ELDORADO

Trilha Ciência de dados com Python

Aula 16

TIC em trilhas

Chamada

TIC em trilhas

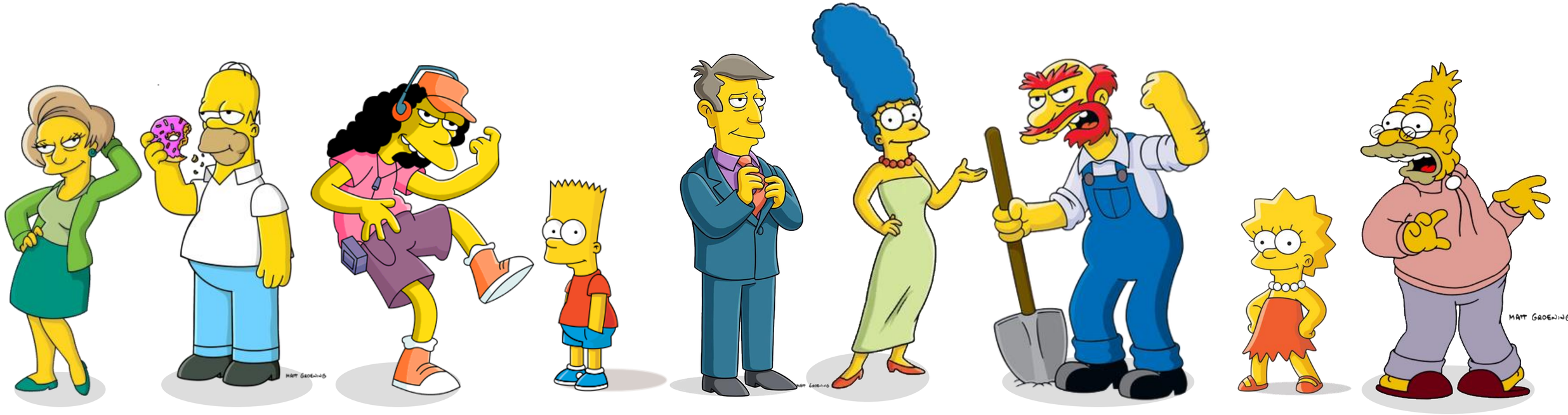
Faísca

TIC em trilhas

Roteiro de hoje!

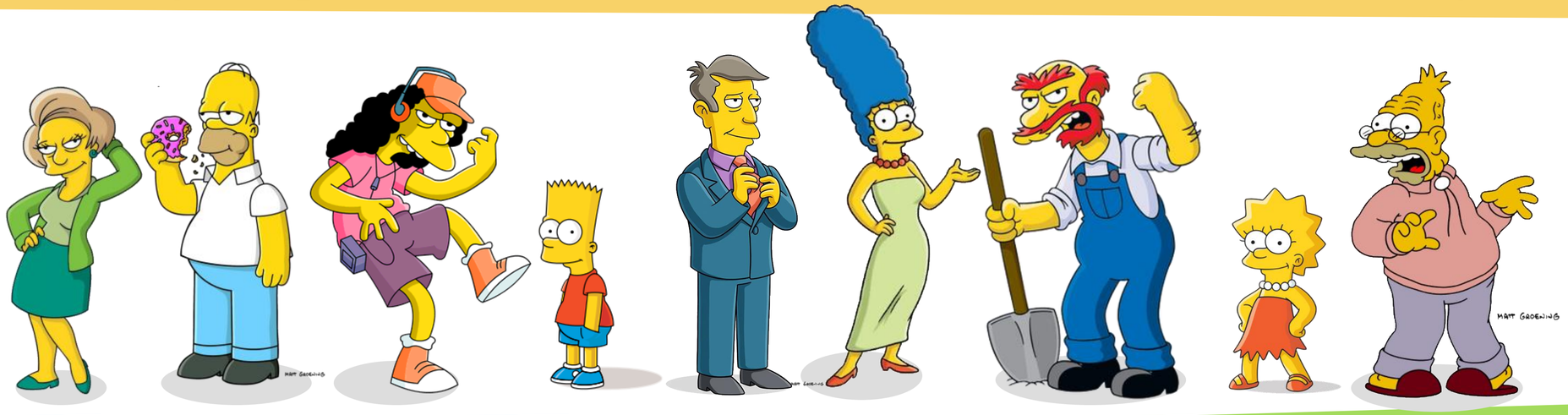
- Definição
- Introdução ao Agrupamento de Dados
  - Definições de Agrupamento
  - Caso de Uso
- Algoritmos Particionais
  - Conceitos básicos
  - Algoritmo
  - Validação
  - Exemplo prático
- Atividade

Qual é o Agrupamento Natural destes Objetos?

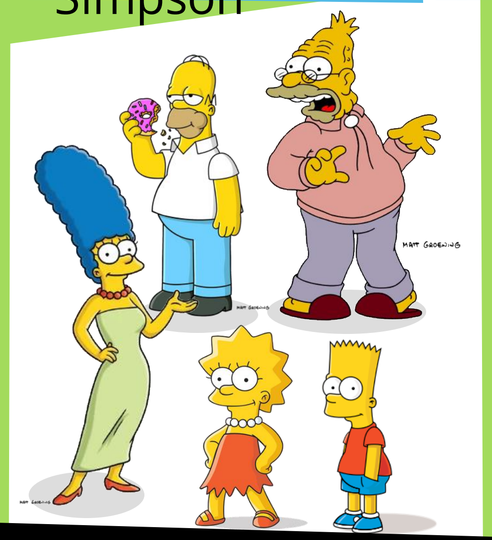


TIC em trilhas

Qual é o Agrupamento Natural destes Objetos?



Família Simpson



Empregados da Escola

Agrupamento

é subjetivo!

Mulheres

Homens

TIC em trilhas

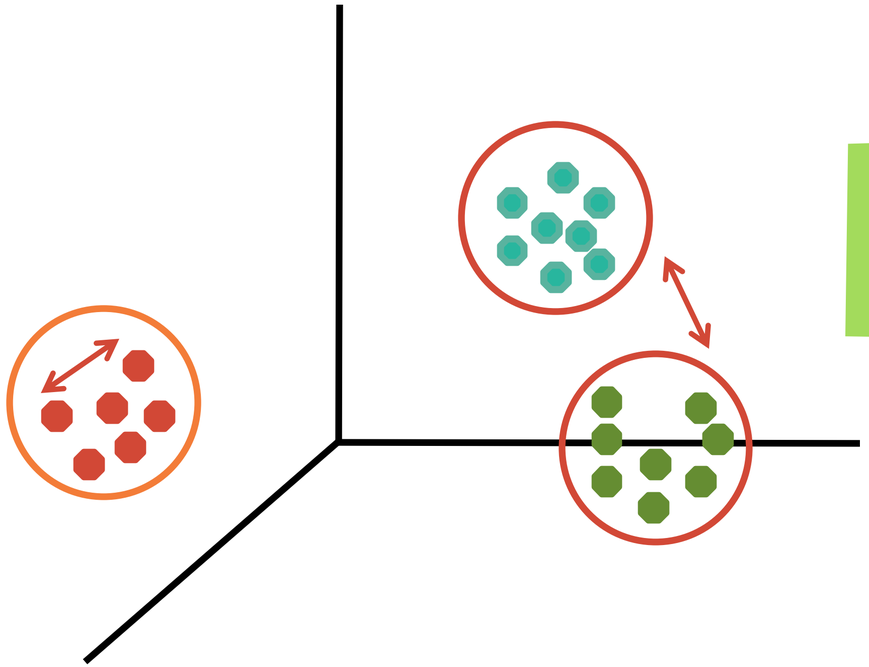
O que é Agrupamento de Dados

- Encontrar os rótulos (classes) implícitos dos dados **sem supervisão**.
- Organizar dados em grupos tal que exista:
  - Alta similaridade intra-grupo.
  - Baixa similaridade entre grupos.

