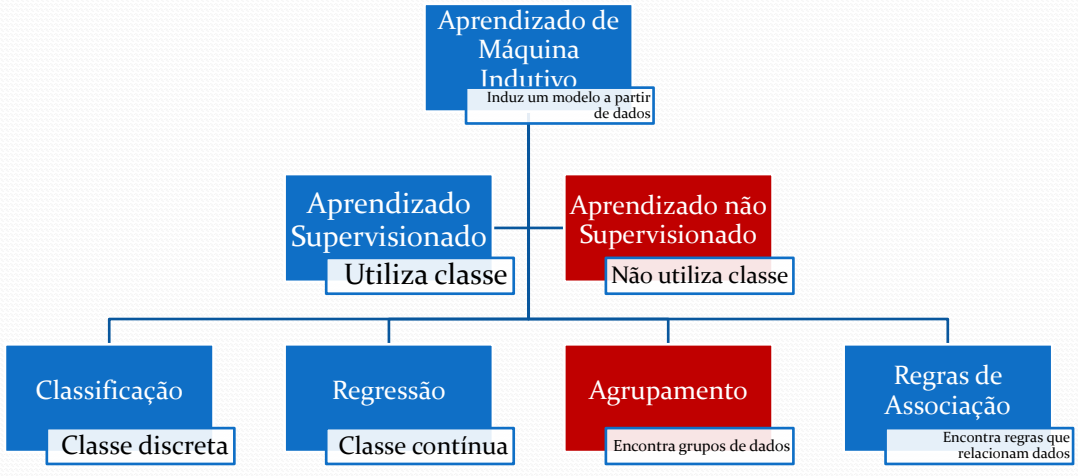
Aprendizado de Máquina

Aprendizado não Supervisionado - Agrupamento

Inteligência Artificial

1

Hierarquia do Aprendizado de Máquina Indutivo



2

Agrupamento (clustering) - Discussão Geral

Tarefa de aprendizado não-supervisionado:

- Exemplos não estão rotulados – não existe um atributo especial conhecido como "atributo meta"

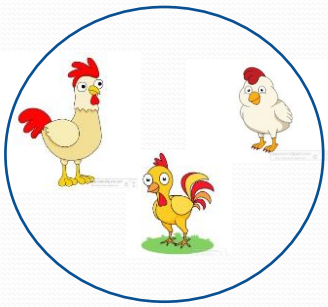
Não conhecida

Cliente 1	renda	dívida	classe
xxx	50	10	bom

exemplo	tempo	temperatura	umidade	vento	classe
E1	sol	2	72	forte	sim

Agrupamento (clustering) - Discussão Geral

- Objetivo: agrupar objetos em *clusters* (grupos) de modo que objetos pertencentes a um mesmo *cluster* são mais similares entre si de acordo com alguma medida de similaridade pré-definida, enquanto que objetos pertencentes a clusters diferentes têm uma similaridade menor.



Agrupamento (clustering) - Aplicações

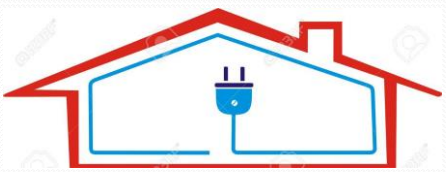
Encontrar grupos de documentos sobre um mesmo assunto

Descobrir funções de genes encontrando grupos de genes com características semelhantes

5

Agrupamento (clustering) - Aplicações

Encontrar grupos de usuários (visitantes) de um site e identificar suas características

