CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Agrupamento

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com

Supervisionado e Não Supervisionado

- Classificação é um processo de aprendizado **supervisionado**, pois as instâncias que fazem parte da base de treinamento já têm seu atributo classe rotulado
- Clusterização é dito um processo de aprendizado **não supervisionado**, pois as instâncias que fazem parte da base de entrada não têm o seu grupo rotulado
- Muitas vezes nem mesmo o número de grupos é previamente definido

Clusterização

- **Clusterização** é a tarefa de identificar um conjunto finito de categorias (ou grupos - clusters) que contêm instâncias similares
- Os ***clusters*** não são previamente definidos
- Coletar e rotular bases de dados pode ser muito caro
- Gravar voz é barato, mas rotular todo o material gravado é caro
- Muitas vezes não se tem conhecimento das classes envolvidas
- Segmentação de mercado, agrupamento de documentos e notícias, perfis de clientes (Netflix), análise de redes sociais

Clusterização - Exemplo

- Deseja-se separar os clientes em grupos de forma que aqueles que apresentam o mesmo comportamento de consumo fiquem no mesmo grupo
- Cada tupla (instância) deste exemplo indica a quantidade média total de produtos consumidos e o preço médio destes produtos relativos a cada consumidor

Consumidor	Qtd.Méd.Tot.Prods.	Preç.Méd.Prods.
1	2	1.700
2	10	1.800
3	2	100
4	3	2.000
5	12	2.100
6	3	200
7	4	2.300
8	11	2.040
9	3	150

Clusterização - Exemplo

Consumidor	Qtd.Méd.	Preç.Méd.
1	2	1.700
2	10	1.800
3	2	100
4	3	2.000
5	12	2.100
6	3	200
7	4	2.300
8	11	2.040
9	3	150

Grupo	Consumidor	Qtd.Méd.	Preç.Méd.
1	1	2	1.700
	4	3	2.000
	7	4	2.300
2	2	10	1.800
	5	12	2.100
	8	11	2.040
3	3	2	100
	6	3	200
	9	3	150

- Cada grupo identificado é caracterizado por consumidores semelhantes em relação à quantidade média total e ao preço médio dos produtos consumidos

Algoritmos de Clusterização

- Algoritmos de clusterização organizam um conjunto de **n** instâncias em (**k**) clusters (grupos) de instâncias semelhantes
- Em um cluster, instâncias devem ser similares entre si e dissimilares (diferentes, distantes) em relação a instâncias de outros *clusters*
- O número **k** de clusters a serem obtidos pode não ser um dados de entrada
- Exemplos de algoritmos: k-means, hierárquico

Clusterização – Entrada de Dados

- Matriz de Dados: contém os valores dos p atributos que caracterizam cada um das n instâncias

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Clusterização – Entrada de Dados

- Matriz de Dissimilaridade (distâncias): contém as distâncias entre cada par de instâncias
- $d_{(i,j)}$ representa a dissimilaridade (diferença) entre as instâncias **i** e **j**
- $d_{(i,j)} \geq 0$
- $d_{(i,j)} = d_{(j,i)}$
- $d_{(i,i)} = 0$

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

Clusterização – Cálculo da Distância

- Distância Euclidiana
  - $\mathbf{d_{(i,j)}} = ((x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2)^{1/2}$

$$d_{(i,j)} = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}$$

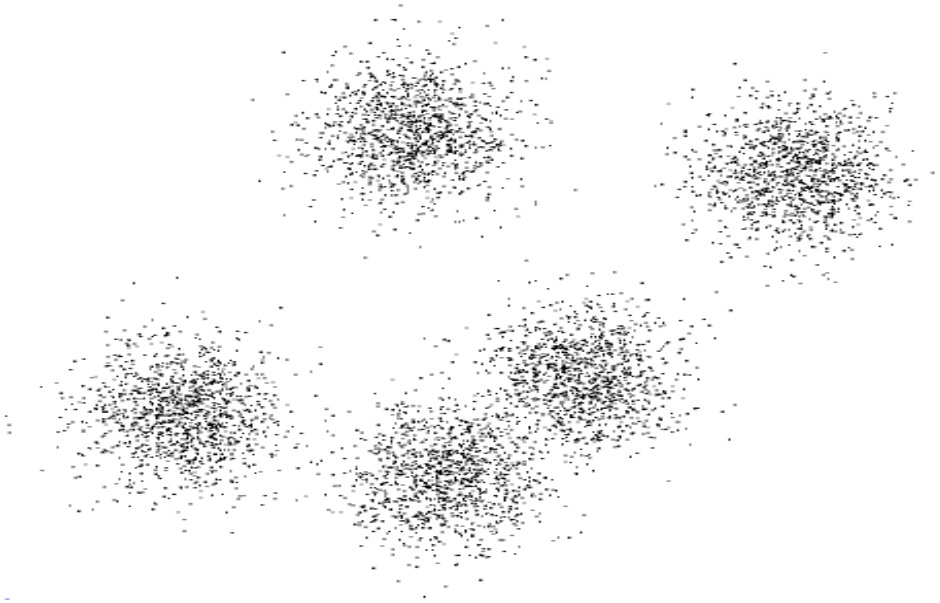
Algoritmo K-Means (MacQueen'67)

- Entrada: **n** instâncias e o número **k** de clusters
- Saída: **n** instâncias organizadas em **k** clusters
- Passos:
  - Passo 1: Defina **k** centroides iniciais, escolhendo **k** instâncias aleatórias da base
  - Passo 2: Associe cada instância para o cluster correspondente ao centroide mais similar
  - Passo 3: Recalcule os centroides dos clusters
  - Passo 4: Repita passo 2 e 3 até atingir um critério de parada (e.g. até um número máximo de iterações ou até não ocorrer alterações nos centroides)

Algoritmo K-Means (MacQueen’67)

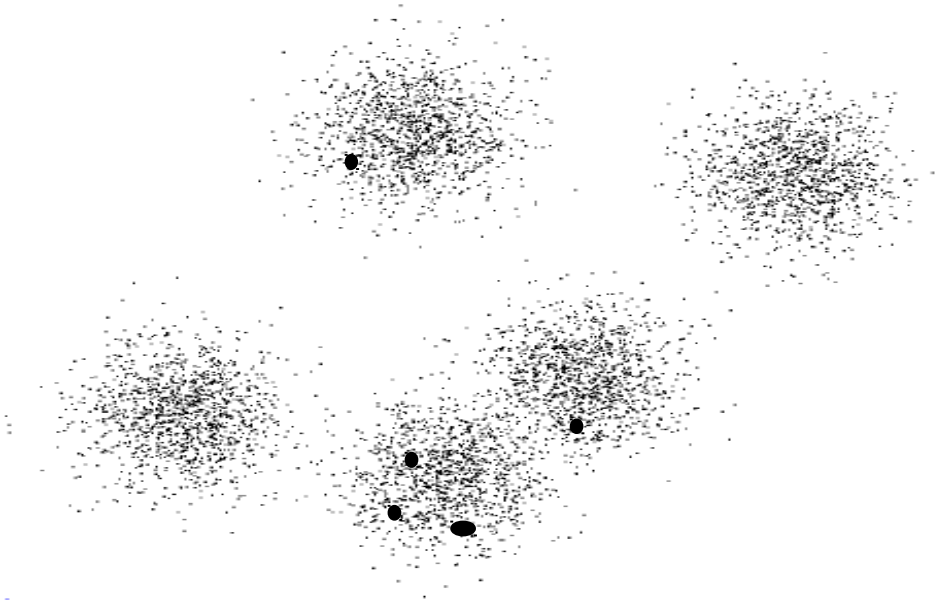
- Exemplo: Conjunto de 8.000 instâncias em 2 dimensões.

k = 5



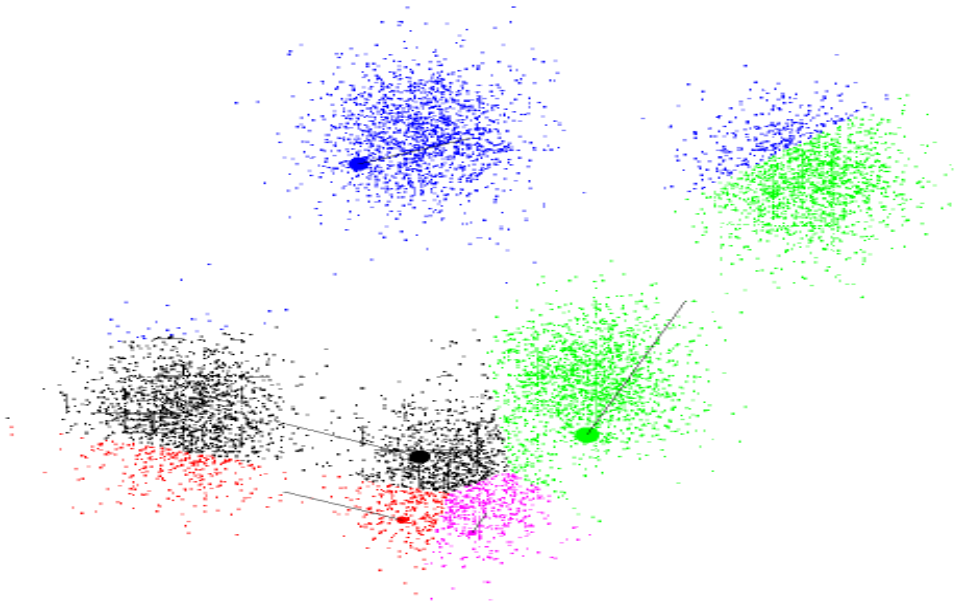
Algoritmo K-Means (MacQueen’67)

Exemplo: Escolher aleatoriamente os ***k*** centroides



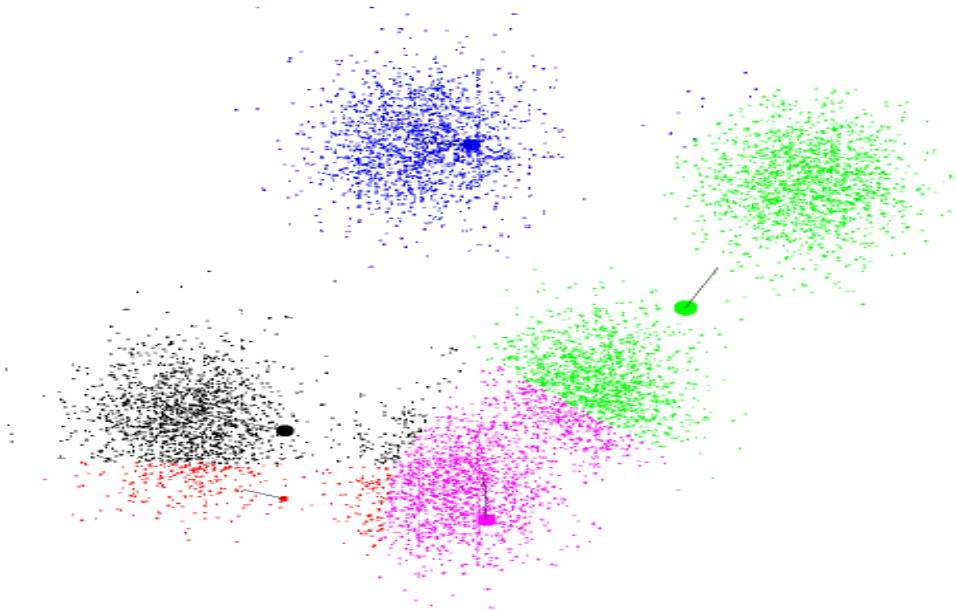
Algoritmo K-Means (MacQueen’67)