Ilustrando Agrupamento

- Agrupamento baseado na Distância Euclidiana no espaço 3-D.

Distâncias intra-grupos são minimizadas



Distâncias entre grupos são maximizadas

O que é Agrupamento de Dados

- Dado um conjunto de objetos, cada um tendo um conjunto de atributos, e uma métrica de similaridade entre eles, achar agrupamentos tal que:
  - Objetos de um agrupamento são muito similares entre si.
  - Objetos de agrupamentos diferentes são menos similares entre si.
- Métricas de Similaridade:
  - Distância Euclidiana, se os atributos são contínuos.
  - Outras métricas específicas a cada problema

Agrupamento: Casos de Uso

Segmentação de Clientes

- Auxilia empresas a melhorar a carta de clientes
- Segmenta os clientes com base no histórico de compras, interesses ou monitoração de atividades
- Paper: Prepaid Telecom Customers Segmentation using the K-means Algorithm
  - Agrupa clientes que possuem planos pré-pagos
    - Identifica padrões em termos de dinheiro gasto em recarga, envio de SMS e navegação na internet
    - Auxilia a empresa a visar padrões específicos de clientes para o direcionamento de campanhas de marketing

Agrupamento: Casos de Uso

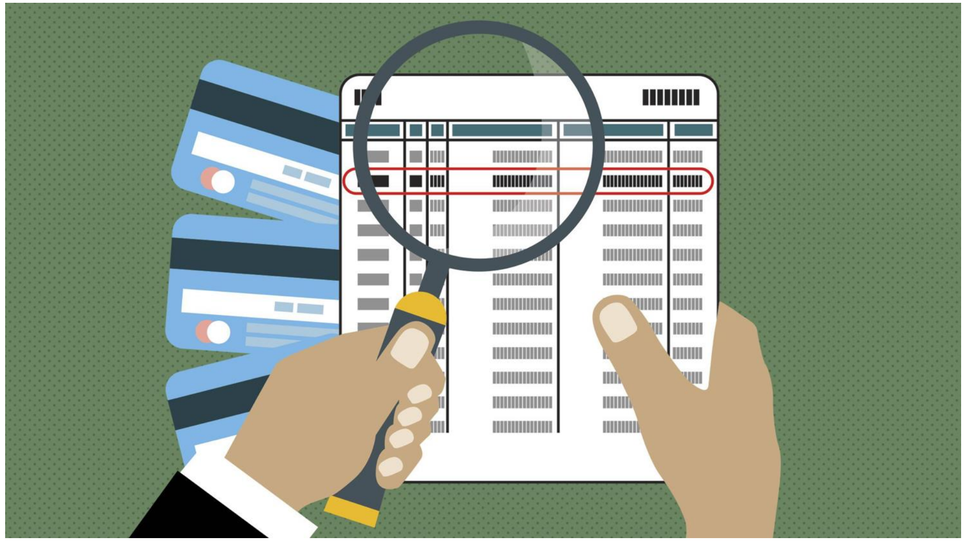
Analise de Jogadores

Fantasy league analysis

- Identificar uma lista de jogadores similares
- O algoritmo de ML aprende o que um “olheiro” realiza
- Criar um “Fantasy team” baseado no resultados dos padrões encontrados pelo algoritmo k-means

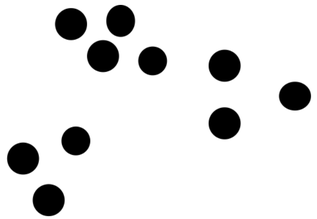
Agrupamento: Casos de Uso

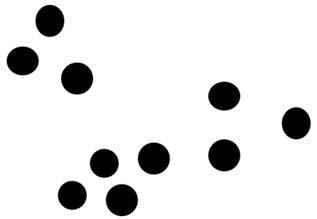
Detecção de Fraude



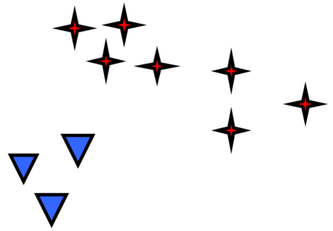
- Utiliza dados históricos sobre informações fraudulentas
- Quando uma nova reclamação aparece é possível isolar com base na sua proximidade de cluster que indicam padrões fraudulentos
- A identificação de fraudes pode ter um impacto muito grande nas empresas

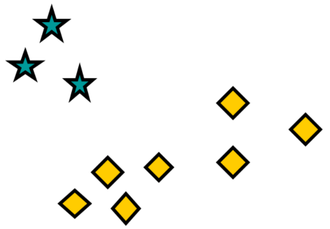
Tipos de Agrupamentos



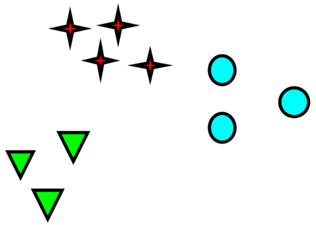


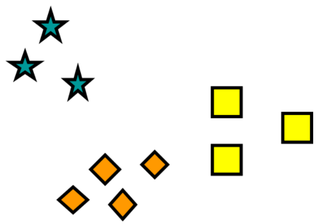
Quantos agrupamentos?