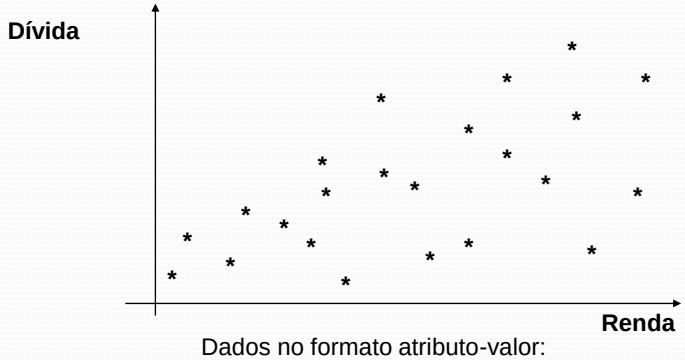


Encontrar sub-populações de consumidores ou padrões de consumo por região

6

Agrupamento (clustering) - Discussão Geral

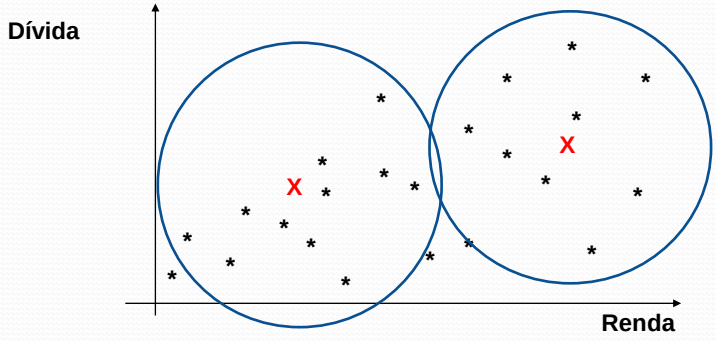
- Representação Gráfica – gráfico de dispersão



Dados no formato atributo-valor:
Renda **Dívida**

Agrupamento (clustering) - Discussão Geral

- Representação Gráfica – gráfico de dispersão



Dados no formato atributo-valor:
Renda **Dívida**

Agrupamento (clustering) - Discussão Geral

- Questões fundamentais para agrupamento:
  - O que é similaridade?
  - Como escolher uma medida de similaridade?
  - Qual o número ideal de grupos?
  - Como escolher um algoritmo?
  - Como validar e interpretar o resultado de um agrupamento?

Questões fundamentais para o agrupamento

- O que é similaridade?

Questões fundamentais para o agrupamento

- Como escolher uma medida de similaridade?
- **Medidas de proximidade**:
- São medidas de similaridade ou de dissimilaridade entre objetos
- Os algoritmos de agrupamento podem utilizar medidas de similaridade ou medidas de dissimilaridade
- Devem ser escolhidas de acordo com:
  - o tipo dos atributos envolvidos (contínuo, categórico)
  - a esparsidade dos dados

11

Questões fundamentais para o agrupamento

- Como escolher uma medida de proximidade?
  - São medidas de dissimilaridade ou similaridade

Entre objetos com atributos contínuos:

Medidas de distância (dissimilaridade)

Medidas de correlação (similaridade)
Correlação de Pearson
Medida de cosseno

Entre objetos com atributos discretos:

Coeficiente de casamento simples (CCS) (similaridade)

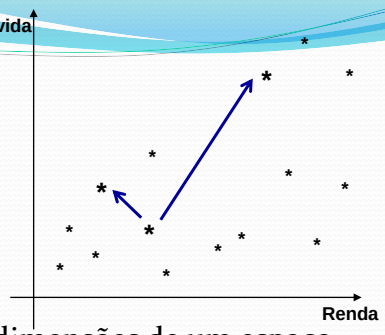
Coeficiente de Jaccard(similaridade)

(São muitas as medidas definidas na literatura)

12

Medidas de dissimilaridade

- Medidas de distância



  - atributos dos exemplos são considerados como dimensões de um espaço multidimensional
  - cada exemplo corresponde a um ponto no espaço
  - similaridade entre dois pontos é inversamente proporcional a distância entre eles

Medidas de dissimilaridade

- Medidas de distância
  - Distância Euclidiana

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

	potência	peso	aceleração	consumo
E1	130	3504	12	18
E2	165	3693	11,5	15
E3	150	3436	11	18

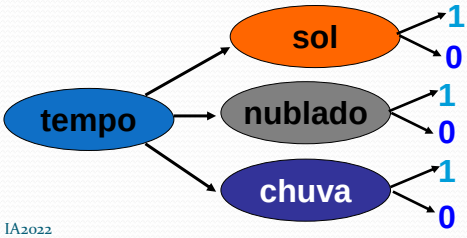
d(E1, E2) = sqrt((130 – 165)^2 + (3504 – 3693)^2 + (12 – 11,5)^2 + (18 – 15)^2)

Sim(E1, E2) = 1- d(E1, E2)

Medidas de dissimilaridade

- Similaridade para variáveis nominais
  - Transformar cada valor do atributo nominal em uma variável binária fictícia.