Exemplo: 1ª iteração



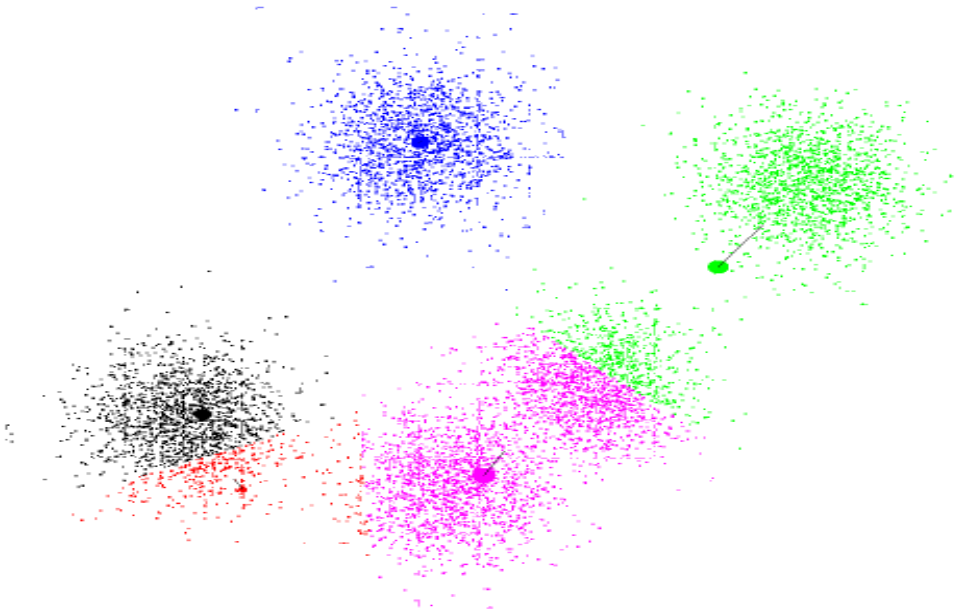
Algoritmo K-Means (MacQueen’67)

Exemplo: 2ª iteração



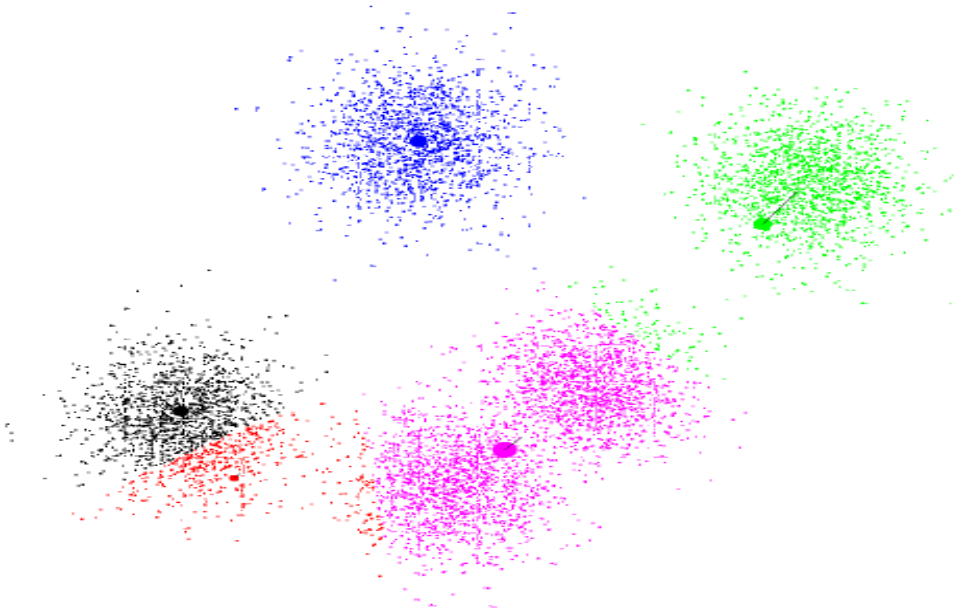
Algoritmo K-Means (MacQueen’67)

Exemplo: 3ª iteração



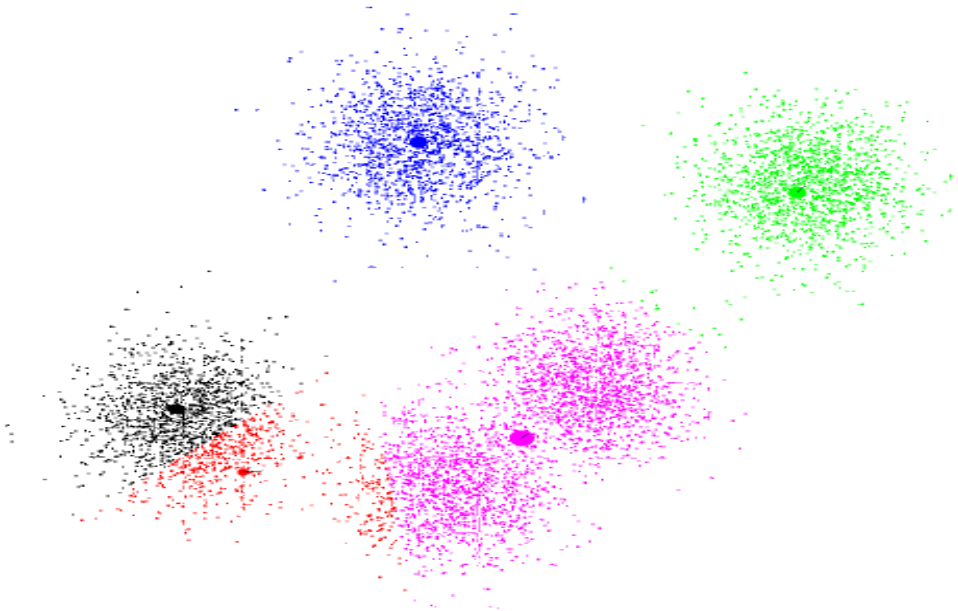
Algoritmo K-Means (MacQueen’67)

Exemplo: 4ª iteração



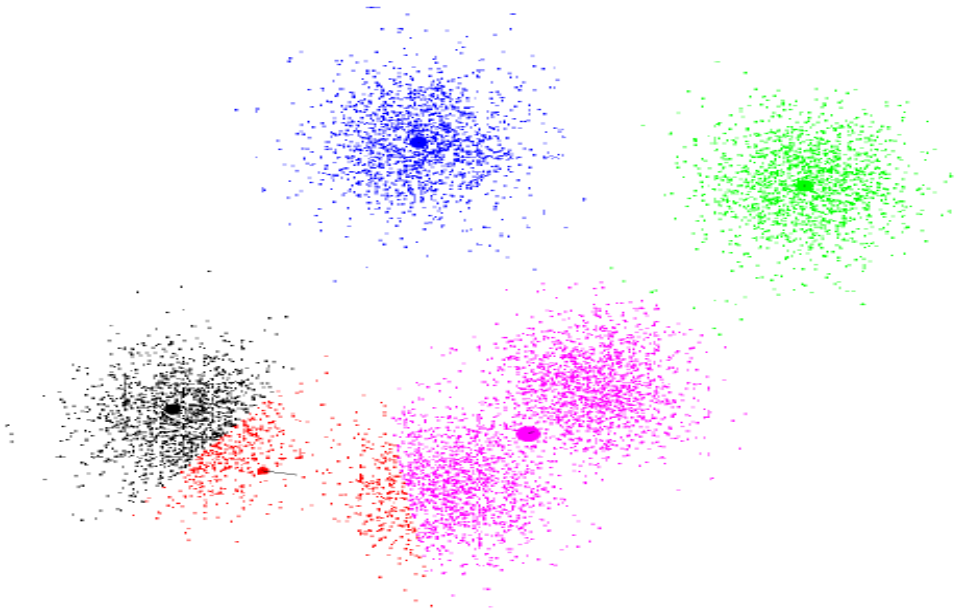
Algoritmo K-Means (MacQueen’67)

Exemplo: 5ª iteração



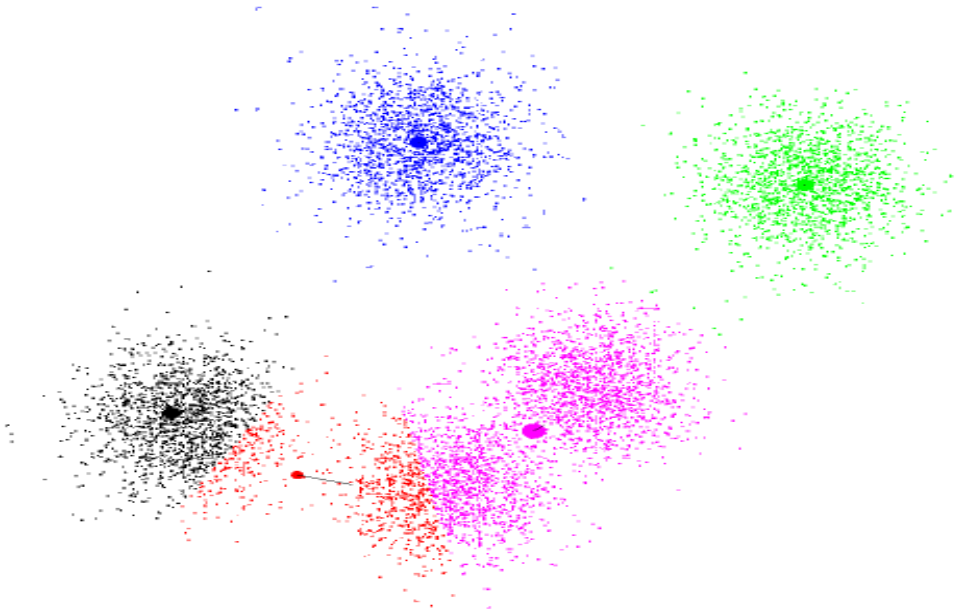
Algoritmo K-Means (MacQueen’67)

Exemplo: 6ª iteração



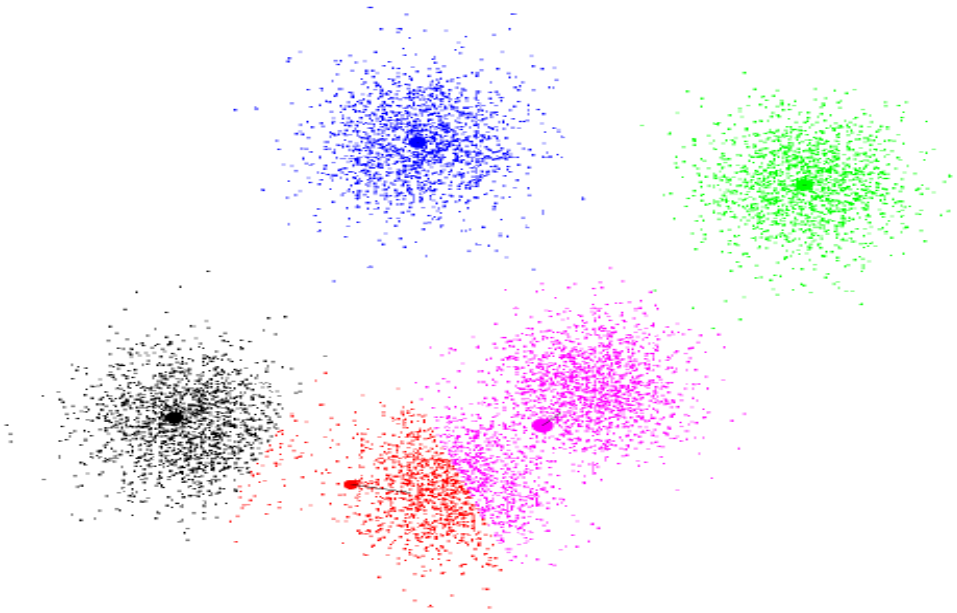
Algoritmo K-Means (MacQueen’67)

