Algoritmos de Clusterização: PAM

Prof. Mateus Mendelson
mendelson.mateus@gmail.com

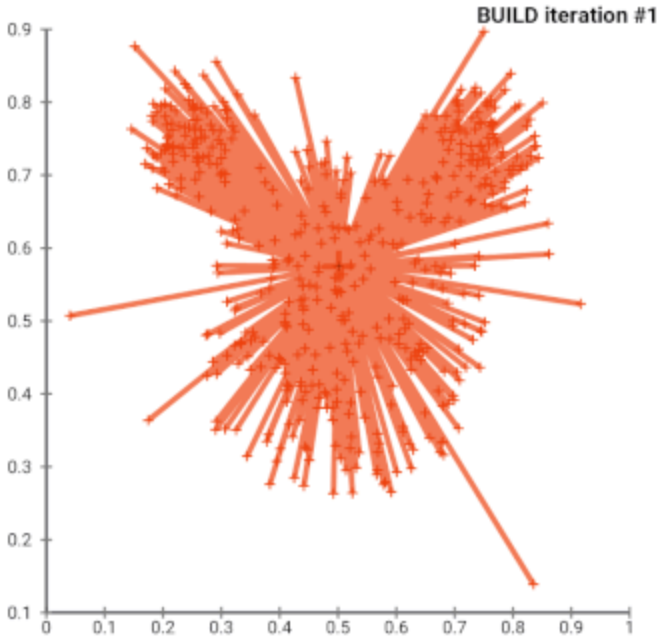
mmendelson.com

1. Introdução

- PAM: Partitioning Around Medoids.
- Esse algoritmo também é conhecido como K-Medoids.
- É um algoritmo de clusterização que relembra o K-Means.
- Ambos separam os dados em grupos e buscam minimizar a distância entre os pontos e seu respectivo centro.
- A grande diferença entre PAM e K-Means é que o PAM escolhe pontos pertencentes aos dados para serem os centros (medoids).
- A vantagem do PAM sobre o K-Means é que o PAM é menos sensível a outliers.

1. Introdução

- O PAM não garante encontrar a solução ótima.
- Também é um algoritmo de rápida execução.



Fonte: https://en.m.wikipedia.org/wiki/K-medoids

2. O algoritmo PAM

1. Definir a quantidade K de clusters que queremos calcular

2. O algoritmo PAM

1. Definir a quantidade K de clusters que queremos calcular.
2. Sortear, aleatoriamente, K pontos dos nossos dados como sendo os medoids.

2. O algoritmo PAM

1. Definir a quantidade K de clusters que queremos calcular.
2. Sortear, aleatoriamente, K pontos dos nossos dados como sendo os medoids.
3. Calcular o custo total de acordo com a métrica MSE (Mean Squared Error) para cada cluster.

$$MSE = \frac{1}{N_i} \sum_{j=1}^{N_i} (X_i - X_j)^2 + (Y_i - Y_j)^2$$

Sendo que N_i é a quantidade de pontos atribuídos ao medoid i e (X_i, Y_i) são as coordenadas do medoid i.

2. O algoritmo PAM

1. Definir a quantidade K de clusters que queremos calcular.
2. Sortear, aleatoriamente, K pontos dos nossos dados como sendo os medoids.
3. Calcular o custo total de acordo com a métrica MSE (Mean Squared Error) para cada cluster.
4. Sortear um ponto que não seja medoid, atribuí-lo como o novo medoid no lugar do medoid mais próximo e recalcular o custo total.

2. O algoritmo PAM

1. Definir a quantidade K de clusters que queremos calcular.
2. Sortear, aleatoriamente, K pontos dos nossos dados como sendo os medoids.
3. Calcular o custo total de acordo com a métrica MSE (Mean Squared Error) para cada cluster.
4. Sortear um ponto que não seja medoid, atribuí-lo como o novo medoid no lugar do medoid mais próximo e recalcular o custo total.
5. Caso o novo custo total seja menor do que o anterior, o novo medoid deve ser mantido; caso contrário, devemos retornar ao medoid anterior.

