SCC0276 - Aprendizado de Máquina
Técnica KNN

Profa. Dra. Roseli Aparecida Francelin Romero
SCC - ICMC - USP

2019

Importância

- O K-Nearest Neighbors é um dos algoritmos de classificação mais básicos, porém essenciais, no Aprendizado de Máquina.
- Pertence ao domínio de aprendizagem supervisionada e encontra intensa aplicação no reconhecimento de padrões, mineração de dados e detecção de intrusões.
- É amplamente aplicável em cenários da vida real, pois é não-paramétrico, ou seja, não faz nenhuma suposição sobre a distribuição de dados (em oposição a outros algoritmos como o GMM, que assume uma distribuição gaussiana dos dados fornecidos).

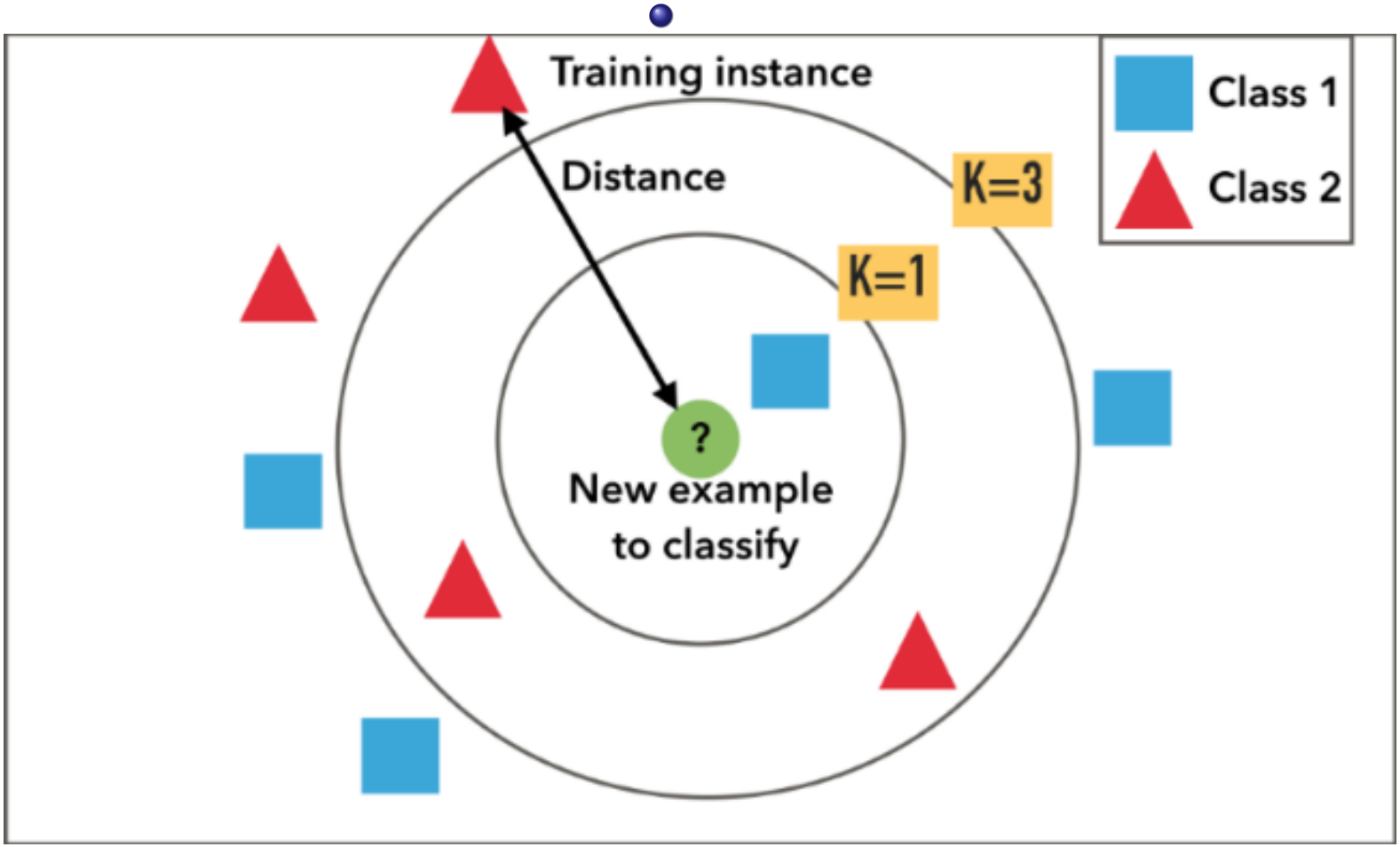
Importância

