K-Vizinhos mais próximos

Thaís de Almeida Ratis Ramos

História

- Foi descrito no início dos anos 1951 por Evelyn Fix e Joseph Hodges;

- Ganhou popularidade, quando um maior poder de computação tornou-se disponível;

- Amplamente utilizado na área de reconhecimento de padrões e estimativa estatística.

Vizinhos mais próximos

- Classifica um novo objeto com base nos exemplos do conjunto de treinamento que são próximos a ele;

- Pode ser utilizado tanto para classificação quanto para regressão;

- Tem variações definidas (principalmente) pelo número de vizinhos considerados.

1- Vizinho mais próximo

- 1-NN, 1-Nearest neighbour
- Calcula distância entre cada 2 pontos
- A métrica mais usual é a distância Euclidiana

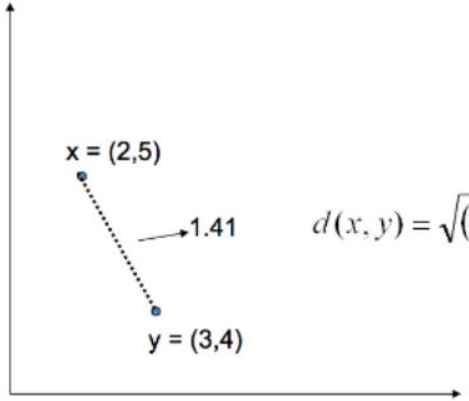
$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Exemplo - Distância Euclidiana

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

PODEM SER USADAS QUAISQUER MÉTRICAS QUE REPRESENTEM SIMILARIDADE ENTRE DOIS OBJETOS, COMO CORRELACÃO, POR EXEMPLO!!

DISTÂNCIA ENTRE STRINGS PODEM SER MEDIDAS PELA DISTÂNCIA DE HAMMING!!



$$d(x, y) = \sqrt{(2-3)^2 + (5-4)^2} = \sqrt{2} = 1.41$$

- CADA DIMENSÃO REPRESENTA UM ATRIBUTO
- X E Y REPRESENTAM INSTÂNCIAS (OBJETOS)

Prática Distância Euclidiana

Distância entre strings

Hamming

Strings de mesmo tamanho

- 1011101 and 1001001 is 2.
- 2173896 and 2233796 is 3.