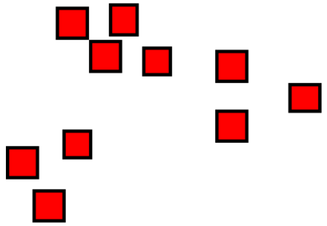


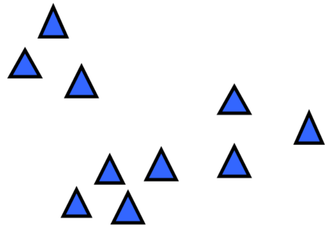
Quatro agrupamentos





Seis agrupamentos

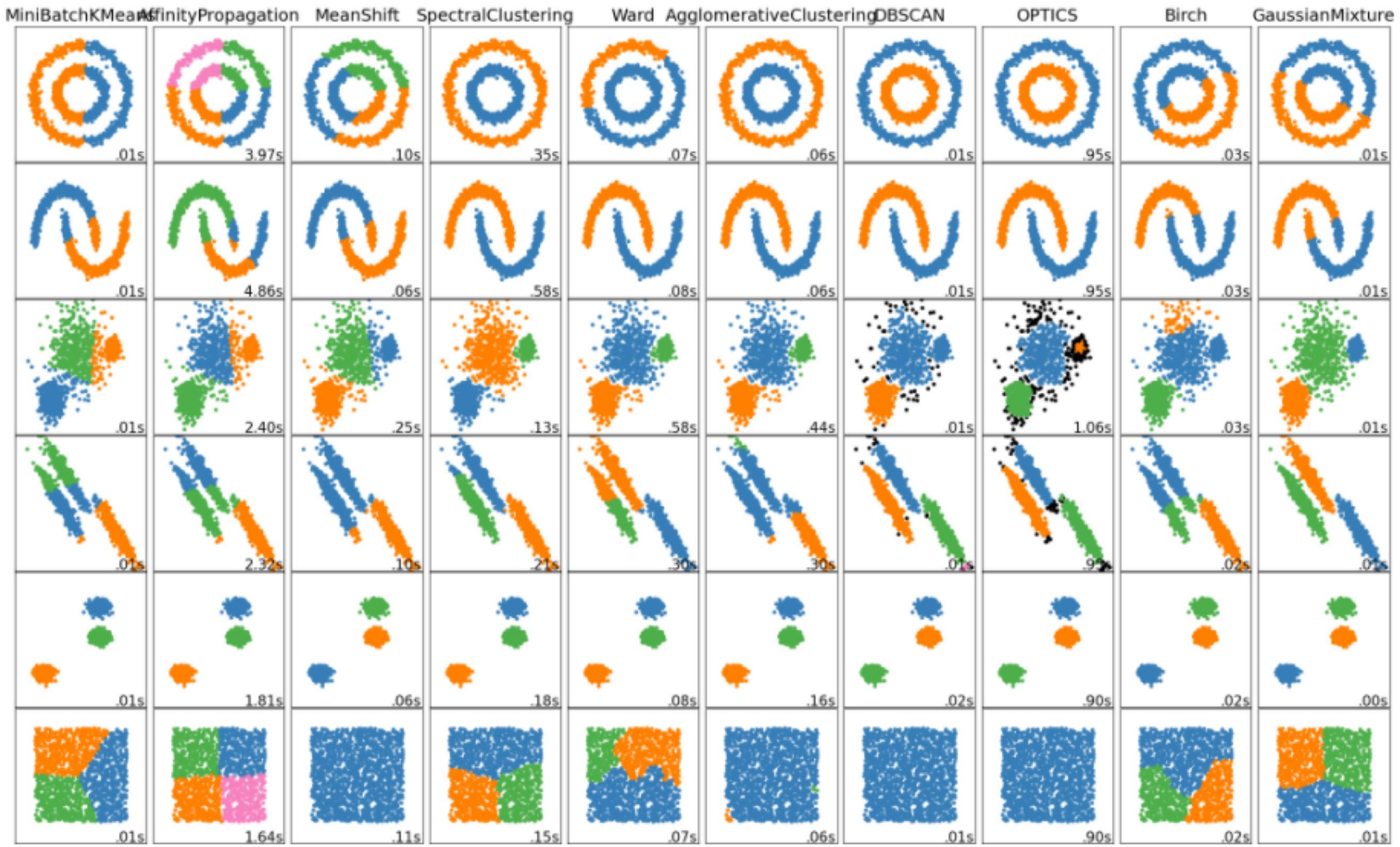




Dois agrupamentos

TIC em trilhas

Tipos de Agrupamentos



A comparison of the clustering algorithms in scikit-learn

Agrupamento Particional

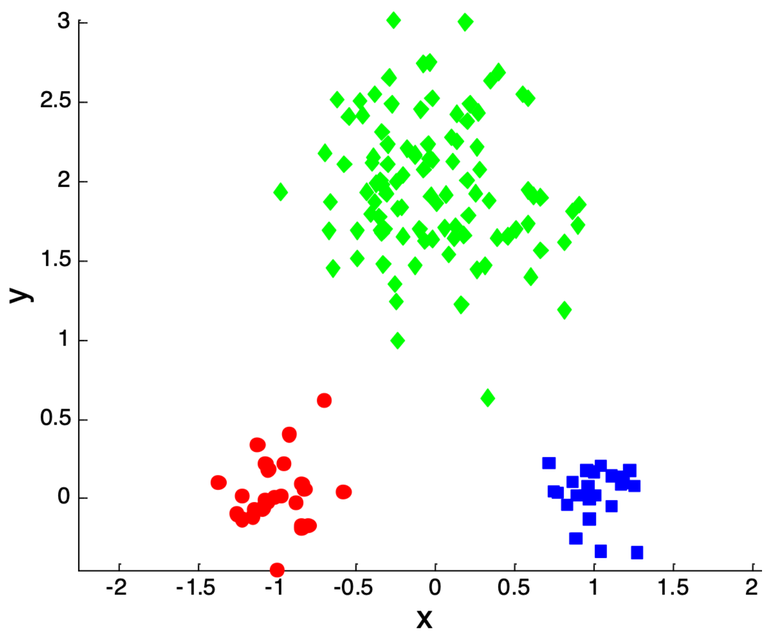
- **k-means:**
  - Cada grupo está associado a um centroide (ponto central).
  - Cada objeto é atribuído ao grupo com o centroide **mais próximo**.
  - O algoritmo é bem **simples**.
  - Número de grupos, ***k***, deve ser especificado.

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

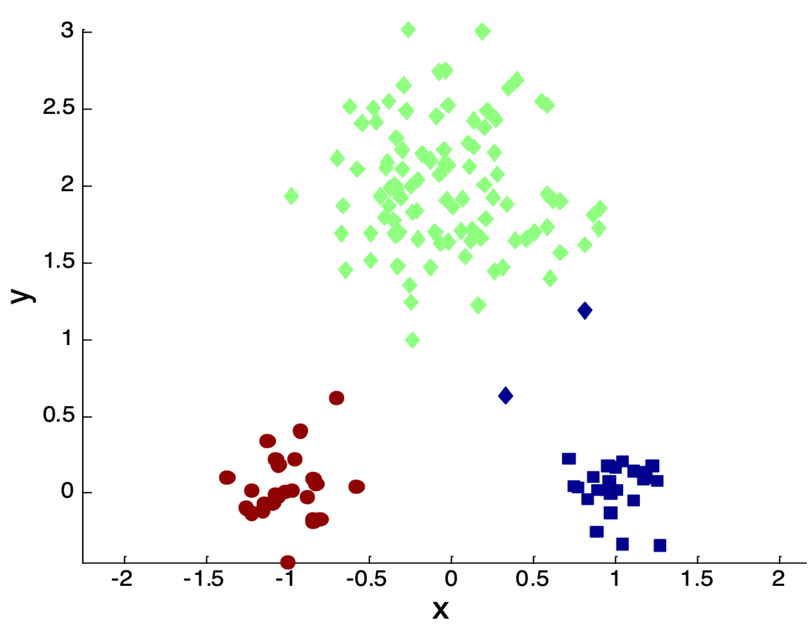
Algoritmo K-means – Detalhamento

- Centroides iniciais são geralmente aleatórios.
  - Agrupamento varia conforme a inicialização.
- Centroide são (tipicamente) a média de todos os objetos do grupo.
- A medida de distância geralmente empregada é a distância Euclidiana.
- *k*-means geralmente converge com poucas iterações.
  - Critério de parada é geralmente modificado para "até que poucos objetos alterem o grupo".
- Complexidade é O(n * k * i * d)
  - **n** = número de objetos
  - **k** = número de grupos
  - **i** = número de iterações
  - **d** = número de atributos

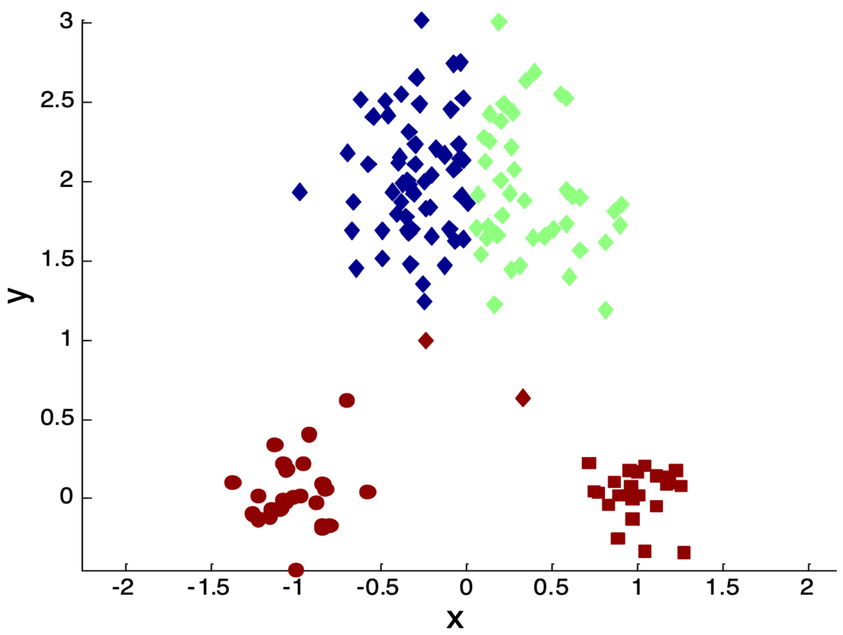
Aonde estão os centroides?



