INTELIGÊNCIA ARTIFICIAL

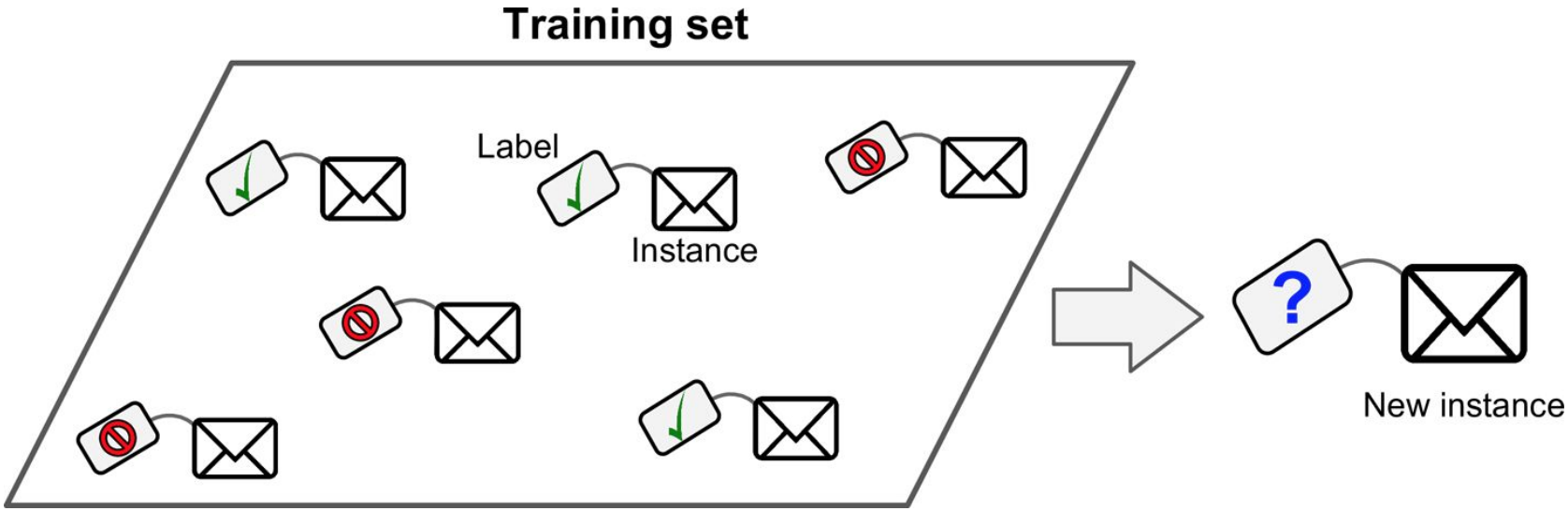
Aprendizado Preditivo
Classificação

Everton Dias
etgdb@cesar.school

Material produzido por
JP Magalhaes
jp@cesar.school

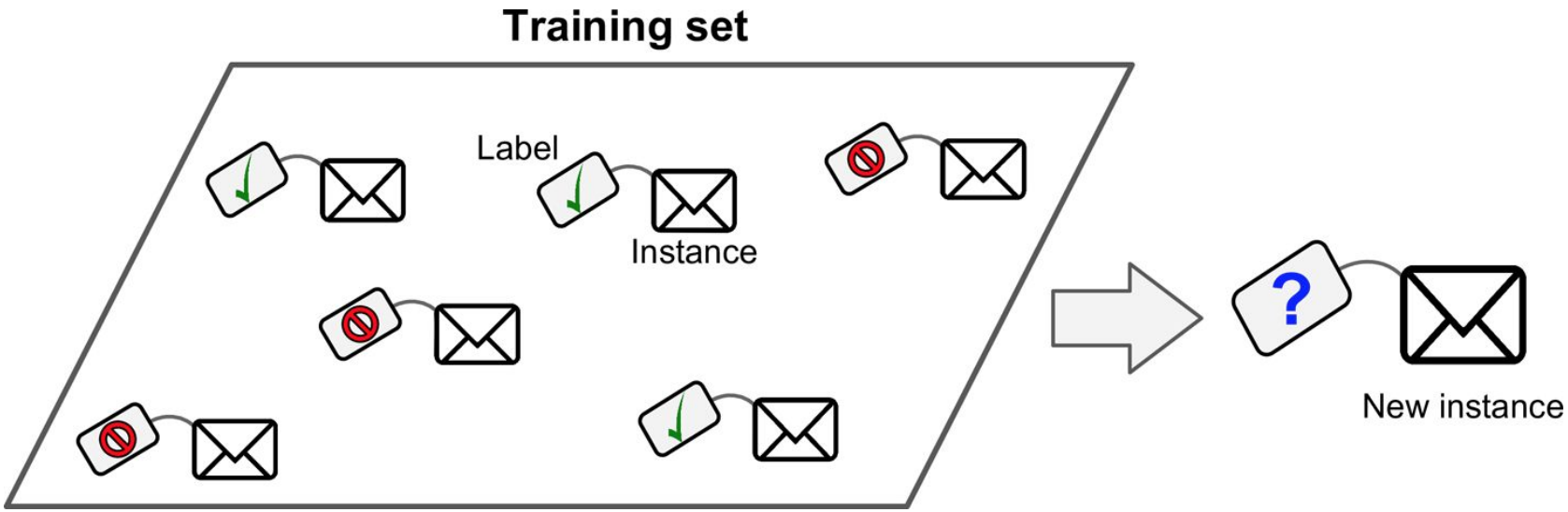
C.E.S.A.R school

Aprendizado Supervisionado (Preditivo)



No aprendizado supervisionado, a base de treinamento é fornecida ao algoritmo de Machine Learning contendo tanto a entrada **X** quanto a saída esperada **y.**

Classificação

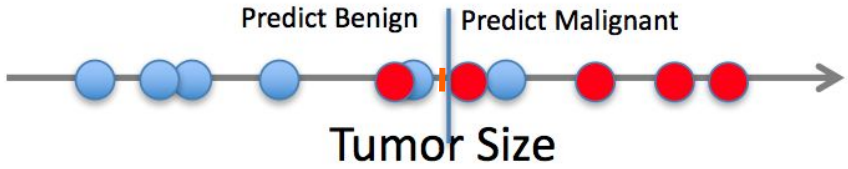


Na classificação, a saída gerada pelo modelo é um valor pertencente a um conjunto discreto, sendo chamada de categoria e, os valores, labels

