Curso 2 – CD, AM e DM

Mineração de Dados

Parte 7
Extração de Padrões
Classificação baseada em Instâncias
Método kNN

Prof. Ricardo M. Marcacini
ricardo.marcacini@icmc.usp.br

Extração de Padrões

Tarefas Preditivas

k-NN

Relembrando...

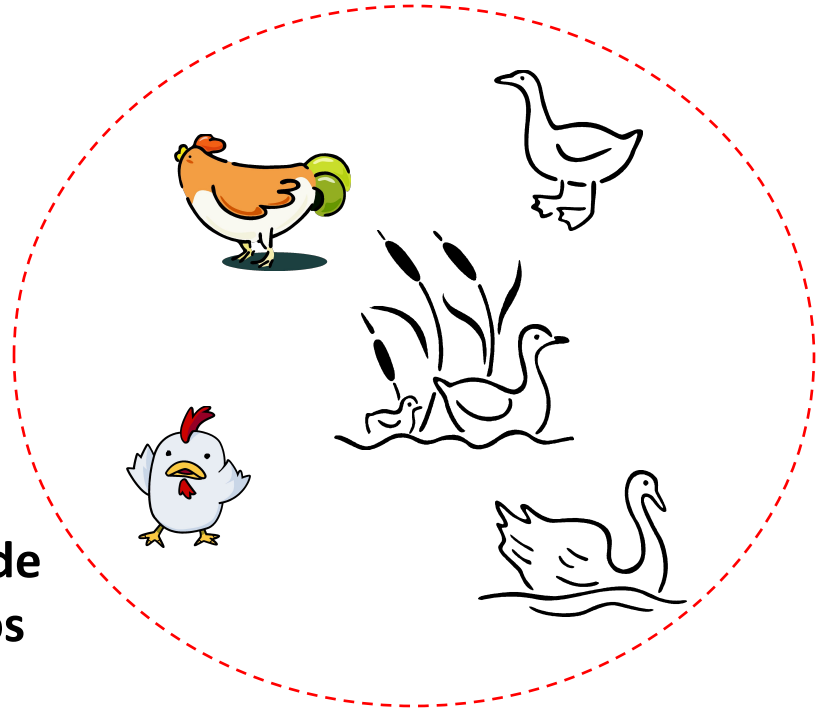
- Medidas de proximidade (Parte 2 de MD)

Distância Euclidiana	Distância de Mahalanobis
Distância de Manhattan	Cosseno
Distância Suprema	Casamento Simples
Distância de Minkowski	Jaccard

O Método kNN (*k-Nearest Neighbors*)

- Ideia principal e motivação:

> *"Se anda como um pato, grasna como um pato, age como um pato, então provavelmente é um pato."*



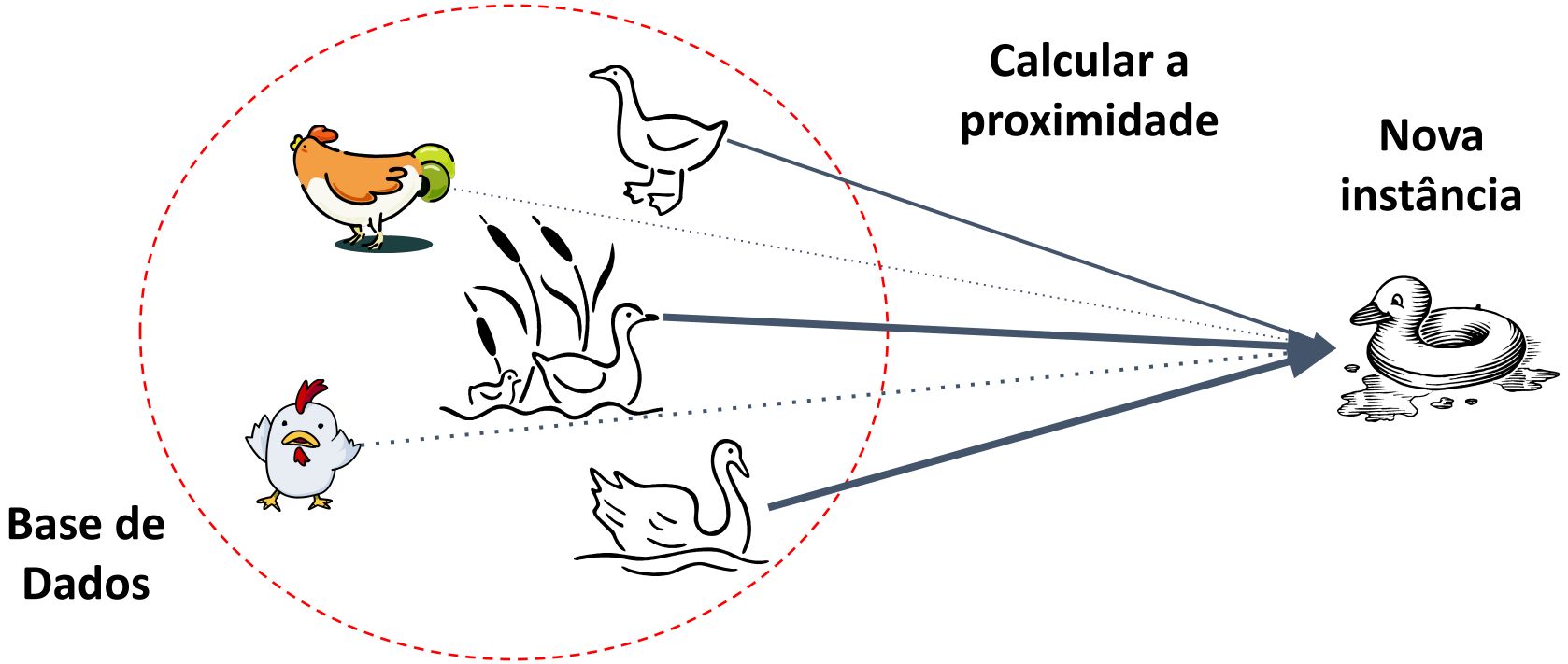
Base de Dados

Nova instância

O Método kNN

- Ideia principal e motivação:

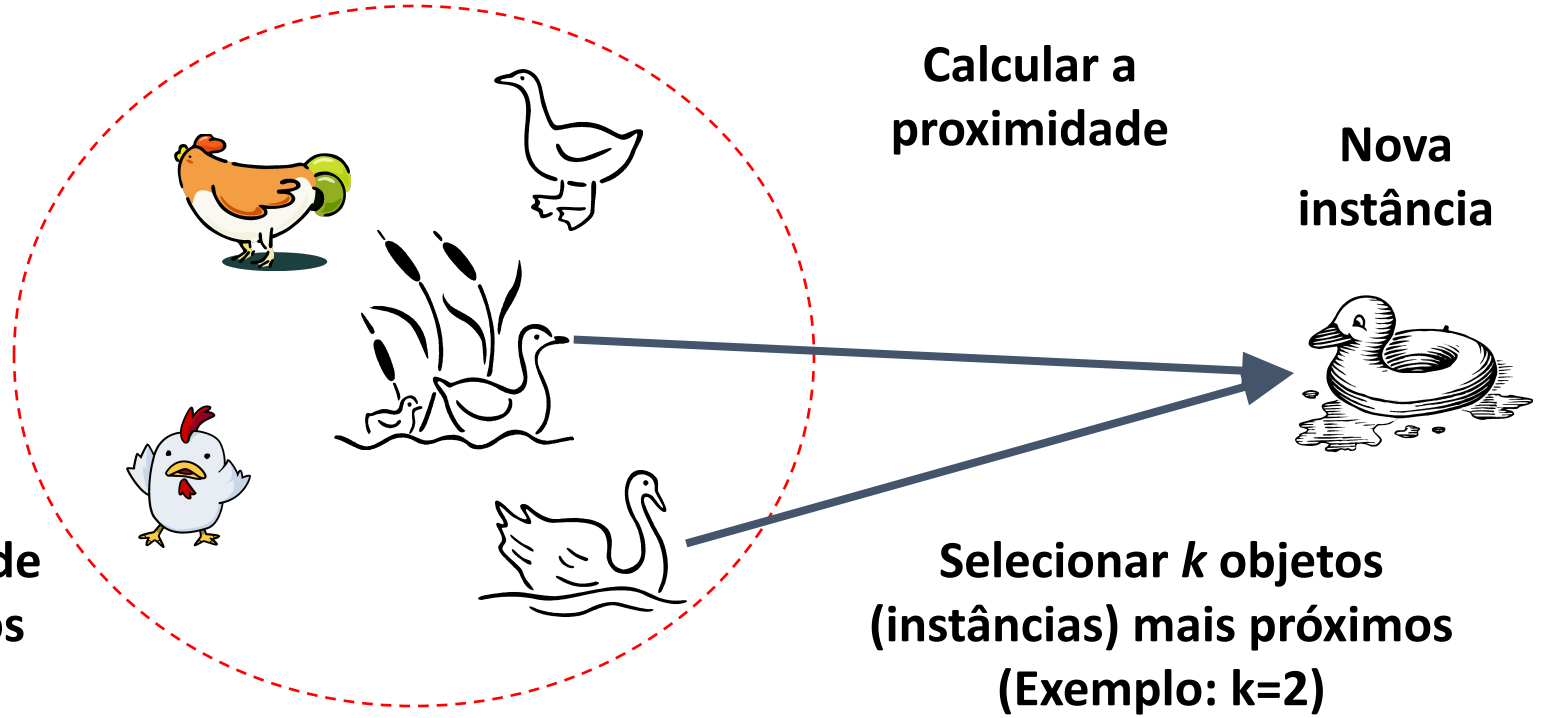
> *"Se anda como um pato, grasna como um pato, age como um pato, então provavelmente é um pato."*



O Método kNN

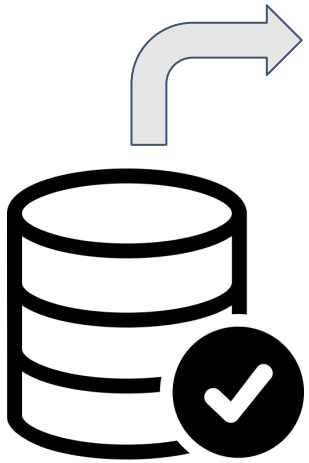
- Ideia principal e motivação:

> *"Se anda como um pato, grasna como um pato, age como um pato, então provavelmente é um pato."*



O Método kNN

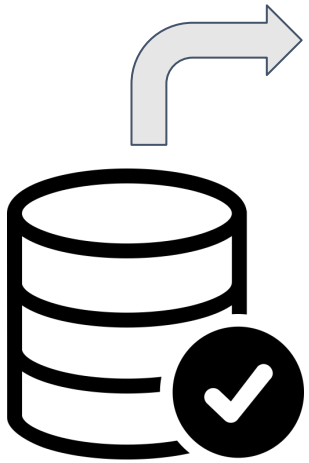
- Objetos (instâncias) com informação de rótulo



	Atributos				Classe
Objeto	Atributo #1	Atributo #2	...	Atributo #d	y
1					
2					
...					
n					