Objetos Originais

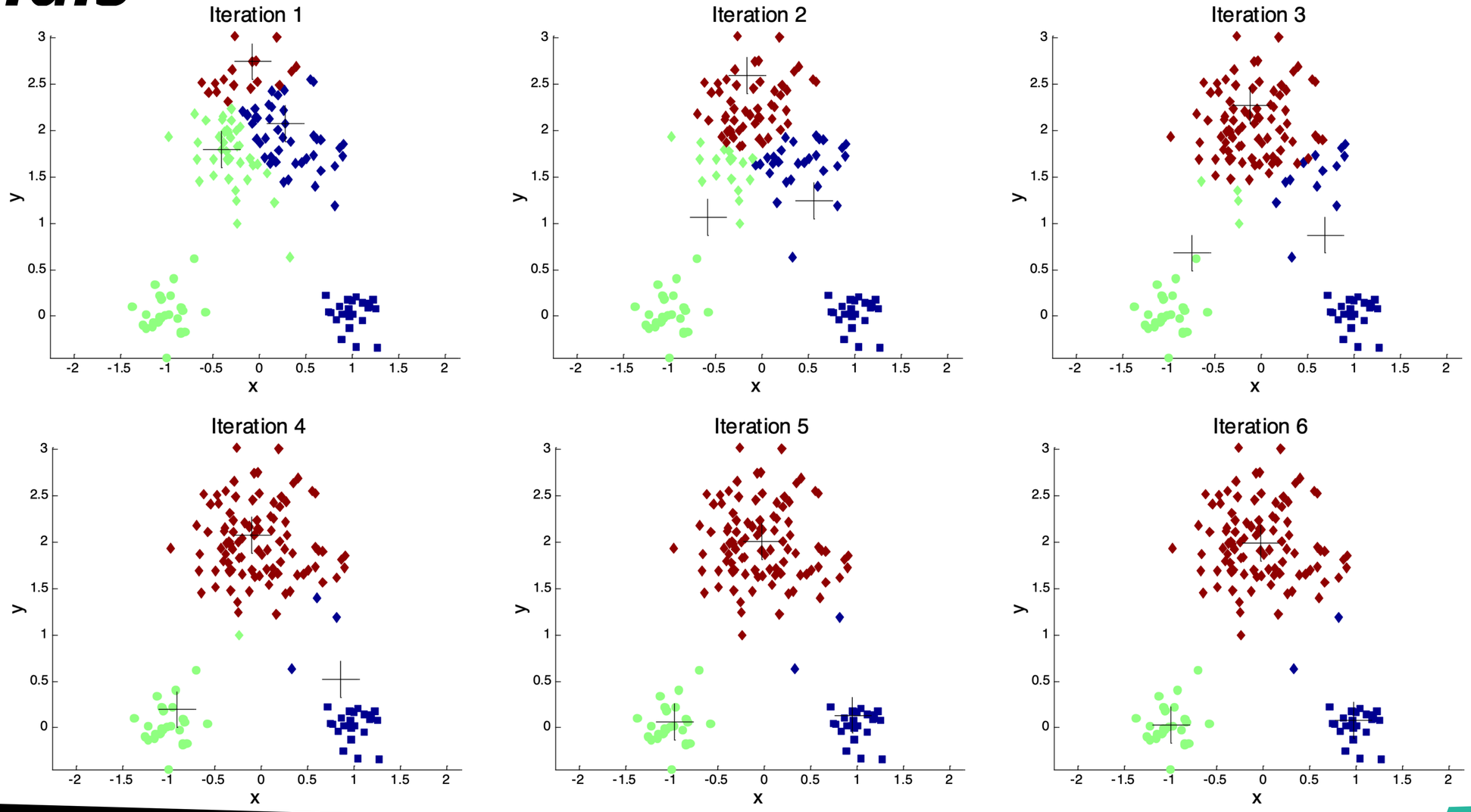


Agrupamento Ótimo



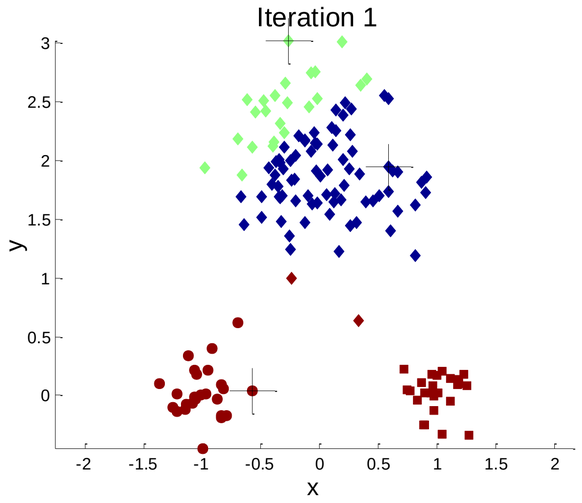
Agrupamento Sub-ótimo

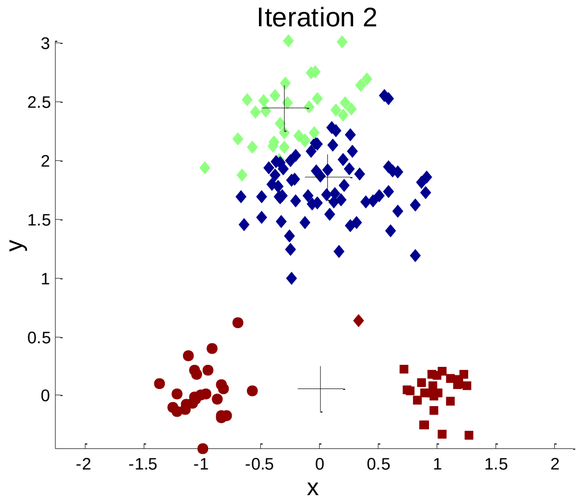
Importância da Escolha dos Centroids Iniciais

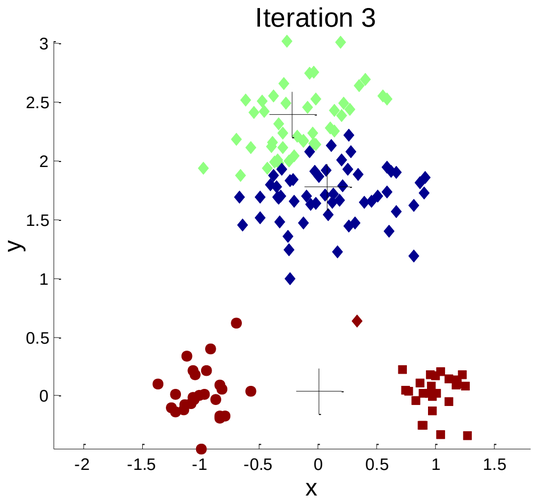


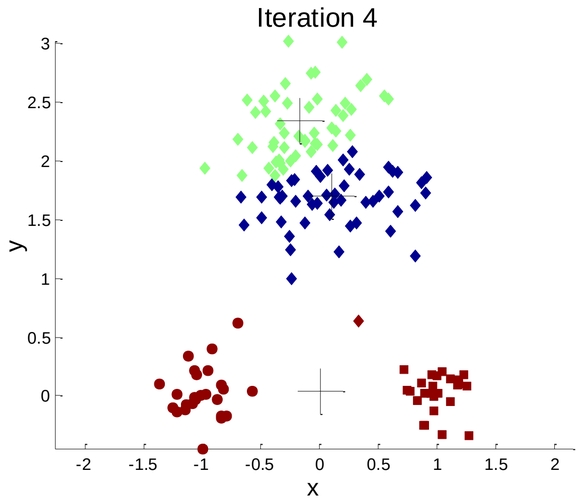
TIC em trilhas

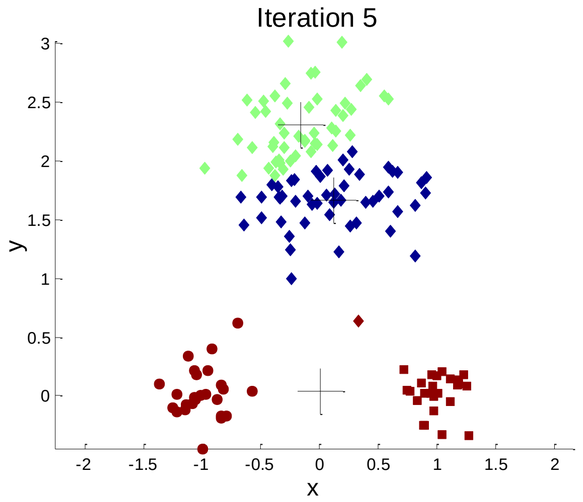
Importância da Escolha dos Centroids Iniciais











TIC em trilhas

Exemplificando...