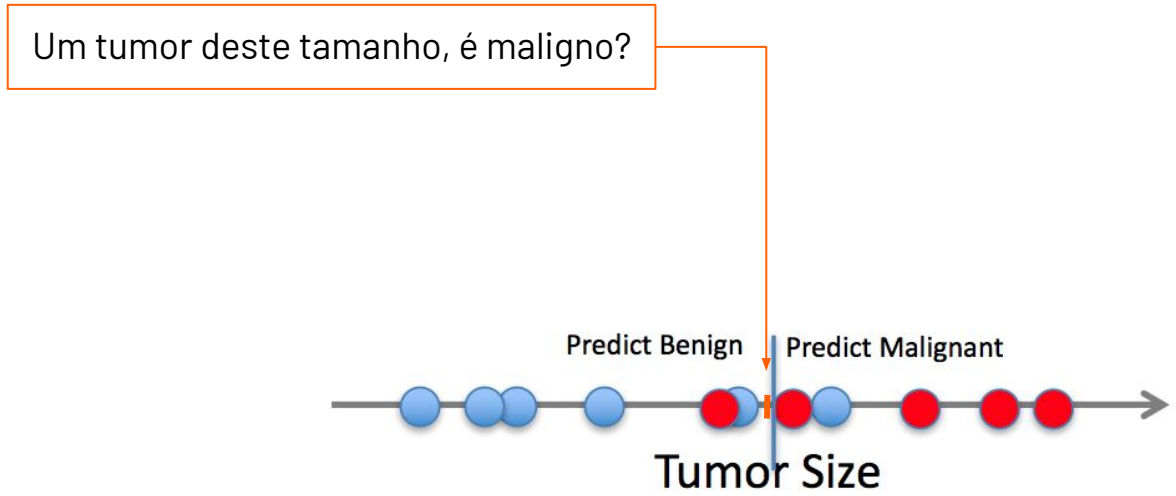
1

Aprendizado Supervisionado
Classificação

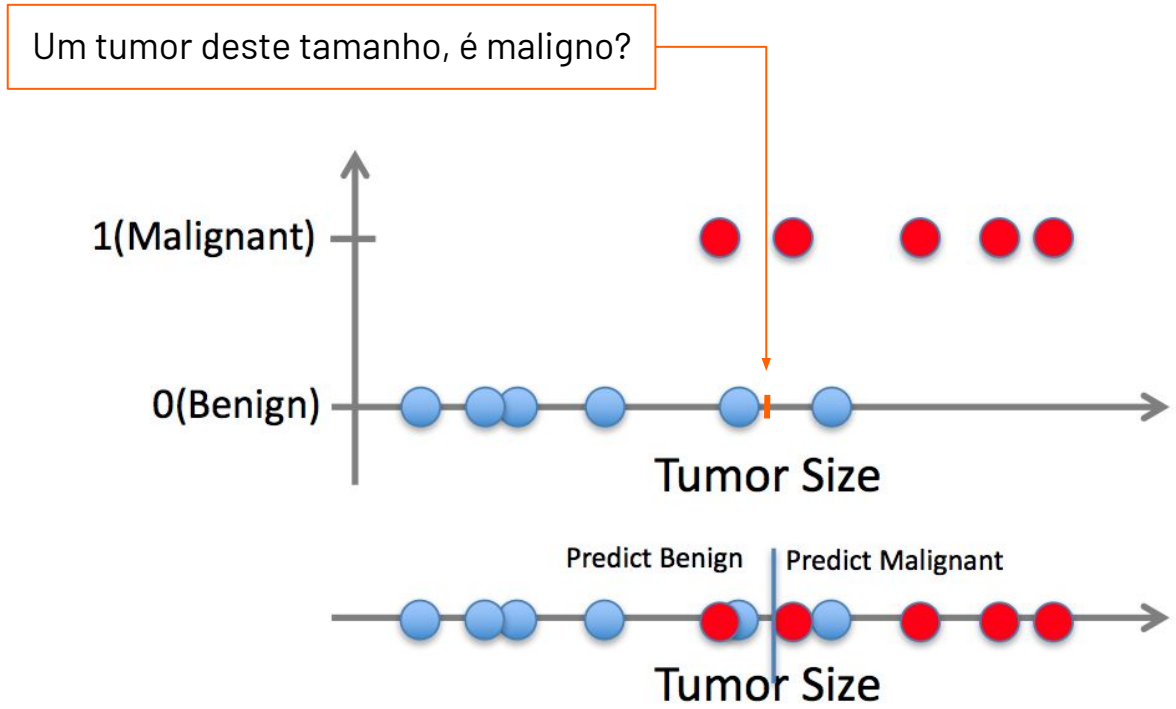
Classificação



Classificação



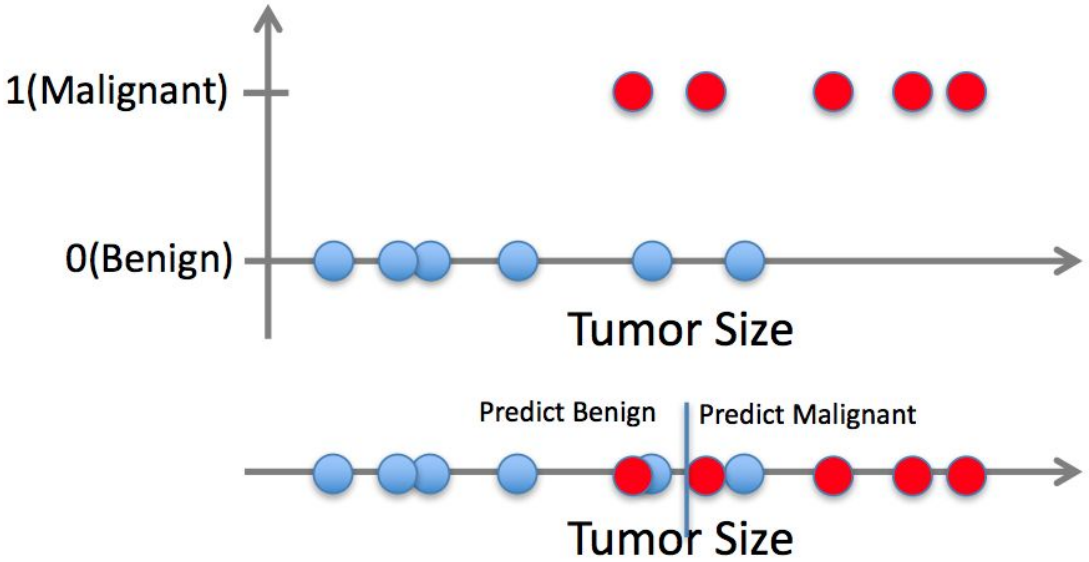
Classificação



Classificação

- Dados $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, $\mathbf{y \in C}$ - um conjunto discreto
  - **y é uma categoria (classe) → classificação**
- Aprende uma função $f(x)$ capaz de predizer y dado $\boldsymbol{X}$

