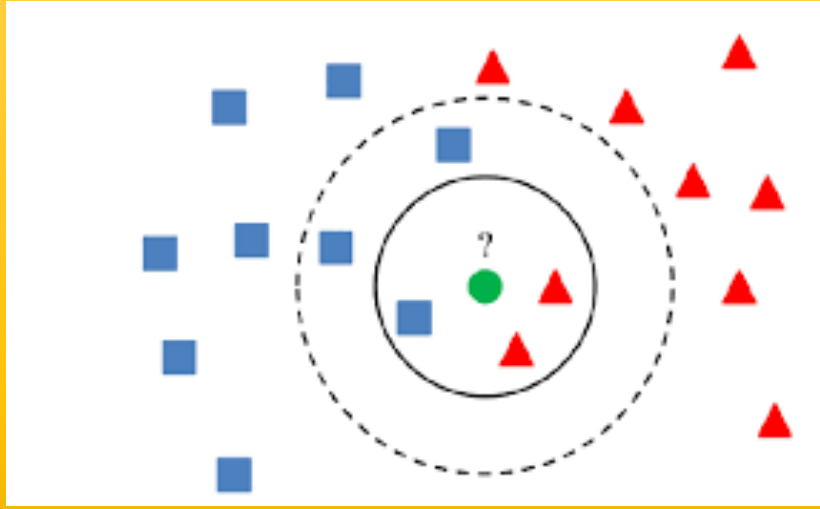
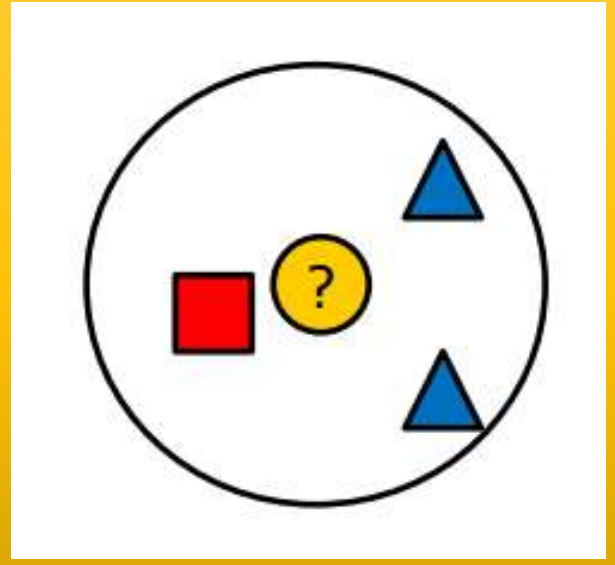
KNN – VIZINHOS MAIS PRÓXIMOS



Prof. Dr. Vladimir Costa de Alencar
UEPB
www.valencar.com

KNN – K-NEAREST NEIGHBOOR

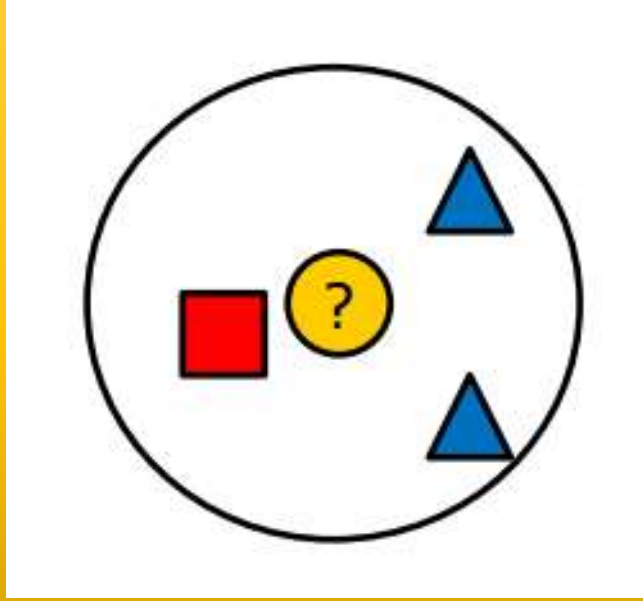
KNN – Vizinhos mais Próximos



É um dos algoritmos de classificação mais simples.

Usado para classificar objetos com base em exemplos de treinamento que estão mais próximos no espaço de características.

