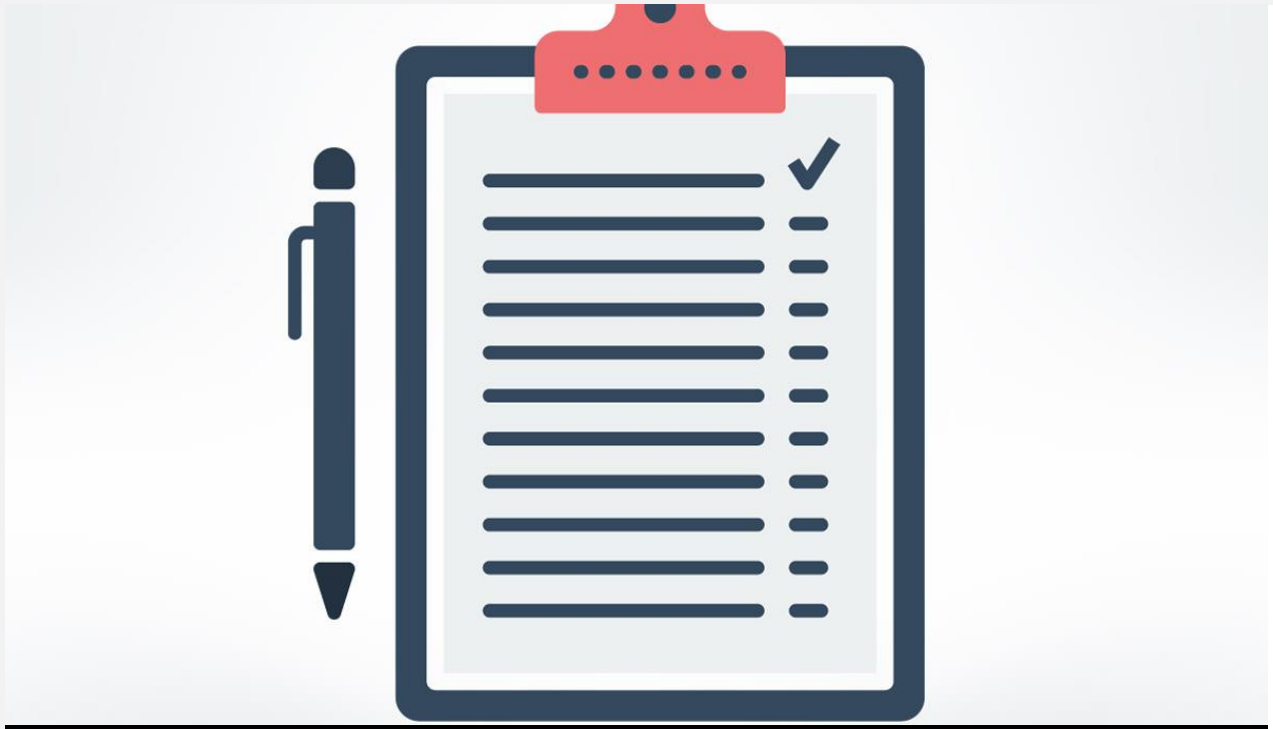
Análise de Agrupamentos

Eduardo Silva – easilva91@gmail.com



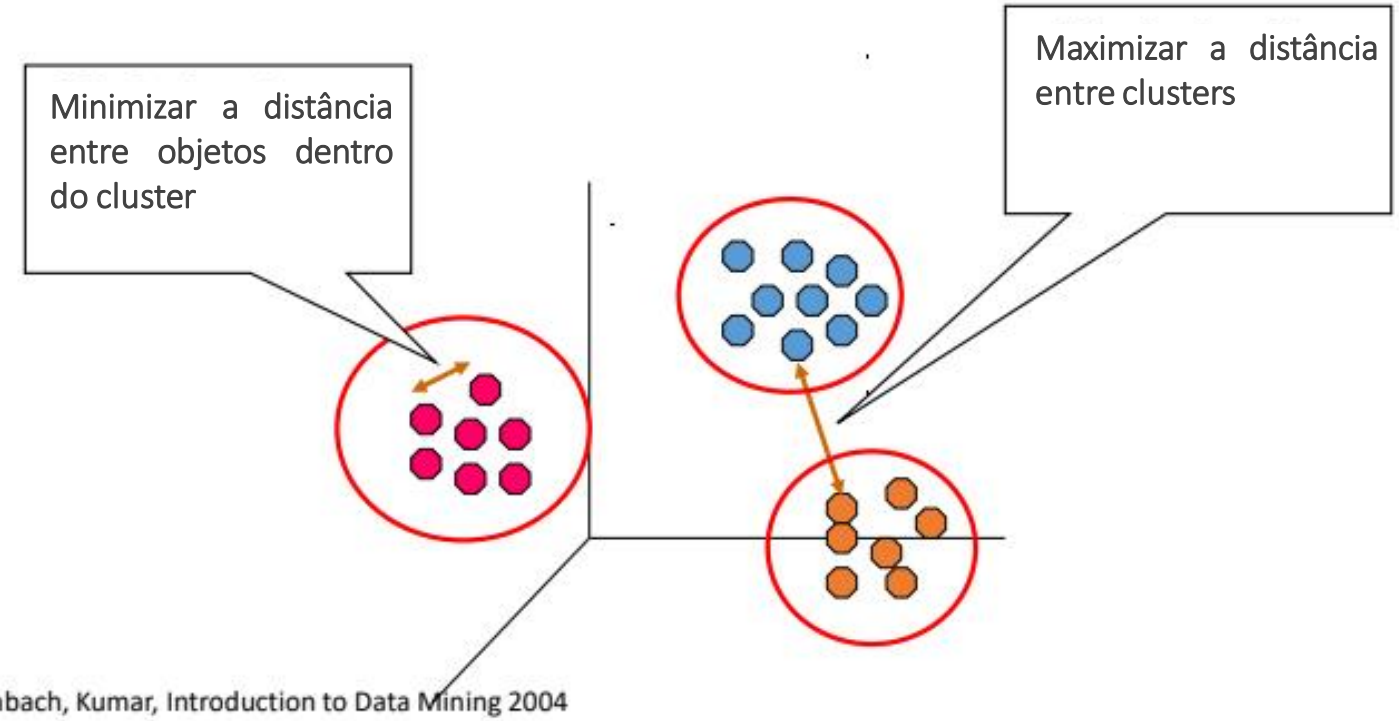
Agenda

1. Análise de Agrupamentos
2. Agrupamento vs classificação
3. Seleção de Variáveis
4. Critério de Similaridade
5. Análise de Agrupamentos: Algoritmos
   I. Métodos Hierárquicos
   II. K-médias/K-means
6. Número de Agrupamentos

Análise de Agrupamentos

- A possibilidade de reduzir a complexidade dos conjuntos reais infinitos de objetos ou fenômenos similares, é uma das ferramentas muito poderosa.
- A análise de agrupamentos/cluster é genérica para uma ampla variedade de metodologias que são usados para agrupar entidades.
- Objetivo: construir grupos de entidades semelhantes uns dos outros.
- A partir de um conjunto de dados (grupo de entidades) organizá-los em grupos homogêneos, determinando uma "estrutura" de semelhanças / diferenças entre as unidades.