2. O algoritmo PAM

1. Definir a quantidade K de clusters que queremos calcular.
2. Sortear, aleatoriamente, K pontos dos nossos dados como sendo os medoids.
3. Calcular o custo total de acordo com a métrica MSE (Mean Squared Error) para cada cluster.
4. Sortear um ponto que não seja medoid, atribuí-lo como o novo medoid no lugar do medoid mais próximo e recalcular o custo total.
5. Caso o novo custo total seja menor do que o anterior, o novo medoid deve ser mantido; caso contrário, devemos retornar ao medoid anterior.
6. Voltar ao passo 4 até que, após α tentativas consecutivas, o melhor custo total não diminua.

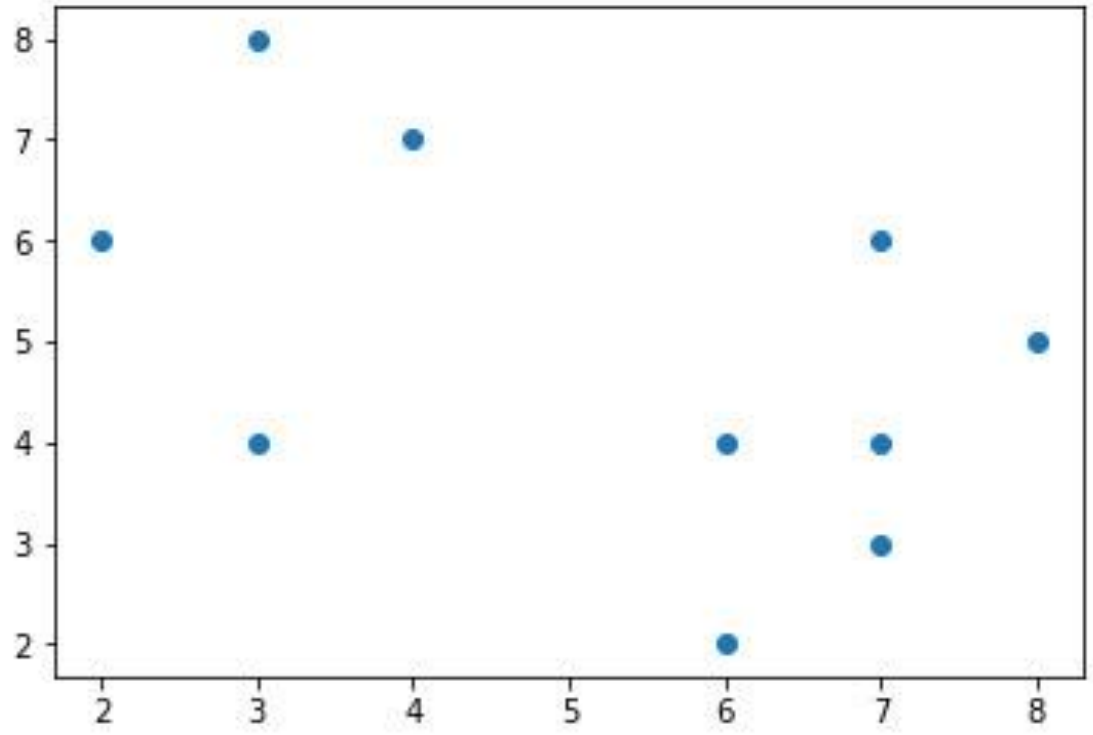
2. O algoritmo PAM

- Considere o conjunto de pontos abaixo.

	X	Y	Custo M1	Custo M2	Medoid associado
0	7	6			
1	2	6			
2	3	8			
3	8	5			
4	7	4			
5	4	7			
6	6	2			
7	7	3			
8	6	4			
9	3	4			

