Aprendizagem automática

KNN
K vizinhos-mais-próximos - classificação e regressão

Luís Rato, Universidade de Évora, 2022/23

Sumário

- Algoritmo
- Fronteira de decisão
- Vantagens, desvantagens e parâmetros
- Regressão
- Algoritmos de regressão
  - KNN
  - Regressão linear, polinomial, SVM, Redes Neuronais, Árvores de decisão, etc…

Algoritmo KNN

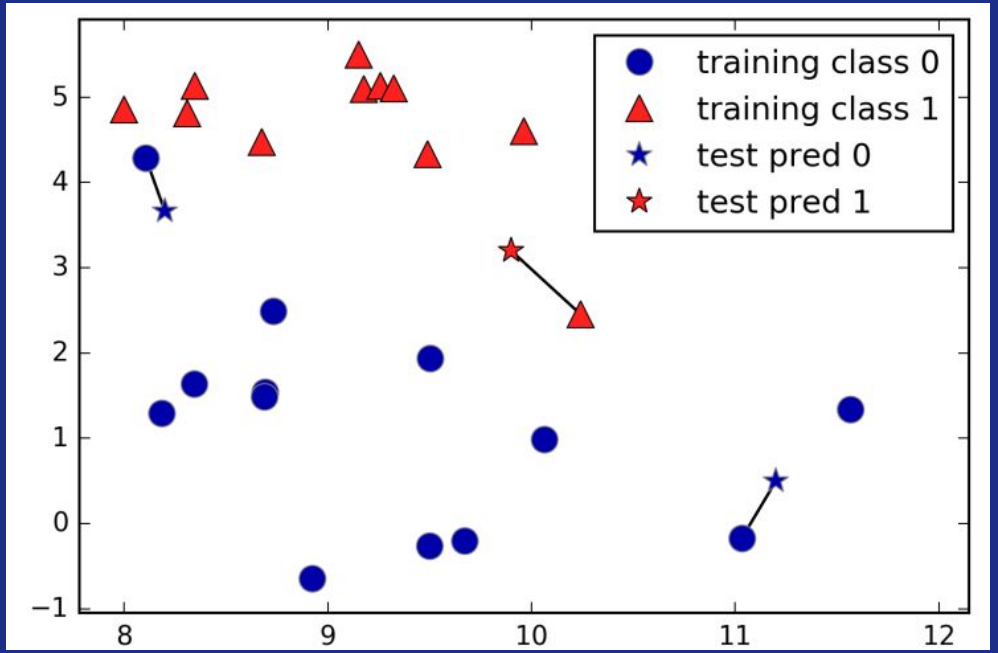
K vizinhos-mais-próximos - classificação

- Construção do modelo
  - Guardar o conjunto de treino
- Previsão de um exemplo
  - Encontrar os K exemplos mais próximos no conjunto treino
  - Atribuir a etiqueta da maioria

