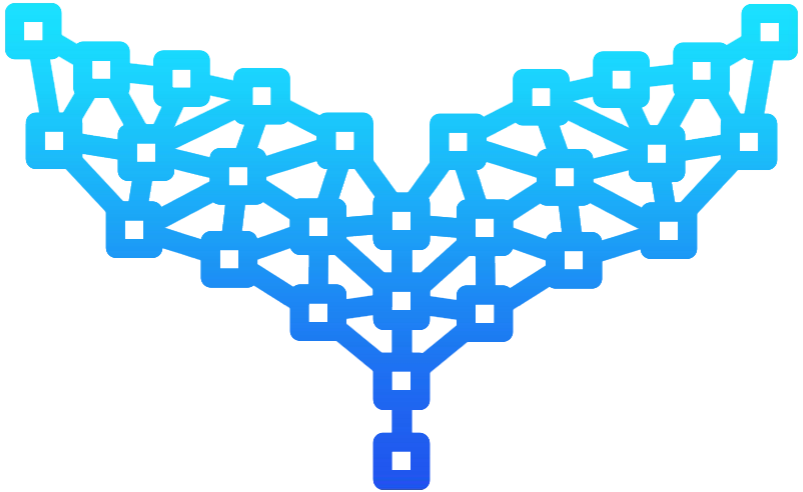
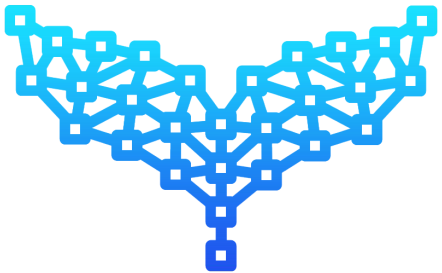
# TAIL

Technology and Artificial Intelligence League



# K-NEAREST NEIGHBORS ALGORITHM (KNN)

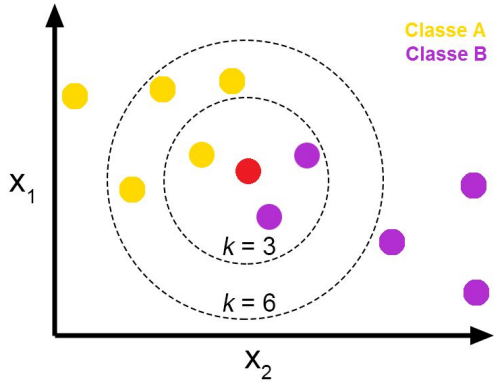


# 1. Introdução

O KNN consiste em um método não paramétrico usado para classificação ou regressão, proposto inicialmente por Thomas Cover. O objetivo desse método é classificar uma determinada variável, atribuindo a ela o rótulo ao qual ele mais se assemelha baseado nas características conhecidas pelo algoritmo.

Alguns exemplos em que pode ser aplicado:

- Resolvendo problemas que focam em encontrar similaridades em observações;
- Datasets pequenos e não muito generalizados, à fim de evitar overfitting.




Exemplo gráfico do KNN, onde o algoritmo tem conhecimento de bolas lilás e amarelas e deve rotular a vermelha onde ela melhor se encaixa.

# 2. Distância Euclidiana

O algoritmo faz uso da fórmula de distância euclidiana, que consiste em calcular a distância entre dois pontos, encontrando a distância geométrica no espaço multidimensional.

$$E(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

Fórmula da distância Euclidiana

# 3. Escalonamento de atributos

A fim de utilizar o cálculo de distância euclidiana é necessário deixar os atributos na mesma escala, para que valores maiores não se sobressaiam.

Existem dois recursos usados no escalonamento sendo eles:

A **padronização** que usa o desvio padrão e a média.

$$x = \frac{x - média(x)}{desvio\ padrão(x)}$$

Fórmula da padronização

A **normalização** que usa valores mínimos e máximos.

$$z = \frac{x - min(x)}{max(x) - min(x)}$$

Fórmula da normalização

# 4. Exemplo

Considerando a seguinte tabela:

| Índice | Renda Anual | Idade | Empréstimo | Classe |
|---|---|---|---|---|
| 1 | 36801 | 43 | 5406 | Não Pagou |
| 2 | 68811 | 24 | 4211 | Não Pagou |
| 3 | 30483 | 34 | 4514 | Pagou |
| 4 | 44930 | 20 | 7708 | Pagou |

Tabela com dados do exemplo

# 5. Exemplo

A partir da tabela apresentada anteriormente, tentaremos classificar a amostra abaixo.