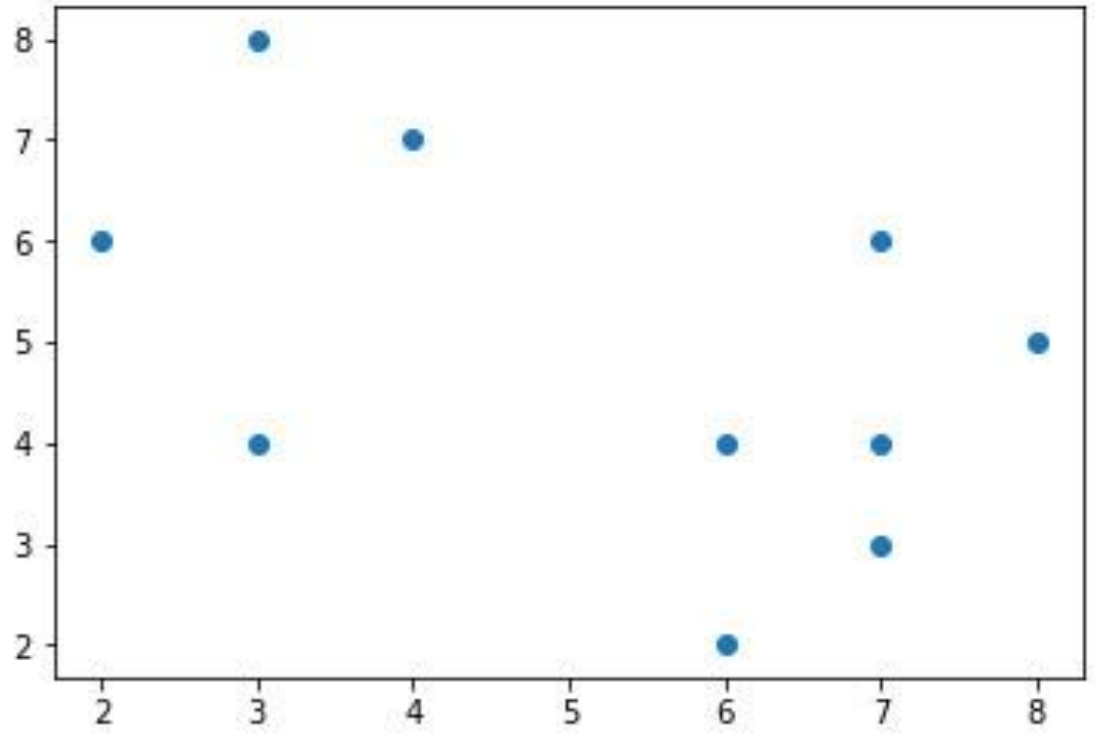
2. O algoritmo PAM

- Considere o conjunto de pontos abaixo.



2. O algoritmo PAM

- Iremos utilizar K = 2, ou seja, sorteamos dois pontos como medoids: M1 (3, 4) e M2 (7, 4).



2. O algoritmo PAM

- Vamos calcular o custo entre cada ponto e cada medoid.

	X	Y	Custo M1	Custo M2	Medoid associado
0	7	6			
1	2	6			
2	3	8			
3	8	5			
4	7	4			
5	4	7			
6	6	2			
7	7	3			
8	6	4			
9	3	4			

2. O algoritmo PAM

- Vamos calcular o custo entre cada ponto e cada medoid.

- Ponto 0: (7, 6)

  ✓ Custo M1 (3, 4):

$(3 - 7)^2 + (4 - 6)^2 = 16 + 4 = 20$

  ✓ Custo M2 (7, 4):

$(7 - 7)^2 + (4 - 6)^2 = 0 + 4 = 4$

2. O algoritmo PAM

- Vamos calcular o custo entre cada ponto e cada medoid.

	X	Y	Custo M1	Custo M2	Medoid associado
0	7	6	20	4	M2
1	2	6			
2	3	8			
3	8	5			
4	7	4			
5	4	7			
6	6	2			
7	7	3			
8	6	4			
9	3	4			

2. O algoritmo PAM

- Vamos calcular o custo entre cada ponto e cada medoid.

	X	Y	Custo M1	Custo M2	Medoid associado
0	7	6	20	4	M2
1	2	6	5	29	M1
2	3	8	16	32	M1
3	8	5	26	2	M2
4	7	4	---	---	---
5	4	7	10	18	M1
6	6	2	13	5	M2
7	7	3	17	1	M2
8	6	4	9	1	M2
9	3	4	---	---	---

2. O algoritmo PAM

- Vamos calcular o custo total (MSE) para cada medoid:
  - ✓ M1: (5 + 16 + 10)/3 = 10.33
  - ✓ M2: (4 + 2 + 5 + 1 + 1)/5 = 2.6
  - ✓ No fim das contas, o custo total médio é de (10.33 + 2.6)/2 = **6.465**

2. O algoritmo PAM

- Agora, iremos sortear um ponto que não é medoid para substituir um dos medoids existentes.
- O ponto sorteador é (7, 3).
- Por estar mais próximo do medoid M2, ele irá o substituir nessa rodada.

2. O algoritmo PAM

- Vamos calcular o custo entre cada ponto e cada medoid.