KNN



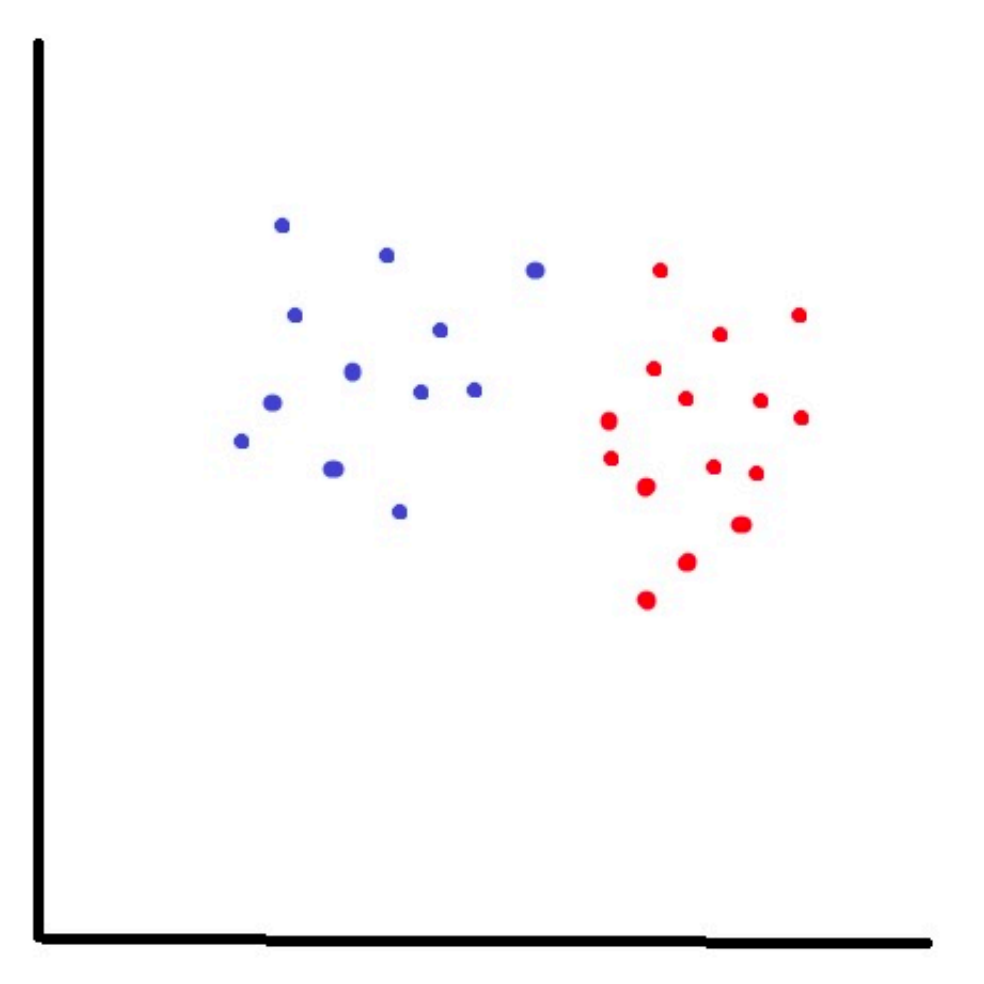
1- Dados de Treinamento

2- Definir a métrica para cálculo da distância

3- Definir o valor de K (número de vizinhos mais próximos que serão considerados pelo algoritmo)

4 - É importante normalizar os dados (mesma escala)

