Clustering

Luciano Barbosa e Everaldo Neto

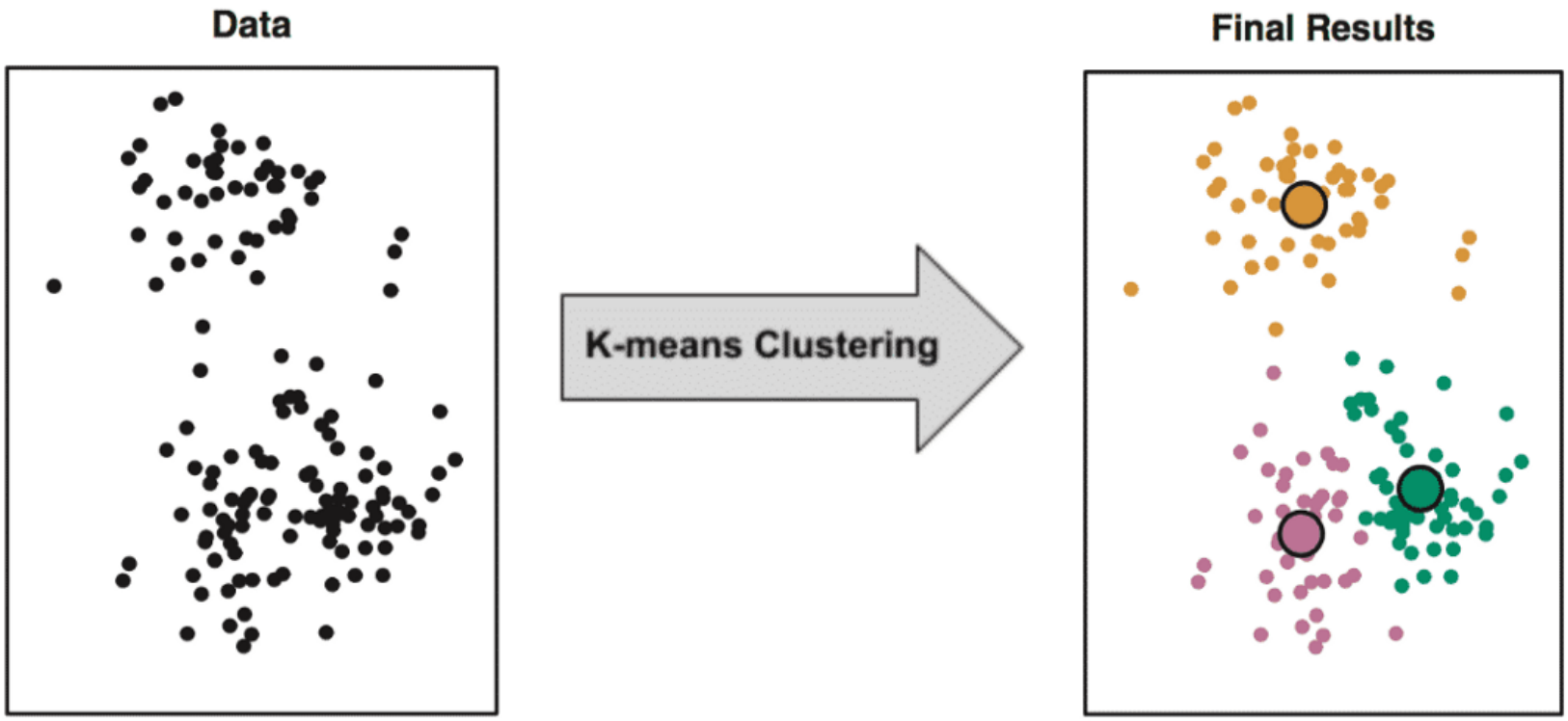
UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Contexto

- No mundo real, nem sempre temos acesso a dados rotulados
- Muitas vezes temos muitos dados e precisamos categorizá-los de alguma forma
- Aprendizado não supervisionado: Clustering

Contexto

- Algoritmos de clustering permitem detectar padrões de forma não supervisionada em conjuntos de dados

