UNIDADE III: APRENDIZAGEM NÃO SUPERVISIONADA

- Sumário:
  - Introdução
  - Agrupamento (*clustering*)
  - Algoritmo *k – médias*

Objectivos

- Aprofundar a noção de aprendizagem não supervisionada
- Adquirir a noção de agrupamento
- Descrever o algoritmo *k – médias*

Aprendizagem não supervisionada

- Em muitas circunstâncias não se conhecem as classes existentes num problema ou os exemplos de treino não estão etiquetados
- Se utilizam técnicas de agrupamento (*clustering*)
- Têm como objectivo a determinação das classes mais prováveis e a divisão das instâncias em grupos "naturais"

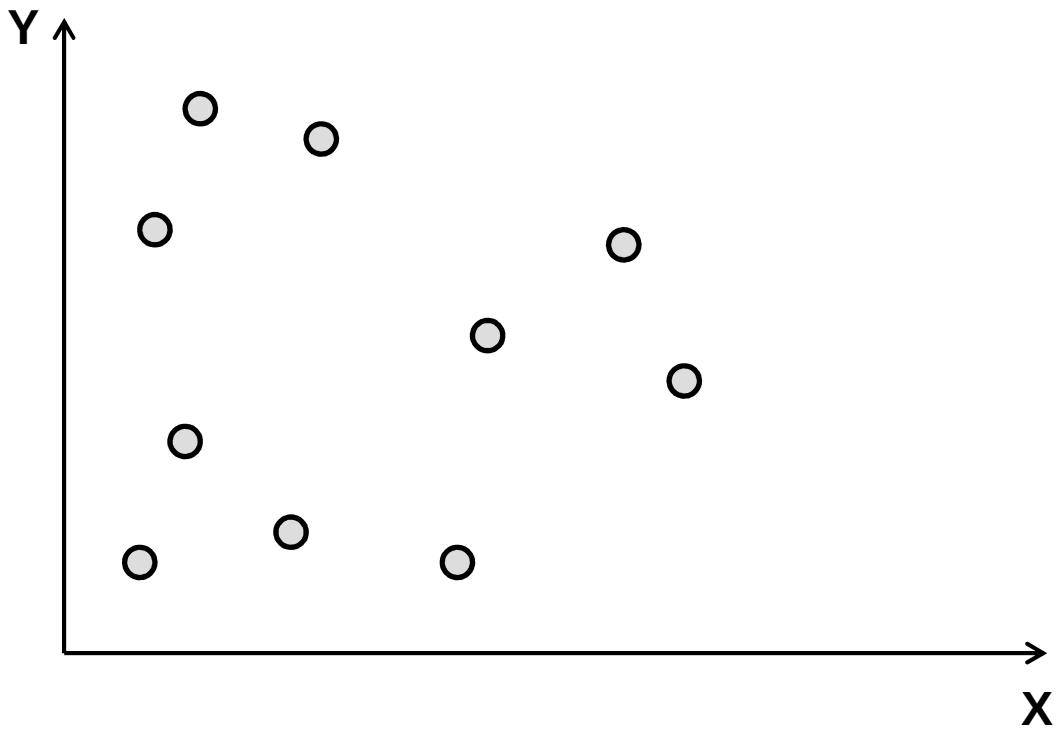
Aprendizagem não supervisionada

- Existem várias formas de saída para este tipo de algoritmos
  - Partição vs. hierarquia
    - Partição: agrupam as instâncias num conjunto de classes mutuamente excluentes
    - Hierarquia: geram uma hierarquia de classes na qual umas classes incluem a outras
  - Exclusivos vs. não exclusivos vs. Fuzzy
    - Exclusivos: cada objeto pertence a um único cluster
    - Não exclusivos: existem objetos que são associados a diferentes clusters
    - Fuzzy : objetos são associados a um cluster com um certo grau de pertinência