- O KNN também é um algoritmo Lazy (em oposição a um algoritmo eager). Isso significa que o KNN significa que ele não usa os pontos de dados de treinamento para fazer qualquer generalização.
- Em outras palavras, não há fase de treinamento explícito ou é muito mínima. Isso também significa que a fase de treinamento é muito rápida.
- KNN mantém todos os dados de treinamento. Para ser mais exato, todos (ou a maioria) dos dados de treinamento são necessários durante a fase de testes.

Classificação

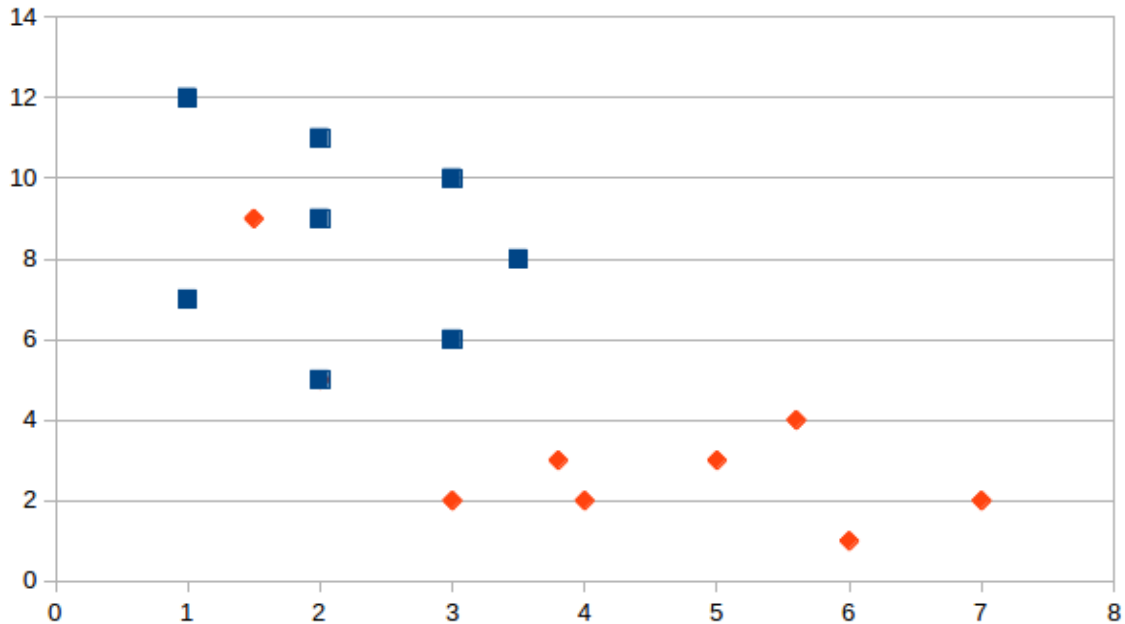
- Um objeto é classificado por uma votação majoritária de seus vizinhos, com o objeto sendo atribuído à classe mais comum entre seus k vizinhos mais próximos.