Exemplo: 7ª iteração



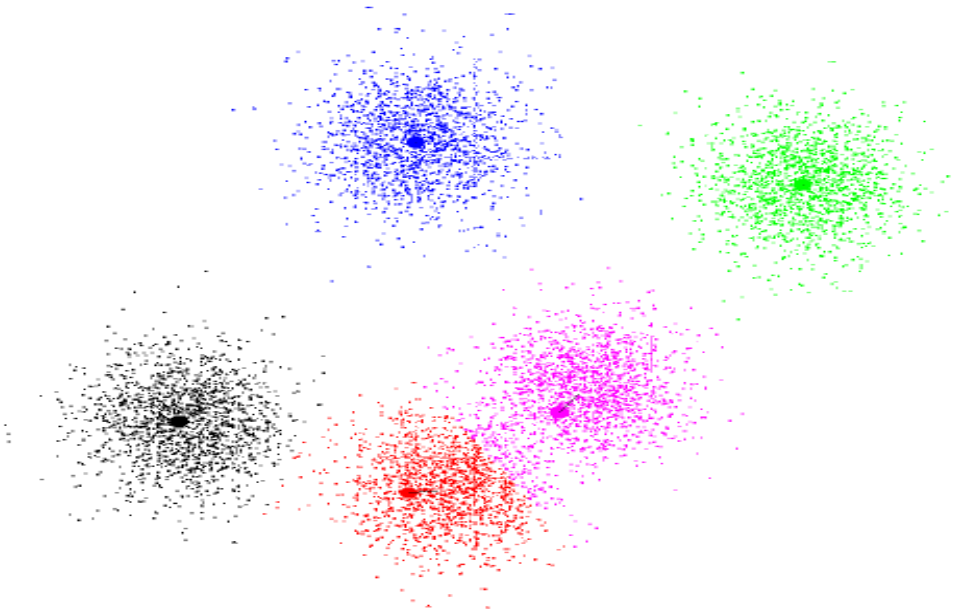
Algoritmo K-Means (MacQueen’67)

Exemplo: 8ª iteração



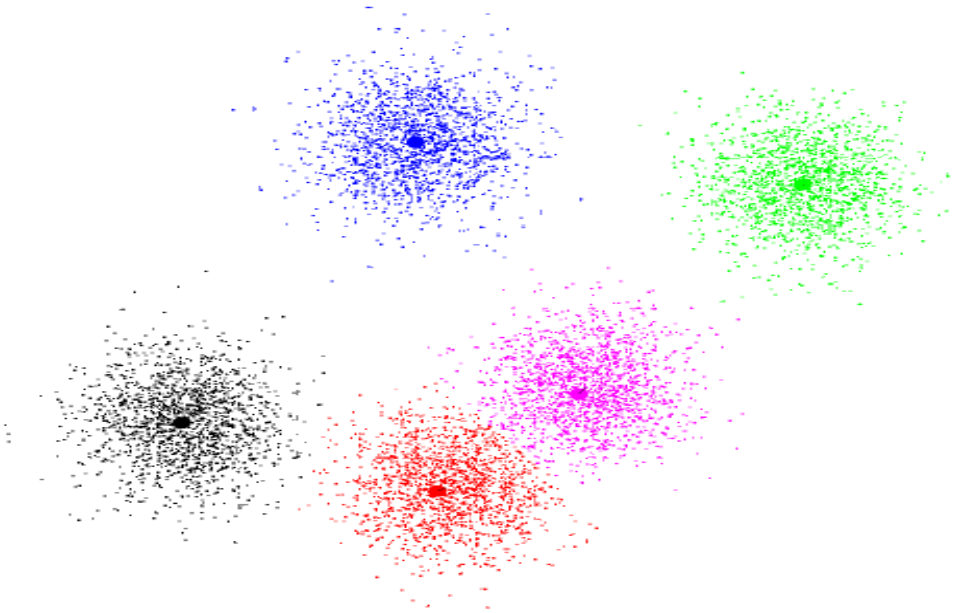
Algoritmo K-Means (MacQueen’67)

Exemplo: 9ª iteração

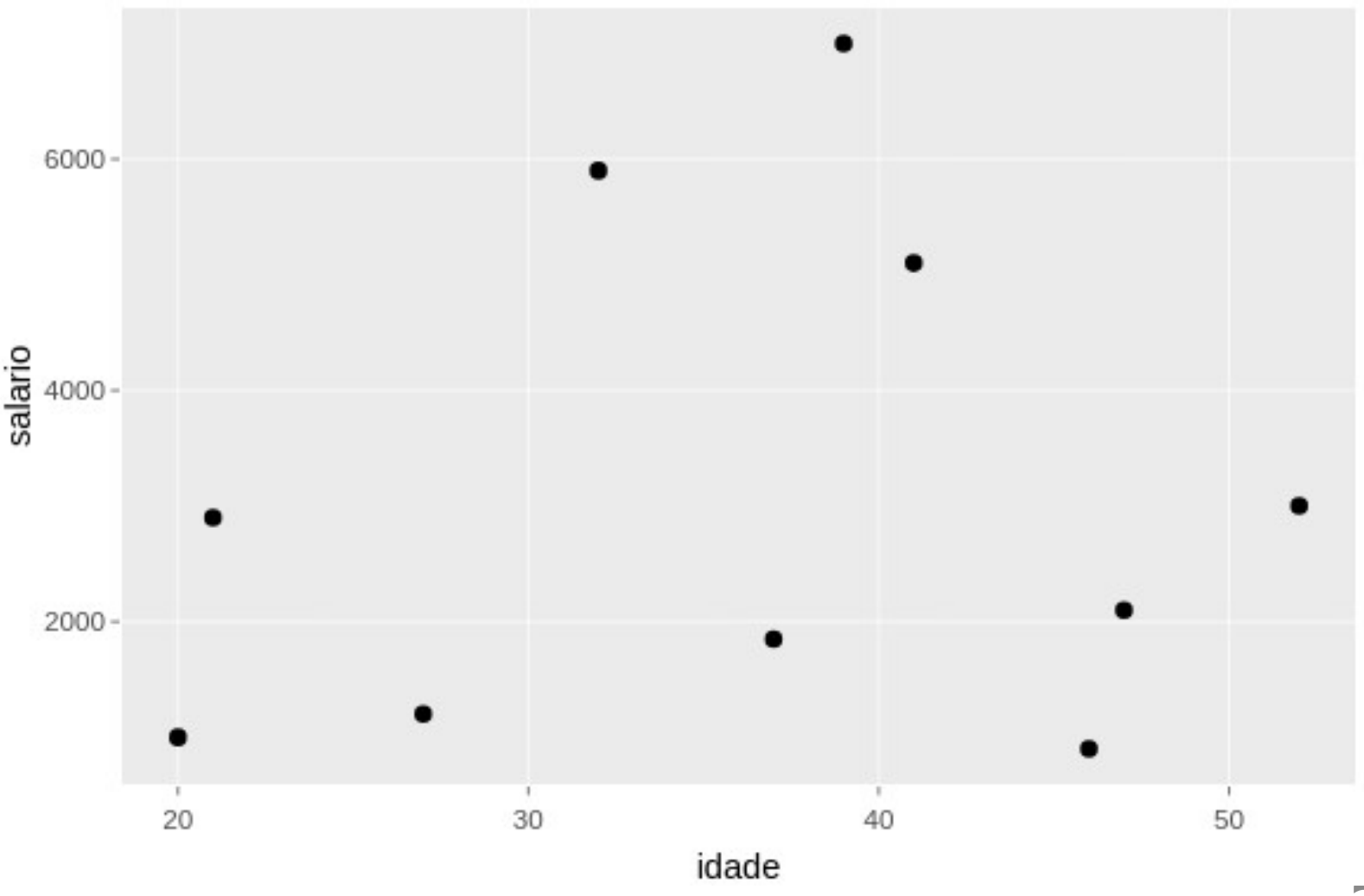


Algoritmo K-Means (MacQueen'67)