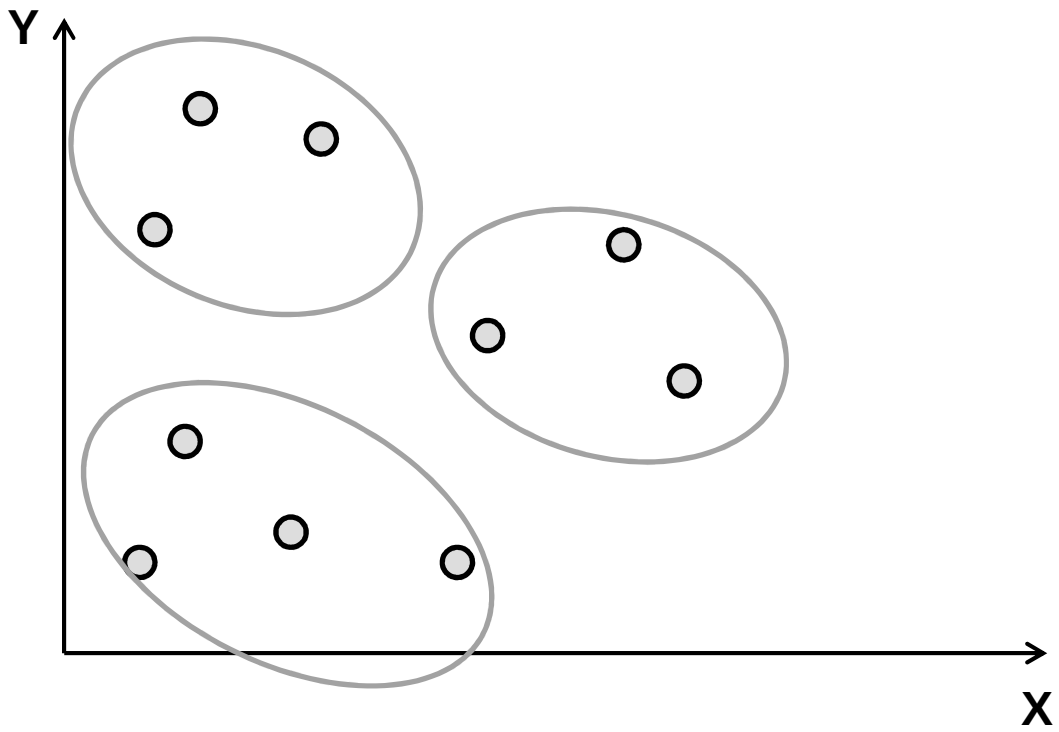
Aprendizagem não supervisionada

- Existem várias formas de saída para este tipo de algoritmos
  - Completos vs parciais
    - Completos: cada objeto pertence a algum cluster
    - Parciais: existem objetos que não estão associados a nenhum cluster (outliers, ruidos, sem interesse)