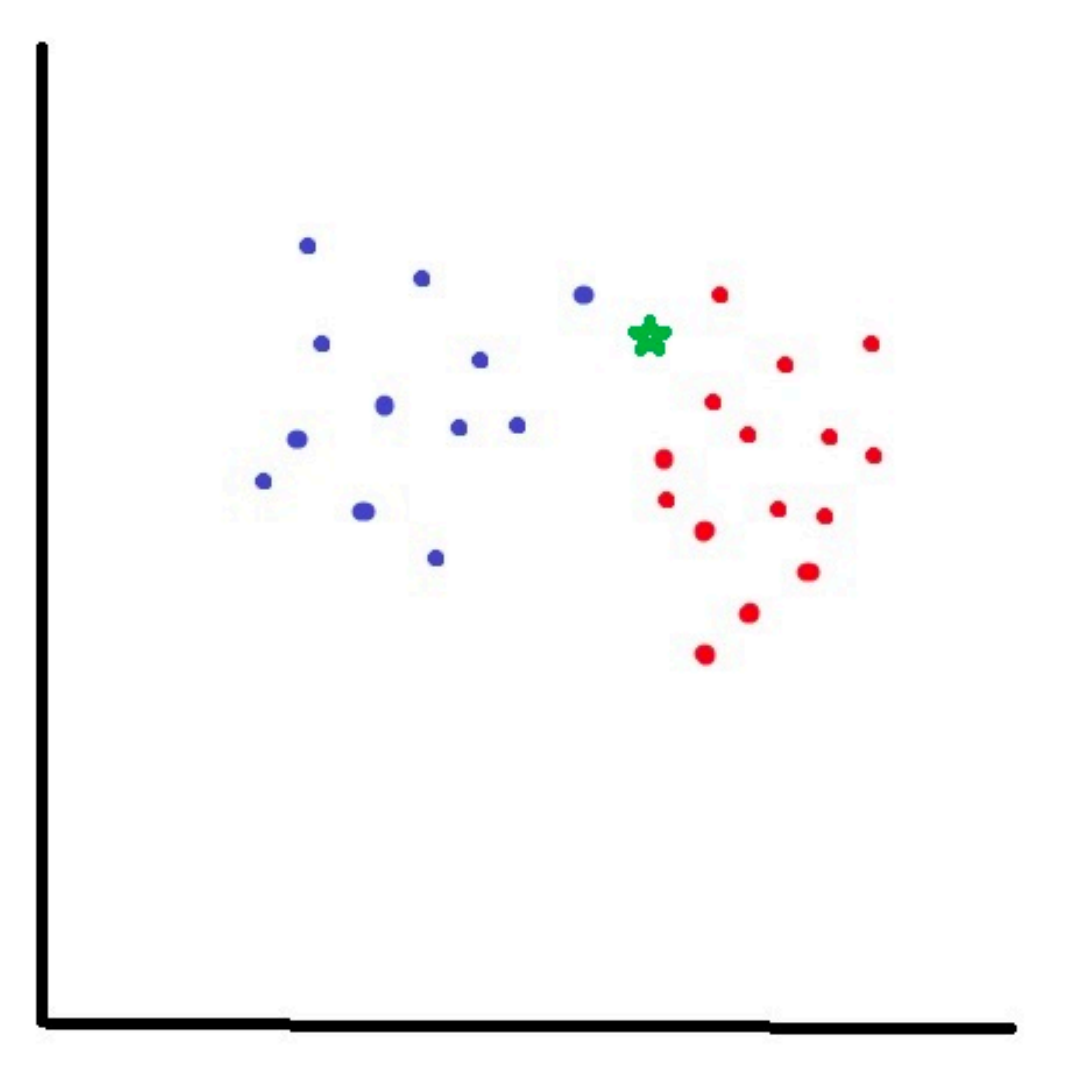
KNN



2 Grupos

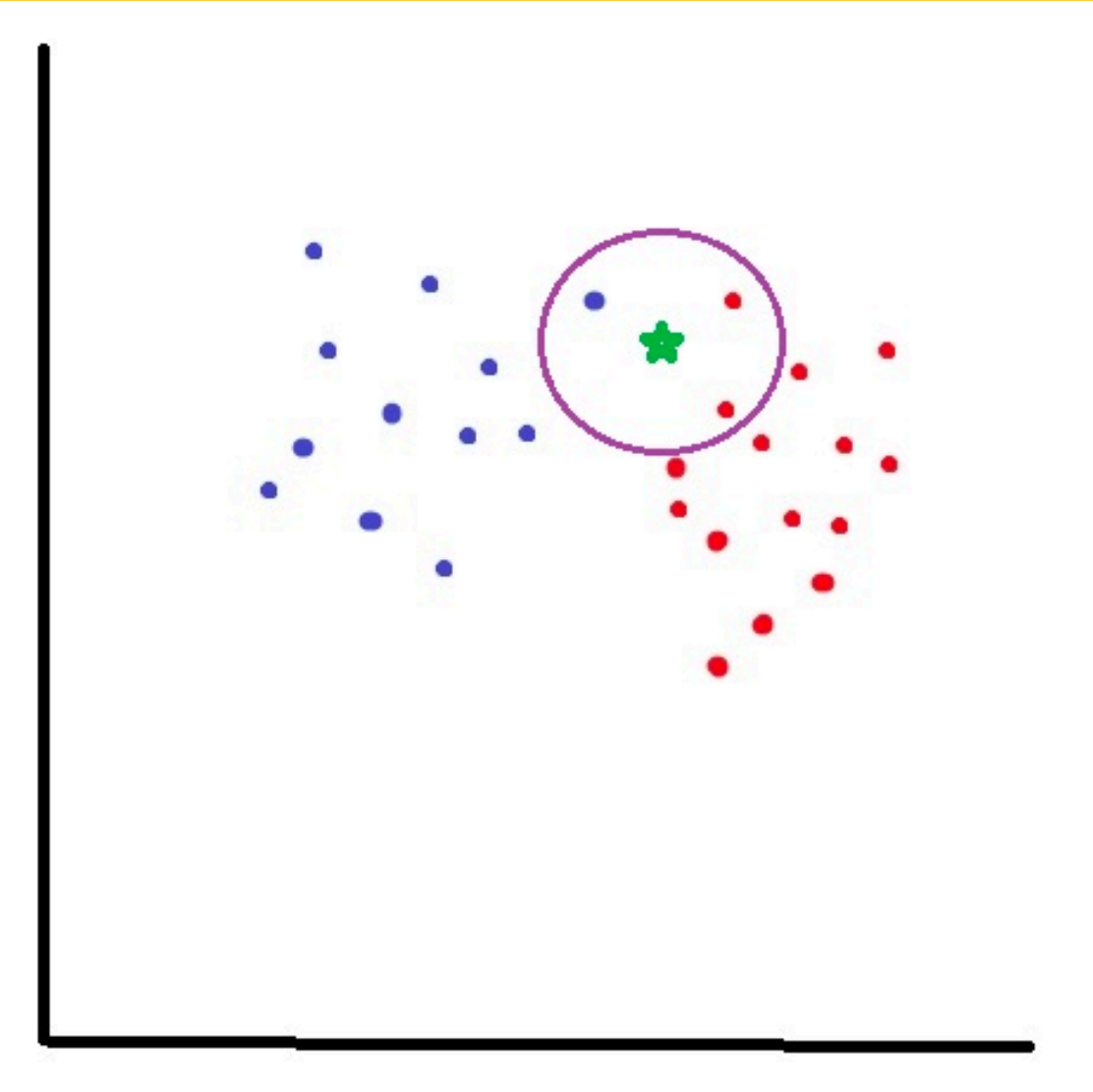
Azul = Clientes Vip
Vermelho = Clientes Normais

KNN



Como classificar o Ponto Verde?

KNN



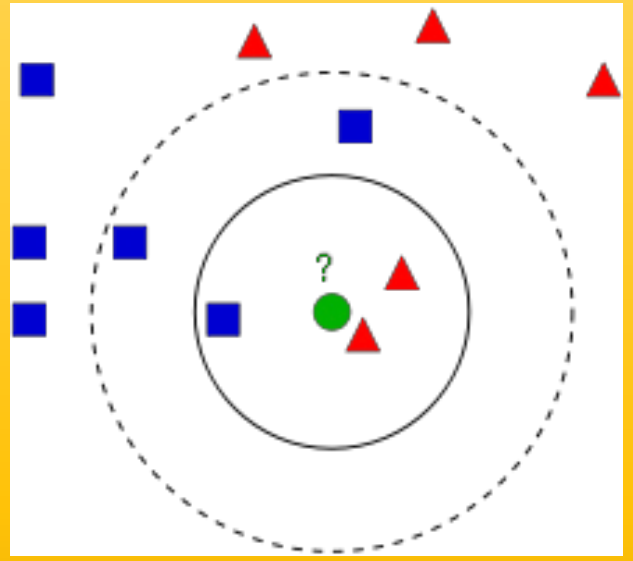
KNN-3

Considera-se o voto majoritário entre os rótulos de classe dos K vizinhos mais próximos

Nesse casso, os 3 Vizinhos mais próximos Determinarão a classe

-> O ponto Verde é Cliente **Normal**

KNN - MÉTRICAS DE DISTÂNCIA USADAS:



$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Distância Euclidiana

Outras métricas de distância:

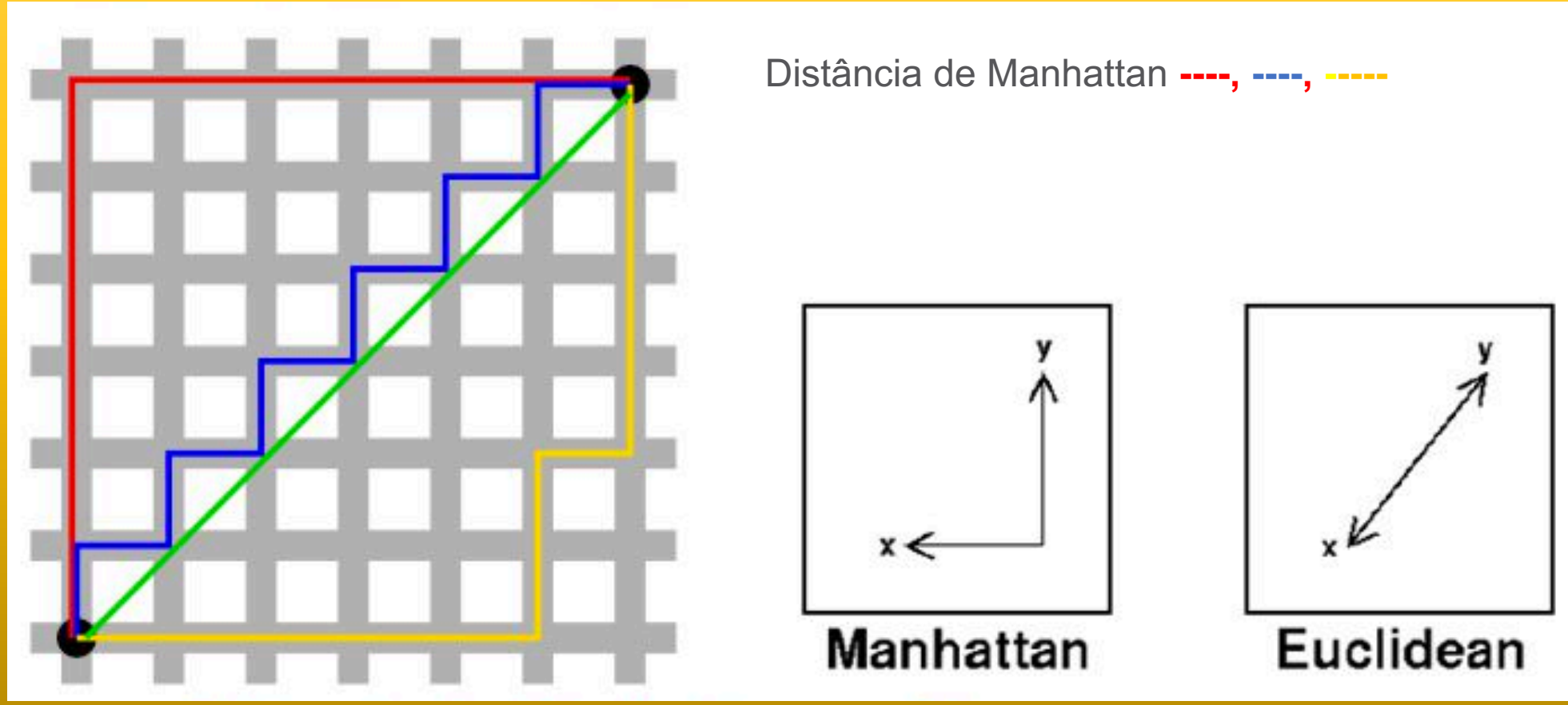
Distância Manhattan

Distância de Minkowsky (Generalização de Euclidiana e Manhattan)