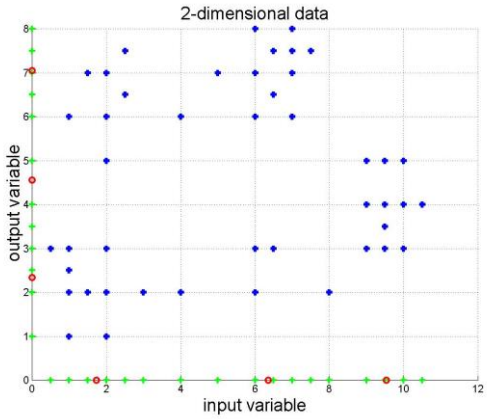
	tempo	temperatura	umidade	vento
E1	sol	amena	alta	forte
E2	nublado	frio	média	forte
E3	sol	frio	alta	fraco



Questões fundamentais para o agrupamento

- Qual o número ideal de grupos?
  - A noção de grupo não é determinística
  - Para um mesmo conjunto de dados, podemos encontrar números diferentes de grupos

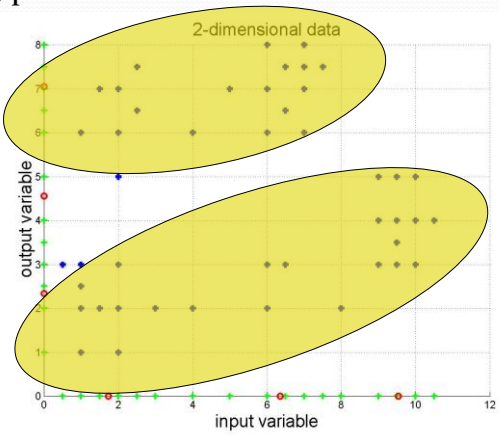
Em quantos grupos esses dados podem ser separados?



Questões fundamentais para o agrupamento

- Qual o número ideal de grupos?
  - A noção de grupo não é determinística
  - Para um mesmo conjunto de dados, podemos encontrar números diferentes de grupos

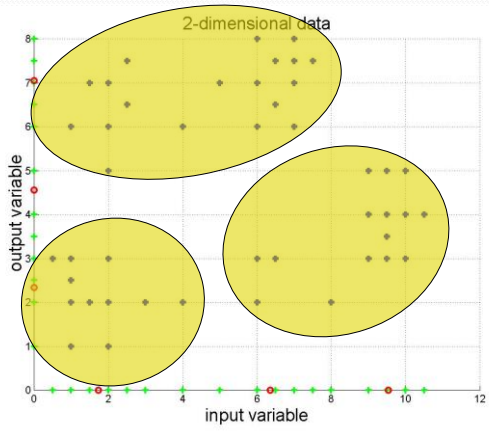
Em quantos grupos esses dados podem ser separados?



Questões fundamentais para o agrupamento

- Qual o número ideal de grupos?
  - A noção de grupo não é determinística
  - Para um mesmo conjunto de dados, podemos encontrar números diferentes de grupos

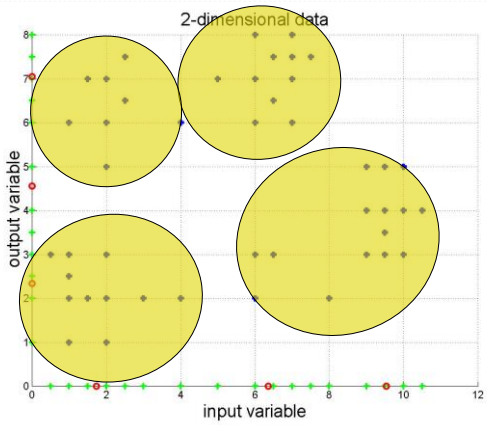
Em quantos grupos esses dados podem ser separados?



Questões fundamentais para o agrupamento

- Qual o número ideal de grupos?
  - A noção de grupo não é determinística
  - Para um mesmo conjunto de dados, podemos encontrar números diferentes de grupos

Em quantos grupos esses dados podem ser separados?