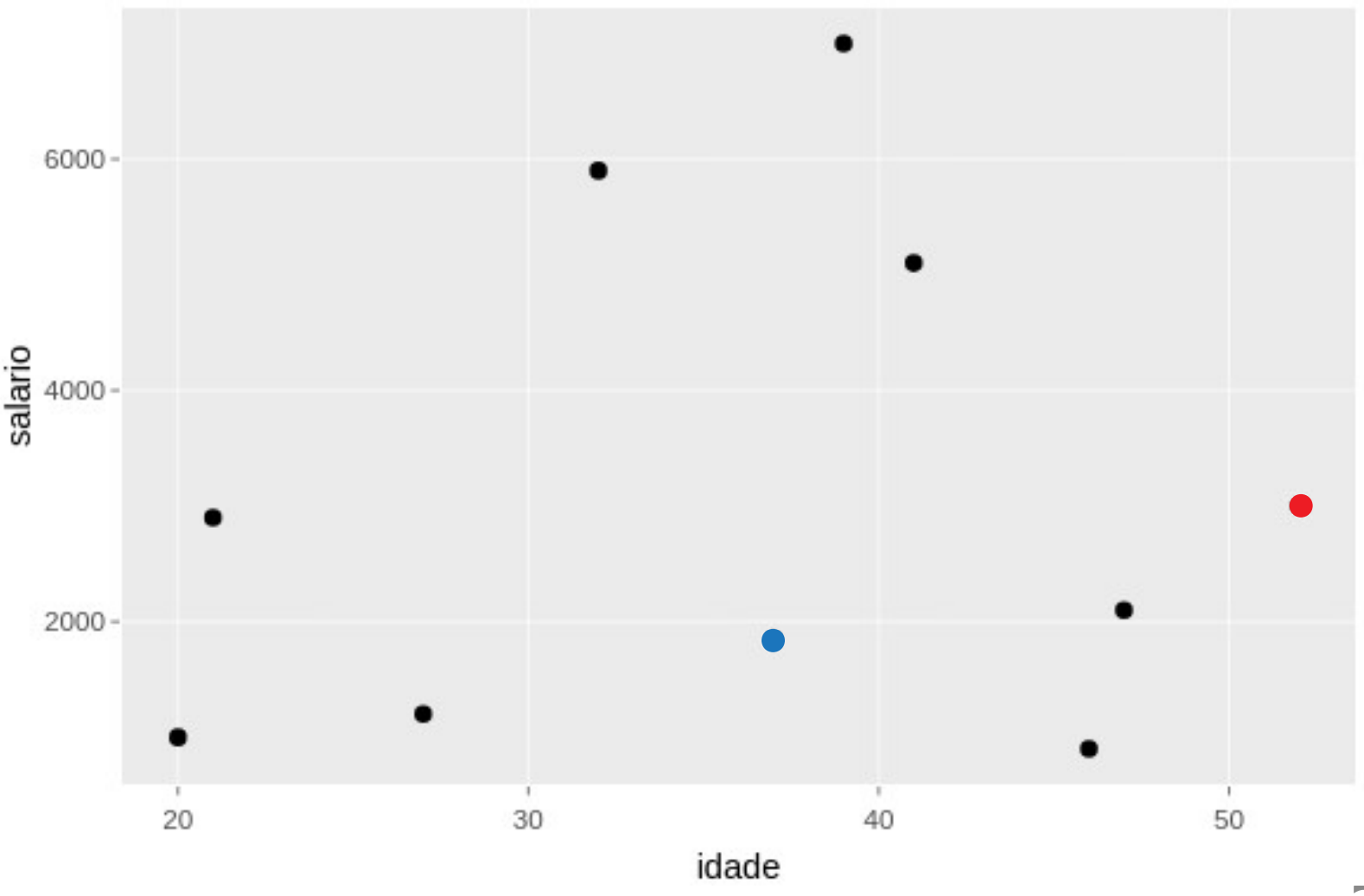
Exemplo: 10ª iteração



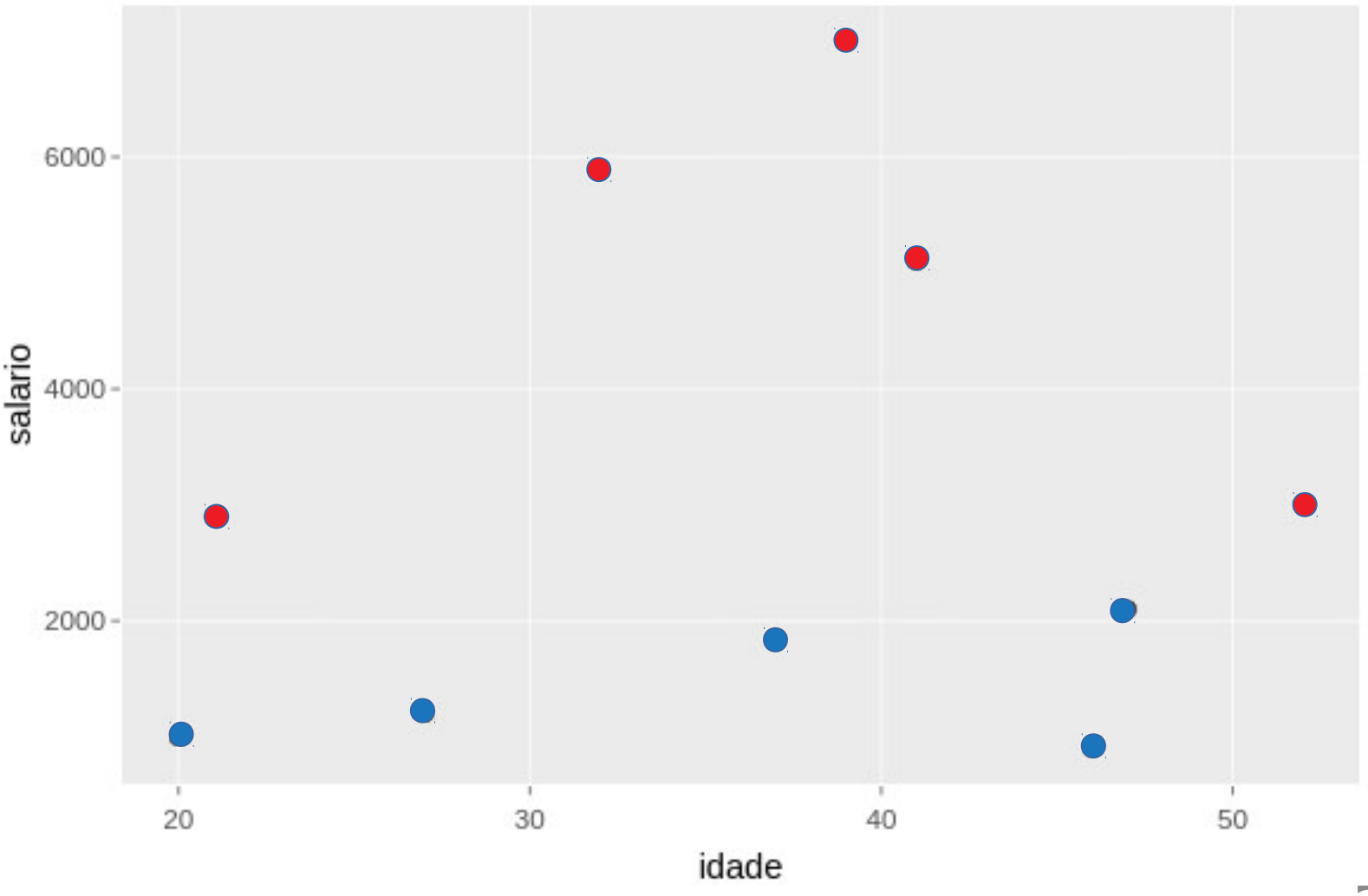
Exemplo – K-Means (k = 2)



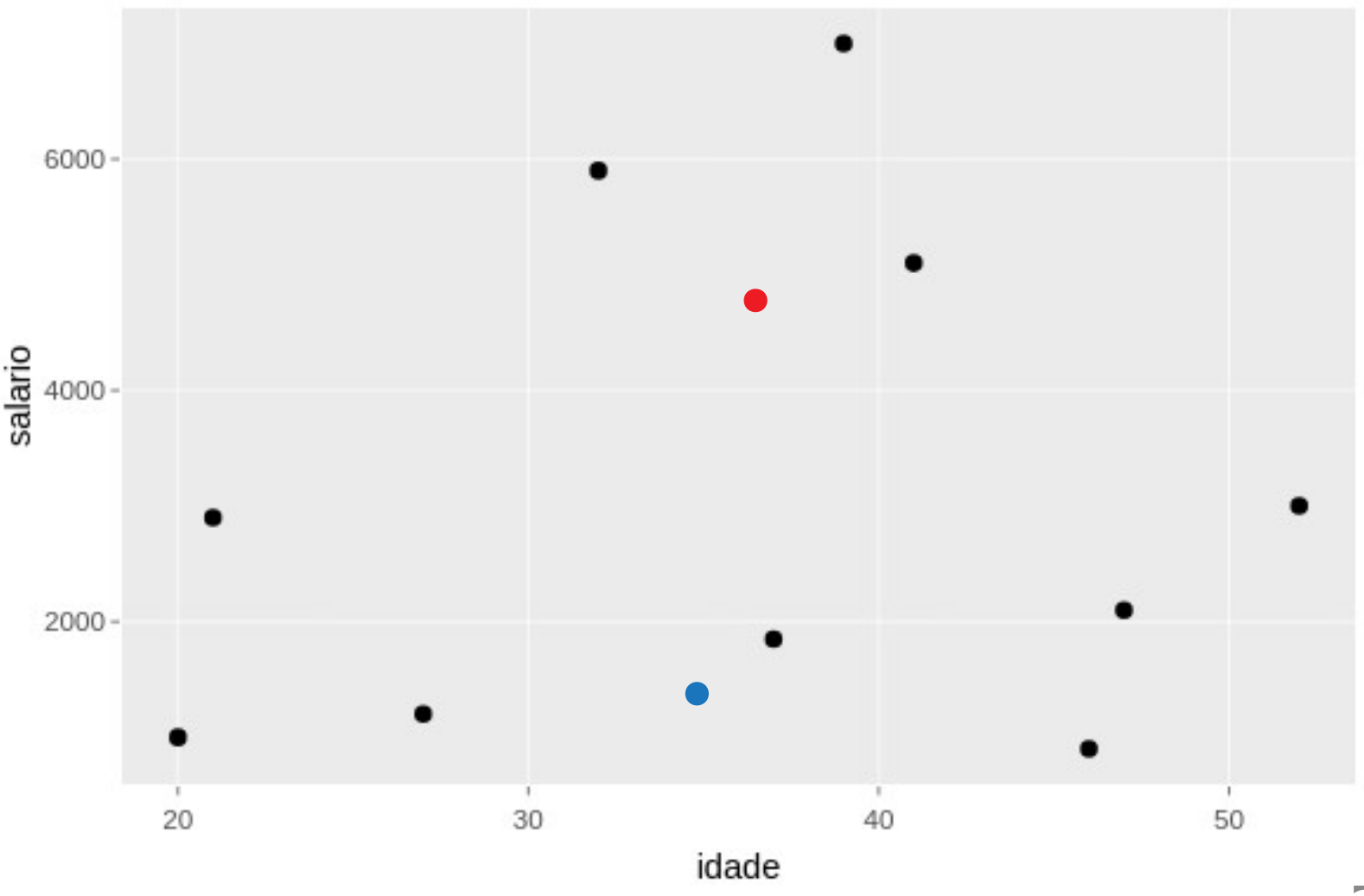
1ª Iteração – Centroides (k = 2)



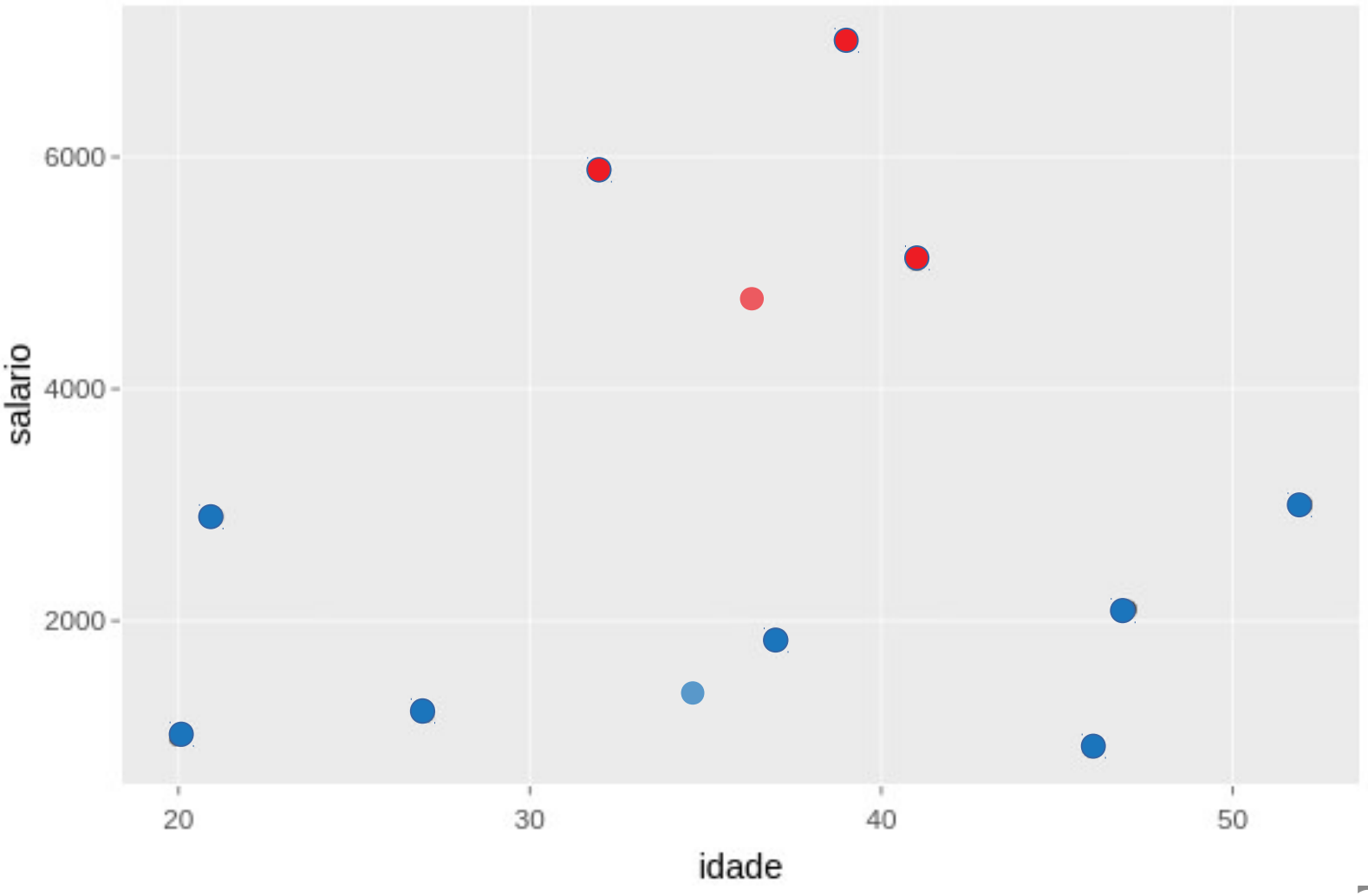
1ª Iteração – Grupos (k = 2)