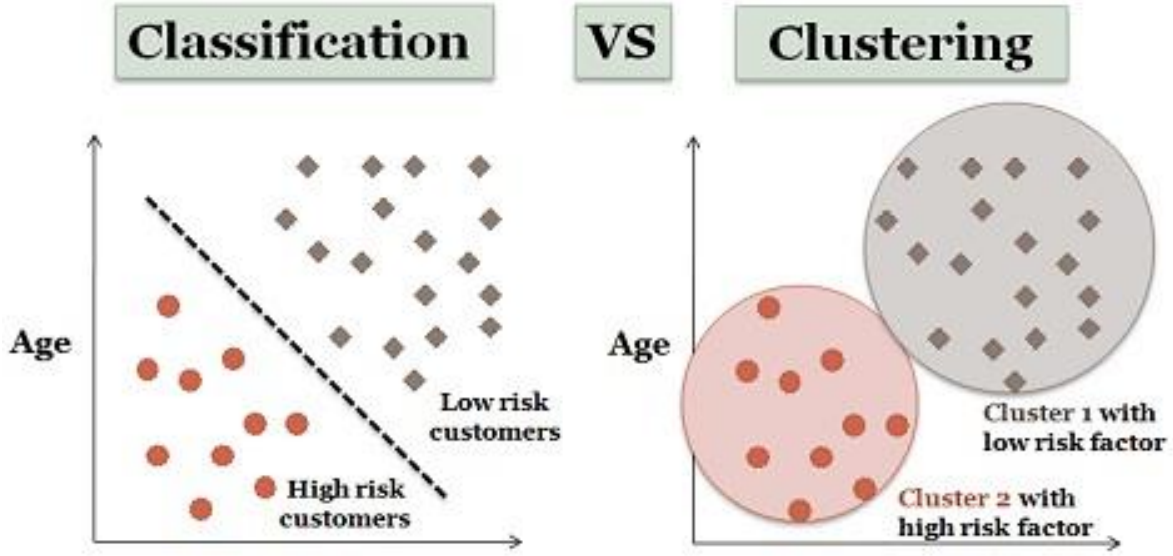
Análise de Agrupamentos



Source: Tan, Steinbach, Kumar, Introduction to Data Mining 2004

Agrupamento vs Classificação

- Classificação
  - Começa com um conjunto de dados pré-classificados, ou o método que usa um conjunto de dados está contido não apenas onde as variáveis usadas para classificar, como também a classe à qual pertence cada um dos registros.
  - Tentamos desenvolver um modelo capaz de prever como um novo registro será classificado.



Risk classification for the loan payees on the basis of customer salary

Agrupamento vs Classificação

- Clusterização/agrupamentos
  - Os dados não estão pré-classificados.
  - Busca por grupos (clusters) de indivíduos que são similares entre si.
  - A crença subjacente é que indivíduos semelhantes em termos das variáveis usado terá comportamentos semelhantes.

Análise de Agrupamento

- Existem 4 estágios básicos que caracterizam os estudos que envolvem análise de agrupamentos/clusters.

1. Definir variáveis
   - definir um conjunto de variáveis sobre as quais avaliará a similaridade / dissimilaridade das entidades.
2. Critério de Similaridade
   - definição de um critério de semelhança ou dissimilaridade entre as entidades (dados normalização).
3. Algoritmo
   - um algoritmo usando análise de cluster para criar grupos de entidades semelhantes.
4. Perfilamento/Profiling
   - Validação da solução resultante.

Seleção de Variáveis

- Segmentação
  - Univariada
  - Multivariada
- Segmentação
  - Valores
  - Necessidades
  - Comportamento
  - Características socio económicas.

Critério de Similaridade

- As medidas geométricas baseadas em espaços euclidianos dominaram o análise de relações de similaridade.
- Estas distâncias representam objetos como pontos no espaço multidimensional, de modo que as semelhanças entre os objetos correspondem às suas distâncias.
- Assim, as metodologias de cluster usam índices de métricas de similaridade que satisfazem as propriedades.
- Cálculos para distinguir similaridade/dissimilaridade:
  - Distância euclidiana (dissimilaridade).
  - Distância de Manhattan.
  - Distância de Minkowski.
  - Correlação de Pearson (similaridade).

Critério de Similaridade

Medida de Correlação

Individuo	Altura(cm)	Peso(kg)	Idade(anos)
A	180	79	15
B	175	75	28
C	170	70	50
D	167	63	25
E	180	71	80
F	165	60	31

Transpor

A	B	C	D	E	F
180	175	170	167	180	165
79	75	70	63	71	60
15	28	50	25	80	31

Coeficiente de correlação

	A	B	C	D	E	F
A	1	-	-	-	-	-
B	0,997	1	-	-	-	-
C	0,972	0,987	1	-	-	-
D	0,991	0,998	0,994	1	-	-
E	0,892	0,924	0,974	0,944	1	-
F	0,982	0,994	0,999	0,999	0,961	1

Análise de Agrupamentos: Algoritmos

- Tradicionalmente, as técnicas de análise de cluster são divididas em dois grandes grupos:

- Métodos Hierárquicos
- Métodos otimizados ou de partição

Métodos Hierárquicos
(reunião de pares semelhantes)

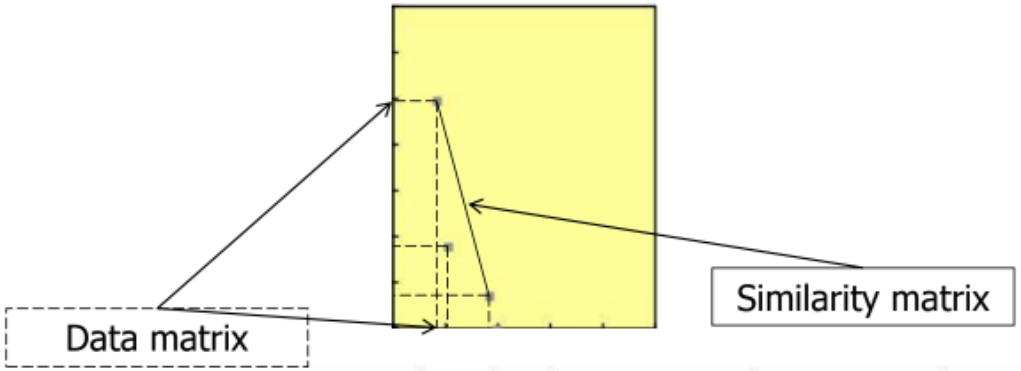
- Média das Distâncias (MMD)
- Centroide
- Ward
- Simple linkage (vizinho + próximo)
- Average Linkage (vizinho + distante)
- Average Linkage

Métodos otimizados ou de partição
(reunião de pares semelhantes)

K-médias ou K-means

Métodos Hierárquicos

- Uma estrutura de cluster em forma de árvore (dendrograma) é criada através da combinação recursiva ou da divisão de registros.
- Os métodos aglomerativos inicializam cada observação como um pequeno cluster próprio e combinam clusters existentes para criar a árvore.
- Os métodos divisivos começam com todos os registros em um grande cluster e separam os registros mais diferentes em um cluster separado.