Comumente, estaremos interessados em estimar as probabilidades de X pertencer a categoria C



2

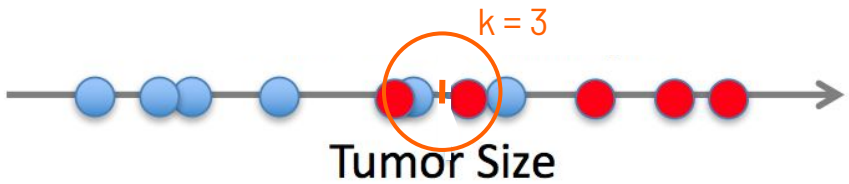
k-NN
k - Vizinhos mais Próximos

k-NN (Vizinhos mais próximos)

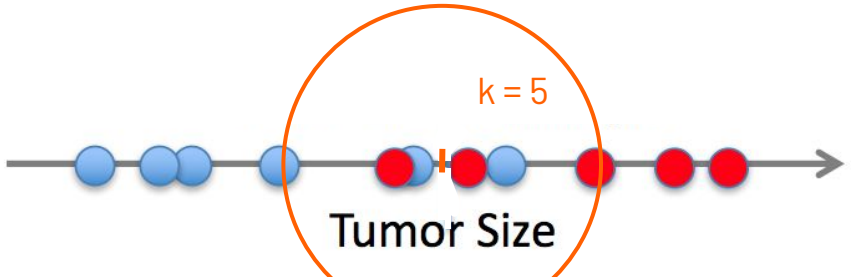
É razoável assumir que pessoas "próximas" pertencem ao mesmo grupo?

k-NN (Vizinhos mais próximos)

É razoável assumir que pessoas "próximas" pertencem ao mesmo grupo?