O Método kNN

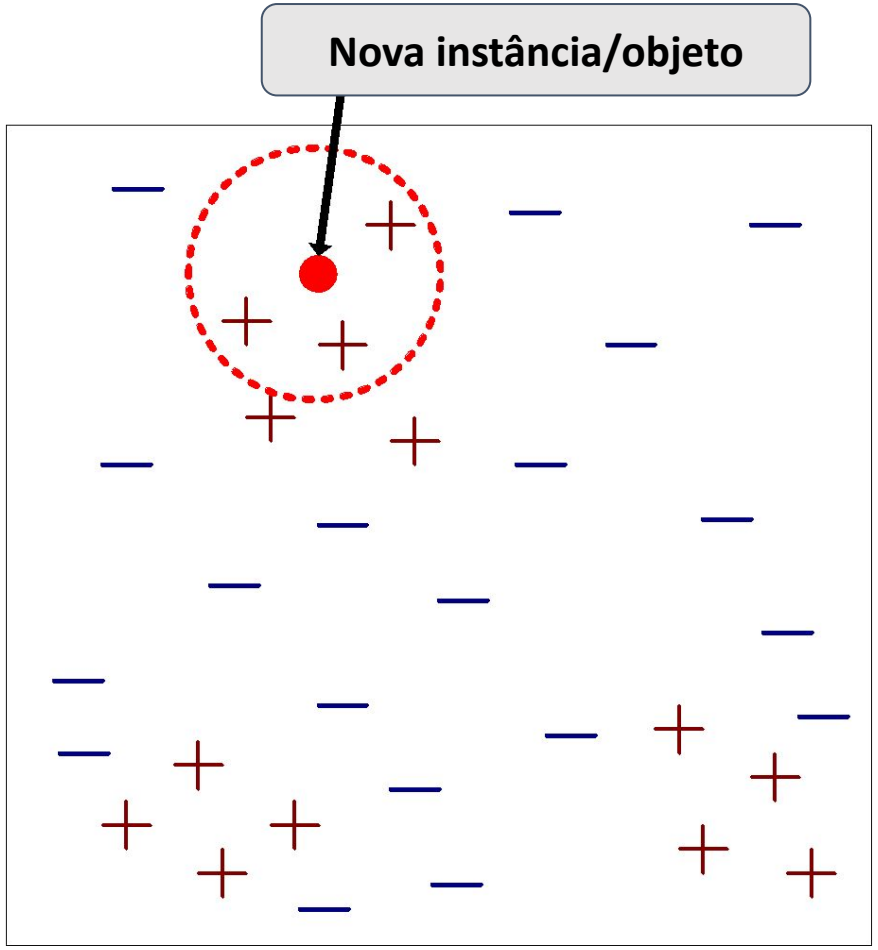
- Objetos (instâncias) com informação de rótulo



	Atributos				Classe
Instância	**Sintoma #1**	**Sintoma #2**	**...**	**Sintoma #d**	**Doença**
Paciente 1					+
Paciente 2					+
Paciente 3					-
Paciente 4					-