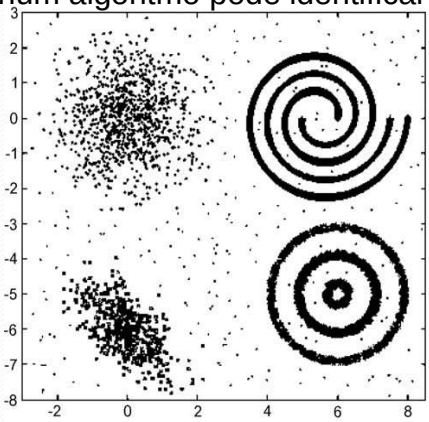
Questões fundamentais para o agrupamento

- Qual o número ideal de grupos?
- O número de grupos encontrados, muitas vezes deve ser definido antes da aplicação do algoritmo de agrupamento (é um parâmetro do algoritmo).
- A forma do grupo está relacionada com a medida de similaridadeou dissimilaridade escolhida.
- O melhor número de grupos de um agrupamento em geral é encontrado por meio de experimentos

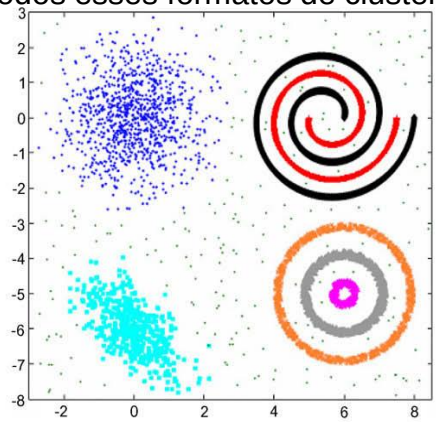
Questões fundamentais para o agrupamento

- Como escolher um algoritmo?

Nenhum algoritmo pode identificar todos esses formatos de clusters



(a) Input data



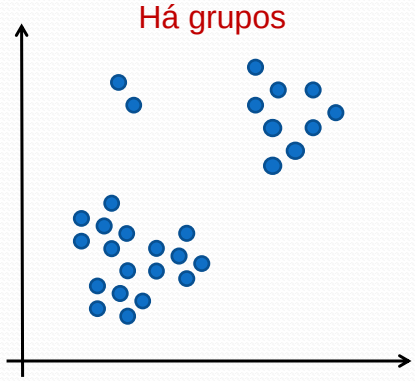
(b) Desired clustering

Questões fundamentais para o agrupamento

- Como validar e interpretar o resultado de um agrupamento?
- Todos os algoritmos de agrupamento, quando aplicados a dados, vão produzir grupos, independente dos dados possuírem grupos ou não.

Não há grupos

Há grupos



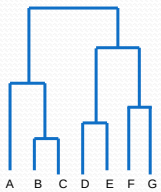
Questões fundamentais para o agrupamento

- Como validar e interpretar o resultado de um agrupamento?
  - Uma estrutura é considerada válida se pode-se afirmar que não foi obtida por acaso ou por meio de artifícios do algoritmo;
  - Na interpretação dos grupos é feita a rotulação dos clusters, definindo sua natureza por meio da análise de seus objetos típicos;
  - Na interpretação, o papel do especialista é fundamental.

Tipos de agrupamento

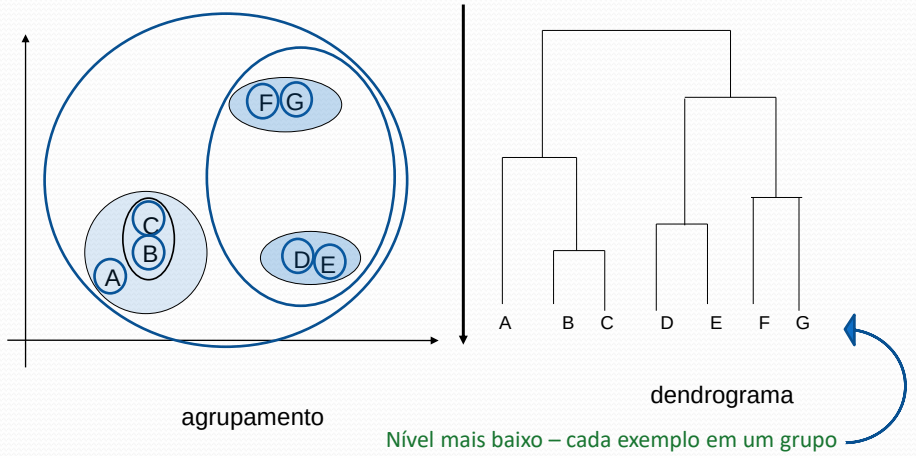
- **Agrupamento particional** – divide os dados em grupos sem sobreposição, de tal forma que cada dado pertence a apenas um grupo.