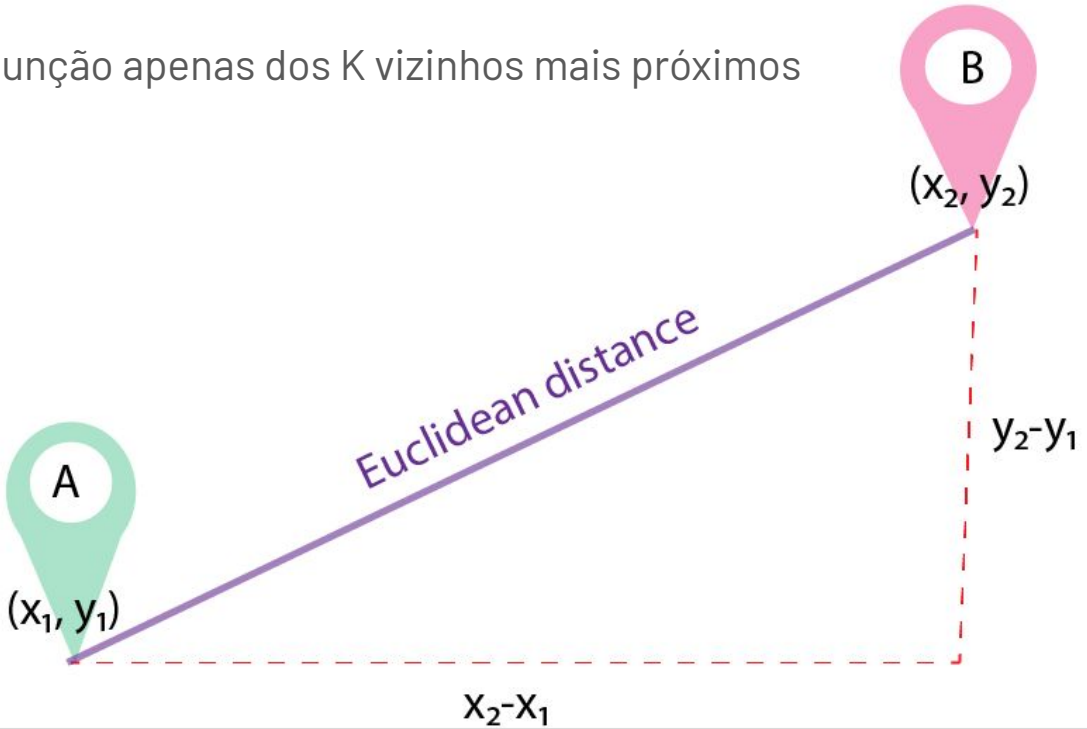
k-NN (Vizinhos mais próximos)



k-NN (Vizinhos mais próximos)

- Para cada nova amostra, ordena todos os exemplos conhecidos de acordo com uma medida de similaridade - Função Distância
- Porém, a saída é determinada em função apenas dos K vizinhos mais próximos
  - Classificação - votação
  - Regressão - média

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$



k-NN (Vizinhos mais próximos)

Exemplo

Cliente	Idade	Salário	# cartões	Classe
George	35	35	3	0
Paulo	22	50	2	1
Raquel	63	200	1	0
Pedro	59	170	1	1
Ana	25	40	4	0
Joao	37	50	2	?

- No kNN, características (diferentes dimensões) são colocadas na mesma equação
  Logo, terão impacto proporcional a sua magnitude
- Para sanar este problema, deve-se 'normalizar' os dados
  - Todas as características devem ter mesma magnitude
  - Utilizamos sklearn.preprocessing -> StandardScaler

$$z = \frac{x - \mu}{\sigma}$$

k-NN (Vizinhos mais próximos)

Exemplo: Cálculo das Distâncias

Desvio-Padrão	19,11	79,40	1,30
Média	40,8	99	2,2

* Não considera o exemplo de teste (João)

Cliente	Idade	Salário	# cartões	Classe
George	35	35	3	0
Paulo	22	50	2	1
Raquel	63	200	1	0
Pedro	59	170	1	1
Ana	25	40	4	0
Joao	37	50	2	?

k-NN (Vizinhos mais próximos)

Exemplo: Cálculo das Distâncias

Cliente	Idade	Salário	# cartões	Classe
George	-0,30	-0,81	0,61	0
Paulo	-0,98	-0,62	-0,15	1
Raquel	1,16	1,27	-0,92	0
Pedro	0,95	0,89	-0,92	1
Ana	-0,83	-0,74	1,38	0
Joao	-0,20	-0,62	-0,15	?

$$z = (x - u) / s$$

- u - média
- s - desvio-padrão

k-NN (Vizinhos mais próximos)

Exemplo: Cálculo das Distâncias

Cliente	Idade	Salário	# cartões	Classe
George	-0,30	-0,81	0,61	0
Paulo	-0,98	-0,62	-0,15	1
Raquel	1,16	1,27	-0,92	0
Pedro	0,95	0,89	-0,92	1
Ana	-0,83	-0,74	1,38	0
Joao	-0,20	-0,62	-0,15	?