	X	Y	Custo M1	Custo M2	Medoid associado
0	7	6			
1	2	6			
2	3	8			
3	8	5			
4	7	4			
5	4	7			
6	6	2			
7	7	3			
8	6	4			
9	3	4			

2. O algoritmo PAM

- Vamos calcular o custo entre cada ponto e cada medoid.

- Ponto 0: (7, 6)

  ✓ Custo M1 (3, 4):

$(3 - 7)^2 + (4 - 6)^2 = 16 + 4 = 20$

  ✓ Custo M2 (7, 3):

$(7 - 7)^2 + (3 - 6)^2 = 0 + 9 = 9$

2. O algoritmo PAM

- Vamos calcular o custo entre cada ponto e cada medoid.

	X	Y	Custo M1	Custo M2	Medoid associado
0	7	6	20	9	M2
1	2	6			
2	3	8			
3	8	5			
4	7	4			
5	4	7			
6	6	2			
7	7	3			
8	6	4			
9	3	4			

2. O algoritmo PAM

- Vamos calcular o custo entre cada ponto e cada medoid.

	X	Y	Custo M1	Custo M2	Medoid associado
0	7	6	20	9	M2
1	2	6	5	34	M1
2	3	8	16	41	M1
3	8	5	26	5	M2
4	7	4	16	1	M2
5	4	7	10	25	M1
6	6	2	13	2	M2
7	7	3	---	---	---
8	6	4	9	2	M2
9	3	4	---	---	---

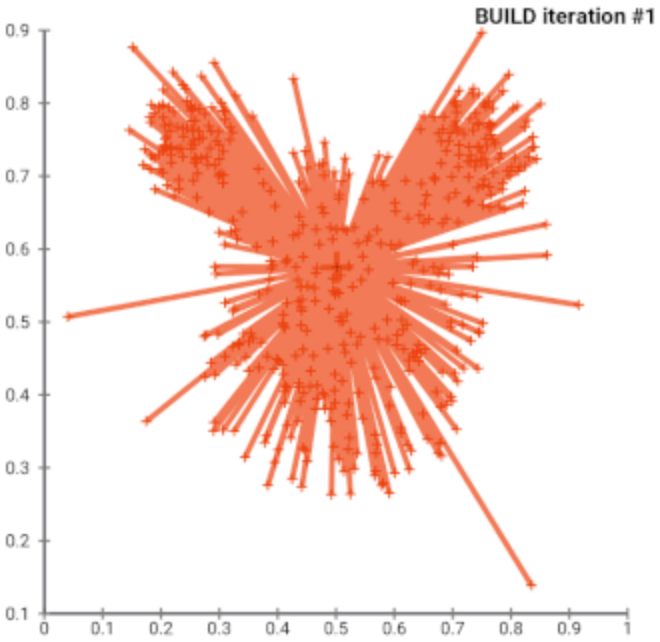
2. O algoritmo PAM

- Vamos calcular o custo total (MSE) para cada medoid:
  - ✓ M1: (5 + 16 + 10)/3 = 10.33
  - ✓ M2: (9 + 5 + 1 + 2 + 2)/5 = 3.8
  - ✓ No fim das contas, o custo total médio é de (10.33 + 3.8)/2 = **7.065**
  - ✓ Como esse valor é maior do que o custo total obtido com os medoids anteriores, iremos reverter essa substituição e continuar sorteando outros pontos como medoids (que ainda não tenham sido escolhidos).