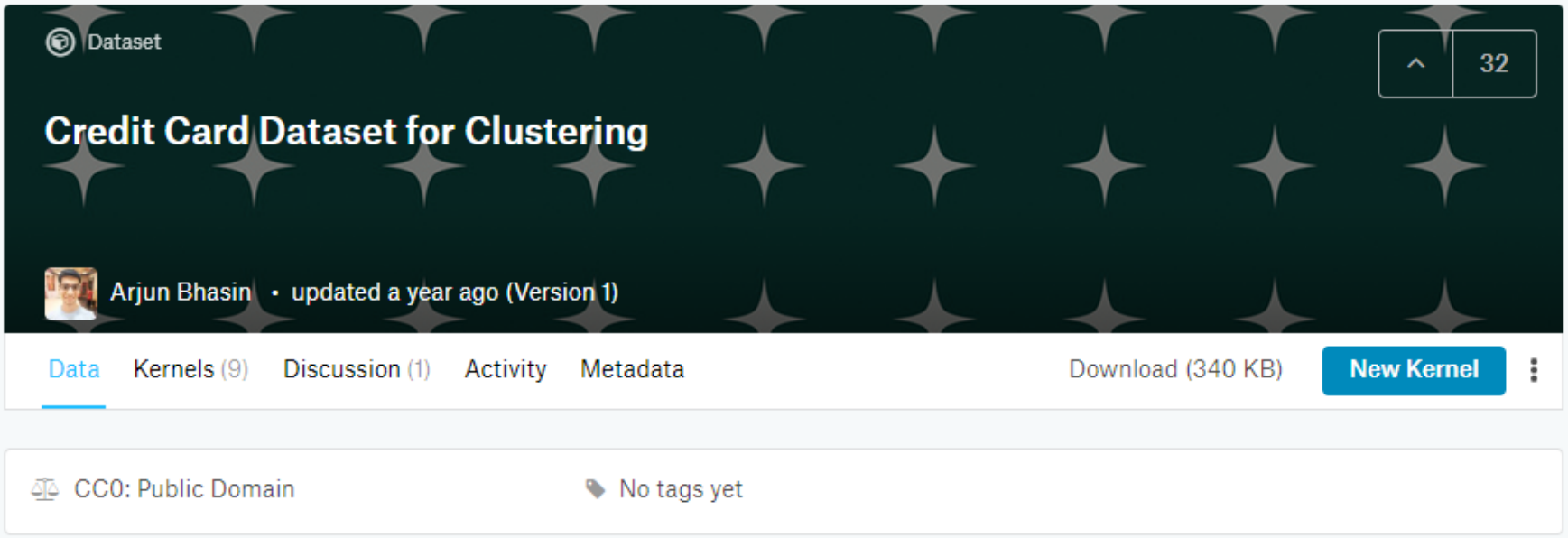


Aplicações

- Marketing – identificar grupos de clientes com perfil de compra similiar
- Recuperação da Informação – agrupar documentos similares para melhorar resultados em engenhos de busca
- Biologia – identificar o grau de semelhança entre as formas ou organismos (filogenética)
- Rotular bases de dados
- (...) entre outras!!!

Competição Kaggle

This case requires to develop a customer segmentation to define marketing strategy. The sample Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables (...)



...resumindo!!!

- Clustering pode ser visto como a tarefa de separar objetos em grupos
  - baseia-se nas *características* que estes objetos possuem
  - o agrupamento dos objetos é feito de acordo com algum *critério similaridade/distância*

Desafio

- Definir a noção do que constitui um cluster
  - *melhor definição depende da natureza dos dados e dos resultados desejados*.