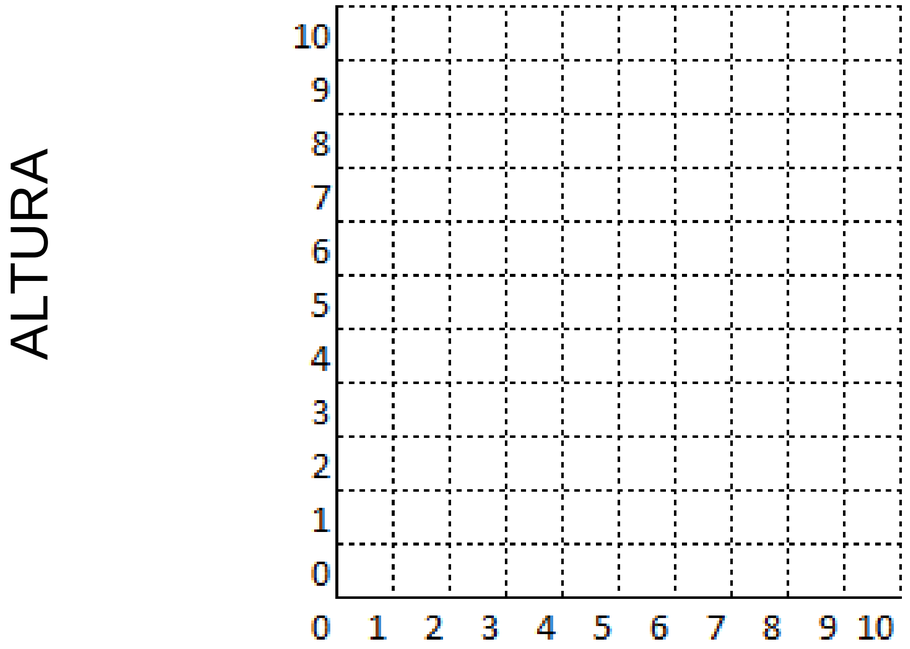
Considere o conjunto de dados abaixo, o qual possui 6 registros de peso e altura normalizados de 0 a 10.

Supondo que o Objeto 1 é o centroide inicial do agrupamento 1 e o Objeto 2 é o centroide inicial do agrupamento 2, quais serão os valores dos centroides dos dois agrupamentos, ao final da execução do algoritmo?

Centroide 1

Centroide 2

Objeto	Peso	Altura
1	2	8
2	8	2
3	6	8
4	2	7
5	8	4
6	2	6



Exemplificando...

1. Selecione k objetos como centroides iniciais.
2. Repita
3. Forme k agrupamentos vinculando todos os objetos aos centroides mais próximos.
4. Recalcule o centroide de cada agrupamento.
5. Até que os centroides não mudem.

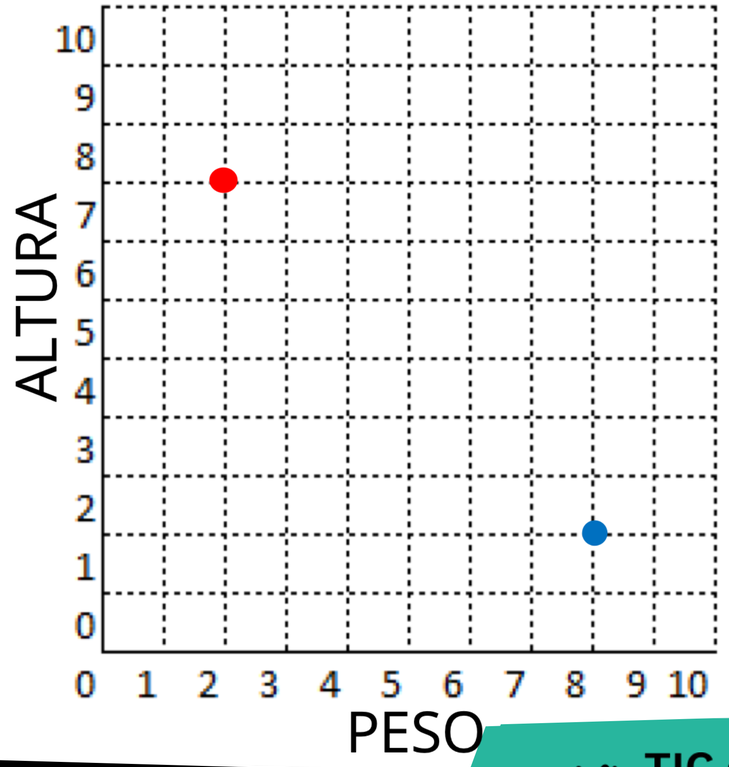
Passo 1: marcar os centroides.

Centroide 1

Centroide 2

Objeto	Peso	Altura
1	2	8
2	8	2
3	6	8
4	2	7
5	8	4
6	2	6

Distancia euclidiana: $\sqrt{(x2-x1)^2+(y2-y1)^2}$



Exemplificando...

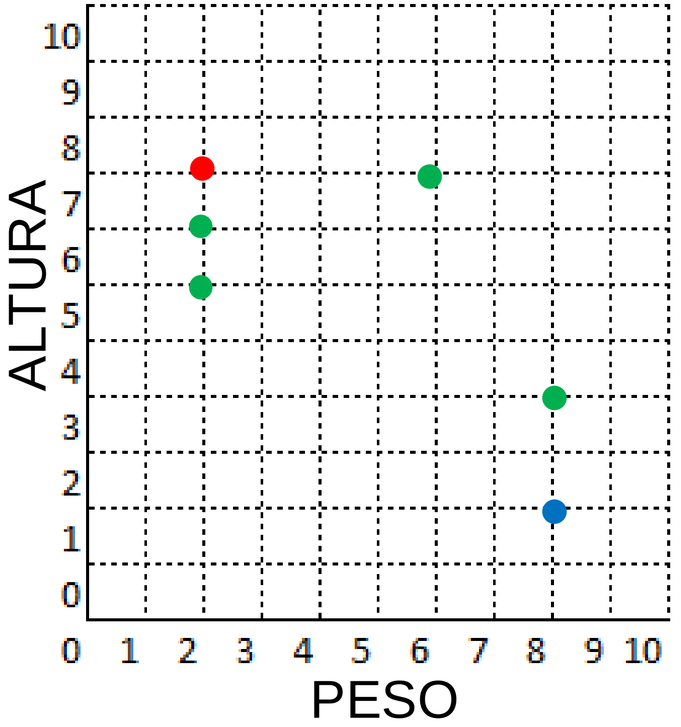
1. Selecione k objetos como centroides iniciais.
2. Repita
3. Forme k agrupamentos vinculando todos os objetos aos centroides mais próximos.
4. Recalcule o centroide de cada agrupamento.
5. Até que os centroides não mudem.

Passo 2: marcar os demais objetos.

Centroide 1

Centroide 2

Objeto	Peso	Altura
1	2	8
2	8	2
3	6	8
4	2	7
5	8	4
6	2	6



Distancia euclidiana: $\sqrt{(x2-x1)^2+(y2-y1)^2}$

Exemplificando...

1. Selecione k objetos como centroides iniciais.
2. Repita
3. Forme k agrupamentos vinculando todos os objetos aos centroides mais próximos.
4. Recalcule o centroide de cada agrupamento.
5. Até que os centroides não mudem.