- **Agrupamento hierárquico** – divide os dados em grupos aninhados, que podem ser representados em uma estrutura de árvore chamada dendrograma



Agrupamento hierárquico

Nível mais alto – todos os exemplos em um só grupo

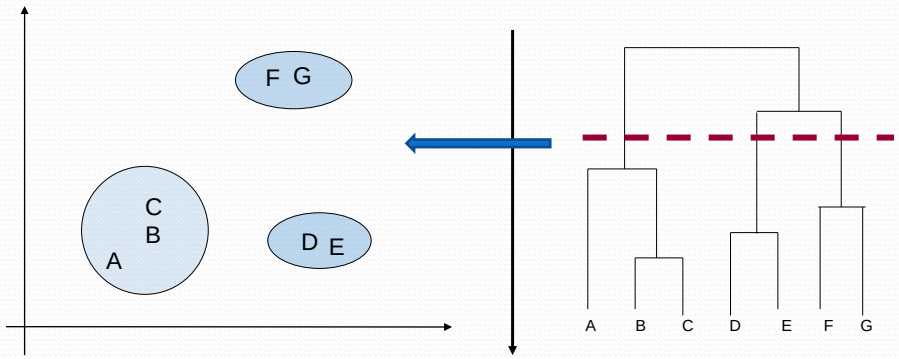


Nível mais baixo – cada exemplo em um grupo

Agrupamento hierárquico

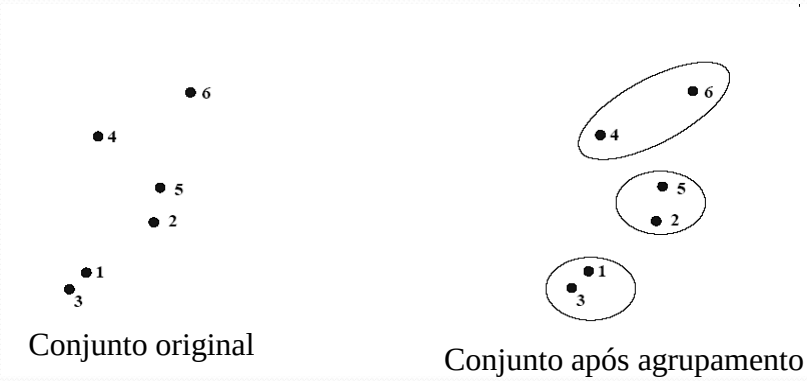
- Seleção de um agrupamento

O corte do dendrograma em um determinado nível define um agrupamento em particular



Agrupamento particional

- Dado um conjunto de dados finito X o problema de agrupamento em X consiste em encontrar vários centros de grupos (clusters) que possam caracterizar adequadamente categorias relevantes de X.



Conjunto original

Conjunto após agrupamento

Agrupamento particional

- Gera uma única partição nos dados.
- Vantagem:
  - mais eficiente para conjunto de dados grande.
- Desvantagem:
  - é necessário definir previamente o número de grupos desejável.
- Gera grupos pela otimização de uma função critério (objetivo).

Agrupamento particional de erro quadrático

- A função critério mais utilizada em algoritmos de agrupamento particionais é o erro quadrático
- O erro quadrático de um agrupamento C com k clusters de um conjunto de padrões E é:

$$err2(E,C) = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left\|x_i^{(j)} - c_j\right\|^2$$

- onde $x_i^{(j)}$ é o i-ésimo exemplo pertencente ao j-ésimo cluster e c_j é o centroide do j-ésimo cluster

Agrupamento k-means (k-médias)

- Mais simples e mais conhecido algoritmo de erro quadrático.
- É fácil de implementar e sua complexidade é *O(n)* com *n* sendo o número de exemplos;
- O usuário define previamente o número de grupos *k*;
- Na versão original e na maioria das aplicações usa distância euclidiana quadrática para calcular similaridade entre os elementos;
- Problema: é sensível à partição inicial e pode convergir para um mínimo local do valor da função critério se a partição inicial não for escolhida apropriadamente.