O Método kNN

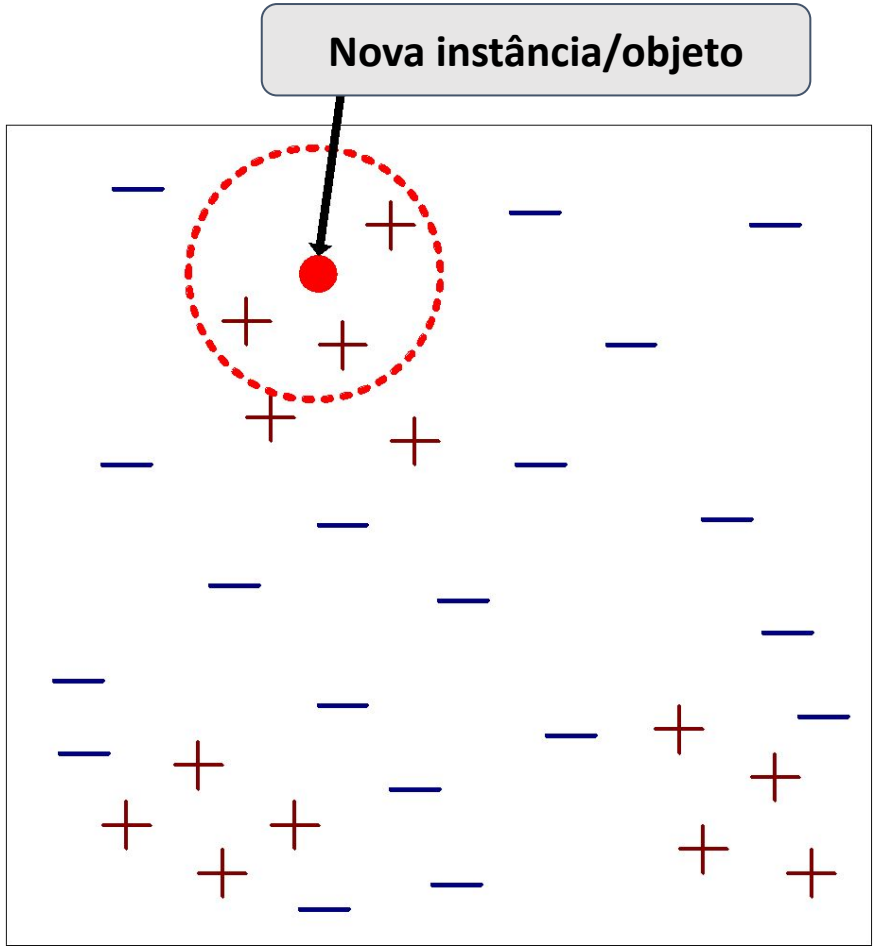
- Pré-requisitos e parâmetros



  - Uma base de instâncias/objetos com <u>informação de classe</u>

O Método kNN

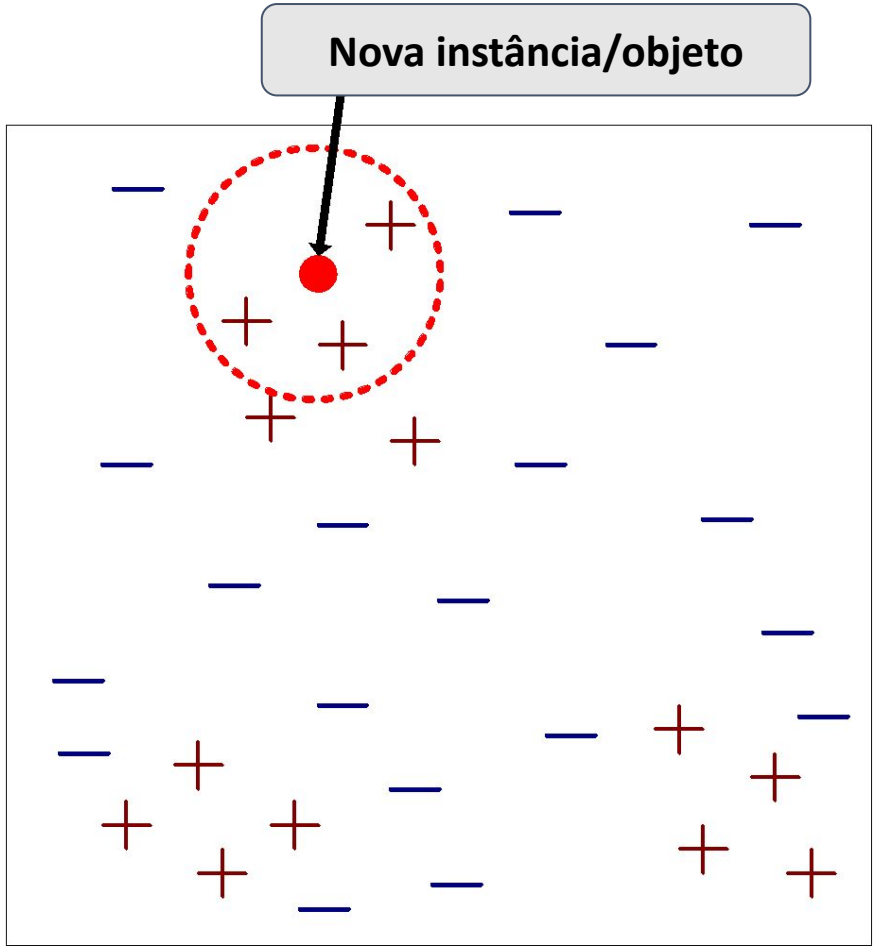
- Pré-requisitos e parâmetros



- Uma base de instâncias/objetos com informação de classe
- Escolher um medida de proximidade entre objetos

O Método kNN

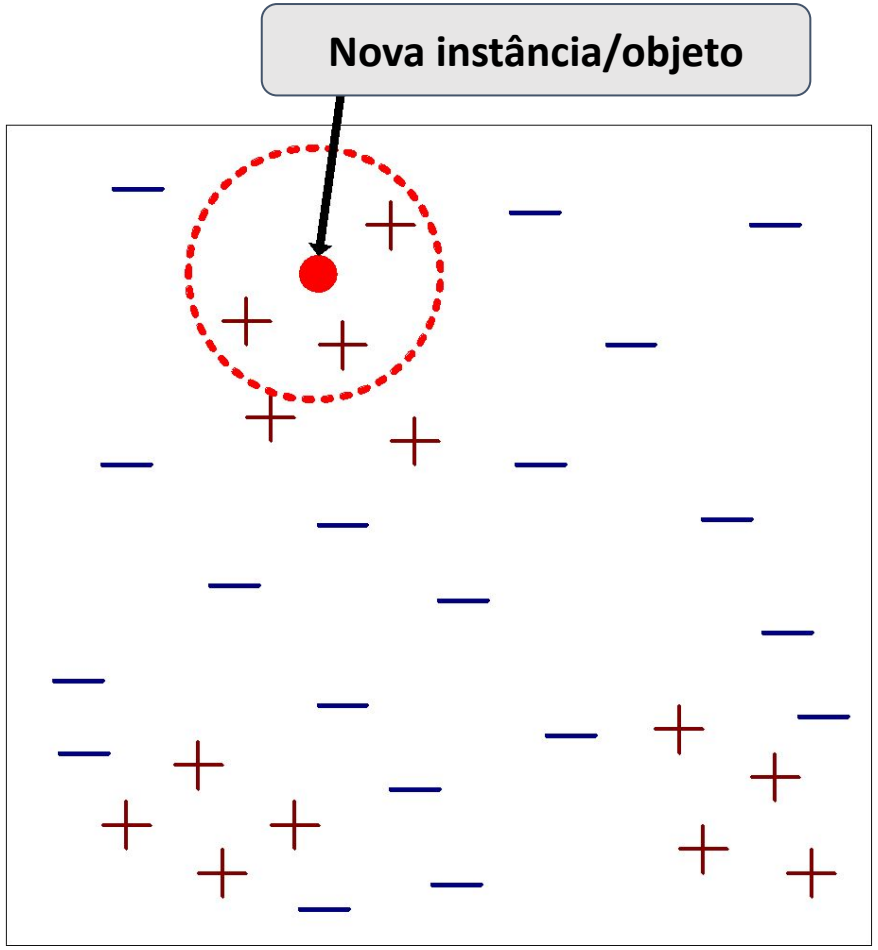
- Pré-requisitos e parâmetros



- Uma base de instâncias/objetos com informação de classe
- Escolher um medida de proximidade entre objetos
- Definir um valor de k (quantidade de vizinhos mais próximos)

O Método kNN

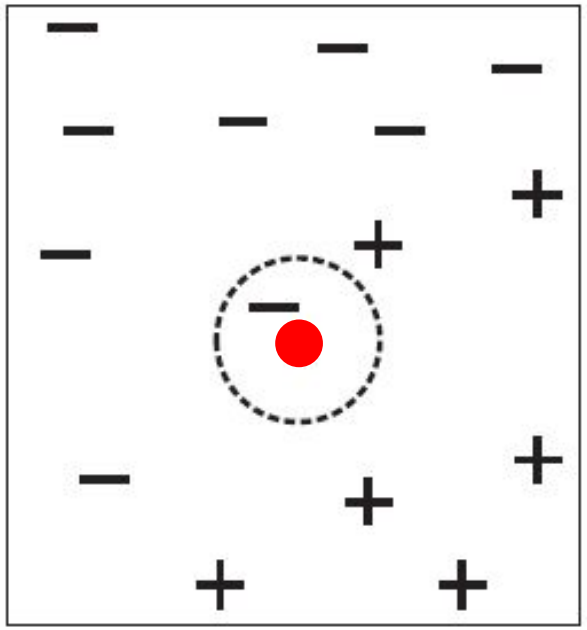
- Pré-requisitos e parâmetros



- Uma base de instâncias/objetos com informação de classe
- Escolher um medida de proximidade entre objetos
- Definir um valor de k (quantidade de vizinhos mais próximos)
- Uma estratégia de votação para determinar a classe

O Método kNN

- O efeito do parâmetro k (votação majoritária)