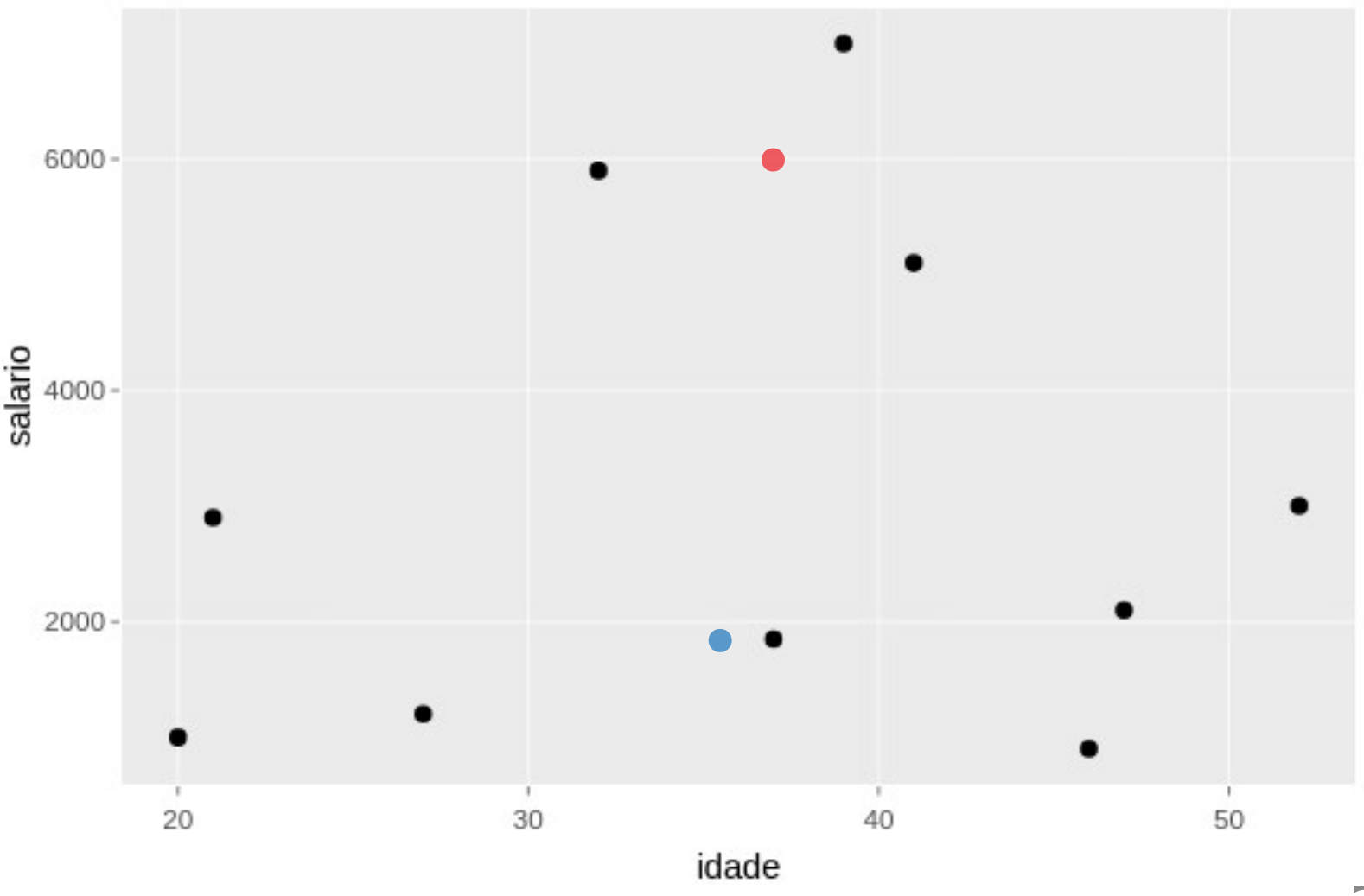
2ª Iteração – Centroides (k = 2)



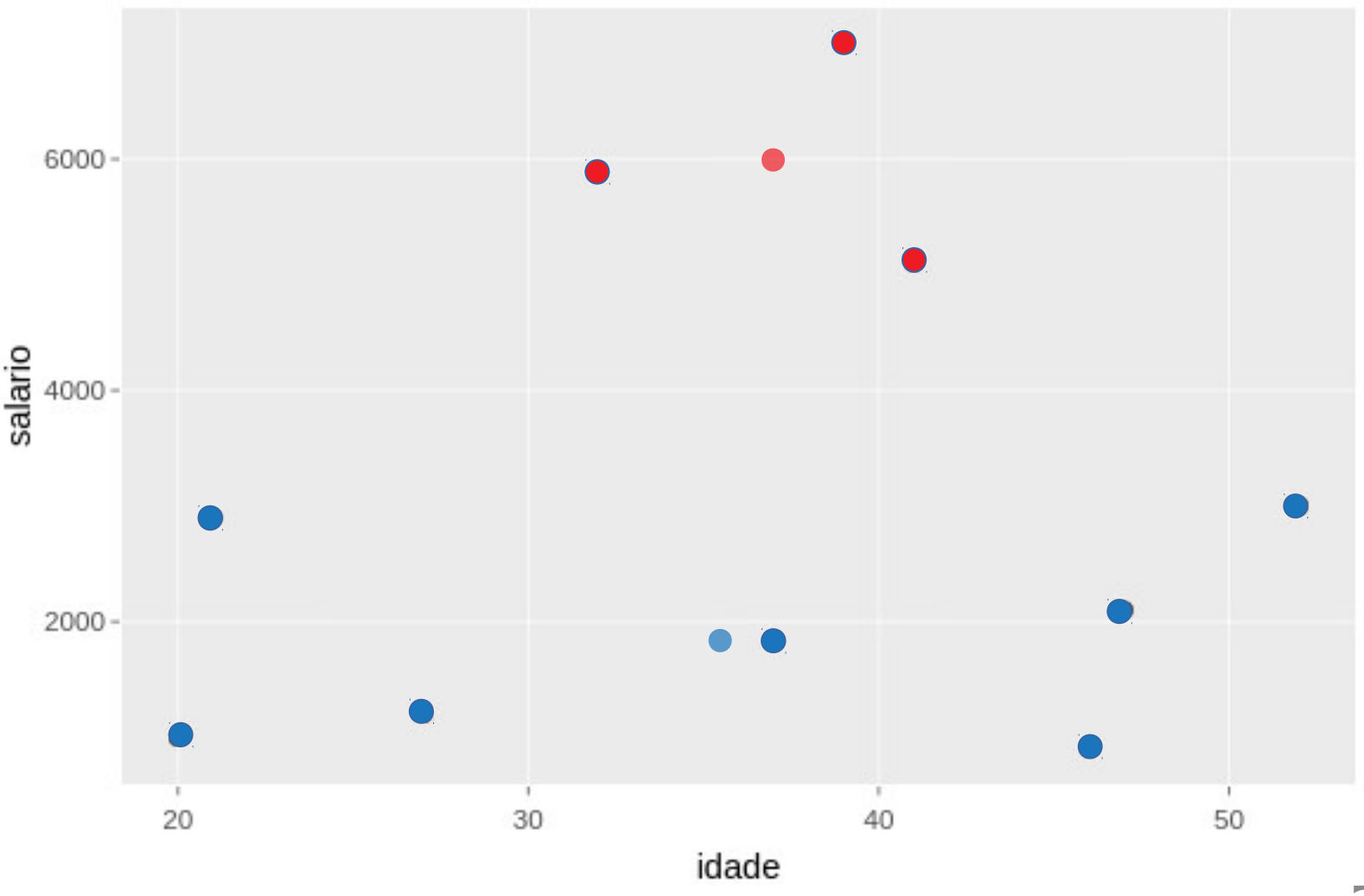
2ª Iteração – Grupos (k = 2)



3ª Iteração – Centroides (k = 2)

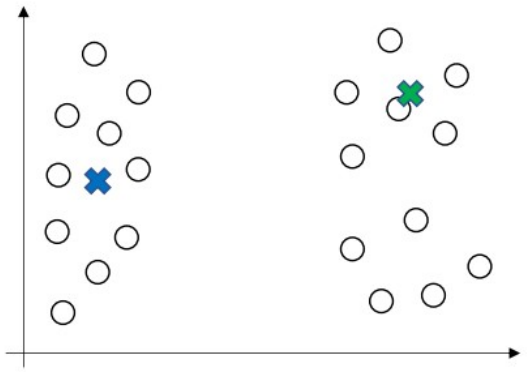


3ª Iteração – Grupos (k = 2)



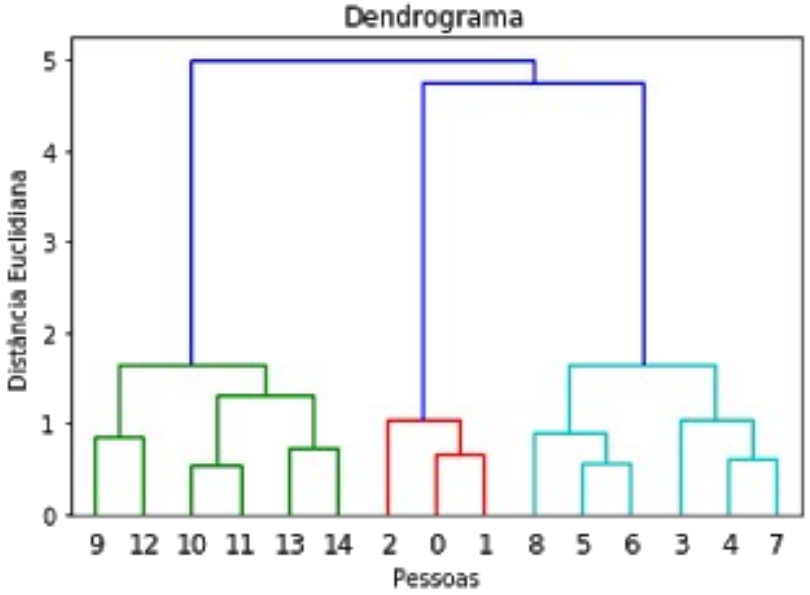
K-Means++

- Reduz a probabilidade de inicializações ruins
- Seleciona os centroides iniciais que estão longes uns dos outros
- O primeiro centroide é selecionado randomicamente. Porém, os outros são selecionados baseado na distância para o primeiro ponto