$z = (x - u) / s$

- u - média
- s - desvio-padrão

Distance
$d\ (Joao\text{-}George) = [(-0,3 + 0,2)^2 + (-0,81 + 0,62)^2 + (0,61 + 0,15)^2\ ]^{1/2} = 0,80$
$d\ (Joao\text{-}Paulo) = [(-0,98 + 0,2)^2 + (-0,62 + 0,62)^2 + (-0,15 + 0,15)^2\ ]^{1/2} = 0,78$

k-NN (Vizinhos mais próximos)

Exemplo: Cálculo das Distâncias

Cliente	Idade	Salário	# cartões	Classe
George	0,30	0,81	0,61	0
Paulo	0,98	0,62	0,15	1
Raquel	1,16	1,27	0,92	0
Pedro	0,95	0,89	0,92	1
Ana	0,83	0,74	1,38	0
Joao	0,20	0,62	0,15	?

Distance	Classe
0,80	0
0,78	1
2,45	0
2,05	1
1,66	0

k-NN (Vizinhos mais próximos)

Exemplo: Cálculo das Distâncias

Cliente	Idade	Salário	# cartões	Classe
George	0,30	0,81	0,61	0
Paulo	0,98	0,62	0,15	1
Raquel	1,16	1,27	0,92	0
Pedro	0,95	0,89	0,92	1
Ana	0,83	0,74	1,38	0
Joao	0,20	0,62	0,15	1

Distance	Classe
0,80	0
0,78	1
2,45	0
2,05	1
1,66	0

k-NN (Vizinhos mais próximos)

Exemplo: Cálculo das Distâncias

Cliente	Idade	Salário	# cartões	Classe
George	0,30	0,81	0,61	0
Paulo	0,98	0,62	0,15	1
Raquel	1,16	1,27	0,92	0
Pedro	0,95	0,89	0,92	1
Ana	0,83	0,74	1,38	0
Joao	0,20	0,62	0,15	0