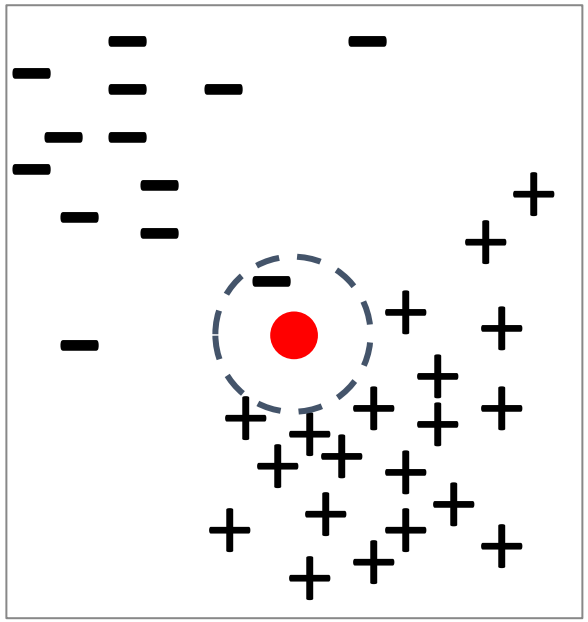


1-NN

Classificação usando a informação de classe do vizinho rotulado mais próximo do novo objeto

O Método kNN

- O efeito do parâmetro k (votação majoritária)



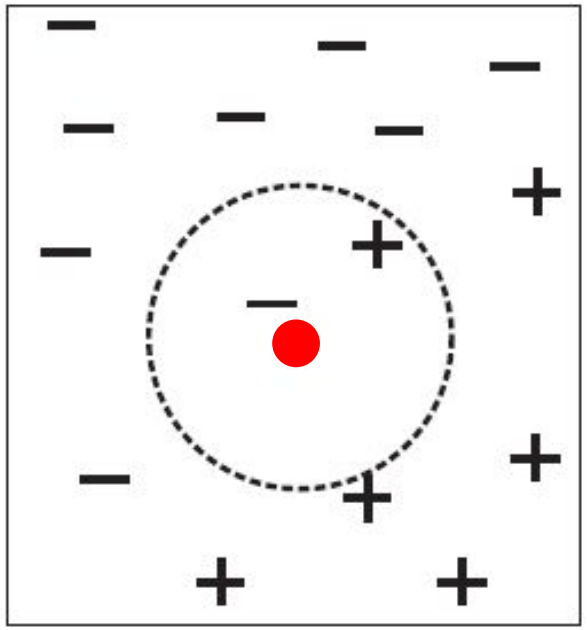
1-NN

Classificação usando a informação de classe do vizinho rotulado mais próximo do novo objeto

Sensível a objetos ruidosos/*outliers*

O Método kNN

- O efeito do parâmetro k (votação majoritária)



2-NN

Valores pares (e pequenos) para k podem ocasionar empates.

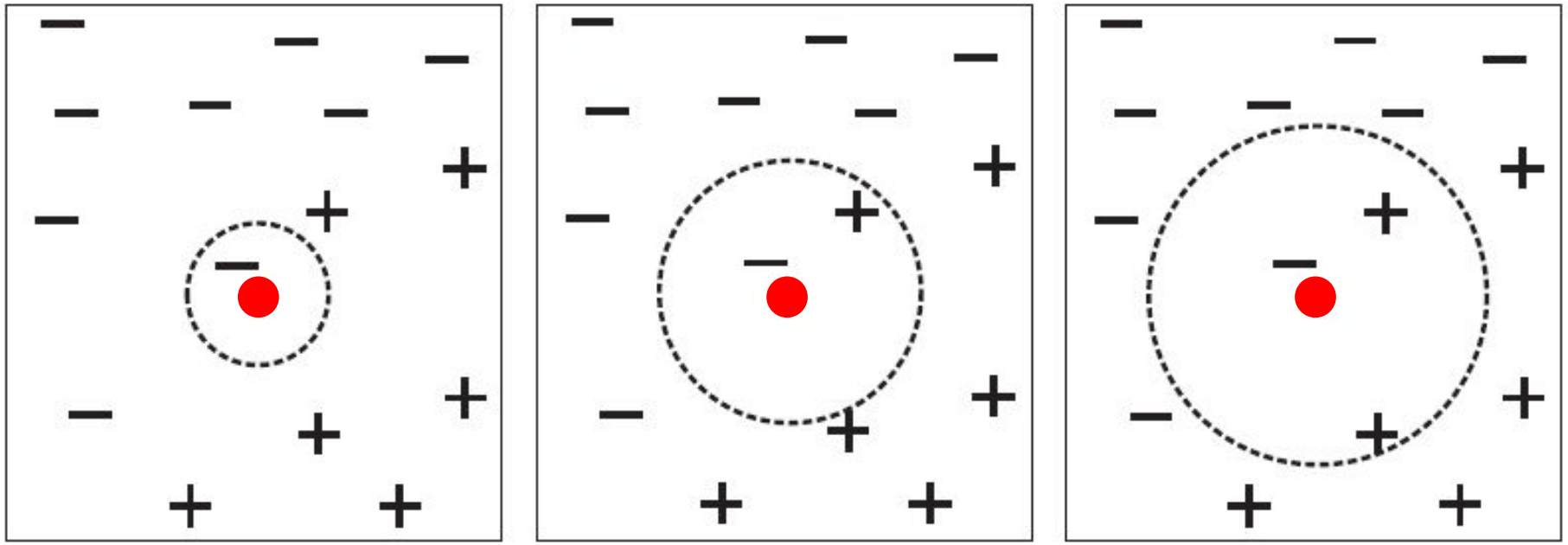
Também pode ocorrer com valores ímpares para k, por exemplo, k=3 e três classes.