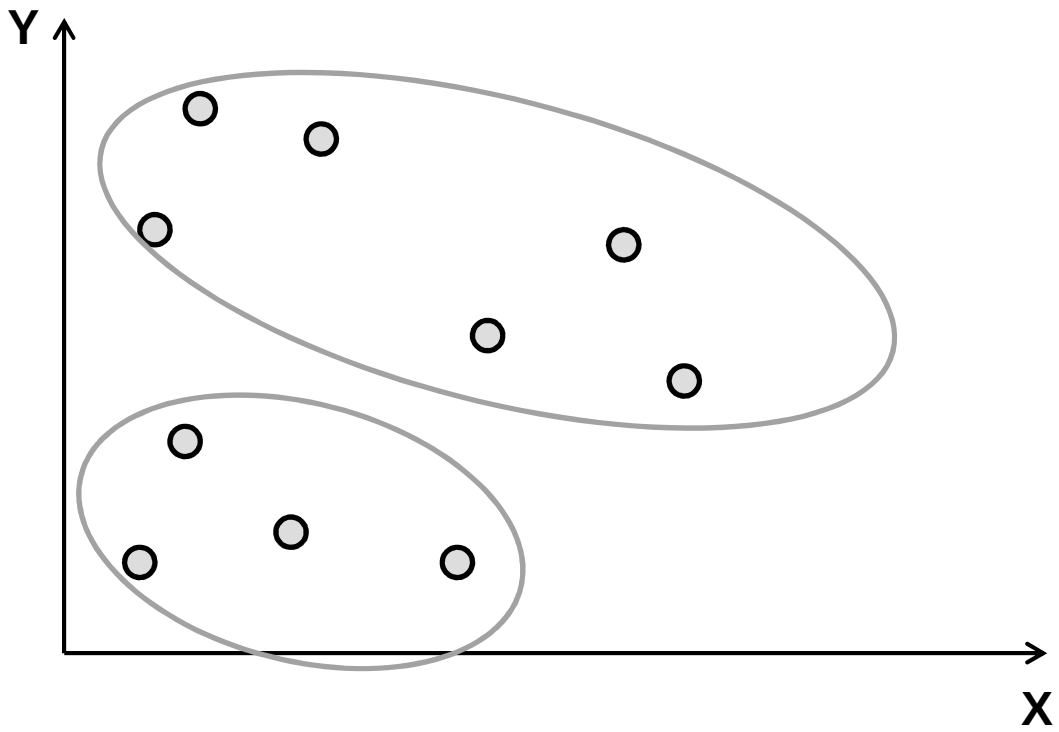
Aprendizagem não supervisionada



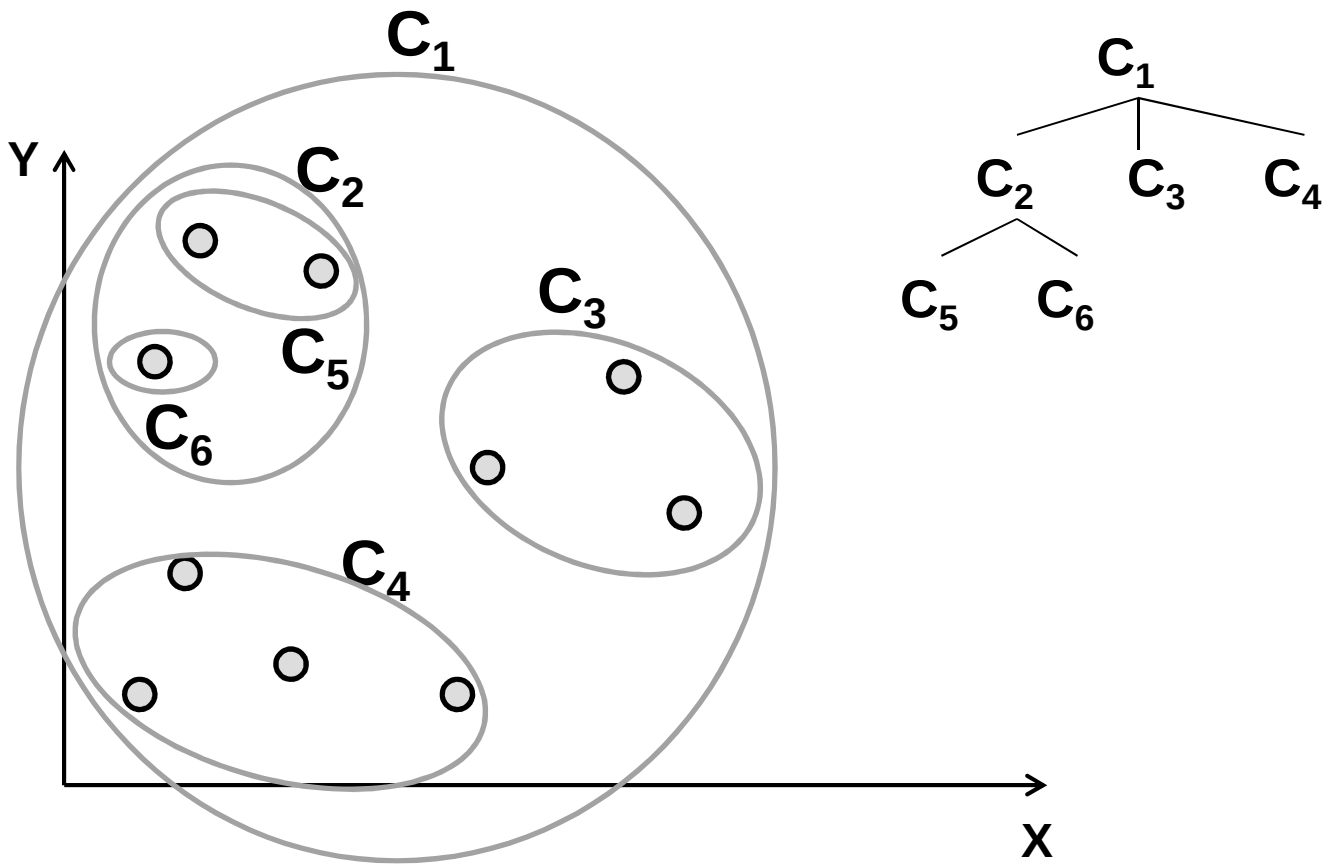
Aprendizagem não supervisionada



Aprendizagem não supervisionada



Aprendizagem não supervisionada



Agrupamento

- Método clássico -> agrupamento em k médias (*k – means*)
- Separa os exemplos de treino num conjunto de classes que agrupam os elementos mais parecidos
- O número de classes a criar é proporcionado pelo utilizador

Algoritmo k – médias

- Tem como objectivo a obtenção de uma partição do conjunto de exemplos de treino de forma tal que se maximize a semelhança entre os elementos de uma classe e minimize a semelhança entre elementos de diferentes classes

Algoritmo k – médias

- Dado
  - Conjunto de exemplos de treino, E
  - Número de classes que se deseja obter, k
  - Medida de semelhança/distância, M
- Determina
  - Conjunto de k classes, C

Algoritmo k – médias

- Selecciona aleatoriamente *k* instâncias como centros de cada classe (*semente*)
- Se atribui cada instância ao *cluster* correspondente ao centro mais próximo
- Recalcula a *semente* de cada classe, através do ponto médio (centroide) dos exemplos agrupados na classe
- Processo repetido com os novos centros calculados para os *clusters*
- Iterações terminam quando os mesmos pontos são associados a cada cluster em rondas consecutivas