Distance	Classe
0,80	0
0,78	1
2,45	0
2,05	1
1,66	0

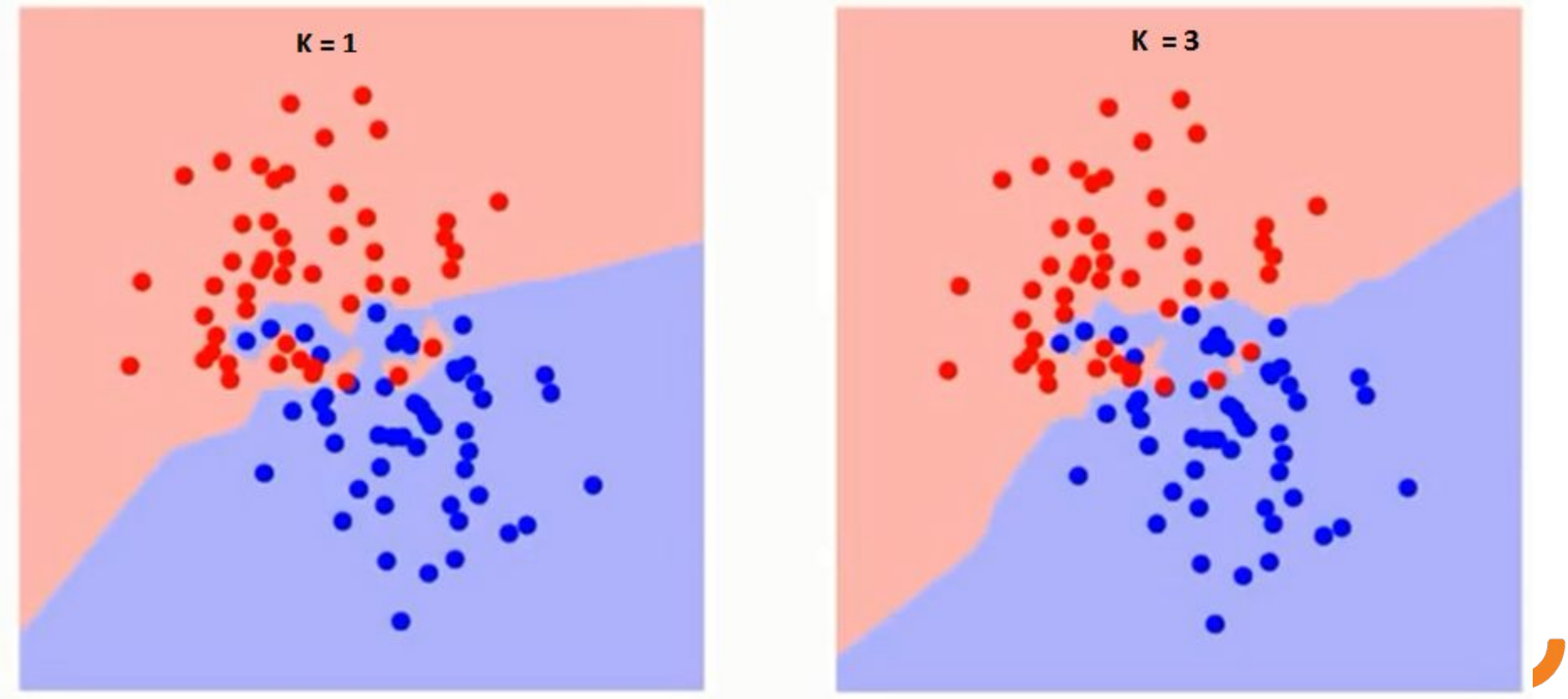
k-NN (Vizinhos mais próximos)

Como escolher o valor de K?

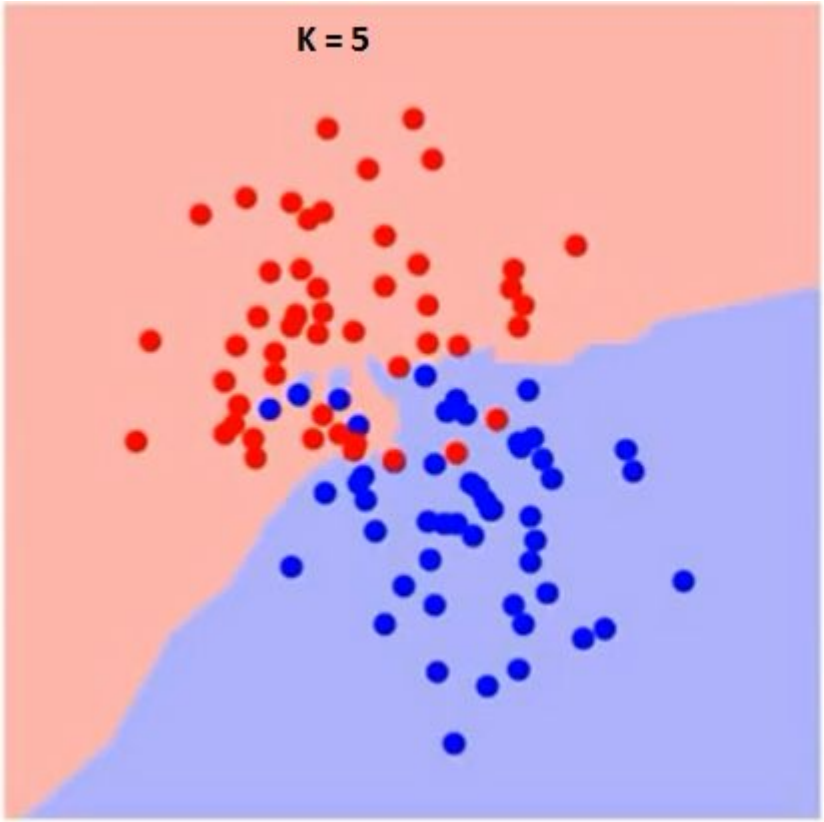
A forma mais básica é rodar o algoritmo várias vezes com diferentes valores de K e escolher o K que minimiza os erros