	X_1	X_2	...	X_p
I_1				
I_2				
...				
I_n				

	I_1	I_2	...	I_n
I_1	0			
I_2	$d(I_2, I_1)$	0		
...	$d(I_{..}, I_1)$	$d(I_{..}, I_2)$	0	
I_n	$d(I_n, I_1)$	$d(I_n, I_2)$	$d(I_n, I_{..})$	0

Métodos Hierárquicos

Input distance matrix:

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

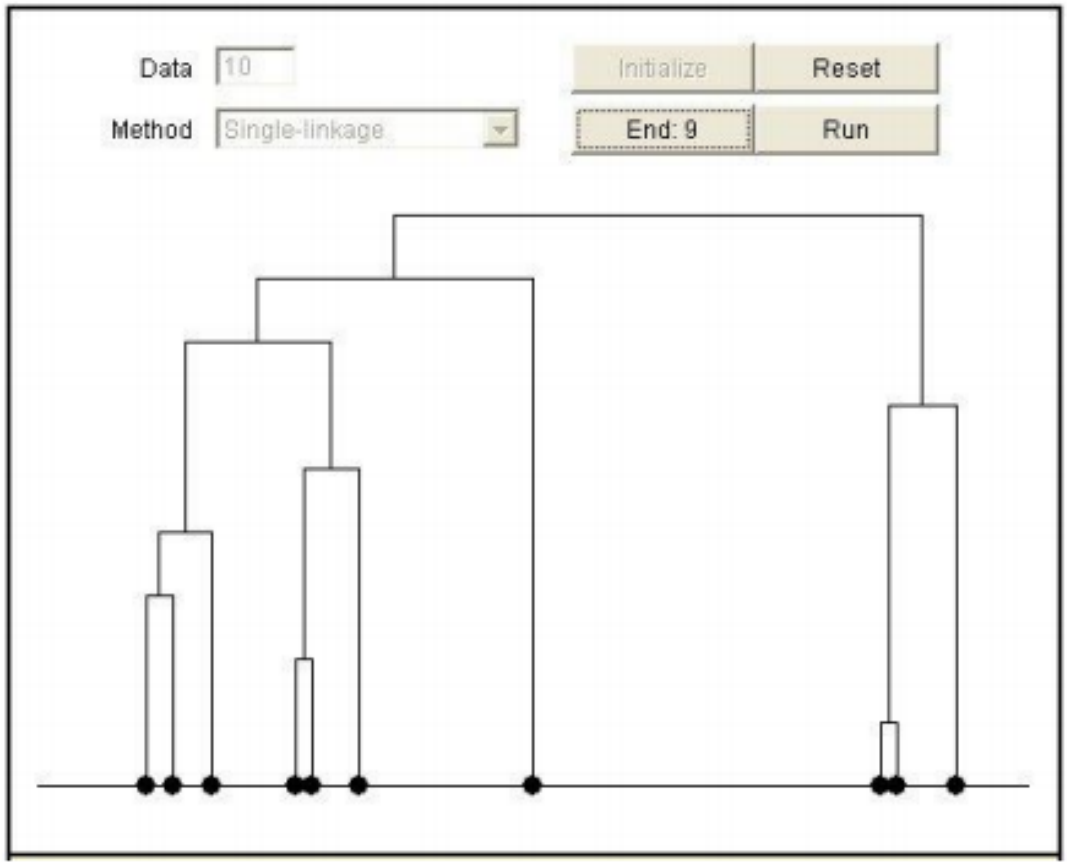
After merging BOS with NY:

	BOS/NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY	0	223	1308	802	2815	2934	2786	1771
DC	223	0	1075	671	2684	2799	2631	1616
MIA	1308	1075	0	1329	3273	3053	2687	2037
CHI	802	671	1329	0	2013	2142	2054	996
SEA	2815	2684	3273	2013	0	808	1131	1307
SF	2934	2799	3053	2142	808	0	379	1235
LA	2786	2631	2687	2054	1131	379	0	1059
DEN	1771	1616	2037	996	1307	1235	1059	0

After merging DC with BOS-NY:

	BOS/NY/DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY/DC	0	1075	671	2684	2799	2631	1616
MIA	1075	0	1329	3273	3053	2687	2037
CHI	671	1329	0	2013	2142	2054	996
SEA	2684	3273	2013	0	808	1131	1307
SF	2799	3053	2142	808	0	379	1235
LA	2631	2687	2054	1131	379	0	1059
DEN	1616	2037	996	1307	1235	1059	0

Métodos Hierárquicos



http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletH.html

K-médias/K-means

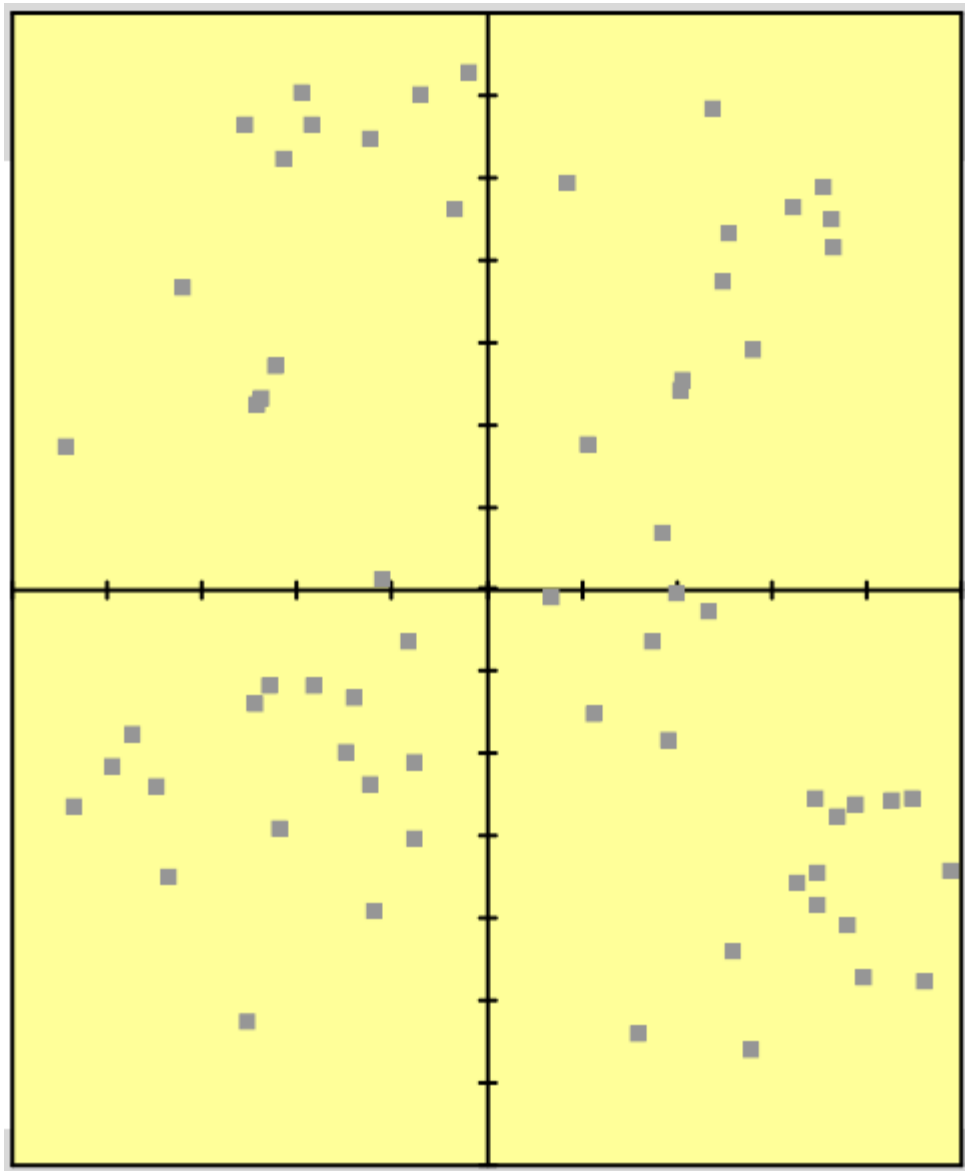
- Métodos de otimização ou partição
  - Dado um banco de dados com **n** objetos,
  - Construa **k** partições, onde cada partição representa um cluster/agrupamento
  - K <= n
  - Classifique os dados em k grupos, satisfazendo as seguintes condições:
    - Cada grupo contém pelo menos um objeto;
    - Cada objeto pertence apenas a um cluster.
  - Dado k, o método cria uma partição inicial (normalmente randomicamente)
  - Então o algoritmo utiliza uma técnica de realocação que tem como objetivo melhorar o particionamento, movendo os objetos de um grupo para o outro.
  - Geralmente, o critério de uma boa partição é que objetos pertençam ao mesmo cluster ou estão próximos uns dos outros.