- *Lazy learning*

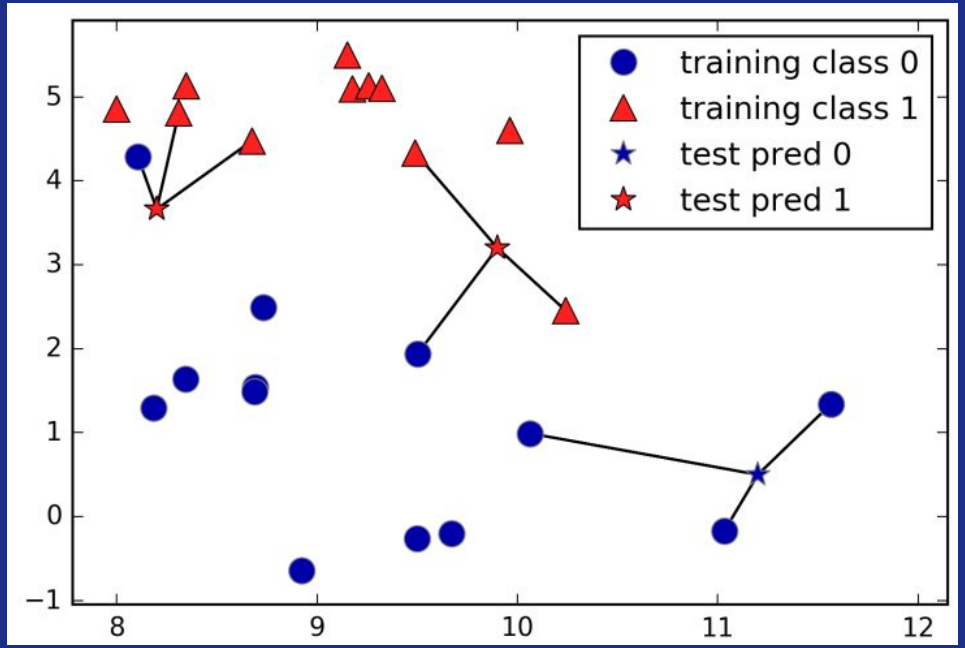
Exemplo, K=1

- Conjunto de dados
  - 26 instâncias
  - 2 atributos (numéricos)
  - 2 classes
- Construção do modelo
  - Guardar os pontos
- Classificação
  - A classe da nova instância é a classe da instância mais próxima



Exemplo, K=3

- Conj dados
  - 26 instâncias
  - 2 atributos (numéricos)
  - 2 classes
- Construção do modelo
  - Guardar os pontos
- Classificação
  - A classe da nova instância é a classe da **maioria** dos 3 vizinhos mais próximos

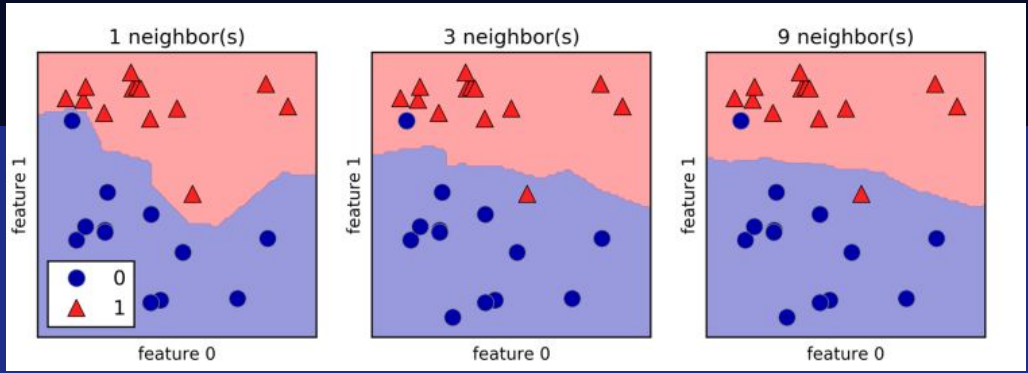


Fronteira de decisão

Fronteira de decisão

- É a fronteira que faz a divisão onde o algoritmo atribui uma classe ou outra
- É calculada através da previsão de todos os possíveis exemplos de teste

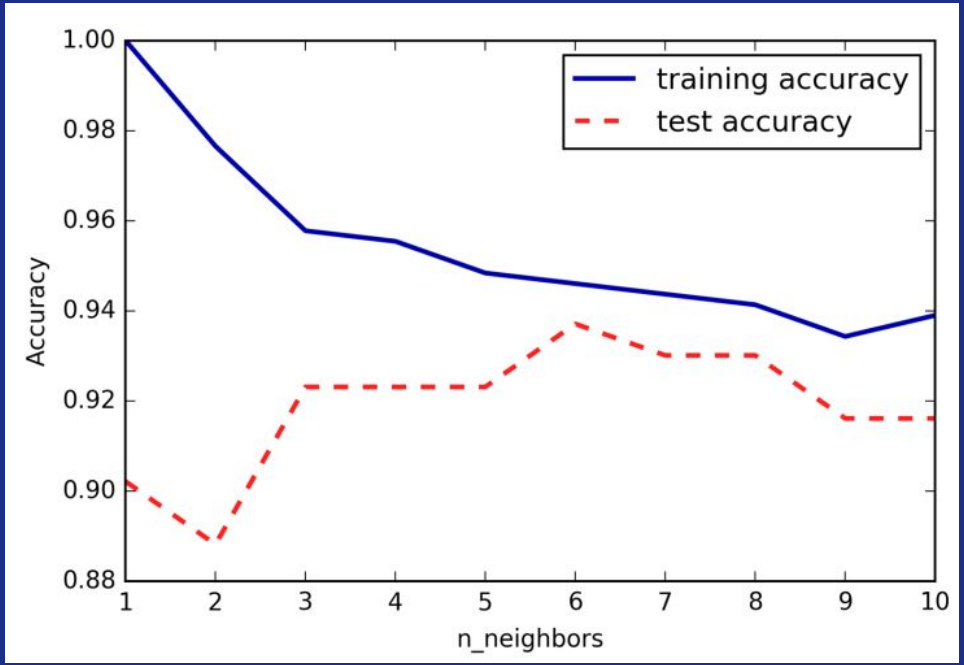
Exemplo



- 1 vizinho
  - A fronteira de decisão segue aproximadamente o conjunto de treino
  - Modelo mais complexo
- Mais vizinhos
  - A fronteira torna-se mais suave
  - Modelo mais simples
- Caso extremo
  - Nº de vizinhos igual o nº exemplos do conj de treino
  - Cada exemplo de teste tem exatamente os mesmos vizinhos
  - Todas as previsões são iguais
    - A classe mais frequente do conj de treino