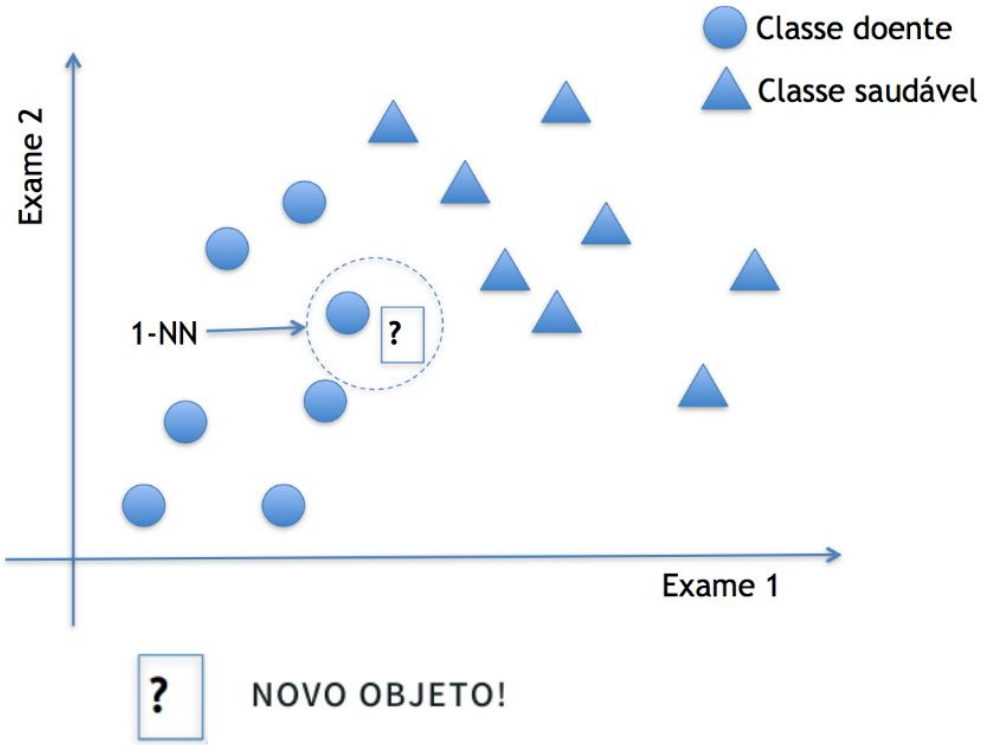
Levenshtein

Strings e tamanhos diferentes

H		O	N	D	A	
H	Y	U	N	D	A	I

H	O		N	D	A	
H	Y	U	N	D	A	I

1-NN


Classe doente
Classe saudável
Exame 2
1-NN
?
Exame 1
?
NOVO OBJETO!

Algoritmo

Entrada: Conjunto de treinamento (D); objeto teste (z); conjunto de classes (L)

Saída: A classe de z

PARA CADA objeto y que pertence a D faça:

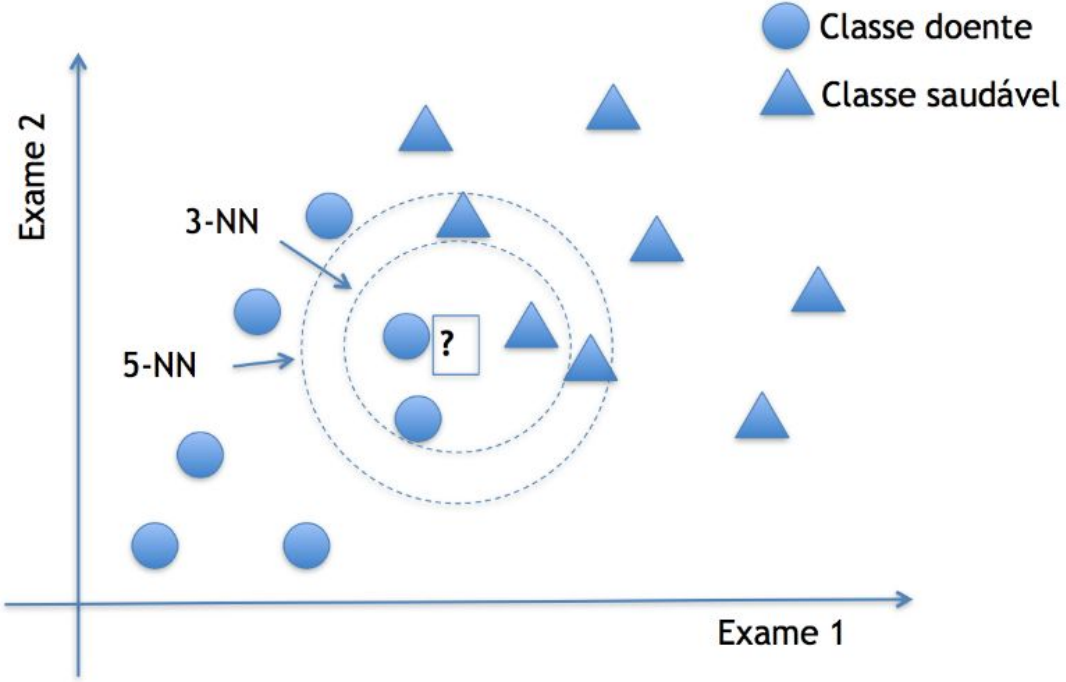
    Compute a distância entre z e y

FIM

Fixe k, o número de vizinhos mais próximos na população (z) a ser considerado

A classe de saída do objeto z será a moda, média ou mediana dos rótulos dos k vizinhos do conjunto selecionado como mais próximo

K-NN



Resumo

Em vez de 1 vizinho mais próximo, os k objetos do conjunto de treinamento mais próximo do ponto de teste X;

Classificação:

➔ O objeto é classificado na classe mais votada: moda;

Regressão:

➔ Minimizar o erro quadrático: média;
➔ Minimizar o desvio absoluto: Mediana

VANTAGENS E DESVANTAGENS

PRÓS:

- O algoritmo é simples e fácil de implementar;
- Não há necessidade de ajustar vários parâmetros;
- Treinamento rápido;
- Ele pode ser usado para classificação, regressão.

CONS:

- A classificação da base de teste é mais lenta e mais custosa em termos de tempo e memória;
- KNN também não é adequado para grandes dados dimensionais.

Exemplo Prático do KNN