K-médias/K-means

- Algoritmo:
  1. Definir semente (para iniciar randomicamente)
  2. Cada individuo é associado com a semente mais próxima
  3. Calcula os centroids dos clusters formados
  4. Volta ao passo 2.
  5. Termina quando os centroids não tem mais alterações.

K-médias/K-means

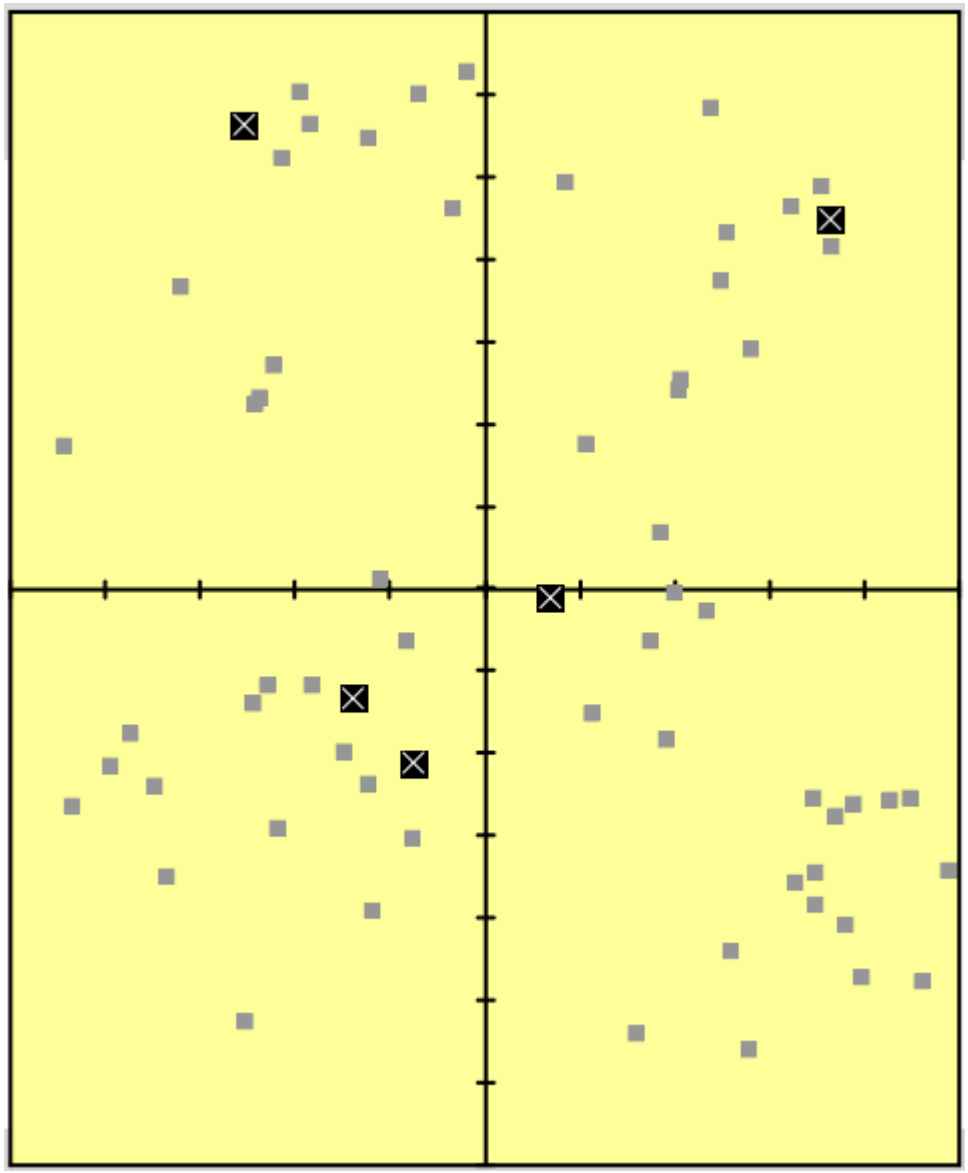
- Dados
  - 2 variáveis
  - Vamos agrupa-los



Fonte: Fiona Cameron, Techniques for Neighbourhood Classification

K-médias/K-means

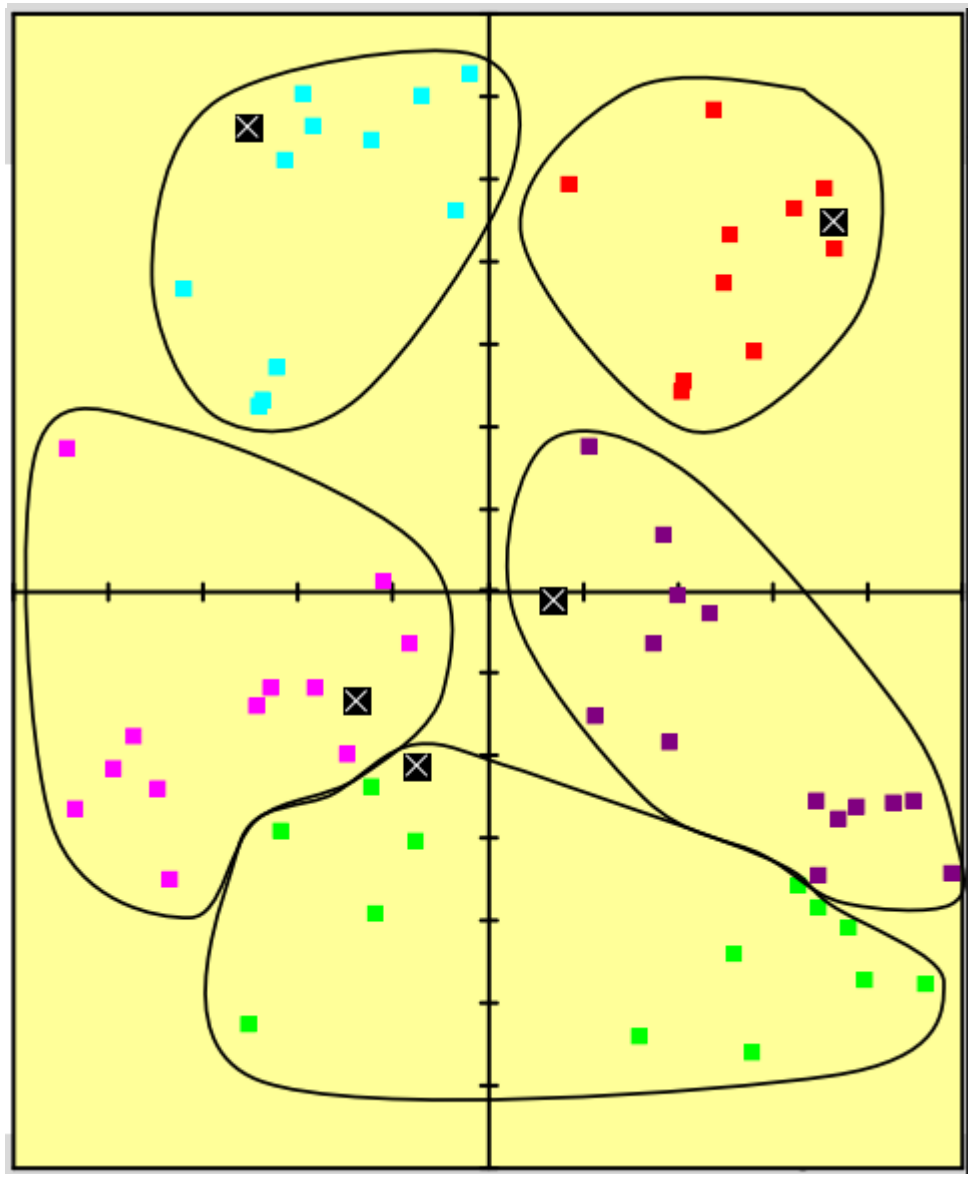
- Inicialização
  - 5 seeds/sementes
  - Definidas aleatoriamente/randomicamente