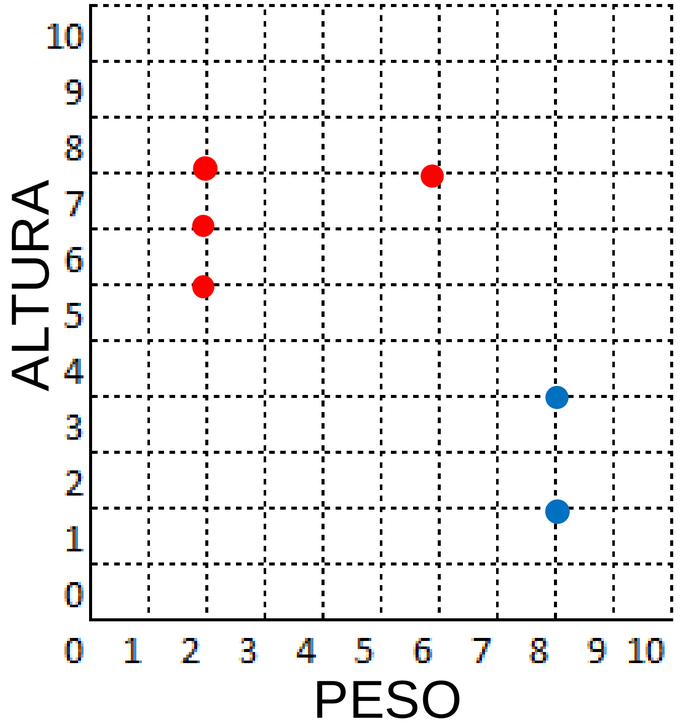
Passo 3: atribuir cada pronto ao centroide mais próximo.

Centroide 1

Centroide 2

Objeto	Peso	Altura
1	2	8
2	8	2
3	6	8
4	2	7
5	8	4
6	2	6

Distancia euclidiana: $\sqrt{(x2-x1)^2+(y2-y1)^2}$



TIC em trilhas

Exemplificando...

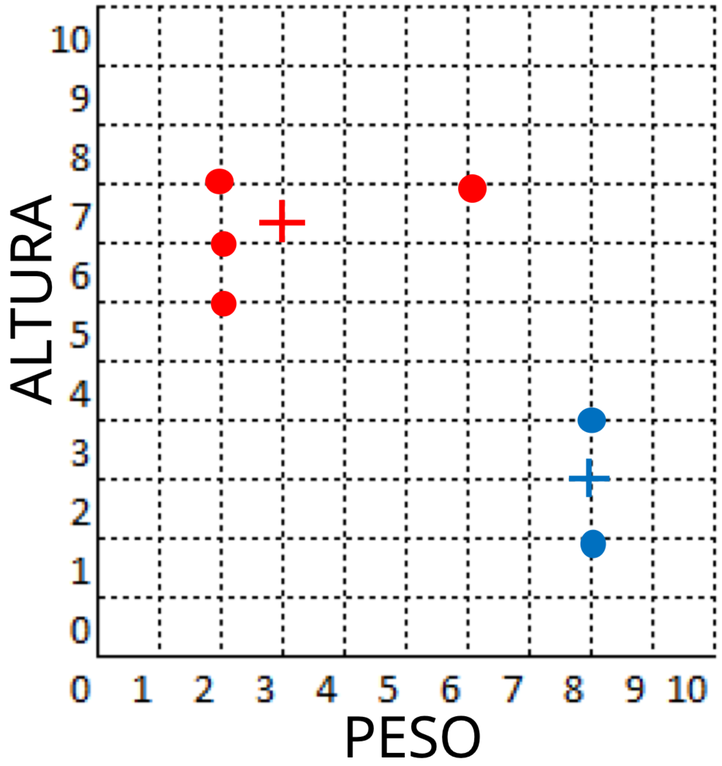
1. Selecione k objetos como centroides iniciais.
2. Repita
3. Forme k agrupamentos vinculando todos os objetos aos centroides mais próximos.
4. Recalcule o centroide de cada agrupamento.
5. Até que os centroides não mudem.

Retorna ao passo 3: marcar os demais objetos.

Centroide 1

Centroide 2

Objeto	Peso	Altura
1	2	8
2	8	2
3	6	8
4	2	7
5	8	4
6	2	6



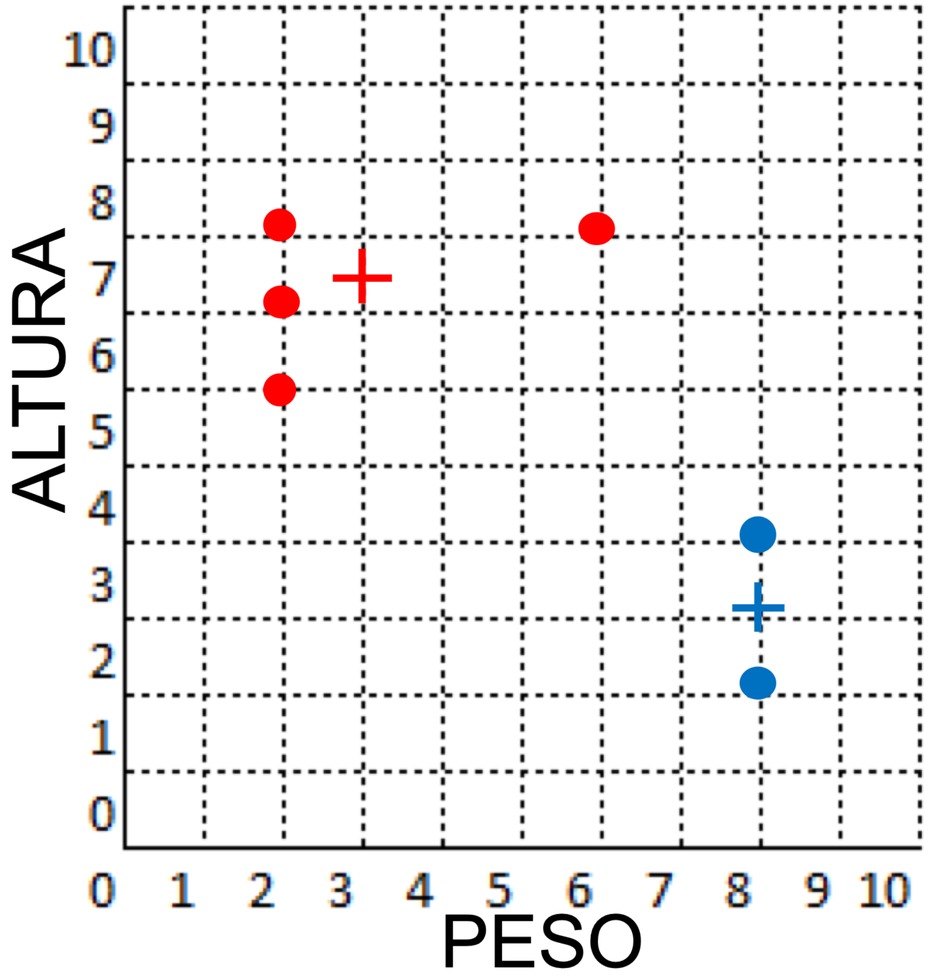
Exemplificando...

Passo 4: recalcular o centroide de cada agrupamento.

Agrupamento 1:
PESO = (2 + 2 + 2 + 6) / 4 = 3
ALTURA = (8 + 7 + 6 + 8) / 4 = 7,25

Agrupamento 2:
PESO = (8 + 8) / 2 = 8
ALTURA = (2 + 4) / 2 = 3

Novos centroides:
Agrupamento 1: (3; 7,25)
Agrupamento 2: (8; 3)



Exemplificando...