Algoritmo Hierárquico

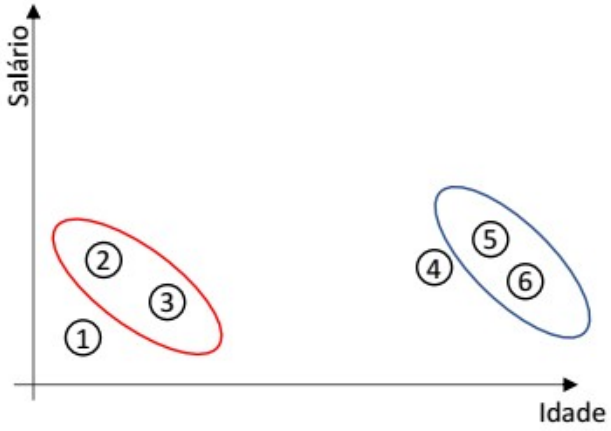
- Não é necessário especificar o número de clusters
- Os clusters são formados pela aglomeração ou divisão das instâncias
- É criada uma estrutura hierárquica em formato de árvore binária que indica o número de clusters
- Os resultados podem ser apresentados em um dendrograma

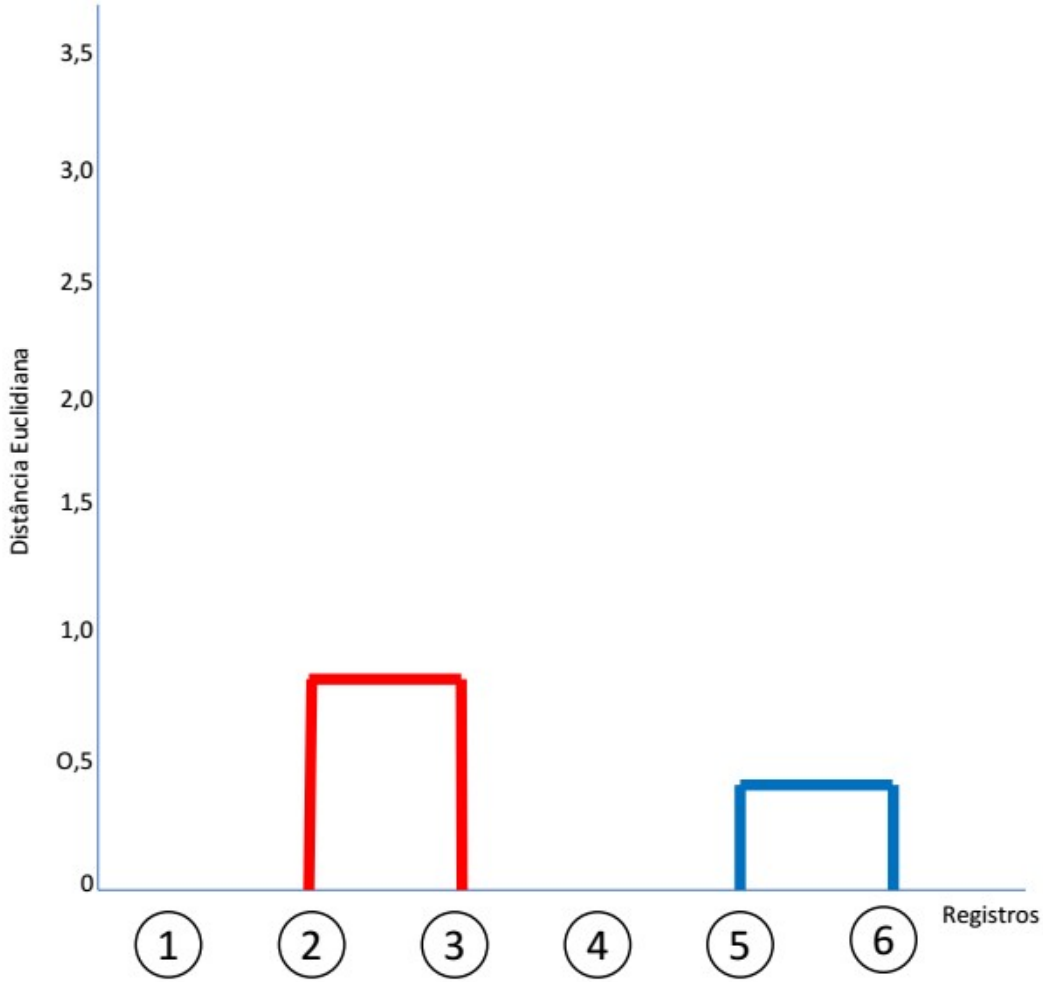


Algoritmo Hierárquico

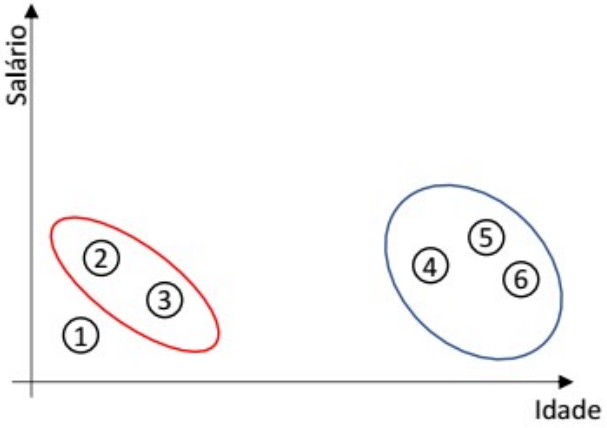
- **Aglomerativo**: é uma abordagem "de baixo para cima".
  - Cada instância é considerada como um grupo individual, e grupos são recursivamente fundidos até produzir um agrupamento final
- **Por divisão**: é uma abordagem "de cima para baixo".
  - Inicialmente, o conjunto de todas as instâncias é considerado como sendo um único grupo e, em seguida, ele é recursivamente dividido para produzir um agrupamento final