2. O algoritmo PAM

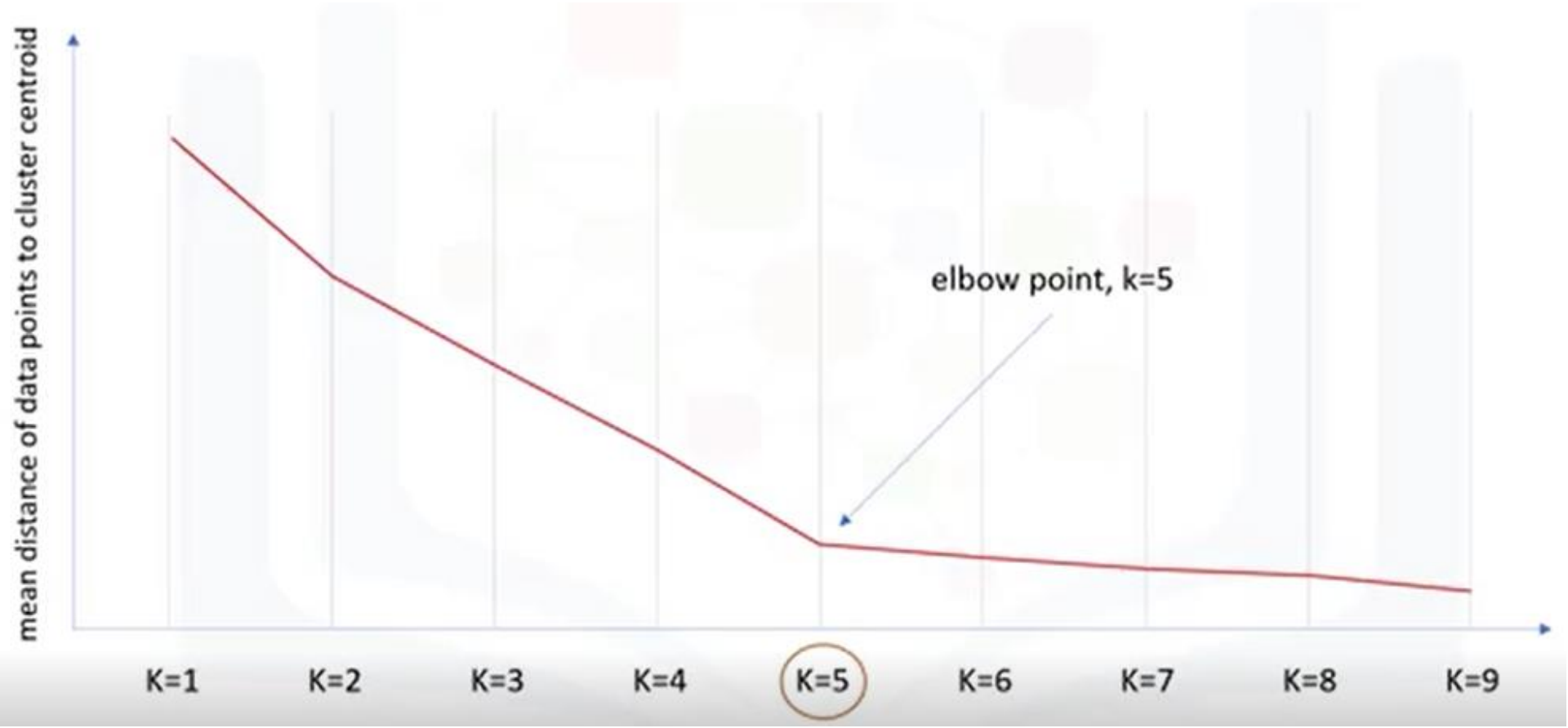
- Continuamos com esse processo até que uma quantidade de tentativas α tenham sido realizadas sem melhoria no custo total médio ou até que todos os pontos tenham sido testados como medoids.

2. O algoritmo PAM



2. O algoritmo PAM

- Como escolher o melhor K?



Fonte: Coursera Machine Learning with Python course

3. Mini-projeto

• Para este projeto, iremos utilizar o dataset de COVID-19 dos casos nos Estados Unidos no período de 21 de janeiro até 09 de abril de 2020 processado.

• Tarefa: implementar a função “fit_pam(pontos, alpha)”, com K fixo e igual a 3. Retorne os centroids finais.

- ✓ pontos: conjunto de pontos 2D (casos x mortes) que serão clusterizados
- ✓ alpha: valor de α que indica a quantidade de tentativas sem melhoria de custo total médio que devem ser realizadas antes da interrupção do algoritmo
- ✓ Desafio: adicionar parâmetro com a quantidade de medoids K variável

3. Mini-projeto

• O seu relatório será o notebook exportado para um arquivo HTML e deve conter:

- ✓ Um scatter plot mostrando os medoids (com marcador x) e seus respectivos pontos (cada cluster deve estar em uma cor distinta)
- ✓ Para cada cluster, também devem ser exibidos seus custos totais, bem como o custo total médio
- ✓ Discorra sobre cada cluster: o que eles indicam?
- ✓ Desafio: implementar uma visualização iterativa do processo de treinamento igual ao gif do início da aula
- ✓ Desafio: plotar o gráfico que permite visualizar o elbow point, variando o valor de K e indicar qual o melhor valor