| | | | | |
|---|---|---|---|---|
| 5 | 48430 | 27 | 5722 | ? |

Amostra do exemplo

# 5. Exemplo

Inicialmente, calcularemos a distância com e sem idade.

$$E(1,5) = \sqrt{(48430 - 36801)^2 + (27 - 43)^2 + (5722 - 5406)^2} = 11.633,30$$

$$E(1,5) = \sqrt{(48430 - 36801)^2 + (5722 - 5406)^2} = 11.633,29$$

---

$$E(2,5) = \sqrt{(48430 - 68811)^2 + (27 - 24)^2 + (5722 - 4211)^2} = 20.436,9344$$

$$E(2,5) = \sqrt{(48430 - 68811)^2 + (5722 - 4211)^2} = 20.436,9342$$

$$E(3,5) = \sqrt{(48430 - 30483)^2 + (27 - 34)^2 + (5722 - 4514)^2} = 17.987,61$$

$$E(3,5) = \sqrt{(48430 - 30483)^2 + (5722 - 4514)^2} = 17.987,60$$

---

$$E(4,5) = \sqrt{(48430 - 44930)^2 + (27 - 20)^2 + (5722 - 7708)^2} = 4.024,207$$

$$E(4,5) = \sqrt{(48430 - 44930)^2 + (5722 - 7708)^2} = 4.024,201$$

# 5. Exemplo

Considerando k=3, k=2 e k=1, a nova amostra seria classificada como, respectivamente, Pagou, Pagou ou Não Pagou, pois temos um empate e Pagou.

Além disso,  deve-se destacar que os valores das distâncias com e sem idade possuem uma diferença quase imperceptível, dessa forma, conclui-se que, sem a normalização, a idade não traz diferença.

# 5. Exemplo

Em seguida, iremos calcular as distâncias com os dados normalizados. Normalizando os dados da tabela, temos que:

| Índice | Renda Anual | Idade | Empréstimo | Classe |
|---|---|---|---|---|
| 1 | 0,164 | 1 | 0,341 | Não Pagou |
| 2 | 1 | 0,173 | 0 | Não Pagou |
| 3 | 0 | 0,608 | 0,086 | Pagou |
| 4 | 0,376 | 0 | 1 | Pagou |
| 5 | 0,468 | 0,304 | 0,432 | ? |

Tabela com dados normalizados

# 5. Exemplo

Calculando as distâncias com e sem idade.

$$E(1,5) = \sqrt{(0,468 - 0,164)^2 + (0,304 - 1)^2 + (0,432 - 0,341)^2} = 0,764$$

$$E(1,5) = \sqrt{(0,468 - 0,164)^2 + (0,432 - 0,341)^2} = 0,317$$

---

$$E(2,5) = \sqrt{(0,468 - 1)^2 + (0,304 - 0,173)^2 + (0,432 - 0)^2} = 0,697$$

$$E(2,5) = \sqrt{(0,468 - 1)^2 + (0,432 - 0)^2} = 0,685$$

---

$$E(3,5) = \sqrt{(0,468 - 0)^2 + (0,304 - 0,608)^2 + (0,432 - 0,086)^2} = 0,656$$

$$E(3,5) = \sqrt{(0,468 - 0)^2 + (0,432 - 0,086)^2} = 0,582$$

---

$$E(4,5) = \sqrt{(0,468 - 0,376)^2 + (0,304 - 0)^2 + (0,4322 - 1)^2} = 0,650$$

$$E(4,5) = \sqrt{(0,468 - 0,376)^2 + (0,432 - 1)^2} = 0,575$$

# 5. Exemplo

Considerando k=3, k=2 e k=1, a nova amostra seria classificada como Pagou. Ademais, os cálculos das distância com e sem idade possuem uma diferença relevante, pois agora estão escalonados.

Dessa forma, além de diminuir o custo computacional, já que os valores estão entre 0 e 1, esse artifício pode fazer uma grande diferença no resultado.

# Referências

KUMAR, Aditya. KNN Algorithm: When? Why? How?. Disponível em: <https://towardsdatascience.com/knn-algorithm-what-when-why-how-41405c16c36f>. Acesso em 26 de Setembro de 2020;

MARQUES, Jair Mendes; NETO, Anselmo Chaves. Análise de Agrupamentos. Disponível em: <https://docs.ufpr.br/~soniaisoldi/ce076/9ANALISEAGRUPAMENTOS.pdf>. Acesso em 26 de Setembro de 2020;