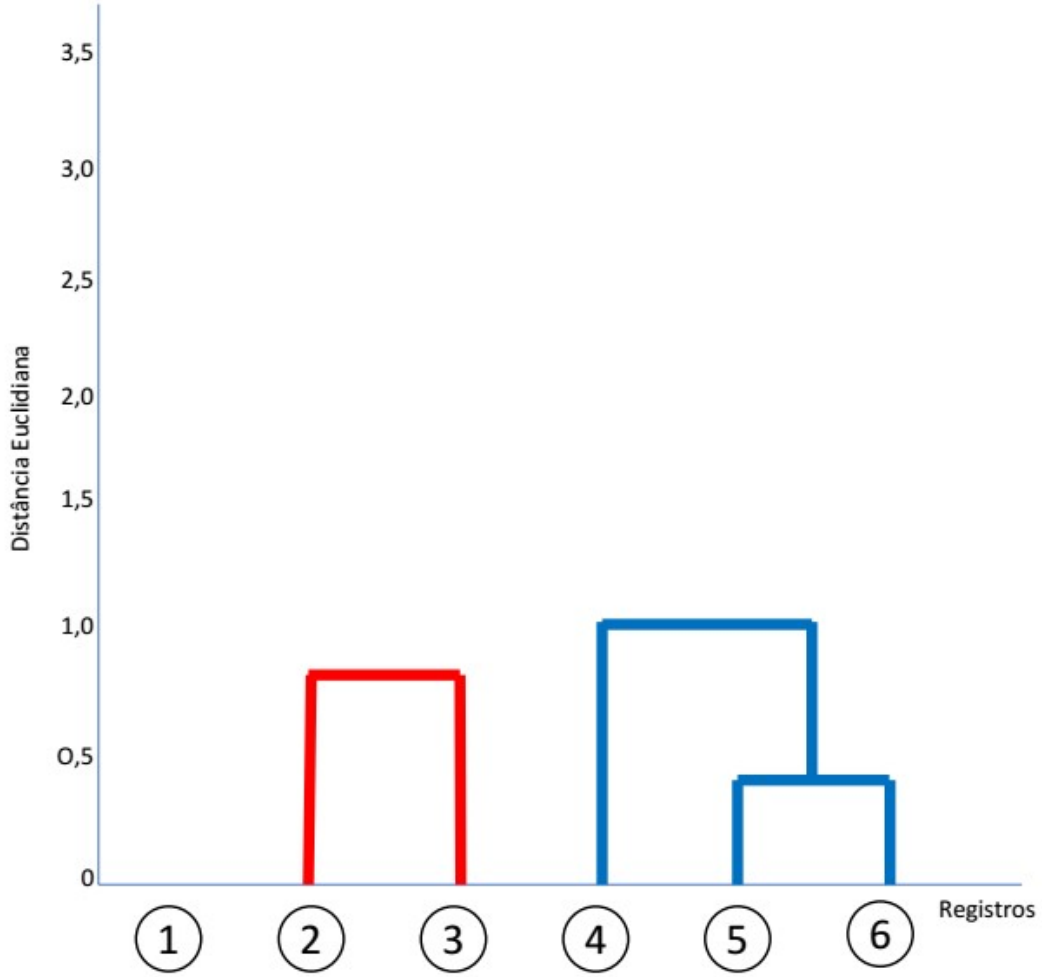
Algoritmo Hierárquico – Exemplo



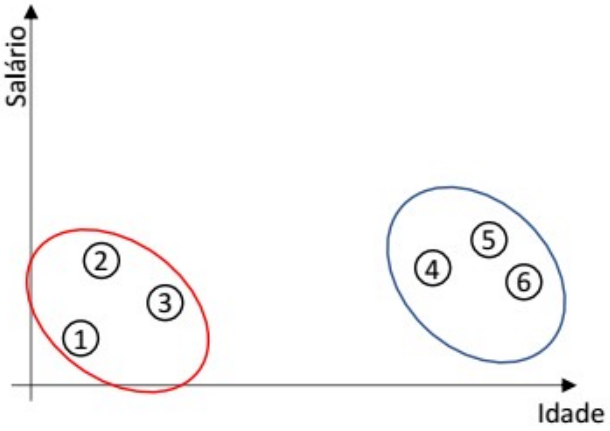


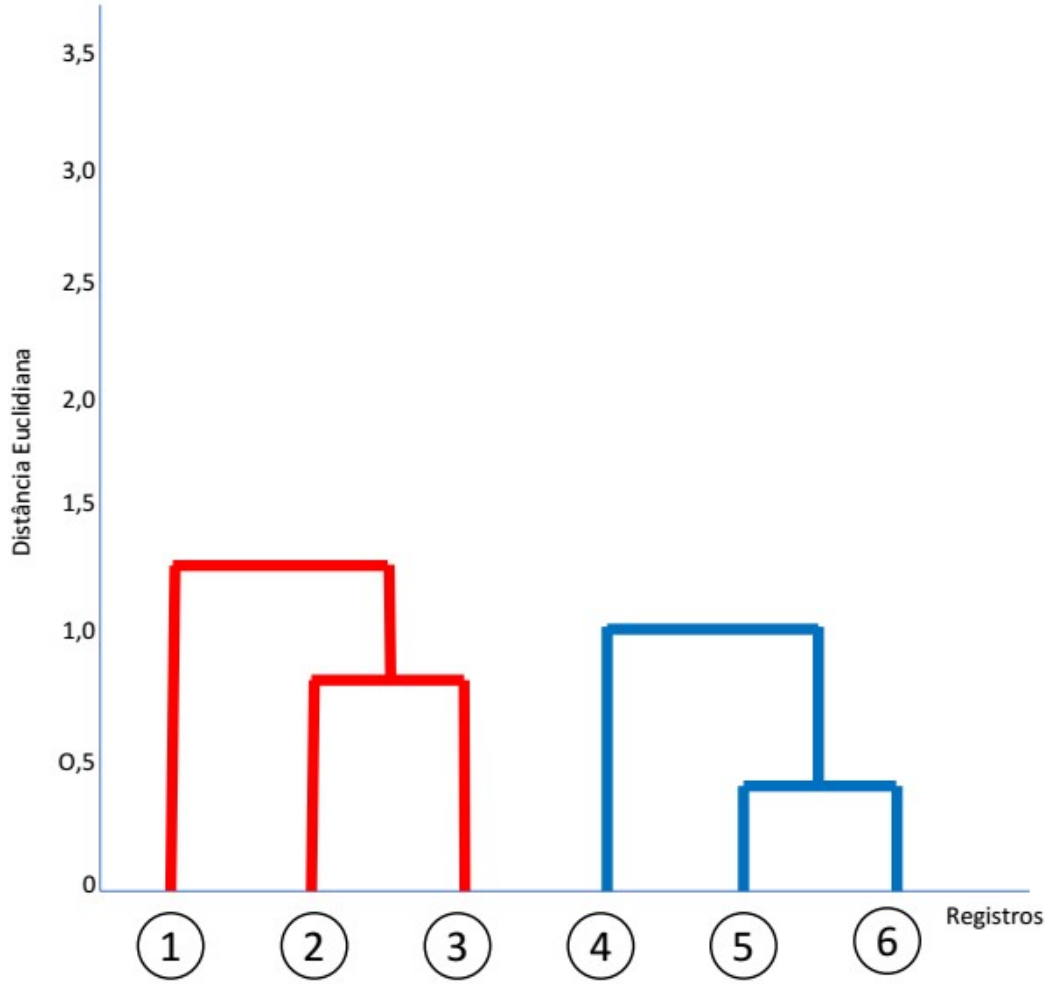
Algoritmo Hierárquico – Exemplo



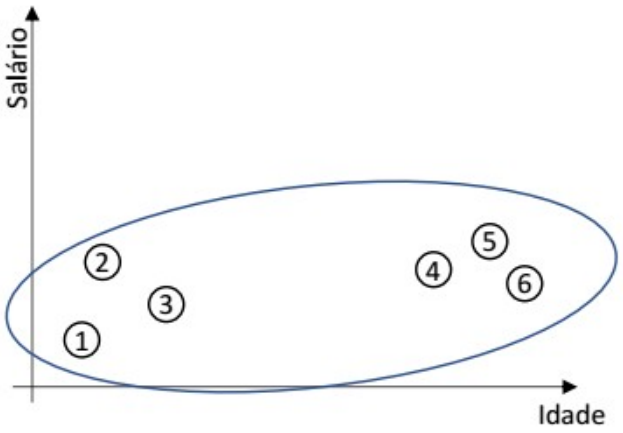


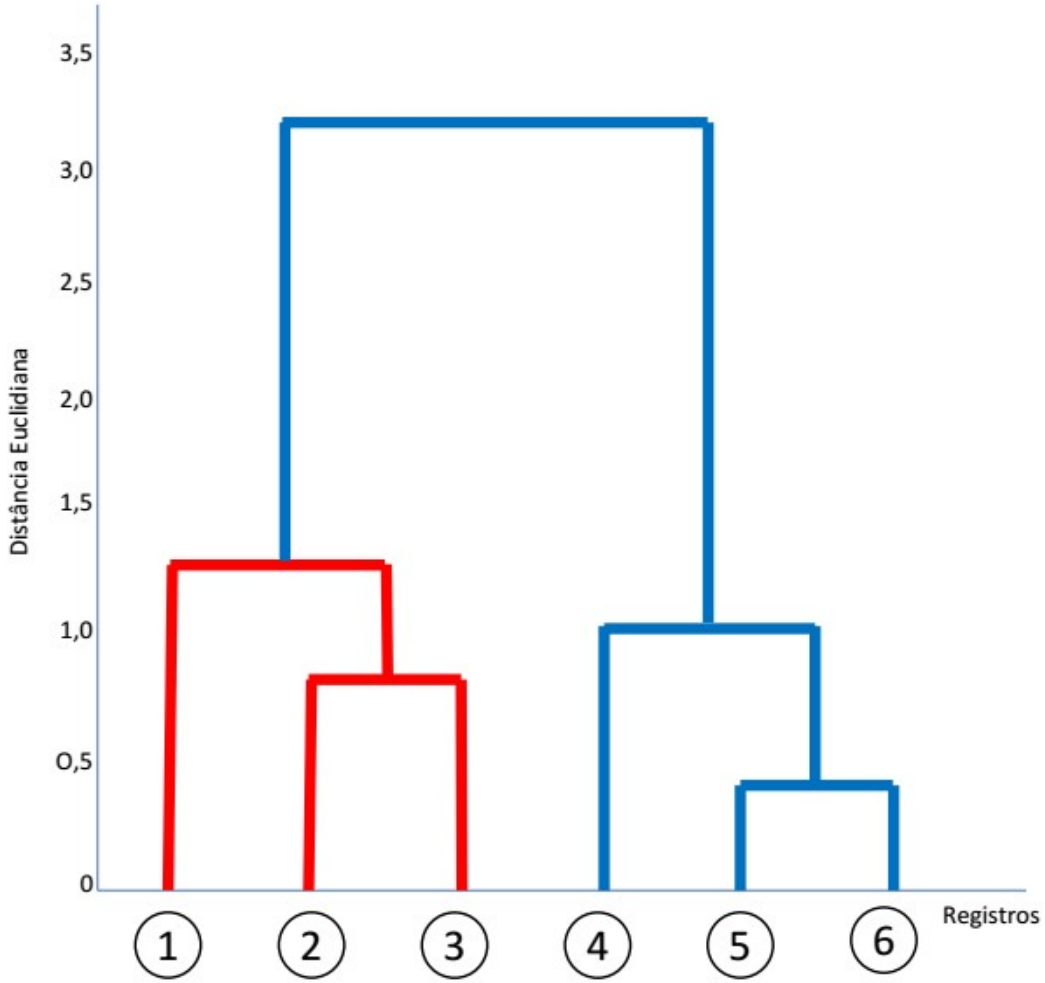
Algoritmo Hierárquico – Exemplo



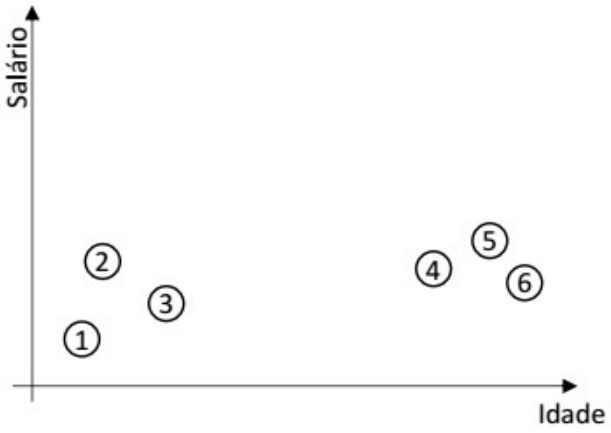


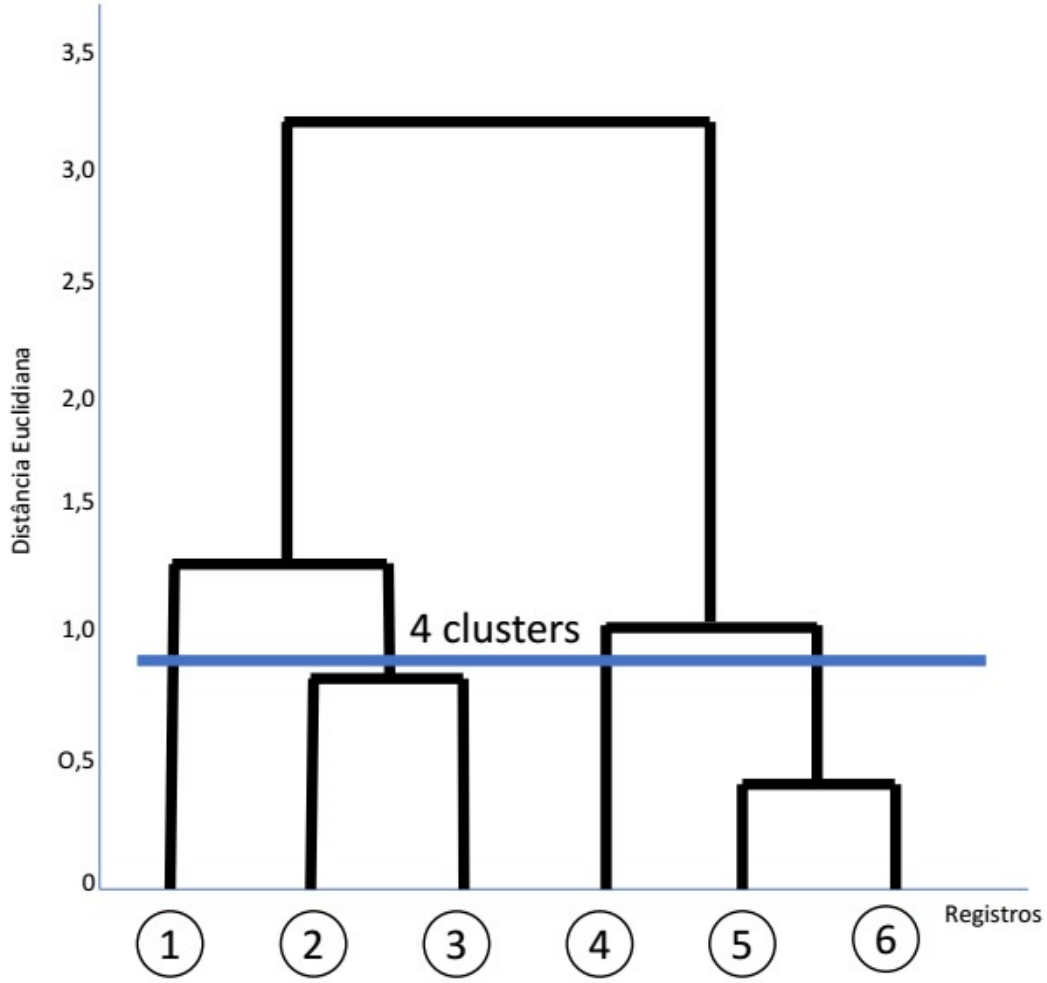
Algoritmo Hierárquico – Exemplo



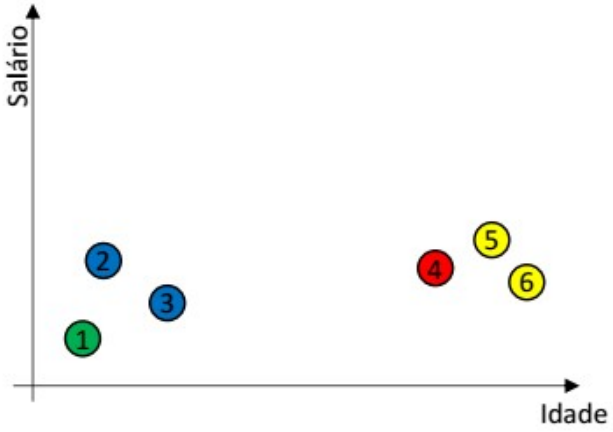