Estratégias comuns para desempatar:

- Utilizar o resultado com “k-1”
- Votação ponderada (em breve)

O Método kNN

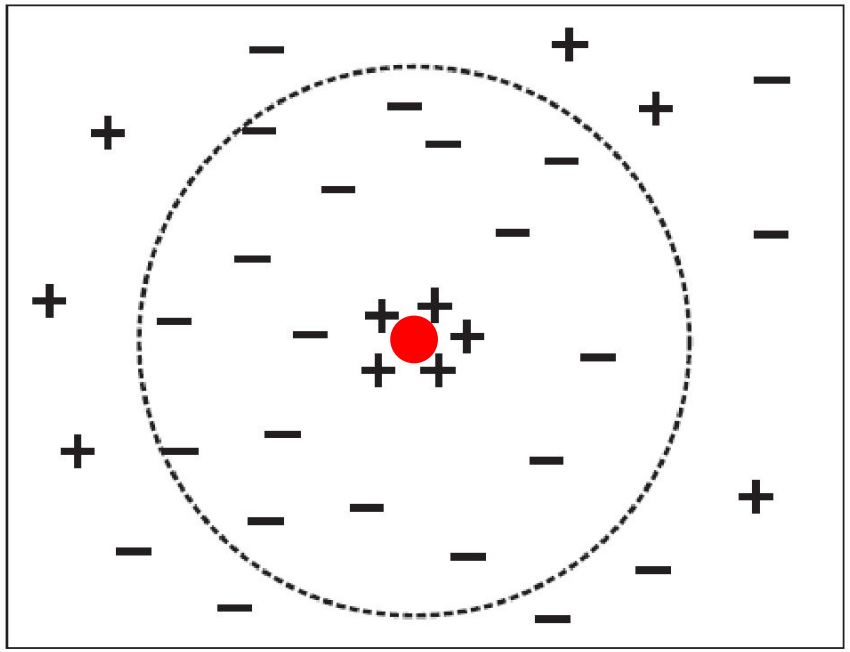
- O efeito do parâmetro k (votação majoritária)



1-NN 2-NN 3-NN

O Método kNN

- O efeito do parâmetro k (votação majoritária)



O uso de valores muito grandes para *k* tende a incluir objetos de outras classes como vizinhos

O Método kNN

- Exemplo (Conjunto de dados IRIS)

ID	Sepal Length	Sepal Width	Petal Length	Petal Width	Class
1	5,1	3,5	1,4	0,2	Iris-setosa
2	4,9	3,0	1,4	0,2	Iris-setosa
3	7,0	3,2	4,7	1,4	Iris-versicolor
4	6,4	3,2	4,5	1,5	Iris-versicolor
5	6,3	3,3	6,0	2,5	Iris-virginica
6	5,8	2,7	5,1	1,9	Iris-virginica

Sepal Length	Sepal Width	Petal Length	Petal Width	Class
5,4	3,1	2,5	1,0	???

- Vamos usar distância euclidiana e testar valores de *k* entre 1 e 6 (votação majoritária).

O Método kNN

- Exemplo (Conjunto de dados IRIS)

Ranking	ID	Distância	Classe
1º	1	1,44	Iris-setosa
2º	2	1,45	Iris-setosa
3º	4	2,29	Iris-versicolor
4º	3	2,68	Iris-versicolor
5º	6	2,80	Iris-virginica
6º	5	3,91	Iris-virginica

- 1-NN = Iris-setosa
- 2-NN = Iris-setosa
- 3-NN = Iris-setosa
- 4-NN = Empate
- 5-NN = Empate
- 6-NN = Empate

O Método kNN

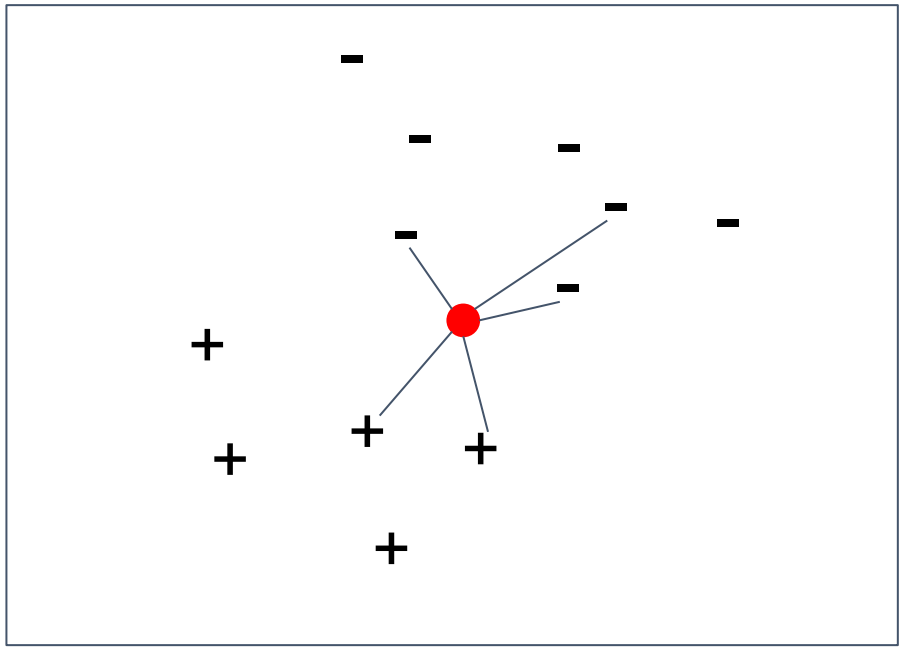
- Voto ponderado
  - Cada objeto vizinho (*x*) recebe um peso conforme sua distância em relação ao *novo* objeto

$$voto = \frac{1}{dist(x, novo)}$$

  - O novo objeto é classificado conforme o somatório (ponderado) de cada voto

O Método kNN

- Voto ponderado
  - O novo objeto é classificado conforme o somatório (ponderado) de cada voto