Exatidão como função do número de vizinhos

- Previsão sobre o conjunto de treino
  - 1 vizinho
    - Previsão perfeita
  - Mais vizinhos
    - O modelo torna-se mais simples e a exatidão decresce
- Previsão sobre conjunto de teste
  - 1 vizinho
    - Menor quando comparada com modelos que usam mais vizinhos
    - O modelo é demasiado complexo
  - 10 vizinhos
    - O modelo é demasiado simples
  - Melhor desempenho
    - 6 vizinhos



Parâmetros e características

Parâmetros

- **Nº de vizinhos**
- Peso dos vizinhos
  - Uniforme
  - Inversamente proporcional à distância
- **Função de distância**
- Cálculo dos vizinhos mais próximos
  - Força bruta (calcula a distância com todos os pontos)
  - Cálculo otimizado (BallTree, KDTree)

Parâmetros

- **Função de distância**
  - Minkowski, p
  - Euclideana (p=2)
  - Manhattan (p=1)
  - Máximo (p=infinito)

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

Função de distância e similaridade

- **Distância e similaridade -** valores numéricos
  - Minkowski, p
  - Euclideana (p=2)
  - Manhattan (p=1)

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- **Distância e similaridade -** Valores lógicos
  - Simple Matching and Jaccard Coefficients
  - SMC = number of matches / number of attributes

    = (f11 + f00) / (f01 + f10 + f11 + f00)
  - J = number of 11 matches / number of non-zero attributes

    = (f11) / (f01 + f10 + f11)