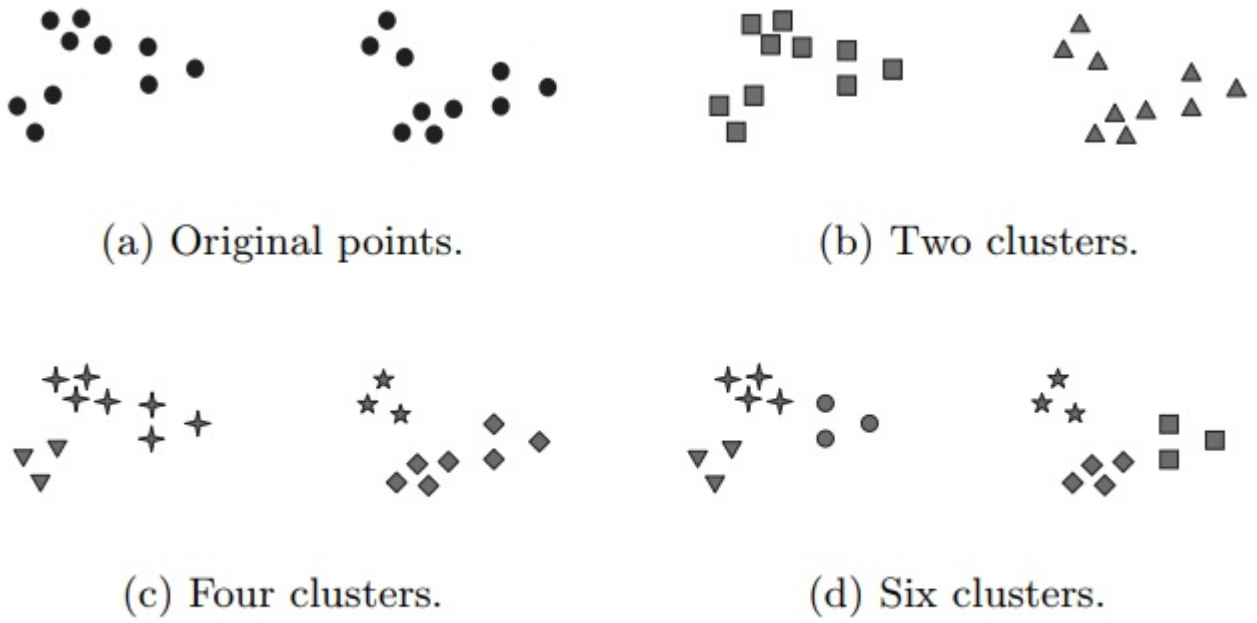
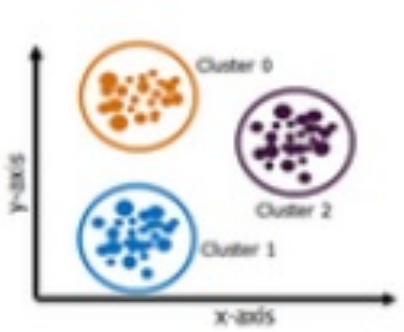
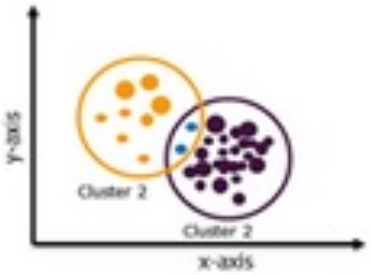


Figure 7.1. Three different ways of clustering the same set of points.

Tipos de clusters



Exclusivos



Sobrepostos

Distância e Similaridade

- Utilizada agrupar objetos
- Diferentes estratégias
  - Euclidiana, Cosseno, Manhattan, Jaccard, ...

Distância

- Medida numérica de quanto dois objetos são diferentes
- Propriedades desejadas:
  1. d(p,p) = 0 (distância mínima)
  2. d(p,q) = d(q,p) para todo p e q (simetria)
- Tipos:

$$L_p(x,y) = [|x_1 - y_1|^p + \cdots + |x_d - y_d|^p]^{1/p}$$ **Minkowski**

$$L_2(x,y) = \sqrt{|x_1 - y_1|^2 + \cdots + |x_d - y_d|^2}$$ **Euclidiana**

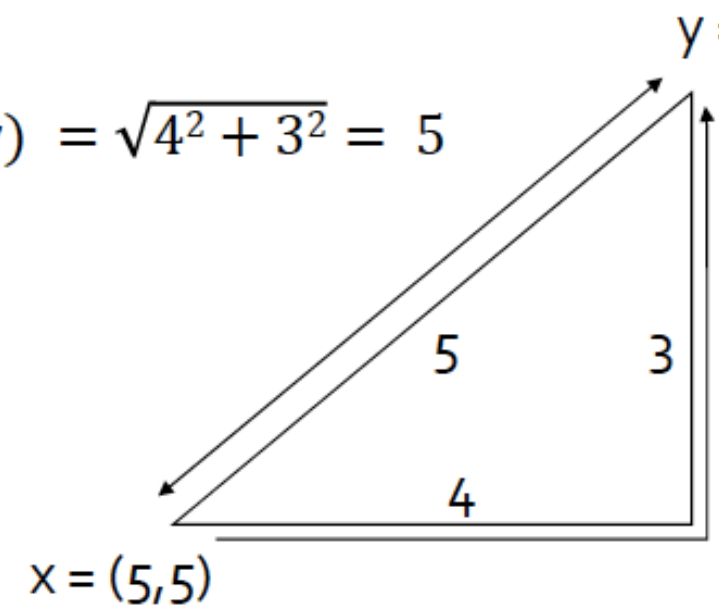
$$L_1(x,y) = |x_1 - y_1| + \cdots + |x_d - y_d|$$ **Manhattan**

$$L_\infty(x,y) = \max\{|x_1 - y_1|, \ldots, |x_d - y_d|\}$$ **Chebyshev**

Exemplo de Distâncias

L_2-norm:

$$dist(x,y) = \sqrt{4^2 + 3^2} = 5$$



L_1-norm:

$$dist(x,y) = 4 + 3 = 7$$

L_∞-norm:

$$dist(x,y) = \max\{3,4\} = 4$$

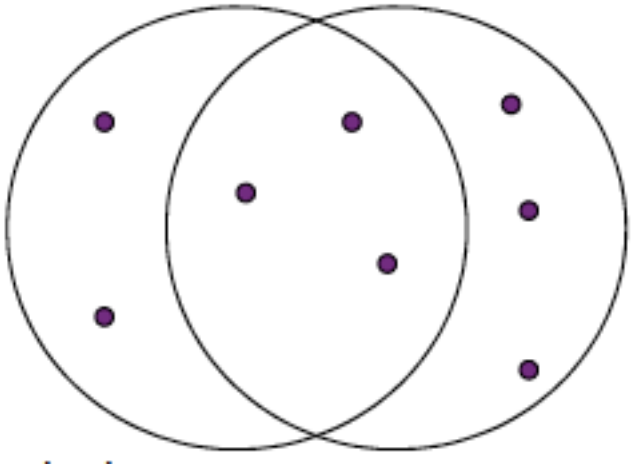
Similaridade

- Medida numérica de quanto dois objetos são parecidos
  - Função que mapeia dois objetos para um número real
  - Usualmente em intervalos [0,1] ou [-1,1]
- Propriedades desejadas:
  1. s(p,p) = 1 (similaridade máxima)
  2. s(p,q) = s(q,p) para todo p e q (simetria)

Similaridade de Jaccard

- Definição: tamanho da intersecção dividido pela união

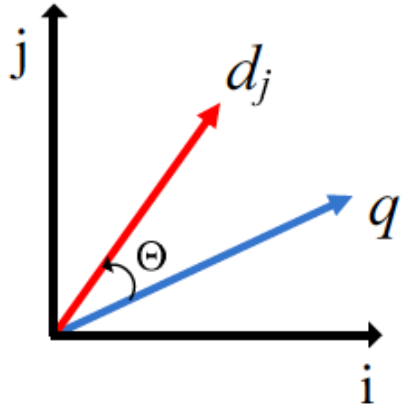
$$JSim(C_1, C_2) = |C_1 \cap C_2| / |C_1 \cup C_2|.$$



