O que é KNN?

## Introdução

KNN, ou também chamado de K-Nearest Neighbors (K vizinhos mais próximos), é um algoritmo de classificação supervisionado. Seu objetivo é, dado um dataset, classificar um ponto *P* baseado na similaridade de características entre ele e os seus K vizinhos através do cálculo de distância entre dois pontos.

## Distância entre dois pontos

Euclidean distance é o cálculo utilizado para medir a distância entre pontos. Dado um ponto e outro , é possível obter sua distância com com o seguinte cálculo:

Ou para pontos em um plano maior ou igual a 3 dimensões:

Onde *i* é a dimensão e *n* um ponto qualquer

No caso de um dataset para treinamento no KNN, o ponto seria o registro e suas coordenadas as características.

Por exemplo, no seguinte dataset hipotético cada pessoa é um ponto no plano cartesiano, enquanto os dados da idade e o valor do empréstimo solicitado são as características. Logo, a posição de cada registro no plano.

| **Nome** | **Idade (X)** | **Valor (Y)** | **Vai Receber? (Label)** |
| --- | --- | --- | --- |
| João | 25 | R$ 9.000,00 | Não |
| Jorge | 60 | R$ 20.000,00 | Sim |
| Paula | 30 | R$ 8.000,00 | Sim |
| Cláudia | 45 | R$ 15.000,00 | Não |

Já a última coluna indica qual classe o dado pertence, ou seja, se a pessoa, baseado na sua idade e valor solicitado, vai ou não receber o empréstimo. Com isso, é possível classificar se o empréstimo da seguinte pessoa seria aprovado, baseado nas características que mais se assemelham com a nova pessoa informada.

| **Nome** | **Idade (X)** | **Valor (Y)** | **Vai Receber? (Label)** |
| --- | --- | --- | --- |
| Maria | 27 | R$ 7.500,00 | ??? |

## O algoritmo

A essência do algoritmo KNN é o cálculo da distância entre os pontos. Com isso, é possível descrever o algoritmo em 6 passos:

1. Definir o valor de K (números de vizinhos que serão analisados).
2. Carregar os dados já classificados.
3. Carregar o dado não classificado.
4. Para cada dado já classificado, calcular a distância entre o dado não classificado e armazenar as distâncias.
5. Pegar as classificações dos registros com as K menores distâncias.
6. Calcular a moda dessas classificações. Esse valor será a predição do registro não classificado.

## O algoritmo na prática

Considerando os dados anteriores, é possível verificar se a Maria receberia ou não o empréstimo solicitado utilizando KNN. Primeiramente, é definido o valor de K, no caso 3. Depois, é calculada a distância das características da Maria para cada registro na primeira tabela:

Calculando a moda o dos registros no array *smallest ordered classifications*, Maria não receberia o valor do empréstimo solicitado.

## KNN em textos e imagens

KNN é um dos algoritmos considerado *estado da arte* para classificação de textos. Isso porque, é possível considerar cada palavra do texto como uma *feature* e, ainda, verificar quais conjuntos de palavras estão relacionadas com suas classificações. Quando se fala em KNN com imagens os resultados não dos melhores, já que estaríamos comparando, para cada pixel das imagens, os valores de seu RGB e não considerando todo o contexto.

Matheus Sena Vasconcelos