Fonte: Fiona Cameron, Techniques for Neighbourhood Classification

K-médias/K-means

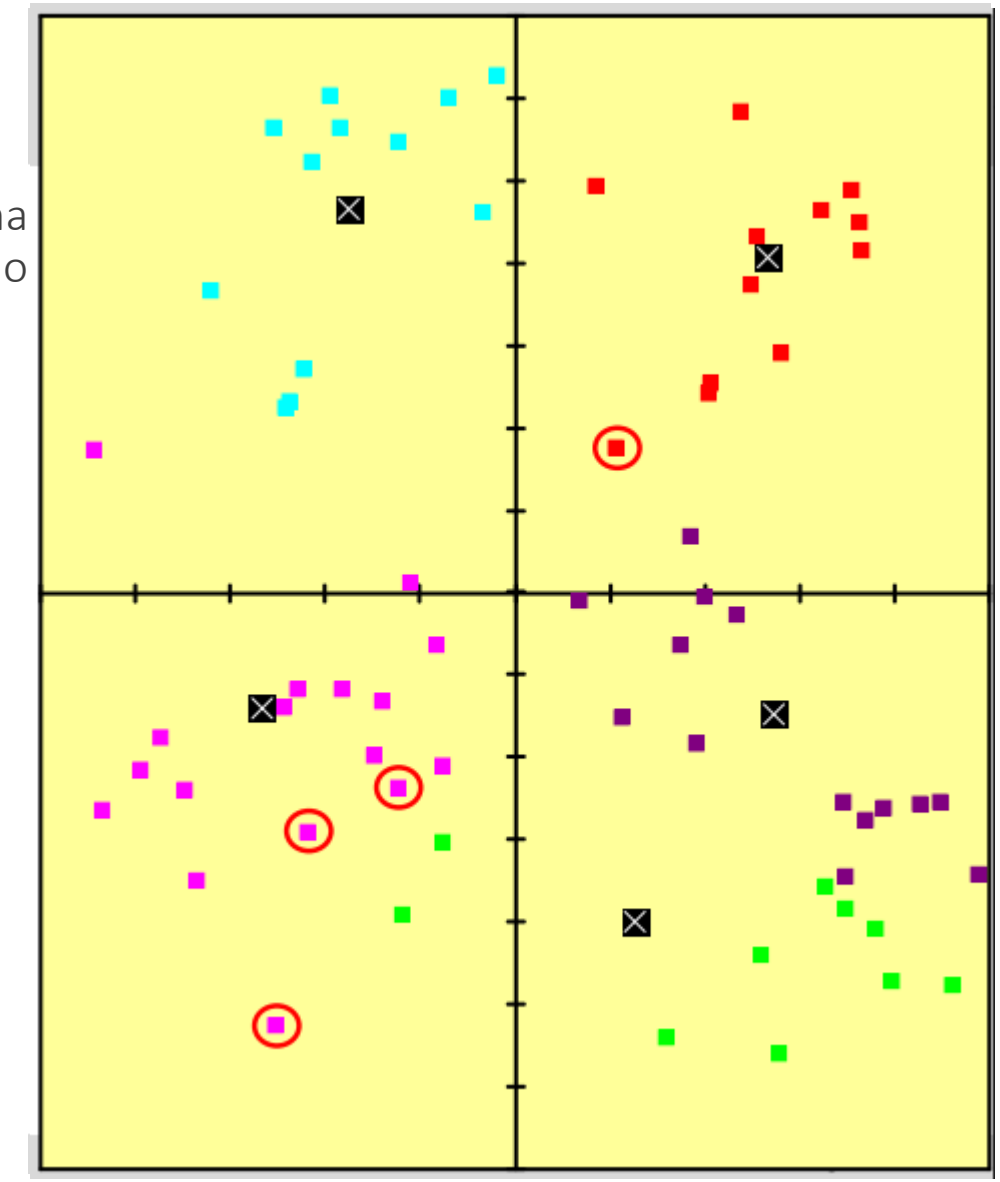
- Iteração 1, primeiro passo
  - Define a seed mais próxima de cada ponto