Agrupamento k-means (k-médias)

Parâmetros de entrada:

- Conjunto de N exemplos não rotulados $\boldsymbol{x}_i$, $i=1,\ldots, N$
- k – número de grupos
- *dist* - medida de distância

Parâmetros de saída:

- k vetores que representam centroides de grupos

Agrupamento k-means (k-médias)

Escolha aleatoriamente um conjunto de vetores distintos para representar os centroides $\boldsymbol{c}_j$, $j = 1,\ldots, k$

Repeat

For i=1 to N

Calcule a distância $dist(\boldsymbol{x}_i, \boldsymbol{c}_j)$ de x_i a cada centroide $\boldsymbol{c}_j$, $i = 1,\ldots, k$

Associe x_i ao centroide $\boldsymbol{c}_j$ que minimiza $dist(\boldsymbol{x}_i, \boldsymbol{c}_j)$

End {For}

For j=1 to k

Atualize os centroides $\boldsymbol{c}_j$, calculando a média dos exemplos que pertencem ao grupo com centro $\boldsymbol{c}_j$

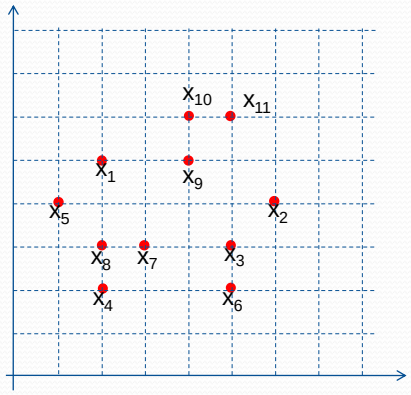
End {For}

Until nenhuma mudança nos c_j ocorra entre duas iterações sucessivas

Agrupamento k-means - Exemplo

Considere os exemplos:

$\mathbf{x}_1 = [2, 5]$
$\mathbf{x}_2 = [6, 4]$
$\mathbf{x}_3 = [5, 3]$
$\mathbf{x}_4 = [2, 2]$
$\mathbf{x}_5 = [1, 4]$
$\mathbf{x}_6 = [5, 2]$
$\mathbf{x}_7 = [3, 3]$
$\mathbf{x}_8 = [2, 3]$
$\mathbf{x}_9 = [4, 5]$
$\mathbf{x}_{10} = [4, 6]$
$\mathbf{x}_{11} = [5, 6]$



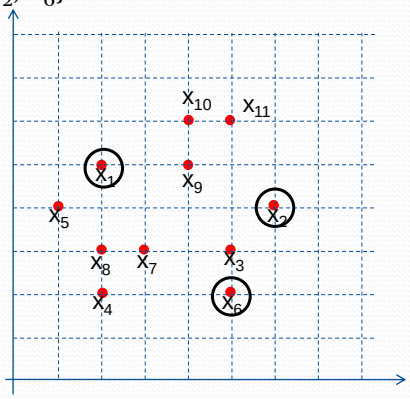
A distância entre exemplos é calculada com a distância euclidiana

Agrupamento k-means - Exemplo

- Aplicar o algoritmo k-means para k = 3
  - Gerar aleatoriamente o vetor de centróides inicial:

$$\Theta = \{x_1, x_2, x_6\}$$

c1	2	5
c2	6	4
c3	5	2



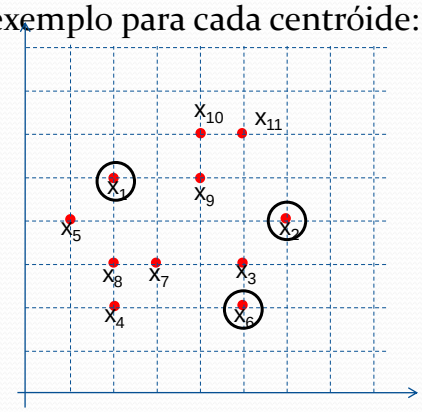
Agrupamento k-means - Exemplo

- Gerar aleatoriamente o vetor de centróides inicial:

$$\Theta = \{x_1, x_2, x_6\}$$

- Calcular a distância de cada exemplo para cada centróide:

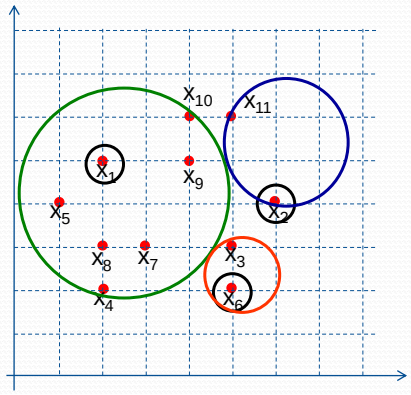
	c1	c2	c3
x1	0	4,1	4,2
x2	4,1	0,0	2,2
x3	3,6	1,4	1,0
x4	3,0	4,5	3,0
x5	1,4	5,0	4,5
x6	4,2	2,2	0,0
x7	2,2	3,2	2,2
x8	2,0	4,1	3,2
x9	2,0	2,2	3,2
x10	2,2	2,8	4,1
x11	3,2	2,2	4,0



Agrupamento k-means - Exemplo

- Atribuir cada exemplo a um cluster, pela menor distância ao centróide:

	c1	c2	c3	grupo
x1	0	4,1	4,2	c1
x2	4,1	0,0	2,2	c2
x3	3,6	1,4	1,0	c3
x4	3,0	4,5	3,0	c1
x5	1,4	5,0	4,5	c1
x6	4,2	2,2	0,0	c3
x7	2,2	3,2	2,2	c1
x8	2,0	4,1	3,2	c1
x9	2,0	2,2	3,2	c1
x10	2,2	2,8	4,1	c1
x11	3,2	2,2	4,0	c2



Agrupamento k-means - Exemplo

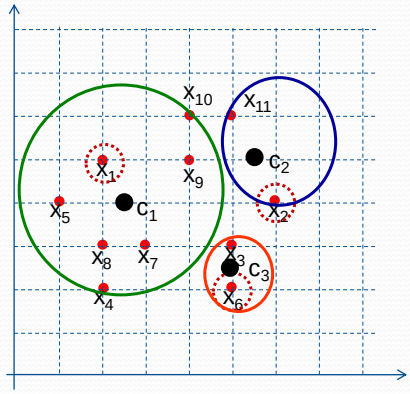
- Recalcular os centróides:

Atual:

c1	2	5
c2	6	4
c3	5	2

Novo:

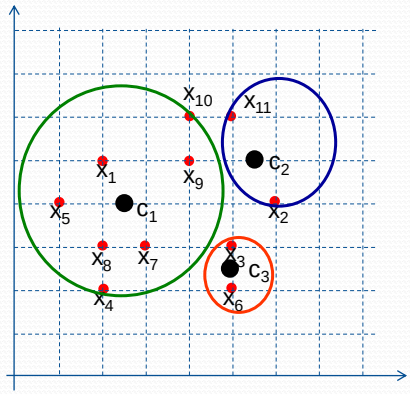
c1	2,6	4
c2	5,5	5
c3	5	2,5



Agrupamento k-means - Exemplo

- Calcular a distância de cada exemplo para cada centróide:
- Atribuir cada exemplo a um cluster, pela menor distância ao centróide:

	c1	c2	c3	
x1	1,2	3,5	3,9	**c1**
x2	3,4	1,1	1,8	**c2**
x3	2,6	2,1	0,5	**c3**
x4	2,1	4,6	3,0	**c1**
x5	1,6	4,6	4,3	**c1**
x6	3,1	3,0	0,5	**c3**
x7	1,1	3,2	2,1	**c1**
x8	1,2	4,0	3,0	**c1**
x9	1,7	1,5	2,7	**c2**
x10	2,4	1,8	3,6	**c2**
x11	3,1	1,1	3,5	**c2**