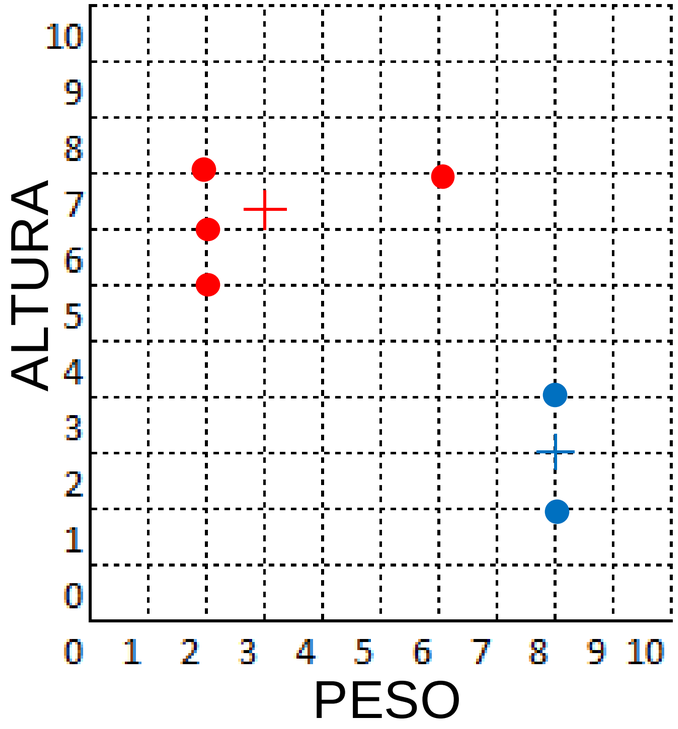
1. Selecione k objetos como centroides iniciais.
2. Repita
3 Forme k agrupamentos vinculando todos os objetos aos centroides mais próximos.
4. Recalcule o centroide de cada agrupamento.
5. Até que os centroides não mudem.

Retorna ao passo 3: atribuir cada objeto ao centroide mais próximo.

Centroide 1

Centroide 2

Objeto	Peso	Altura
1	2	8
2	8	2
3	6	8
4	2	7
5	8	4
6	2	6



Exemplificando...

Passo 4: recalcular centroides.

Neste caso, na 2ª iteração, em diante, os centroides não mudam porque os objetos não trocam de grupos.

Logo, o algoritmo do k-Means termina e a resposta é:

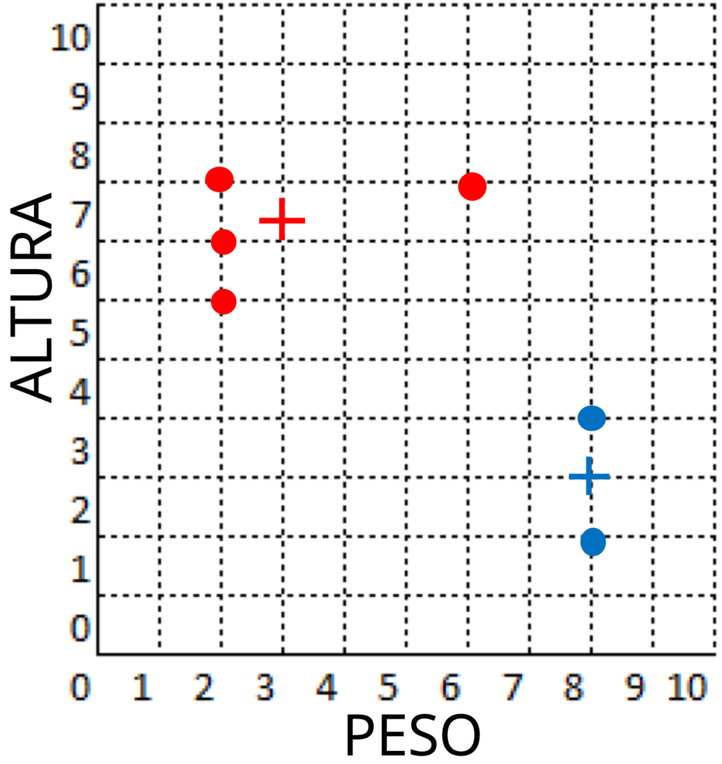
Centroide do Agrupamento 1: (3; 7,25)

Centroide do Agrupamento 2: (8; 3)

Centroide 1

Centroide 2

Objeto	Peso	Altura
1	2	8
2	8	2
3	6	8
4	2	7
5	8	4
6	2	6



Avaliando Agrupamentos

- Como avaliar a qualidade dos resultados do agrupamento?

- Mas os grupos não são subjetivos?

- Então para que avaliá-los?

  Para evitar descobrirmos padrões em ruído.
  Para comparar dois ou mais algoritmos de agrupamento.
  Para comparar dois ou mais resultados de agrupamentos.
  Para comprar dois grupos.

Avaliando Agrupamentos

- Medida mais comum é a **Soma dos Erros Quadráticos (SSE – *sum of squared errors*)**

  - Para cada objeto, o erro é a distância ao grupo mais próximo.
  - Para obter SSE, elevamos os erros ao quadrado e os somamos.

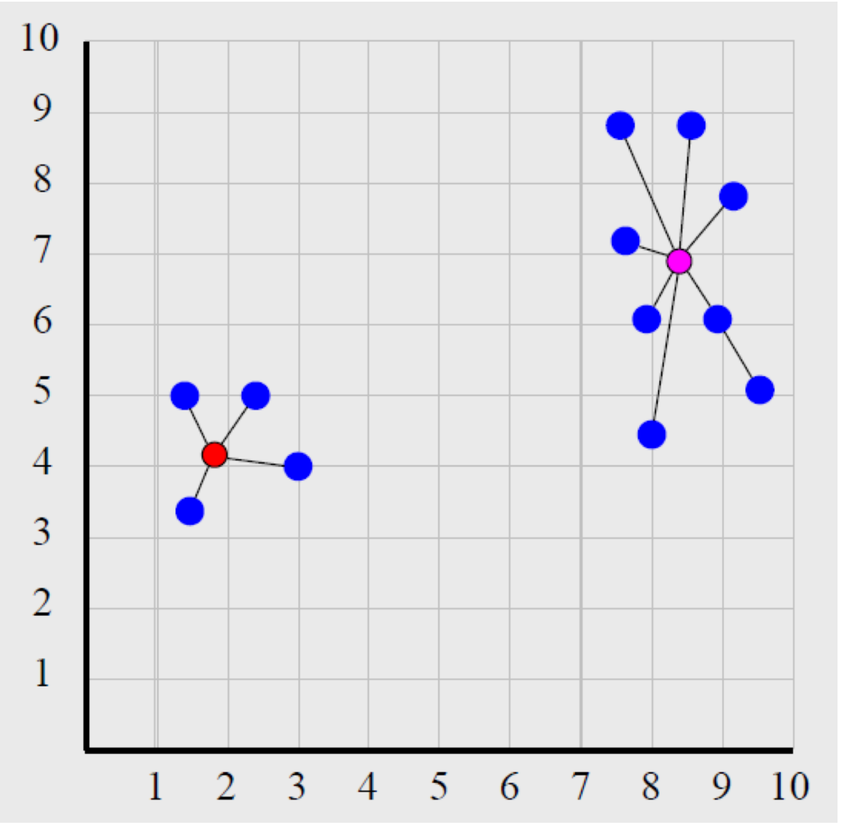
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x é um objeto do grupo C_i e m_i é o centro do grupo C_i.
  - Dados dois agrupamentos, podemos escolher o de menor SSE.
    - **Utilize SSE apenas para comparar agrupamentos de mesmo *k*.**
    - **Quanto maior for *k*, menor será o SSE! Por quê?**

Avaliando Agrupamentos

- Quanto menor o SSE, mais compactos (coesos) são os grupos, pois minimizar o SSE significa minimizar a variância intra-grupo.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$



- Note que o quadrado da distância Euclidiana é o cálculo da distância sem a raiz quadrada. Portanto, SSE pode ser interpretado como a soma das distâncias Euclidianas sem raiz de todos os objetos com relação aos seus centroides.

Exercício

- Execute o algoritmo k-means para a base de dados abaixo, considerando k=2 para os centroides abaixo descritos. Ao final, informe os centroides resultantes e o SSE de cada grupo.