Fonte: Fiona Cameron, Techniques for Neighbourhood Classification

K-médias/K-means

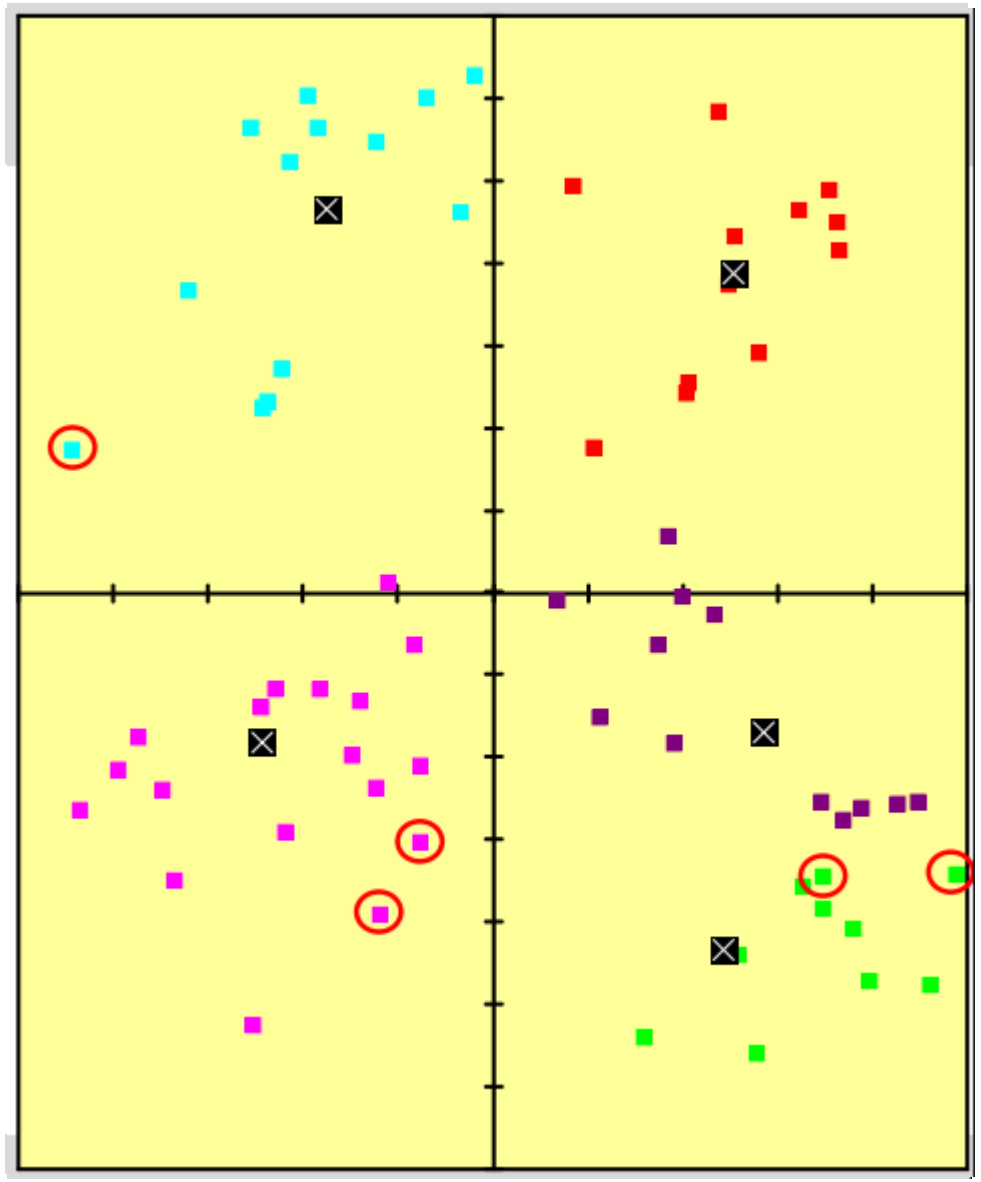
- Iteração 1, segundo passo
  - Recalcule a semente de modo que ela fique na nuvem de pontos representando (denominado centroide) o centro
  - Alguns indivíduos mudam de agrupamento