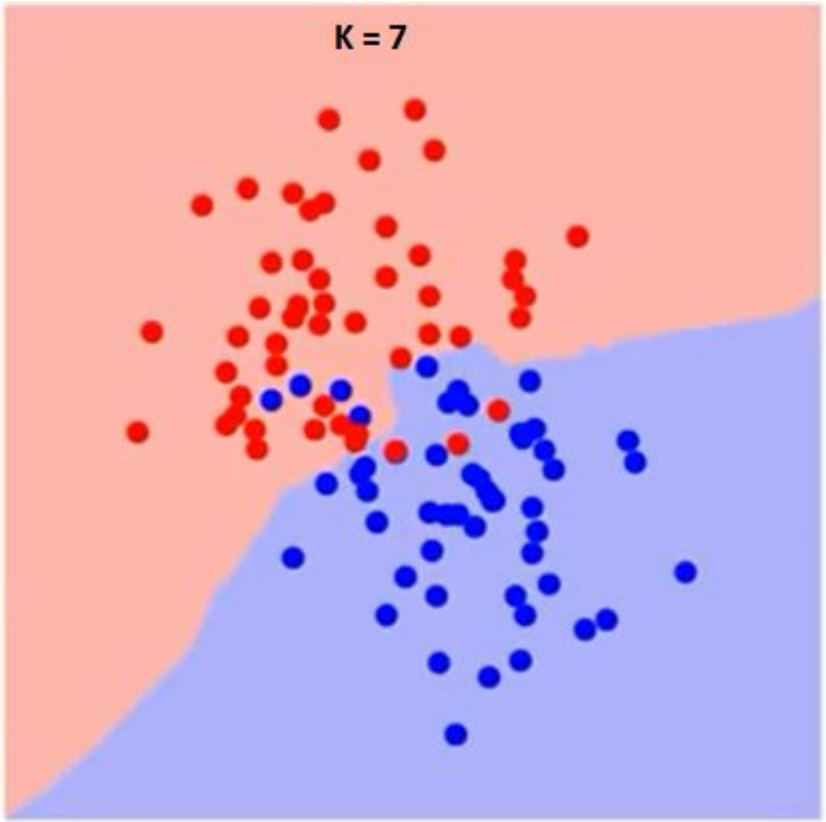
- K muito baixo torna as predições menos estáveis
- K muito alto torna o algoritmo muito rígido

k-NN - Impacto do valor de k na 'superfície de decisão'



k-NN - Impacto do valor de k na 'superfície de decisão'





k-NN

Prós

- Algoritmo preguiçoso - treinamento rápido
- Simples, fácil de entender e implementar
- Aplicável a problemas complexos
- Não é necessário treinamento prévio ou configurar muitos parâmetros
- Pode continuar aprendendo durante o uso
- Muito versátil: Classificação, Regressão, busca, aprendizado por reforço, etc

Contras

- Algoritmo preguiçoso - **predição custosa**
- Processamento e armazenamento cresce linearmente com o número de exemplos e características da base
- Não constrói um modelo explícito, não obtendo uma representação compacta dos dados
- Leva em consideração todas as características igualmente, sendo afetado por atributos redundantes e/ou irrelevantes

3

Prática

C . E . S . A . R

Pessoas impulsionando inovação.

Inovação impulsionando negócios.

NOSSO CONTATO

cesar.org.br

cesar.school