	X	Y
1)	1	1
2)	1	2
3)	2	1
4)	3	2
5)	3	4
6)	4	4

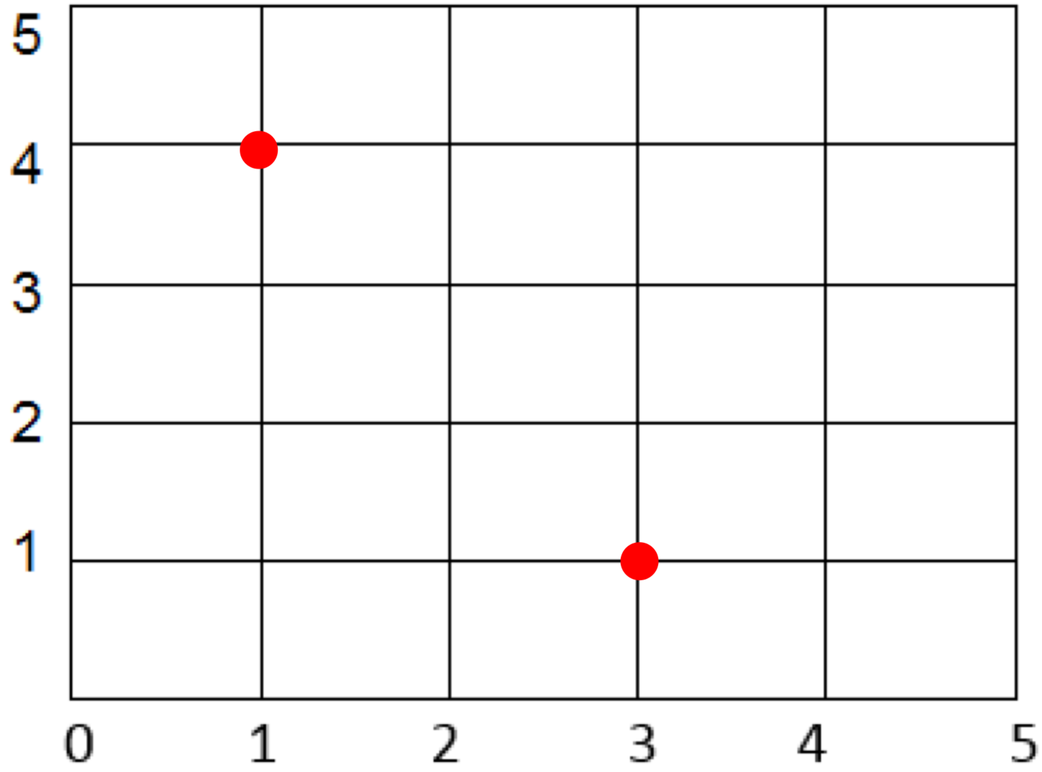
- Centroides Iniciais:
- C_1 (X=3, Y=1)
- C_2 (X=1, Y=4)

Exercício

- Execute o algoritmo k-means para a base de dados abaixo, considerando k=2 para os centroides abaixo descritos. Ao final, informe os centroides resultantes e o SSE de cada grupo.

	X	Y
1)	1	1
2)	1	2
3)	2	1
4)	3	2
5)	3	4
6)	4	4

- Centroides Iniciais:
- C_1 (X=3, Y=1)
- C_2 (X=1, Y=4)

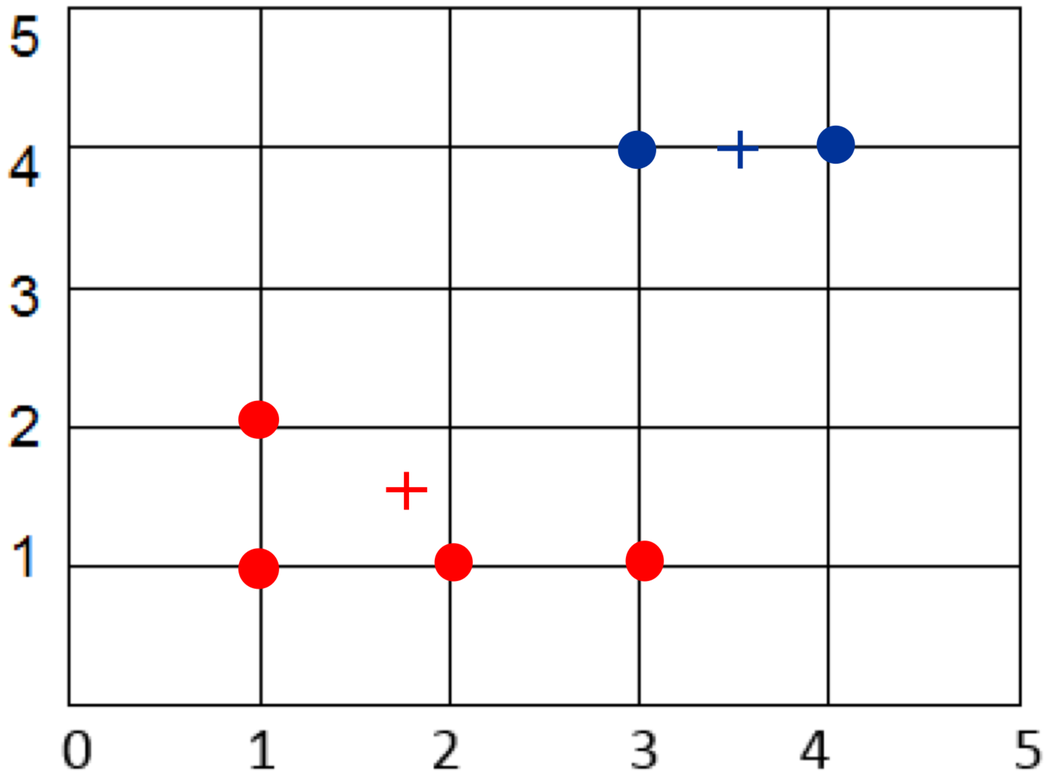


Exercício – Resultados Finais

- Centroides finais:

C_1 (X = 1.75, Y = 1.5)

C_2 (X = 3.50, Y = 4.0) SSE = 3.74 (C_1) + 0.50 (C_2) = 4,24



Prática

- Dataset: movies
  - Encontrar possíveis recomendações de filmes a partir do agrupamento de gêneros.
- Fragmento do DataSet

```
movieId,title,genres
1,Toy Story (1995),Adventure|Animation|Children|Comedy|Fantasy
2,Jumanji (1995),Adventure|Children|Fantasy
3,Grumpier Old Men (1995),Comedy|Romance
4,Waiting to Exhale (1995),Comedy|Drama|Romance
5,Father of the Bride Part II (1995),Comedy
6,Heat (1995),Action|Crime|Thriller
7,Sabrina (1995),Comedy|Romance
8,Tom and Huck (1995),Adventure|Children
9,Sudden Death (1995),Action
10,GoldenEye (1995),Action|Adventure|Thriller
11,"American President, The (1995)",Comedy|Drama|Romance
12,Dracula: Dead and Loving It (1995),Comedy|Horror
13,Balto (1995),Adventure|Animation|Children
14,Nixon (1995),Drama
15,Cutthroat Island (1995),Action|Adventure|Romance
16,Casino (1995),Crime|Drama
17,Sense and Sensibility (1995),Drama|Romance
18,Four Rooms (1995),Comedy
19,Ace Ventura: When Nature Calls (1995),Comedy
```

Referências

- Adaptação dos slides de Pang-Ning Tan
  - Michigan State University
  - http://www.cse.msu.edu/~ptan/
  - ptan@cse.msu.edu
- Adaptação dos slides de Eamon Keogh
  - University of California at Riverside
  - http://www.cs.ucr.edu/~eamonn/
  - eamonn@cs.ucr.edu

ELDORADO

Trilha Ciência de dados com Python

Aula 16

TIC em trilhas