O Método kNN

- Exemplo com voto ponderado

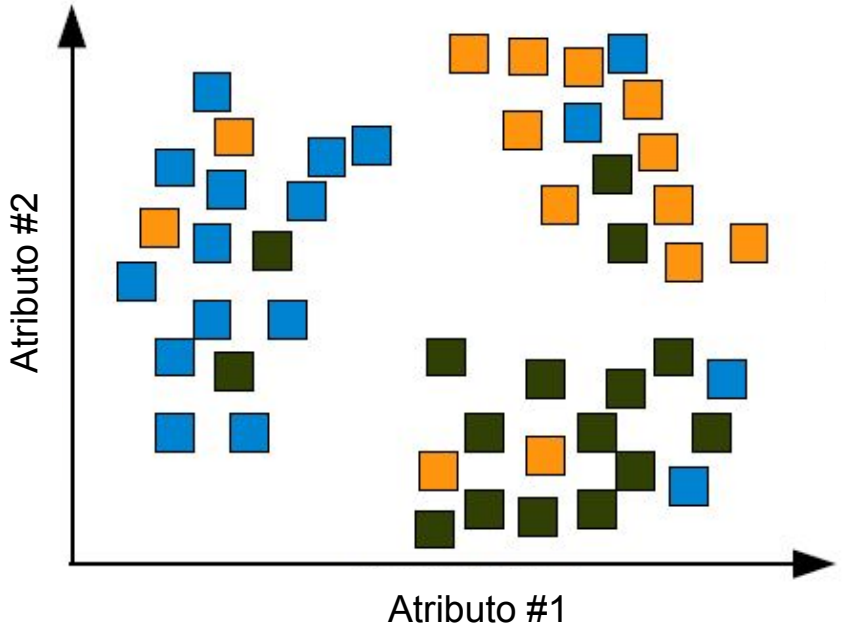
Ranking	ID	Distância	Classe
1º	1	1,44	Iris-setosa
2º	2	1,45	Iris-setosa
3º	4	2,29	Iris-versicolor
4º	3	2,68	Iris-versicolor
5º	6	2,80	Iris-virginica
6º	5	3,91	Iris-virginica

- 1-NN = Iris-setosa
- 2-NN = Iris-setosa
- 3-NN = Iris-setosa
- 4-NN = Iris-setosa
- 5-NN = Iris-setosa
- 6-NN = Iris-setosa

O Método kNN

- Seleção de instâncias e remoção de *outliers*

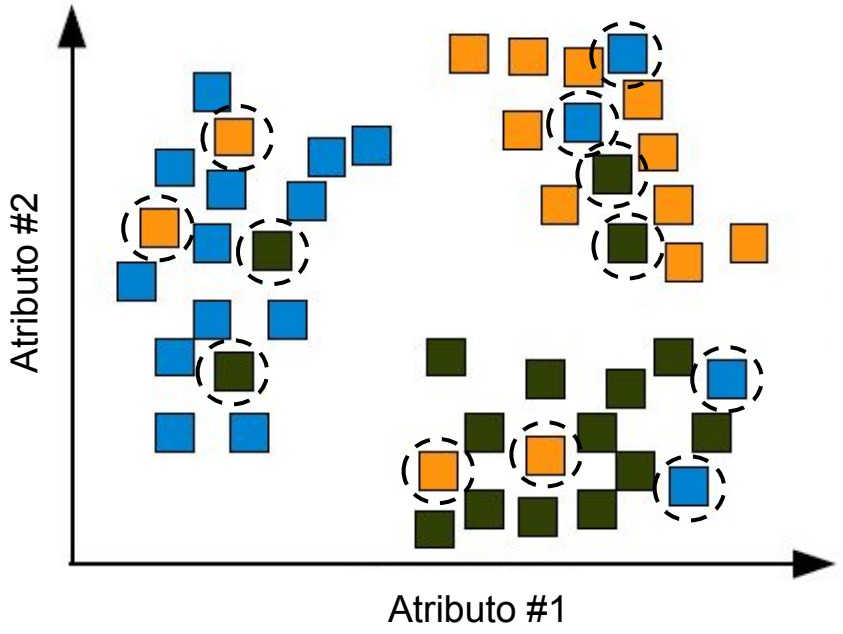
Se uma instância/objeto do conjunto de treinamento possui vizinhos com classes diferentes, então esse exemplo pode ser considerado um *outlier* e ser removido do conjunto de treinamento.



O Método kNN

- Seleção de instâncias e remoção de *outliers*

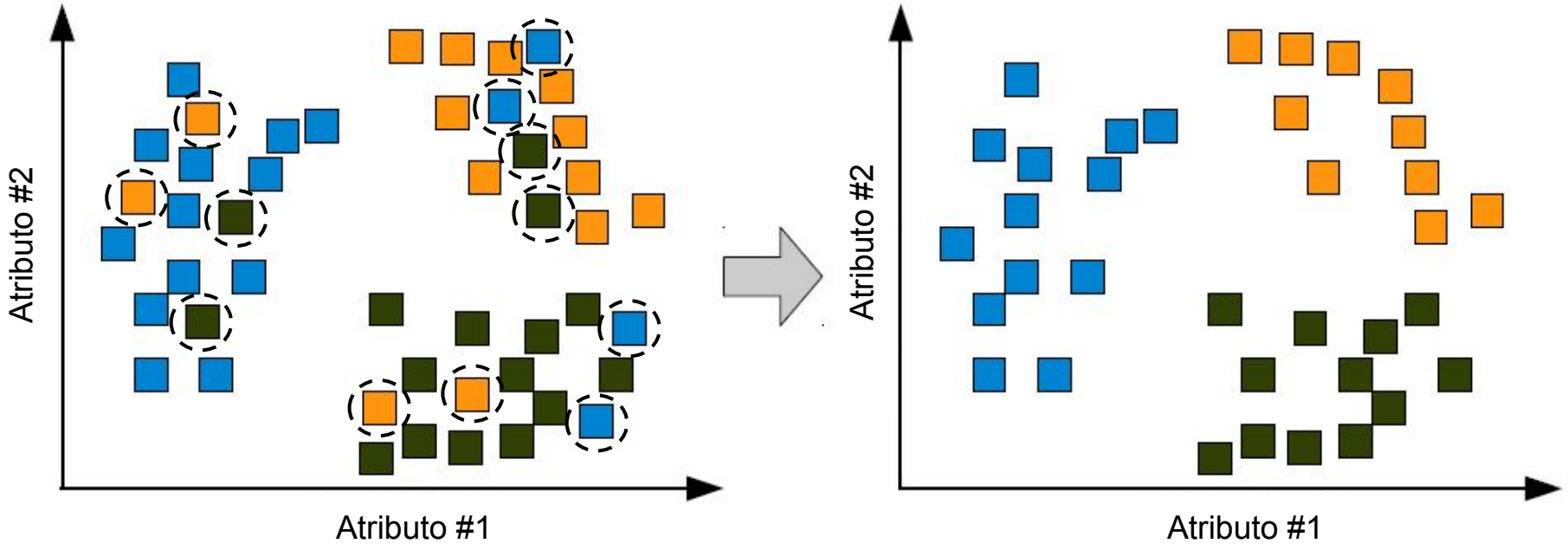
Se uma instância/objeto do conjunto de treinamento possui vizinhos com classes diferentes, então esse exemplo pode ser considerado um *outlier* e ser removido do conjunto de treinamento.