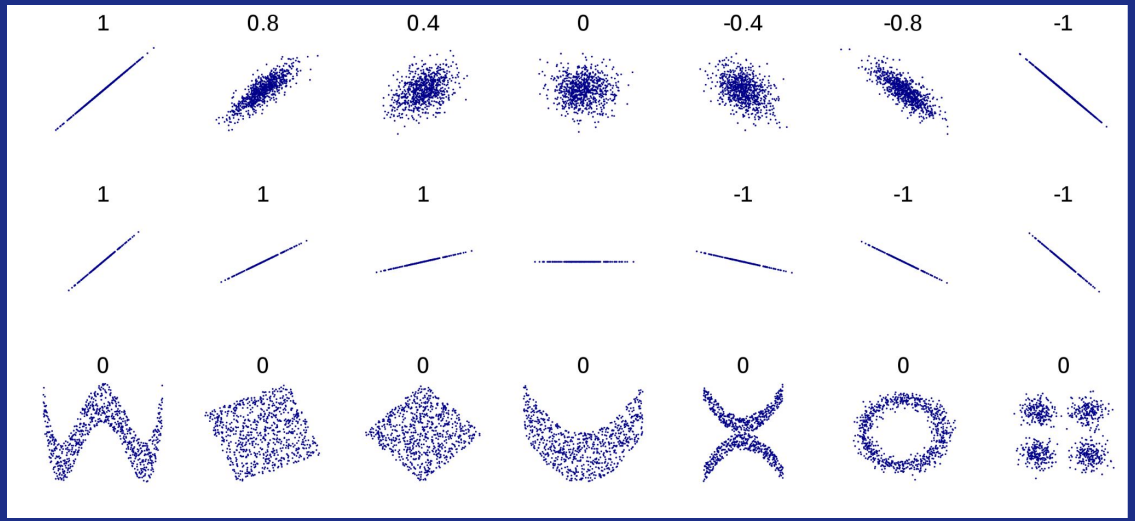
Funções de distância e similaridade

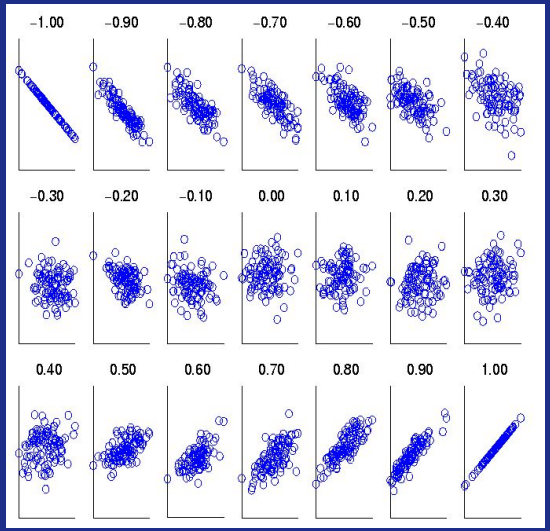
- **Similaridade do coseno**
  - Cos (d1, d2) = <d1,d2> / (||d1|| ||d2||)

- Outras
  - Correlação
  - Covariância
  - Desvio padrão
  - Informação mútua
  - Etc …

Função de distância e similaridade

- **Correlação**
  - Avalia relações lineares
  - Não avalia dependência estatística em geral





Características KNN

- Pontos fortes
  - Fácil de perceber
  - Muitas vezes dá bons resultados sem grandes ajustes
- Pontos fracos
  - A previsão é lenta (cálculo dos vizinhos)
  - Não tem bom desempenho quando existem muitos atributos
- Outras considerações
  - É importante fazer o pré-processamento dos dados
  - É um bom algoritmo baseline

Regressão

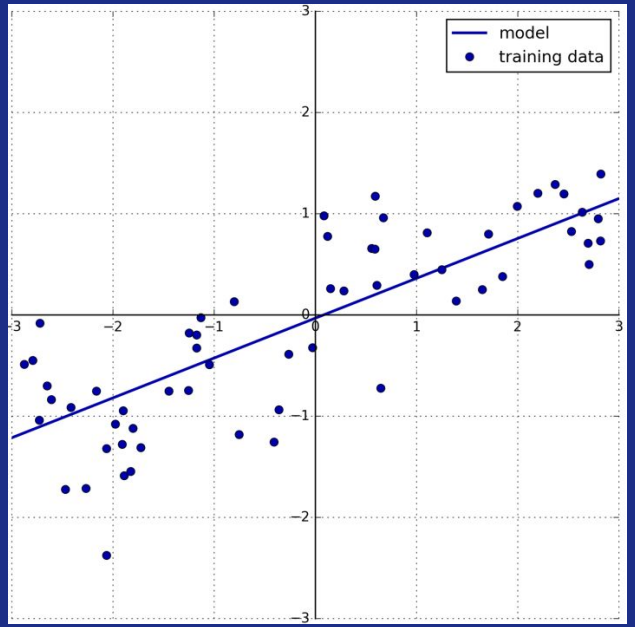
Tarefa

- Prever um **valor numérico contínuo**
- Exemplos
  - prever rendimento anual a partir da educação, idade e onde vive
  - prever a colheita de milho uma plantação, a partir de colheitas anteriores, clima e número de funcionários

Modelos lineares

Modelo linear

- Combinação linear
  - $w_1 * x_1 + \ldots + w_n x_n + b$
  - $x_1, \ldots, x_n$ são atributos, $w_1, \ldots, w_n$, b são coeficientes
- Aprendizagem
  - encontrar pesos $w_1, \ldots w_n$, b que aproximam o conjunto de dados
- Modelo
  - hiperplano, soma pesada dos atributos
- Comparado com KNN parece muito restritivo
  - mas é poderoso para conjuntos com muitos atributos



Modelos lineares

- Existe uma grande variedade de modelos lineares
- Diferenciam-se
  - na forma como os parâmetros são aprendidos
  - como a complexidade é controlada
- Modelos populares
  - regressão linear
  - regressão Ridge
  - Lasso

Regressão linear

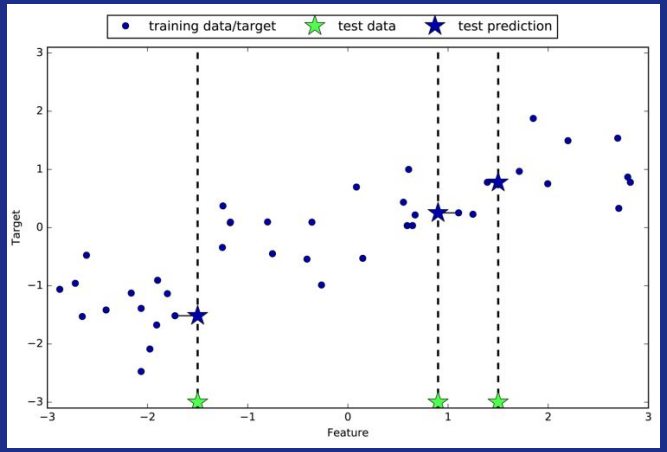
- É o modelo de regressão mais clássico e simples
- Aprendizagem
  - Encontra os parâmetros w e b que minimizam o erro quadrático médio entre as previsões e os valores reais da regressão no conj de treino
- Erro quadrático médio (*MSE - Mean Square Error*)
  - soma dos quadrados das diferenças entre as previsões e os valores reais
- Características
  - Não tem parâmetros
  - Não é possível controlar a complexidade
- Também conhecido como *Ordinary Least Squares* (OLS)