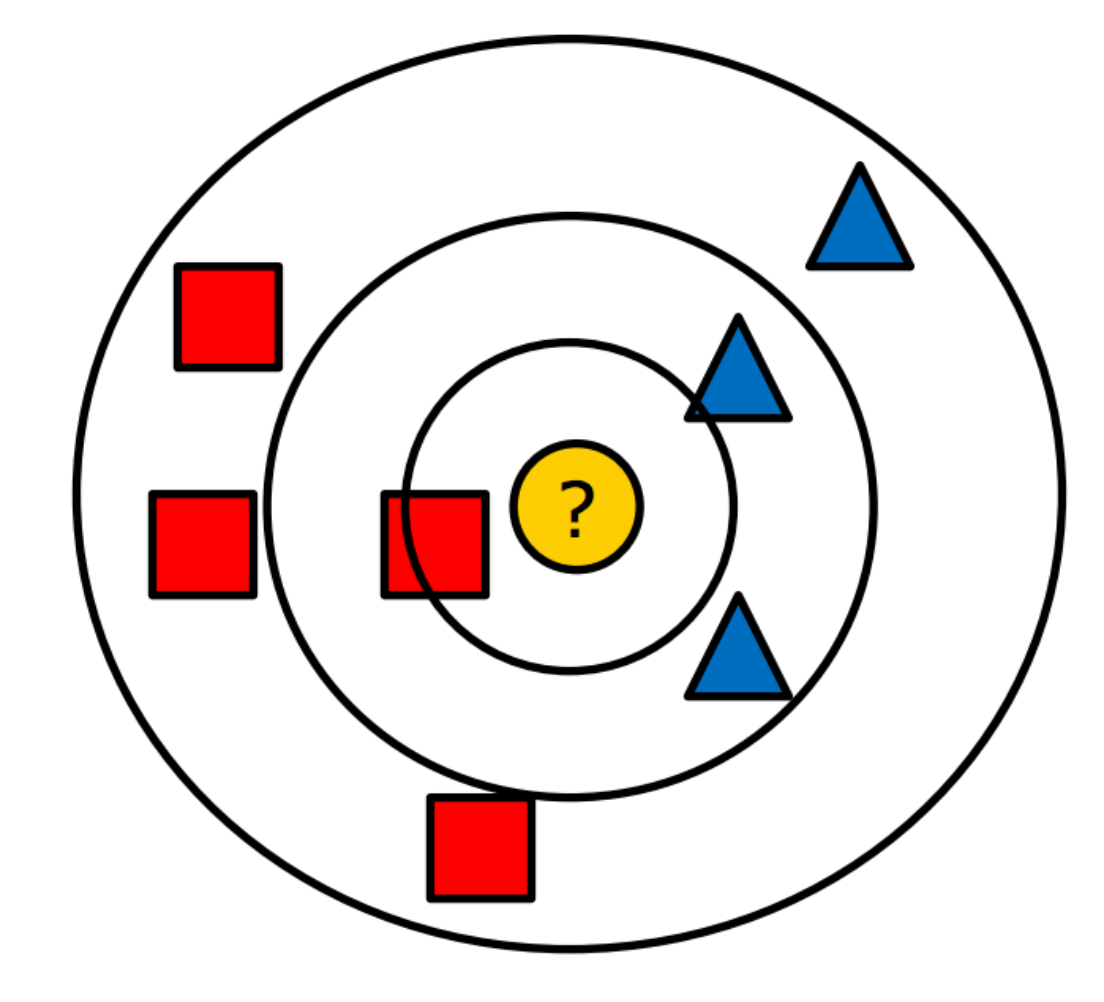
Distância de Hamming (utilizada para sinalizar erros na transmissão de palavras binárias)

KNN

Red: **Manhattan distance**. Green: diagonal, straight-line distance. Blue, yellow: equivalent Manhattan distances.



KNN



K = 1 Pertence a classe de quadrados.

K = 3 Pertence a classe de triângulos.

K = 7 Pertence a classe de quadrados

KNN

Como escolher o valor de K?

Se K for muito pequeno, a classificação fica sensível a pontos de ruído.

Se k é muito grande, a vizinhança pode incluir elementos de outras classes.

Além disso, é necessário sempre escolher um **valor ímpar** para K, assim se evita **empates** na votação.

KNN

A precisão da classificação utilizando o algoritmo KNN depende fortemente do modelo de dados.

Na maioria das vezes os atributos precisam ser normalizados para evitar que as medidas de distância sejam dominado por um único atributo.

Exemplos: Altura de uma pessoa pode variar de 1,20 a 2,10.
Peso de uma pessoa pode variar de 40 kg a 150 kg.
O salário de uma pessoa podem variar de R$ 800 a R$ 20.000.