- Exemplo: Jsim(C_1,C_2) = 3/8

Similaridade de Cosseno

- Cosseno do ângulo entre 2 vetores no espaço de características



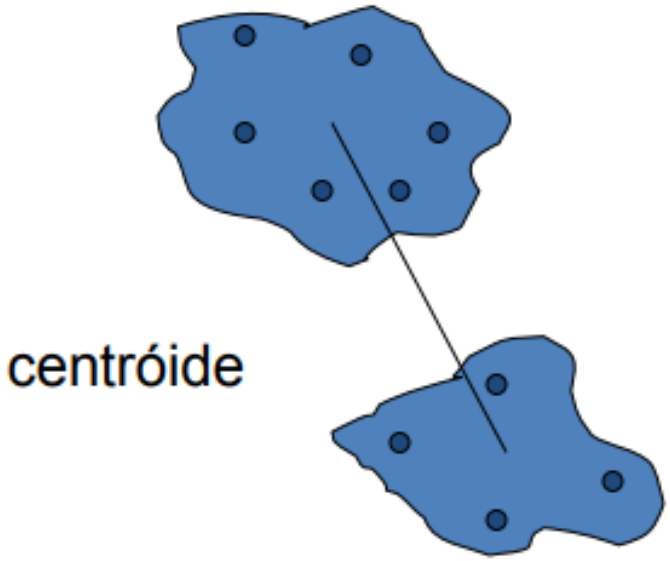
$$\vec{d_j} = (w_{1j}, w_{2j}, \ldots, w_{tj})$$
$$\vec{q} = (w_{1q}, w_{2q}, \ldots, w_{tq})$$

$$cos(\theta) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|}$$

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$$

K-means

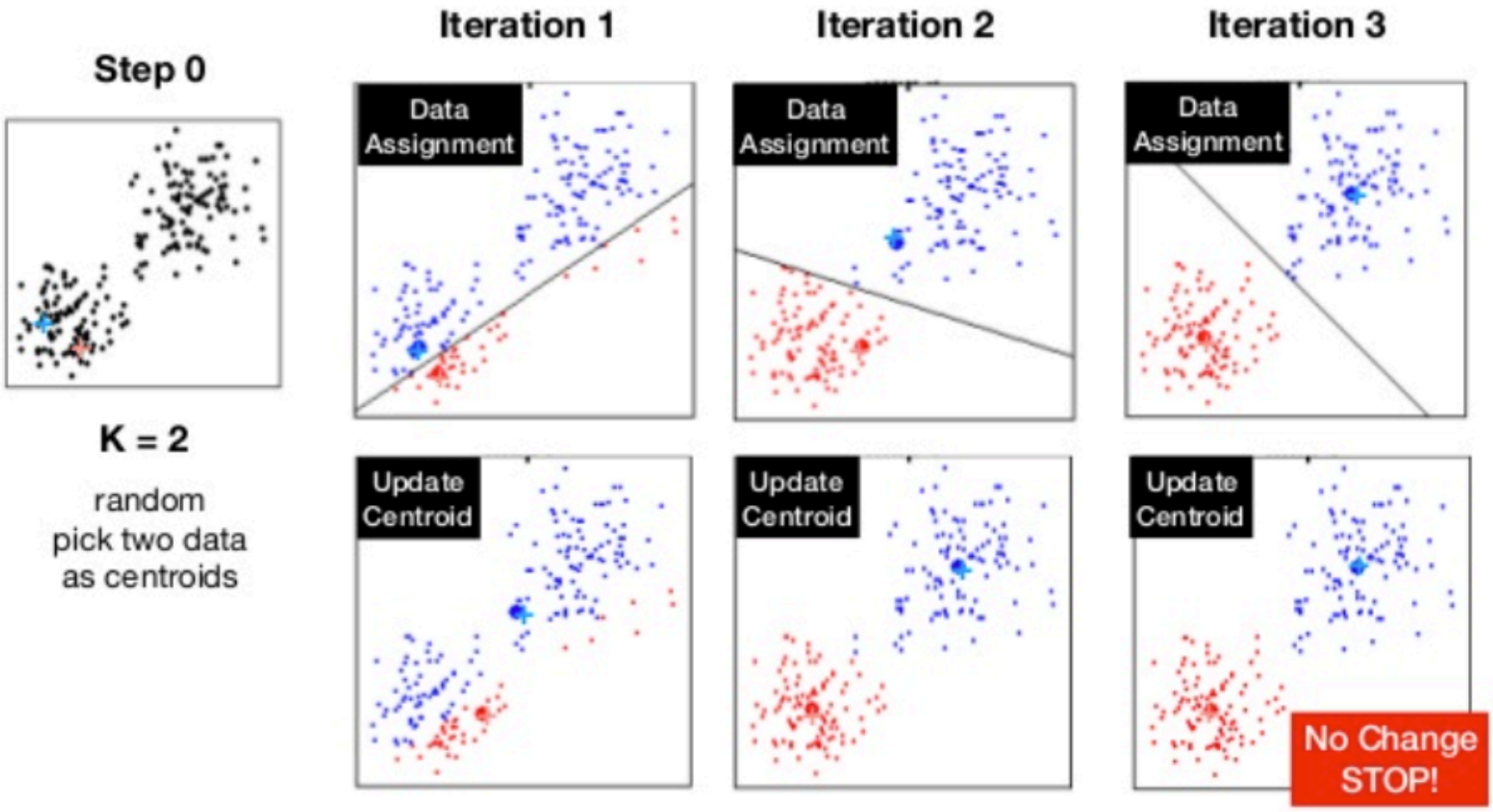
- Divide o conjunto de dados em k partições
- Utiliza o conceito de centróide: vetor médio do grupo



