acessar o ***environment* conda** existente:

- $ conda install scikit-learn

Tutorial completo e documentação:

- http://scikit-learn.org/stable/user_guide.html
- http://scikit-learn.org/stable/tutorial/index.html