KNN - NORMALIZANDO OS DADOS:

Antes

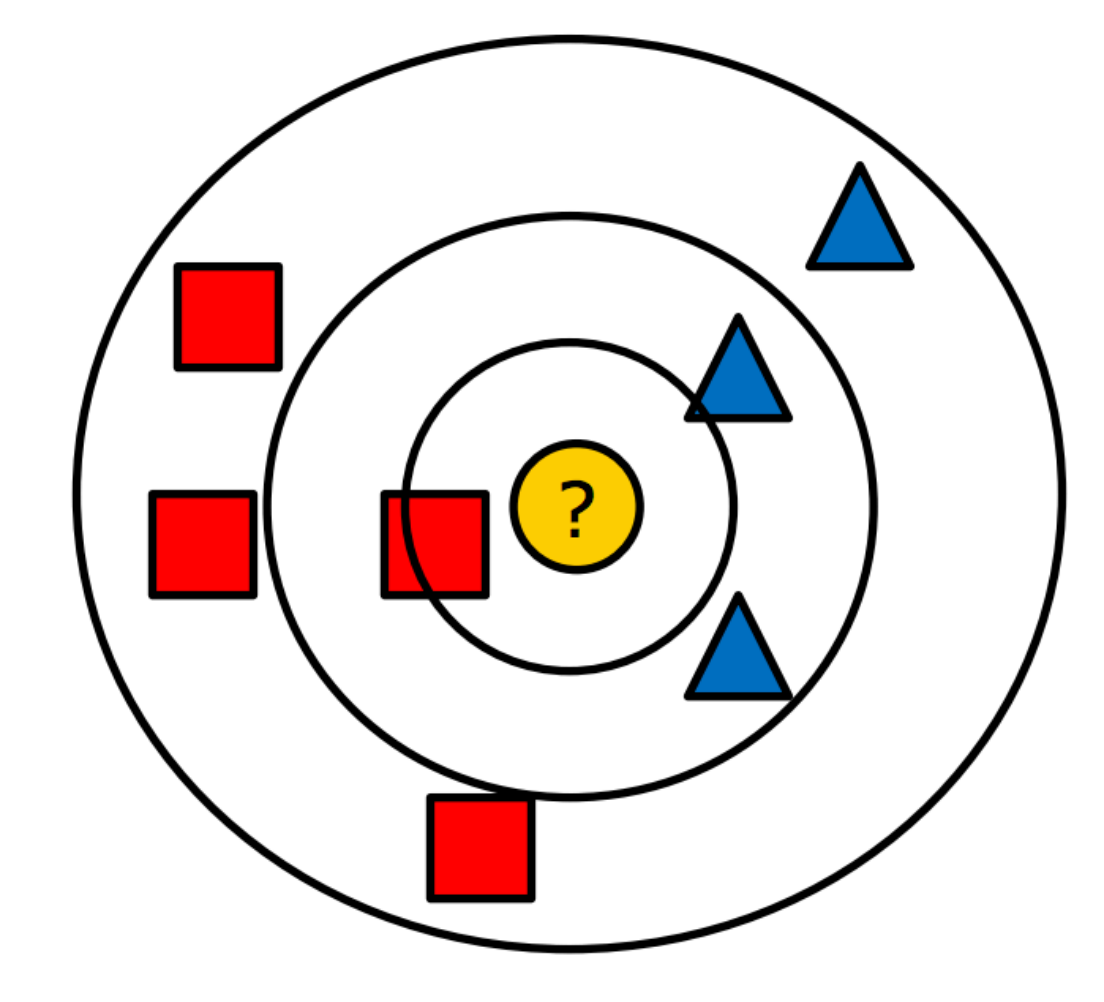
	Altura	Peso	Salario
0	1.77	90	10000
1	1.52	51	990
2	1.62	57	2000
3	1.82	95	3000
4	1.55	53	1200
5	1.93	100	5000

Depois

```
df_norm = (df - df.mean()) / (df.max() - df.min())
```

	Altura	Peso	Salario
0	0.166667	0.319728	0.699408
1	-0.443089	-0.476190	-0.300592
2	-0.199187	-0.353741	-0.188494
3	0.288618	0.421769	-0.077506
4	-0.369919	-0.435374	-0.277284
5	0.556911	0.523810	0.144469

KNN



K = 1 Pertence a classe de quadrados.

K = 3 Pertence a classe de triângulos.

K = 7 Pertence a classe de quadrados

KNN

Vantagens:
Técnica simples e facilmente implementada. Bastante flexível.
Em alguns casos apresenta ótimos resultados.

Desvantagens:
Classificar um exemplo desconhecido pode ser um processo computacionalmente complexo.

Requer um calculo de distância para cada exemplo de treinamento. Pode consumir muito tempo quando o conjunto de treinamento é muito grande.

A precisão da classificação pode ser severamente degradada pela presença de ruído ou características irrelevantes.