Modelos de Classificação
K-Nearest Neighbors (K-NN)

Prof. Gustavo Willam Pereira

INSTITUTO FEDERAL
Sudeste de Minas Gerais

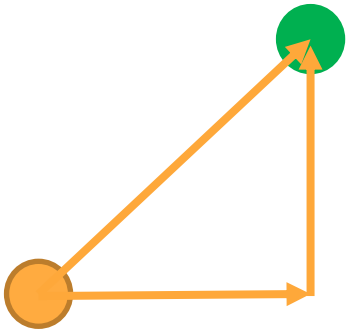
K-NN

- O algoritmo K-NN é uma classificação realizada com base nos vizinhos.
- Então, precisamos utilizar uma métrica de distância.
- Normalmente uma métrica de distância utilizada é a distância euclidiana.
- Vimos um pouco sobre a distância euclidiana na aula de pré-processamento de dados, veja abaixo a equação novamente:

$$d = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

K-NN

O algoritmo K-NN é uma classificação realizada com base nos vizinhos



Métrica de distância euclidiana

$$d = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

$$\mathrm{d} = \left[|x_i - x_j|^p + |y_i - y_j|^p \right]^{1/p}$$

Métrica de distância Manhattan p = 1

Métrica de distância Minkowski p qualquer

Métrica de distância Mahalanobis

INSTITUTO FEDERAL
Sudeste de Minas Gerais

K-NN

- Na distância euclidiana a potência da diferença entre as variáveis das amostras i e j é 2.
- No entanto, poder-se-ia utilizar qualquer outra potência como métrica de distância, no entanto, deixará de ser distância euclidiana.
- Uma vez definida a métrica de distância, podemos descrever o algoritmo, veja abaixo os passos realizados.

1) Escolha o número de vizinhos (K Neighbors); por exemplo k = 5;

2) Escolha os K vizinhos mais próximos do ponto a ser classificado com base na métrica de distância;

3) Dentre os K (no exemplo 5) vizinhos mais próximos, conte quantos foram classificados em cada categoria;

4) Classifique o ponto como a categoria que mais apareceu nos K vizinhos.

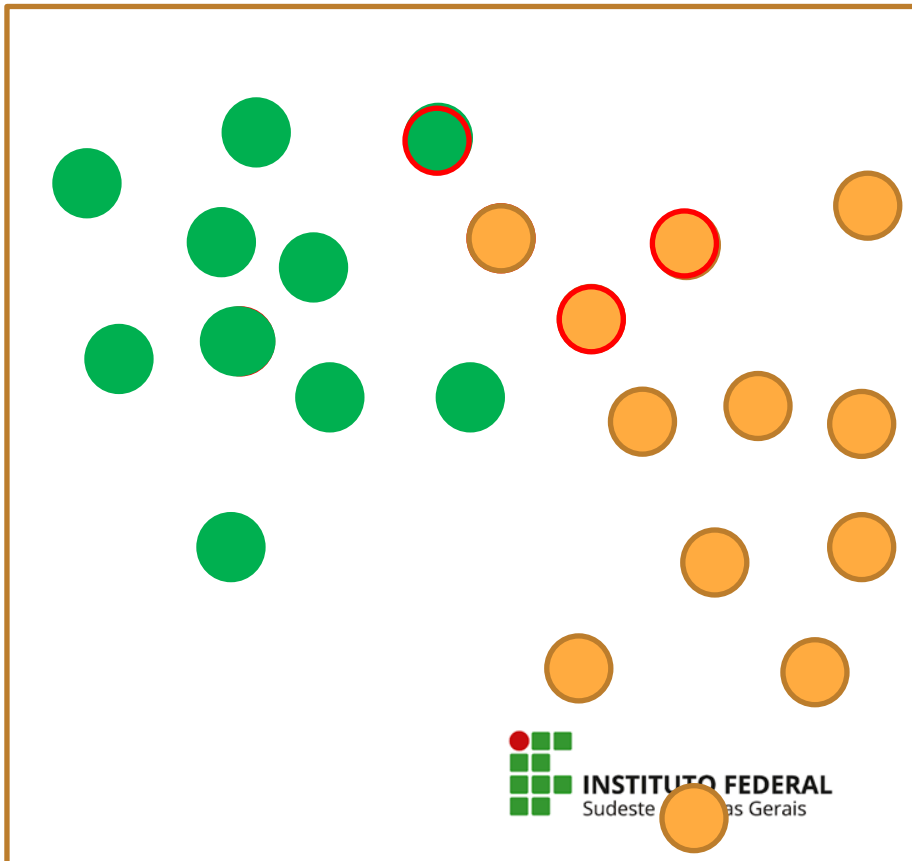
K-NN

1) Escolher um número k de vizinhos, por exemplo k = 3

2) Escolha os k vizinhos mais próximos do ponto a ser classificado

3) Conte quantos foram classificados em cada categoria

4) Classifique o ponto como a categoria que mais apareceu nos K vizinhos.



K-NN

- Na Figura abaixo você poderá verificar um esquema do funcionamento do algoritmo.

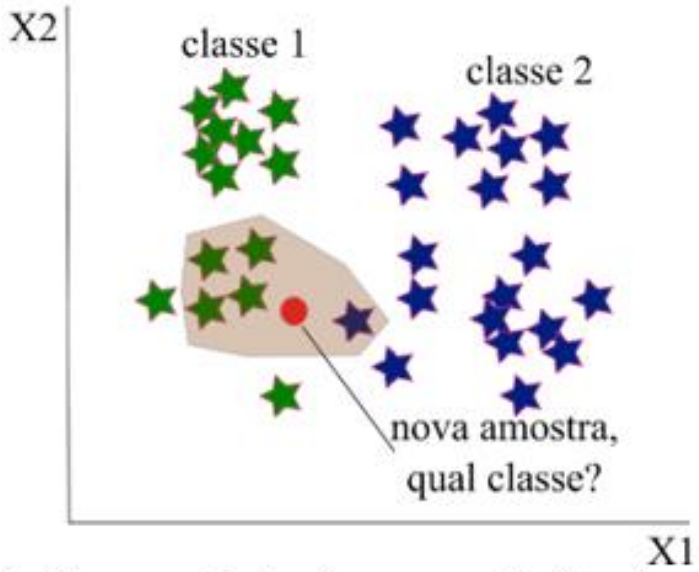


Figura 4 – Esquema de funcionamento do algoritmo K-NN.

K-NN

- Na Figura 4 são apresentadas duas classes.
- Uma instancia em vermelho não tem a classe definida. Para definir essa classe, então escolhemos o número k de vizinhos igual a 5 (passo 1).
- Buscamos os 5 vizinhos mais próximos pela métrica de distância euclidiana (passo 2). Os vizinhos mais próximos estão na área sombreada da Figura 4. Então contamos quantos temos de cada classe. No caso temos quatro da classe 1 e um da classe 2 (passo 3).
- Então a instância será classificada como classe 1 (passo 4).
- O algoritmo é extremamente simples e não é baseado em modelos, mas sim em instâncias. No entanto, apresenta grande capacidade de aprendizagem (overfitting).
- O K-NN gera uma classificação não-linear.

K-NN

- O K-NN se baseia em métricas de distância, ele é influenciado pela amplitude das variáveis, então temos que manter, após a linha de divisão de dados em treinamento e teste, as linhas para padronizar os dados.

INSTITUTO FEDERAL
Sudeste de Minas Gerais