Algoritmo K-means

1. Seleciona k objetos aleatórios como centróide de caga grupo
2. Calcula a distância de cada instância para os k centróides e a atribue ao grupo com menor distância
3. Centróide dos grupos é recalculado
4. Repita 1 a 3 até convergência

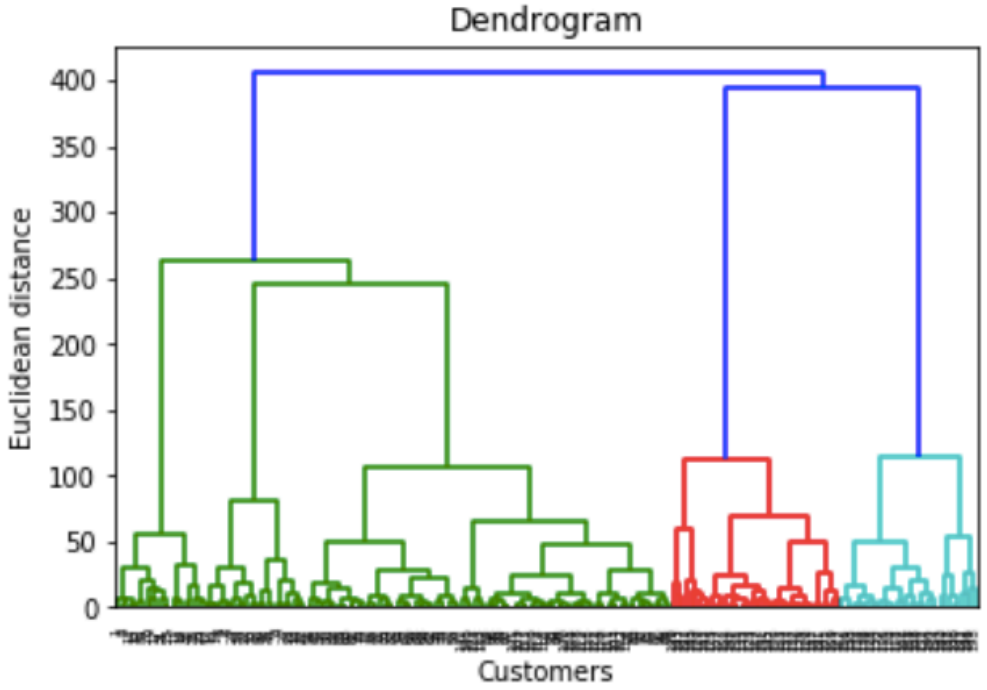
Etapas do K-Means



Fonte: https://www.slideshare.net/radiohead0401/cluster-analysis-assignment-update