Fonte: Fiona Cameron, Techniques for Neighbourhood Classification

K-médias/K-means

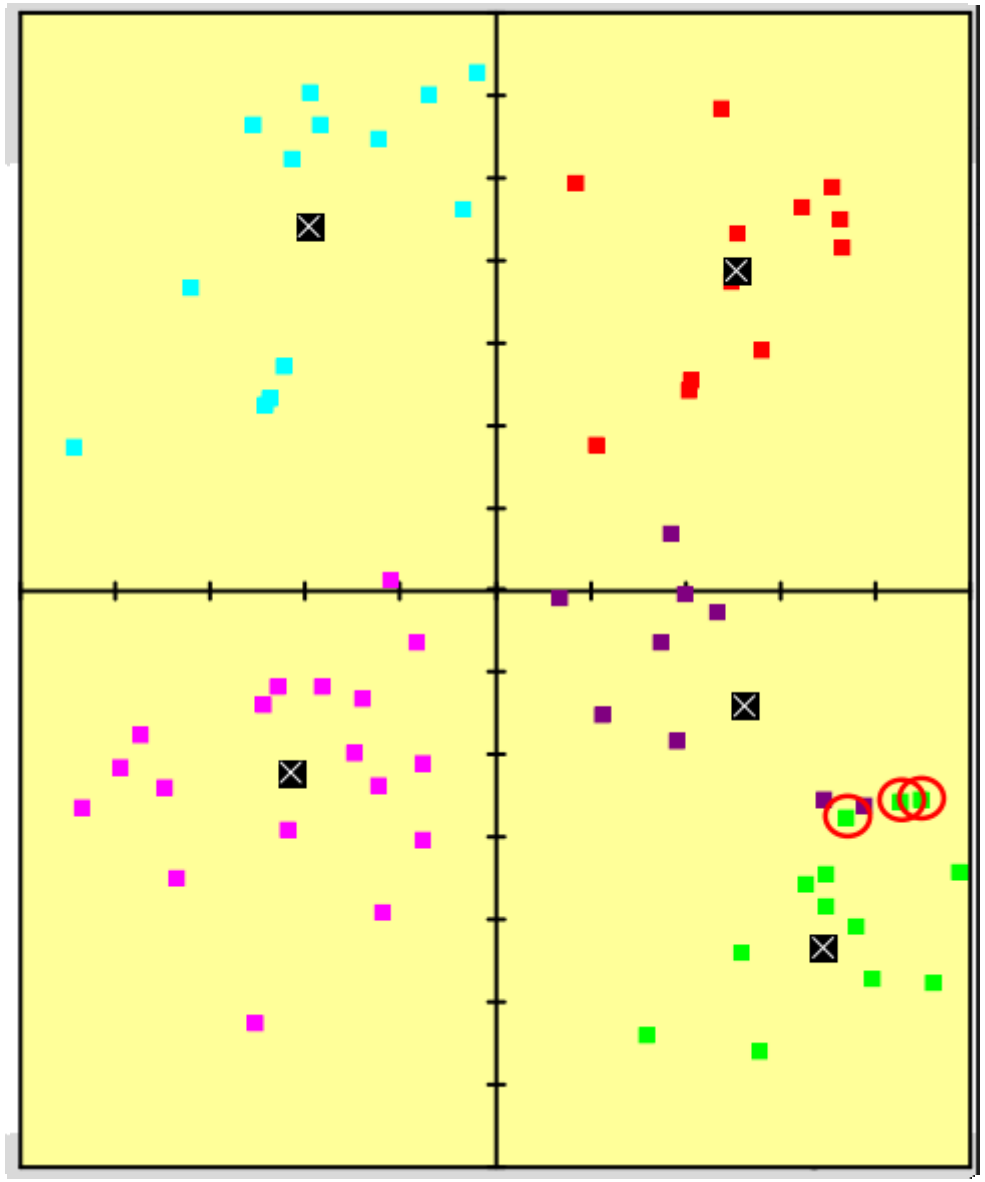
- Iteração 2



Fonte: Fiona Cameron, Techniques for Neighbourhood Classification

K-médias/K-means

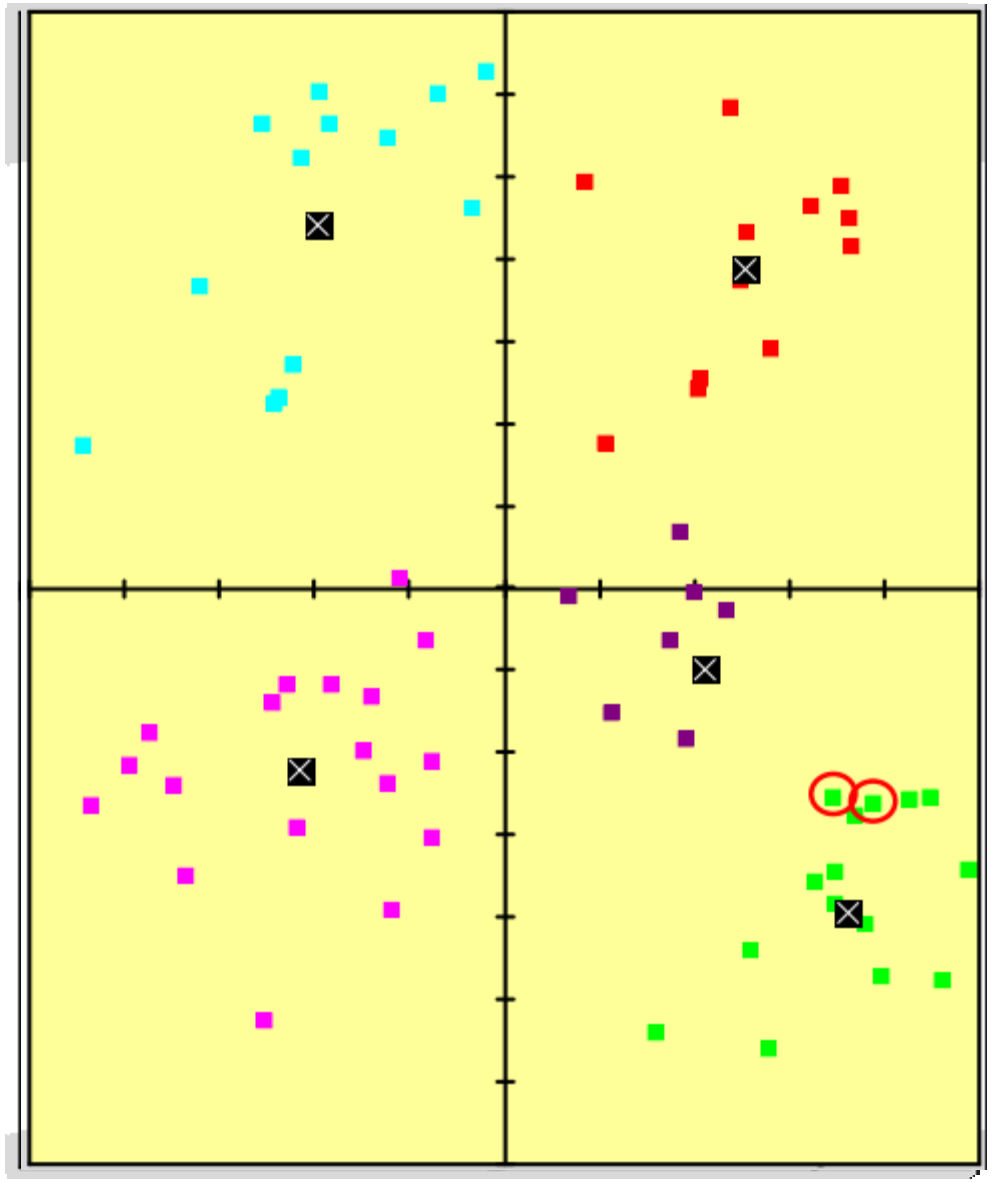
- Iteração 3



Fonte: Fiona Cameron, Techniques for Neighbourhood Classification

K-médias/K-means

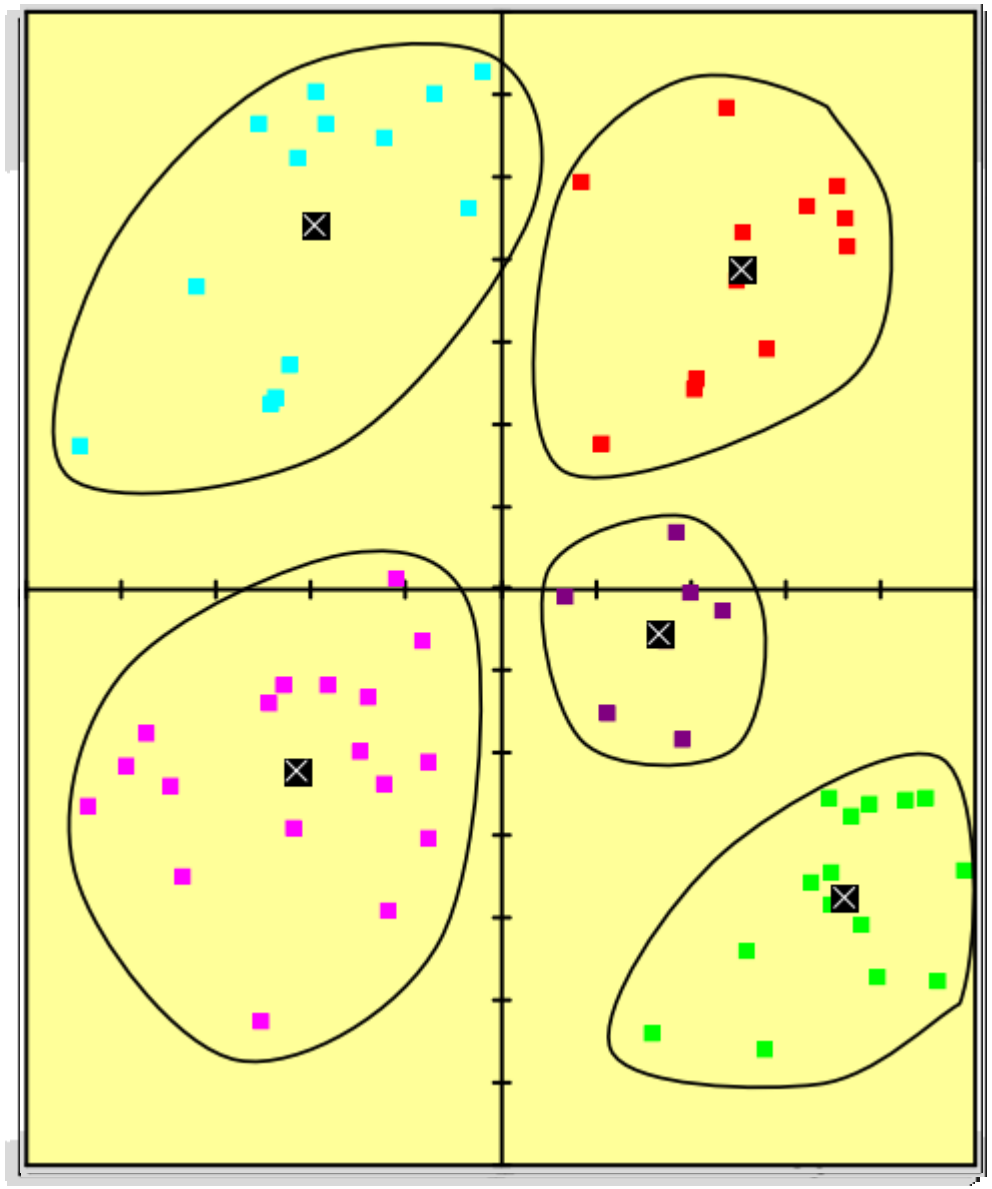
- Iteração 4



Fonte: Fiona Cameron, Techniques for Neighbourhood Classification

K-médias/K-means

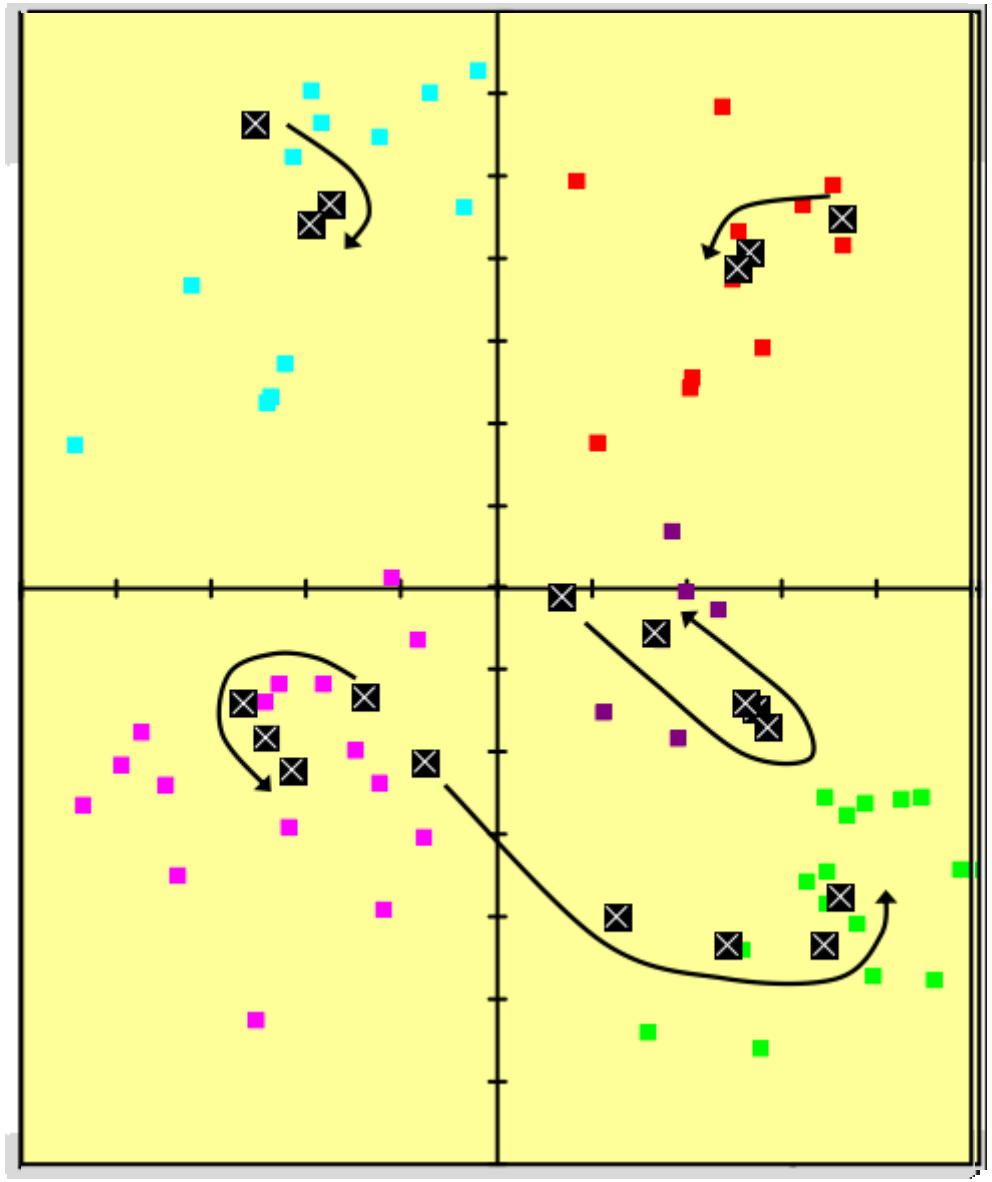
- Solução Final