Classificação



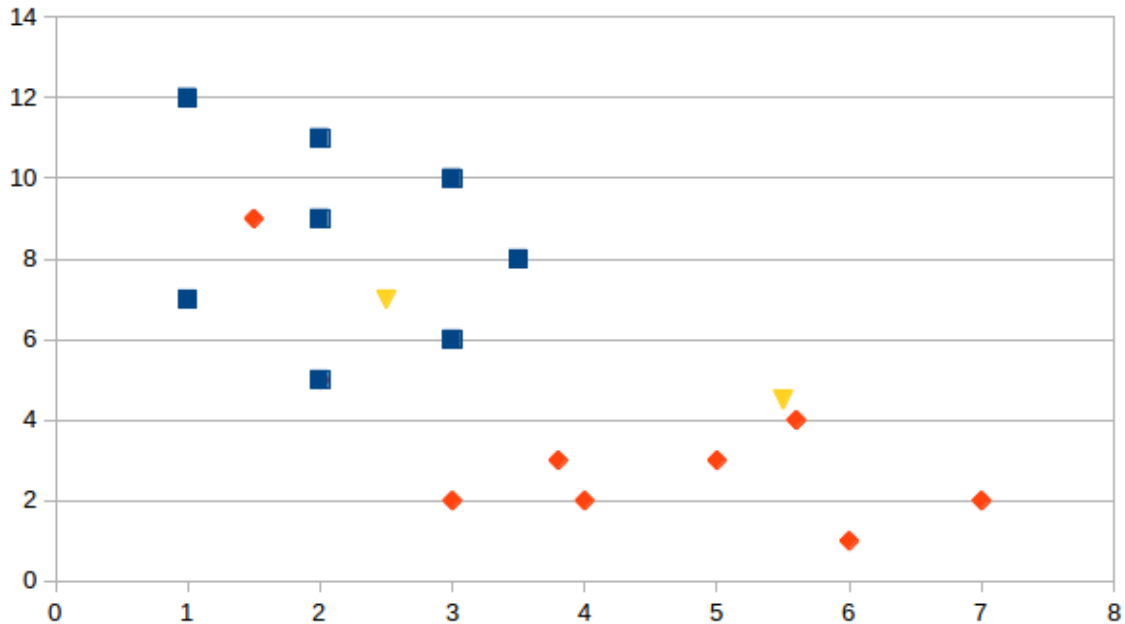
Problema de Classificação

- Dado dois conjuntos de pontos:



Problema de Classificação

- Dado outro conjunto de pontos de dados (também chamado de dados de teste), não classificados, que estão marcados como "amarelos". Desejamos classificá-los:



- Intuitivamente, podemos ver que o primeiro ponto (2.5,7) deve ser classificado como "azul"e o segundo ponto (5.5, 4.5) deve ser classificado como "vermelho".

Algoritmo Resumido

- Um inteiro positivo k é especificado, junto com uma nova amostra
- Selecionamos as k entradas em nosso banco de dados que estão mais próximas da nova amostra
- Encontramos a classificação mais comum dessas entradas
- Esta é a classificação que damos à nova amostra

Algoritmo KNN

Seja m o número de amostras de dados de treinamento. Seja p um ponto desconhecido.

- Armazene as amostras de treinamento em uma matriz de pontos de dados arr []. Isso significa que cada elemento dessa matriz representa uma tupla (x, y).
- para i = 0 a m:
  Calcule a distância euclidiana d (arr [i], p).
- Construa o conjunto S contendo as K menores distâncias obtidas. Cada uma dessas distâncias corresponde a um ponto de dados já classificado.
- Devolva o rótulo majoritário entre S.

Medidas de Distancia

- O tipo de medida de distância a ser usada depende da experiência ou tipo de dados a serem processados.

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan $$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski $$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

Aplicações

- Classificações de crédito - coletando características financeiras versus comparando pessoas com características financeiras semelhantes a um banco de dados. Pela própria natureza de uma análise de crédito, pessoas com detalhes financeiros semelhantes receberiam classificações de crédito semelhantes.
- Portanto, eles gostariam de poder usar esse banco de dados existente para prever a classificação de crédito de um novo cliente, sem precisar executar todos os cálculos.

Aplicações

- EMPRÉSTIMO: O banco deve conceder um empréstimo a um indivíduo?
- Um indivíduo deseja fazer um empréstimo, será que ele vai pagar o empréstimo, de acordo com o esperado?
- Essa pessoa está mais próxima das características das pessoas que não pagaram ou não assumiram seus empréstimos?

Aplicações

- Em ciência política - classificando um eleitor em potencial para um que “votará” ou “não votará”, ou para “votar no partido da direita” ou “votar de esquerda”.
- Exemplos mais avançados podem incluir detecção de manuscrito (como OCR), reconhecimento de imagem e até mesmo reconhecimento de vídeo.

Algumas Características

- A KNN armazena todo o conjunto de dados de treinamento que ele usa como sua representação.
- KNN não aprende nenhum modelo.
- A KNN faz predições just-in-time, calculando a similaridade entre uma amostra de entrada e cada instância de treinamento.