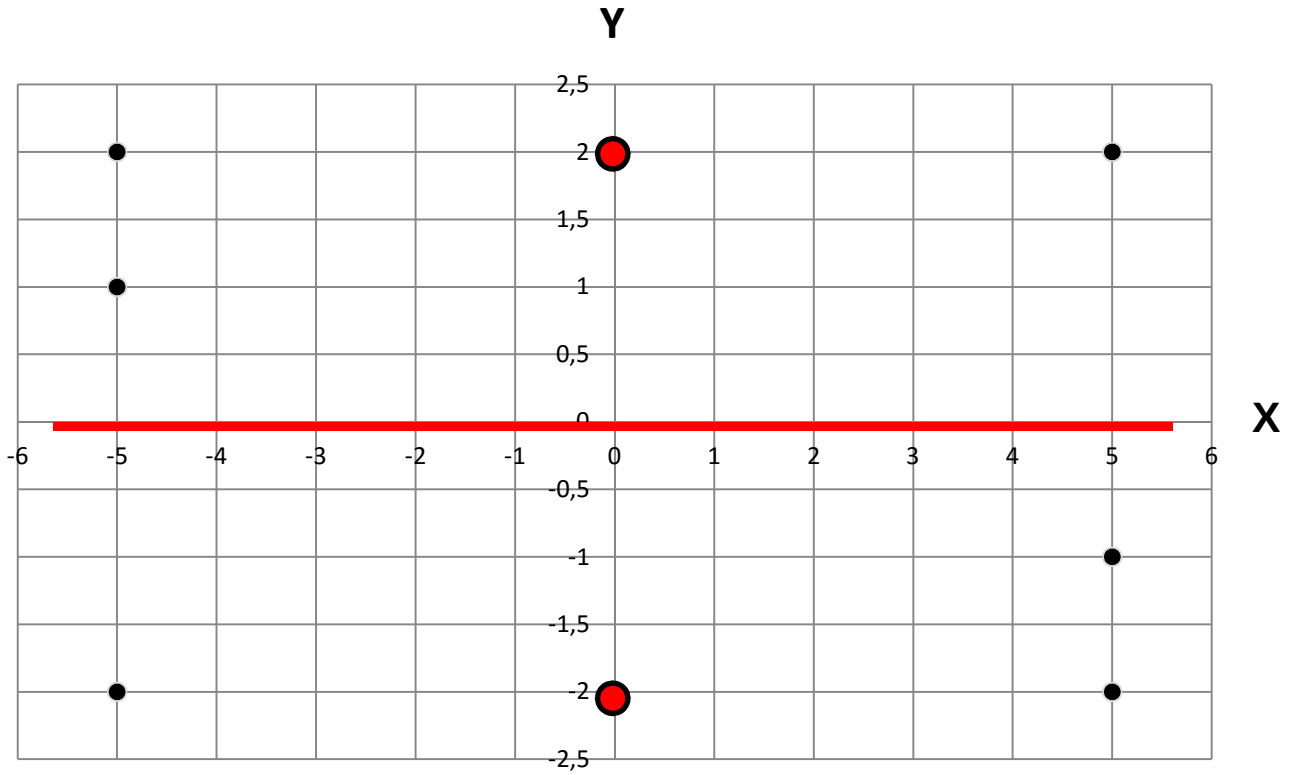
Exemplo

- Dados os exemplos indicados na tabela, caracterizados por 2 atributos, utilize o algoritmo *k médias* para determinar uma partição adequada dos dados em 2 grupos. Realize 3 execuções do algoritmo, partindo de diferente centros iniciais a seguir indicados:
  - $C_1 = (0, -2)$, $C_2 = (0, 2)$
  - $C_1 = (-5, 2)$, $C_2 = (5, -2)$
  - $C_1 = (0, -2)$, $C_2 = (5, 2)$