O Método kNN

- Seleção de instâncias e remoção de *outliers*

Se uma instância/objeto do conjunto de treinamento possui vizinhos com classes diferentes, então esse exemplo pode ser considerado um outlier e ser removido do conjunto de treinamento.



O Método kNN

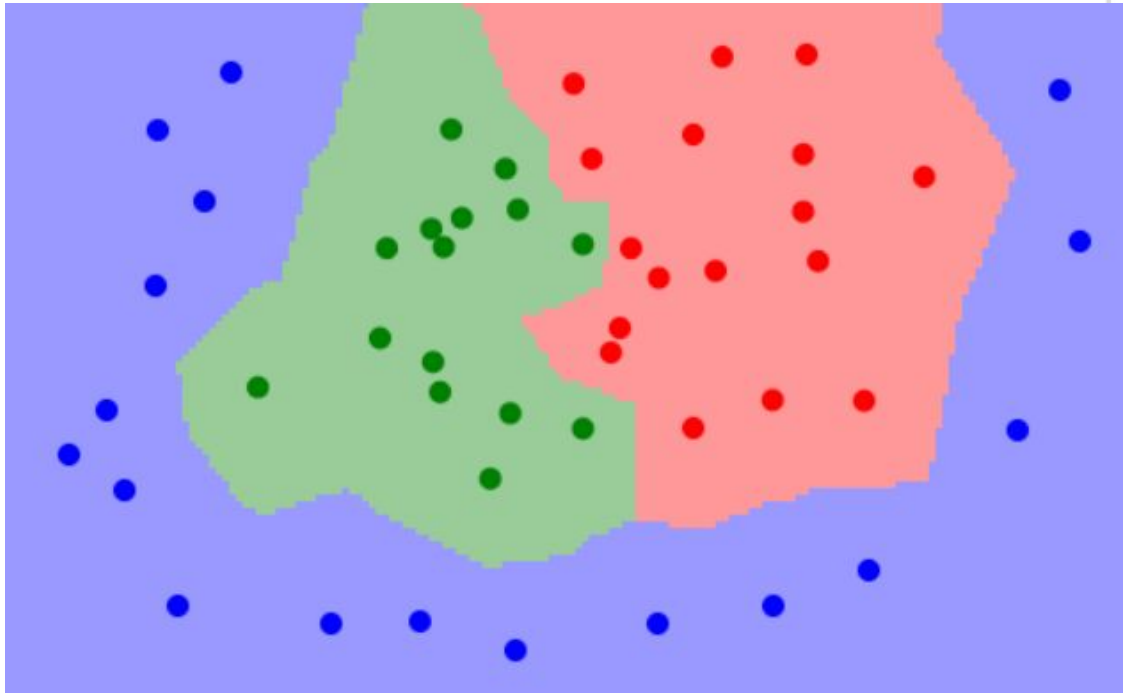
- Propriedades
  - Classificação local
  - Estratégia *lazy* (preguiçosa)

Armazena os objetos do treinamento. Espera um novo objeto de teste para realizar a classificação.

  - Não paramétrico

Não assume qualquer distribuição a respeito dos dados.

  - Capaz de modelar espaços de decisões complexos



http://vision.stanford.edu/teaching/cs231n-demos/knn/

O Método kNN

- kNN para problemas de regressão

O valor (numérico) predito é a média dos valores do atributo classe dos *k*-vizinhos mais próximos

ID	Idade	Anos de Profissão	Salário
1	20	2	2000
2	25	3	2500
3	50	25	8000
4	32	12	5000
5	27	5	3000
6	30	10	2700
7	31	13	????

Qual o salário estimado de ID=7?

O Método kNN

- kNN para problemas de regressão

O valor (numérico) predito é a média dos valores do atributo classe dos *k*-vizinhos mais próximos

ID	Idade	Anos de Profissão	Salário
1	20	2	2000
2	25	3	2500
3	50	25	8000
4	32	12	5000
5	27	5	3000
6	30	10	2700
7	31	13	????

Qual o salário estimado de ID=7?

Resposta:

Usando k=2 e distância euclidiana

$$\frac{5000+2700}{2} = 3850$$

O Método kNN

- kNN para problemas de regressão

Observação: mesma ideia pode ser empregada para tratamento de valores ausentes (imputação de valores ausentes)

ID	Idade	Anos de Profissão	Salário
1	20	2	2000
2	25	3	2500
3	50	25	8000
4	32	12	5000
5	27	5	3000
6	30	10	2700
7	31	13	????

Qual o salário estimado de ID=7?

Resposta:

Usando k=2 e distância euclidiana

$$\frac{5000+2700}{2} = 3850$$

Considerações Finais

- Determinar parâmetros é um problema experimental
  - Valor de *k*
  - Medida de proximidade
  - Votação majoritária ou ponderada

Considerações Finais

- Determinar parâmetros é um problema experimental
  - Valor de *k*
  - Medida de proximidade
  - Votação majoritária ou ponderada
- Padronizar os atributos

Considerações Finais

- Determinar parâmetros é um problema experimental
  - Valor de *k*
  - Medida de proximidade
  - Votação majoritária ou ponderada
- Padronizar os atributos
- Técnicas para reduzir custo computacional
  - Paralelismo
  - Remover objetos redundantes e *outliers*
  - Indexação para acelerar o cálculo de distâncias

Bibliografia

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.

Tan, P.N.; Steinbach, M.; Karpatne, A.; Kumar, V. (2016). *Introduction to Data Mining (2nd Edition)*. Pearson.