Regressão por K-vizinhos

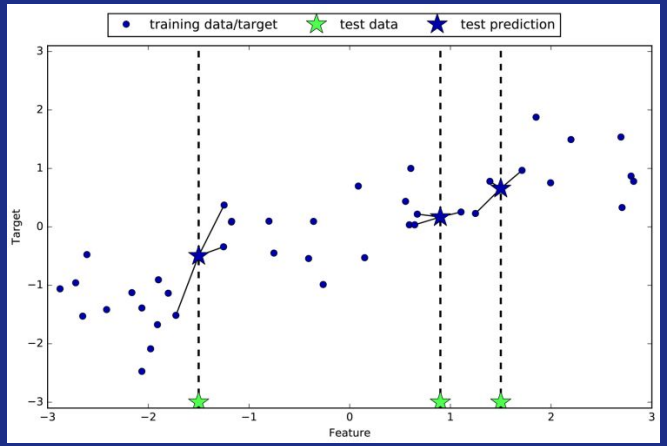
- **Algoritmo**
  - Construção do modelo
  - Guardar o conjunto de treino
- **Previsão de um exemplo**
  - Encontrar os K exemplos mais próximos no conjunto de Treino
  - Atribuir a média dos vizinhos

K vizinhos-mais-próximos

- 1 vizinho
  - a previsão é o valor do vizinho mais próximo

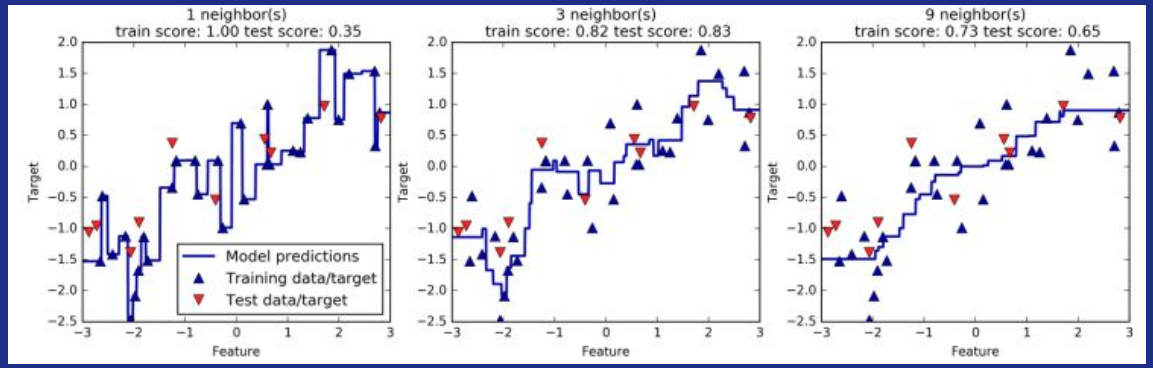


- k vizinhos
  - a previsão é a média dos vizinhos relevantes



Influência do nº de vizinhos

- 1 vizinho
  - previsão pouco estável
    - cada ponto do conjunto **treino** tem influência nas previsões
    - valores previstos percorrem todos os pontos do conjunto de **treino**
- Mais vizinhos
  - previsões mais suaves mas que não se ajustam tanto os dados de **treino**



Regressão por K-vizinhos

- **Algoritmo**
- **Coeficiente R^2** (R = coeficiente de correlação de Pearson)
  - coeficiente de determinação
  - medida estatística que indica quão bem as previsões se aproximam dos dados reais
  - normalmente um valor entre 0 e 1
- **1 - corresponde a uma previsão perfeita**
  - o modelo explica completamente a variabilidade dos dados reais
- **0 - corresponde a um modelo constante que prevê a média do conjunto de treino**
  - **o modelo não explica nenhuma variabilidade dos dados**