Fonte: Fiona Cameron, Techniques for Neighbourhood Classification

K-médias/K-means

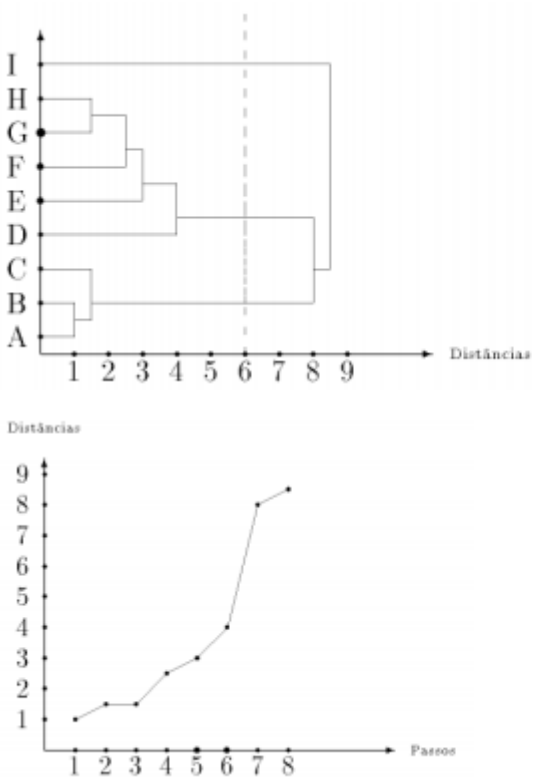
- Movimento dos Centroides

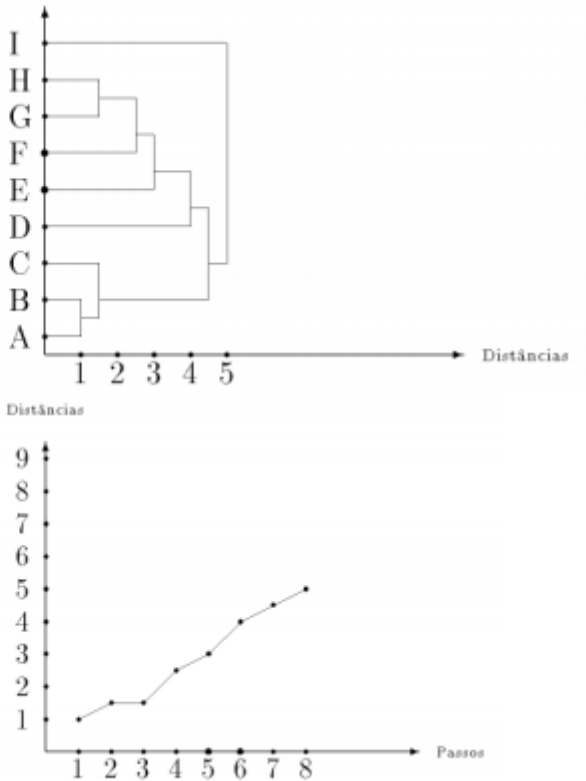


Fonte: Fiona Cameron, Techniques for Neighbourhood Classification

Número de Agrupamentos

- Um problema difícil de resolver.
- Produza vários clusters com diferentes k, e escolha o melhor.
- Use métodos hierárquicos de forma a escolher o número de clusters baseado no dendrogram.





- https://colab.research.google.com/drive/1wAitTnzo7pAQndUG9BDU2jmIR7ZDVzIB

Obrigado

Real Python