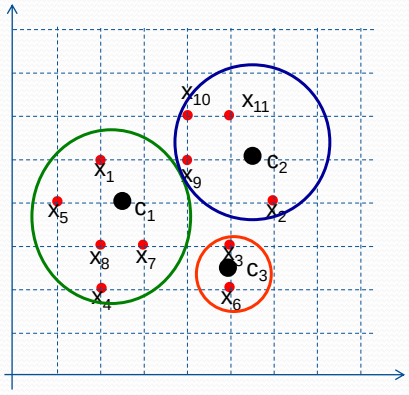


Agrupamento k-means - Exemplo

- Atribuir cada exemplo a um cluster, pela menor distância ao centróide:

x_9 e x_{10} mudaram de c_1 para c_2

	c1	c2	c3	
x1	1,2	3,5	3,9	c1
x2	3,4	1,1	1,8	c2
x3	2,6	2,1	0,5	c3
x4	2,1	4,6	3,0	c1
x5	1,6	4,6	4,3	c1
x6	3,1	3,0	0,5	c3
x7	1,1	3,2	2,1	c1
x8	1,2	4,0	3,0	c1
x9	1,7	1,5	2,7	c2
x10	2,4	1,8	3,6	c2
x11	3,1	1,1	3,5	c2



Agrupamento k-means - Exemplo

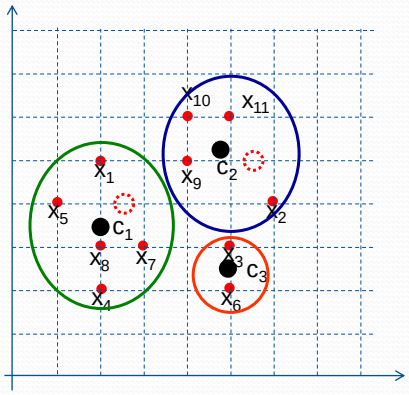
- Recalcular os centróides:

Atual:

c1	2,6	4
c2	5,5	5
c3	5	2,5

Novo:

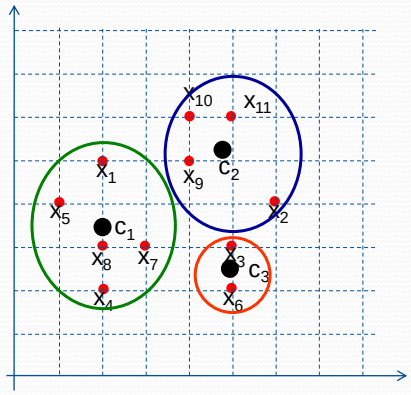
c1	2,0	3,4
c2	4,8	5,3
c3	5	2,5



Agrupamento k-means - Exemplo

- Calcular a distância de cada exemplo para cada centróide:
- Atribuir cada exemplo a um cluster, pela menor distância ao centróide:

	c1	c2	c3	
x1	1,6	2,8	3,9	**c1**
x2	4,0	1,77	1,80	**c2**
x3	3,0	2,3	0,5	**c3**
x4	1,4	4,3	3,0	**c1**
x5	1,2	4,0	4,3	**c1**
x6	3,3	3,3	0,5	**c3**
x7	1,1	2,9	2,1	**c1**
x8	0,4	3,6	3,0	**c1**
x9	2,6	0,9	2,7	**c2**
x10	3,3	1,1	3,6	**c2**
x11	4,0	0,7	3,5	**c2**



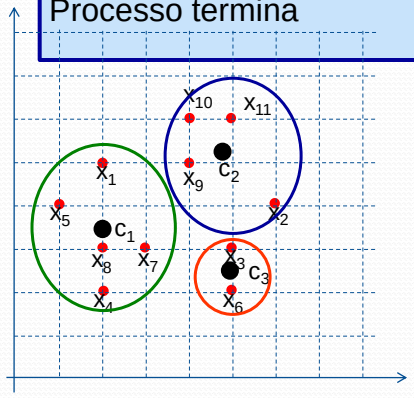
41

Agrupamento k-means - Exemplo

- Atribuir cada exemplo a um cluster, pela menor distância ao centróide:

Não houve mudanças
Processo termina

	c1	c2	c3	
x1	1,6	2,8	3,9	**c1**
x2	4,0	1,77	1,80	**c2**
x3	3,0	2,3	0,5	**c3**
x4	1,4	4,3	3,0	**c1**
x5	1,2	4,0	4,3	**c1**
x6	3,3	3,3	0,5	**c3**
x7	1,1	2,9	2,1	**c1**
x8	0,4	3,6	3,0	**c1**
x9	2,6	0,9	2,7	**c2**
x10	3,3	1,1	3,6	**c2**
x11	4,0	0,7	3,5	**c2**



42

Bons estudos!

Até a próxima aula