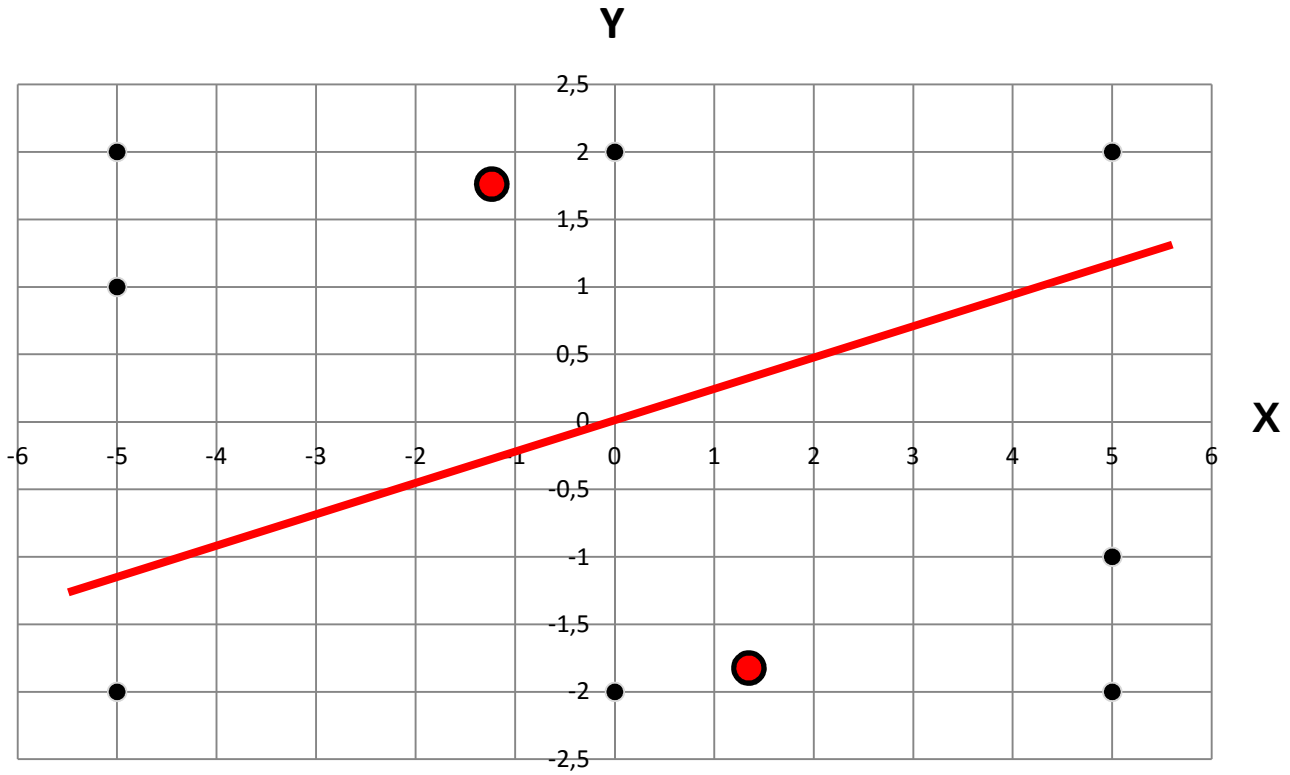
Exemplo

Instância	X	Y
1	-5	2
2	5	-2
3	0	2
4	0	-2
5	-5	1
6	-5	-2
7	5	2
8	5	-1

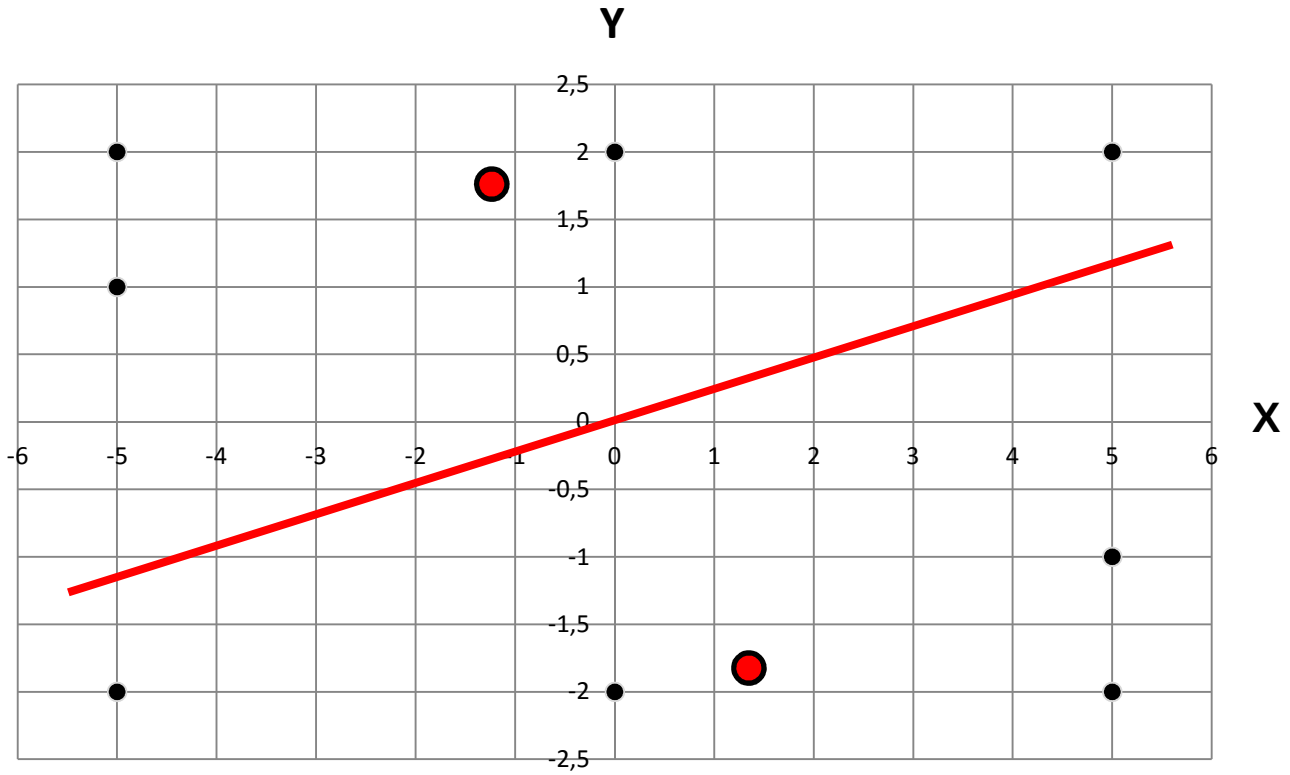
Exemplo



Exemplo



Exemplo



Questões…

- Ao basear-se numa medida de distância possui mesmos inconvenientes do algoritmo kNN
  - Necessária normalização dos dados
  - Medidas de distância/semelhança pouco adequadas para atributos nominais
- Necessidade de proporcionar a priori o número de classes
  - Determinação experimental do melhor número de classes

Bibliografia

- Witten Frank, pg. 138 – 139, 141
- Borrajo Millán