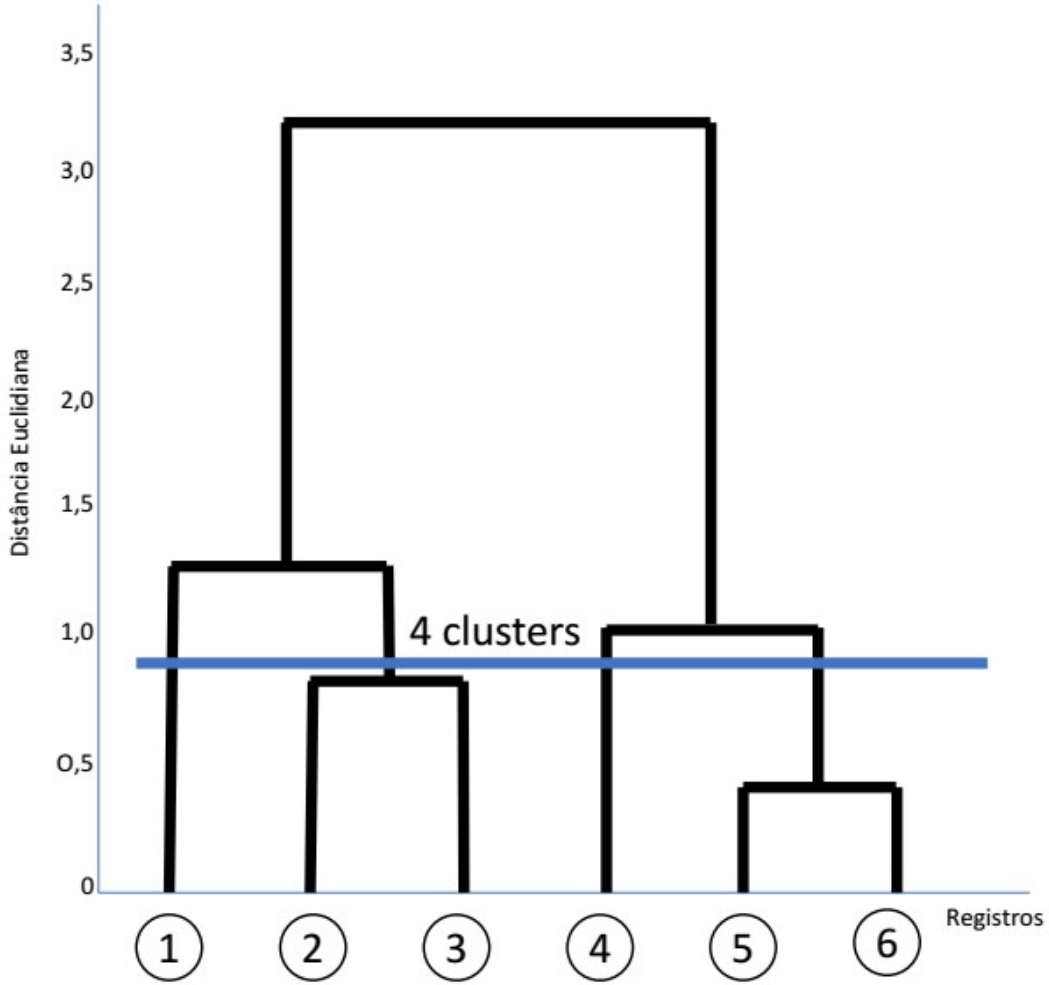
Algoritmo Hierárquico – Exemplo



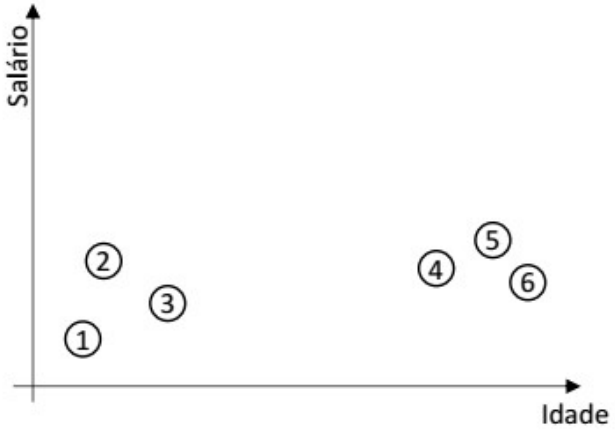


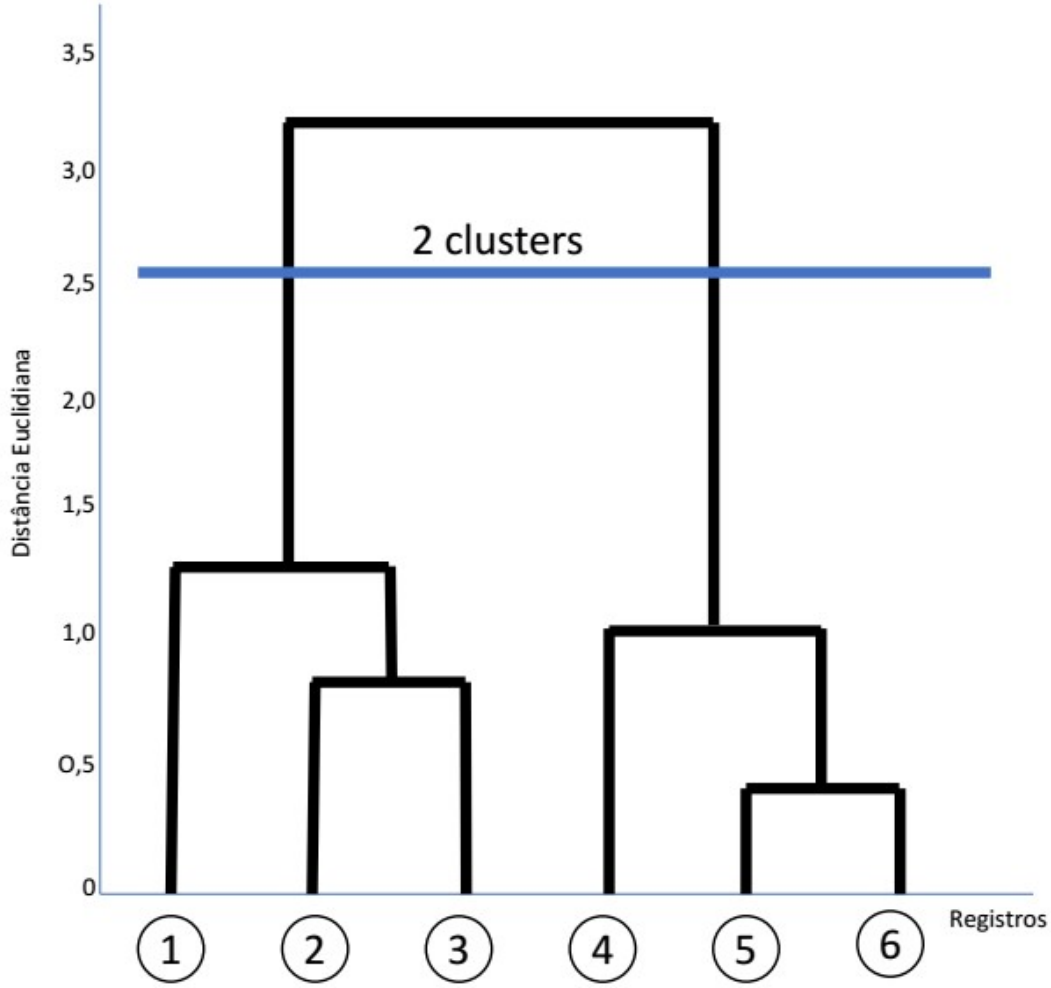
Algoritmo Hierárquico – Exemplo



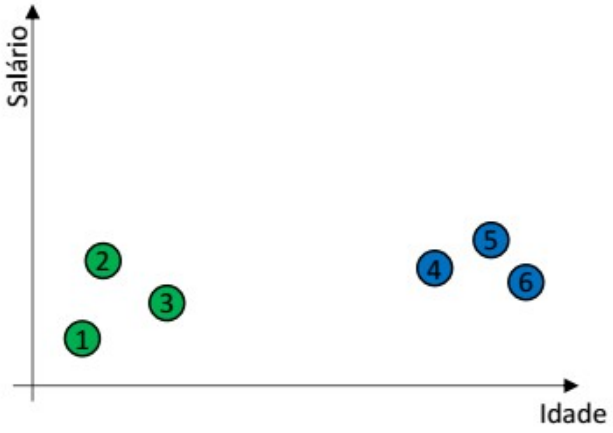


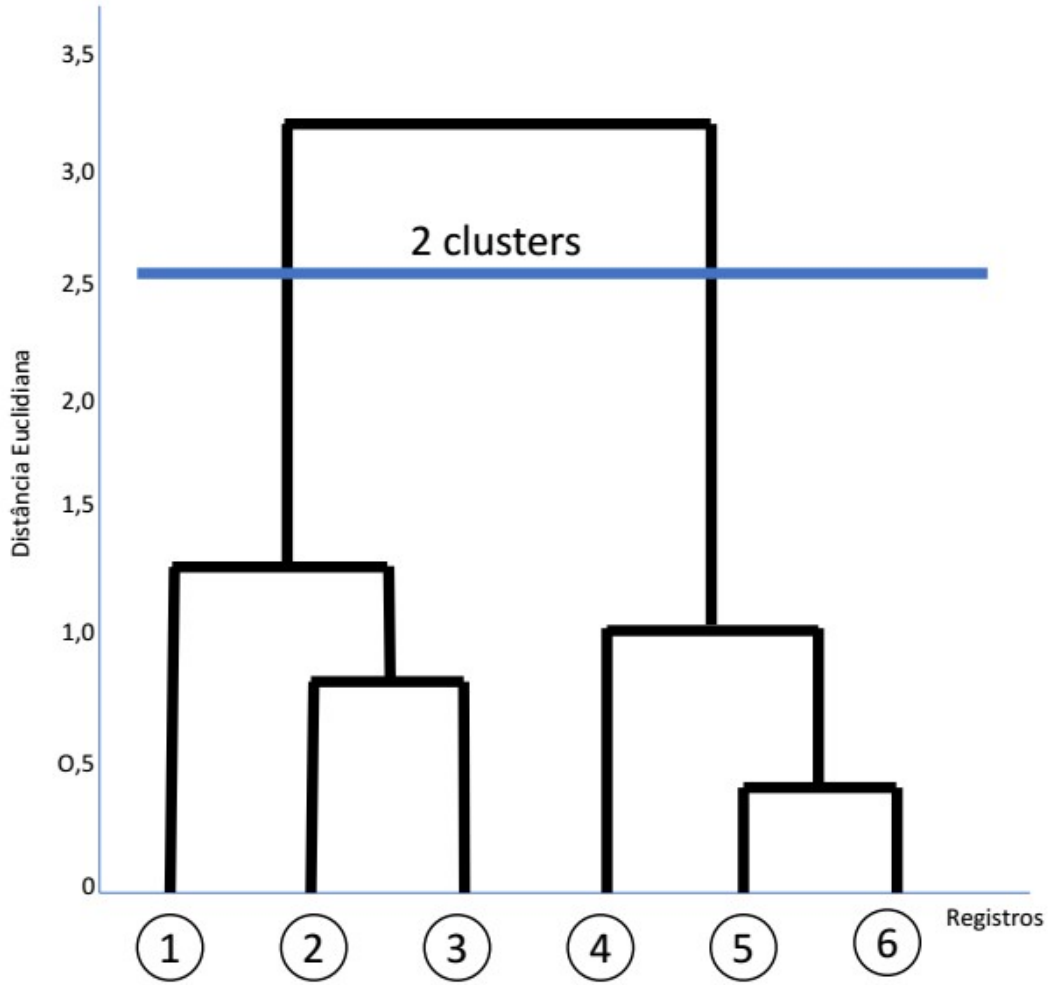
Algoritmo Hierárquico – Exemplo





Algoritmo Hierárquico – Exemplo





CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Agrupamento

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com