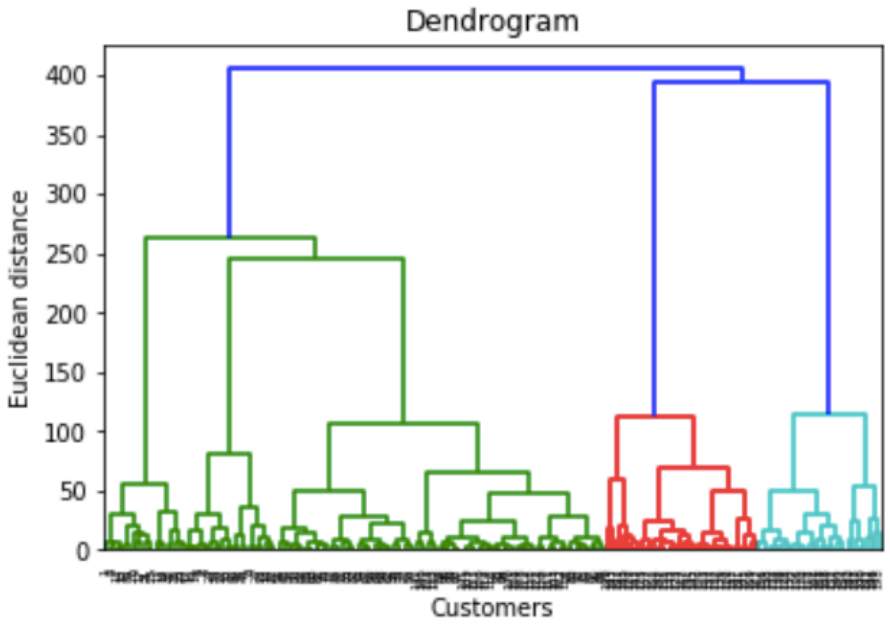
Hierarchical Agglomerative Clustering - HAC

- Algoritmo aglomerativo (bottom-up)
- Não requer definir o número de clusters
- Gera um dendograma que permite analisar proximidade entre os objetos



Algoritmo HAC

1. Calcule a similaridade entre todos as instâncias
2. Cada instância é um cluster
3. Encontre o par mais similar
4. Junte o par em um elemento
5. Repetir passos 1 a 4 até formar um único grupo ou ter k grupos

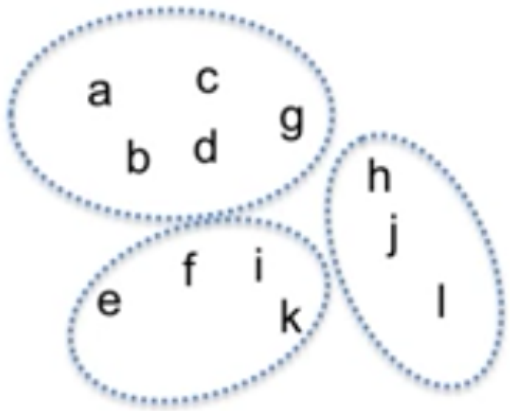


Métricas de avaliação

- É esperado que os clusters gerados sejam homogêneos e isolados
- Maximizar similaridade dentro do grupo e minimizar similaridade intergrupo
- Considera o rótulo dos dados
  - Rand Index
  - V-score ou V-Measure
- Não considera rótulos
  - Silhouete Score

Rand Index

- Similar a acurácia em classificação
- Pares de instâncias rotuladas
  - Ex: a,b=sim
    c,d=não
    e,h=sim
    g,h=não



- TP: pares do mesmo grupo que foram clusterizados no mesmo grupo
- TN: pares de grupos diferentes que foram clusterizados em grupos diferentes
- Rand Index = TP+TN/Total

V-score

- Similar a F-score
- Homogeneity - verifica se todos os objetos de um cluster pertence a mesma classe (rótulo)
- Completeness - verifica se todos os objetos de uma classe estão no mesmo cluster
- V-score = 2 * (homogeneity * completeness) / (homogeneity + completeness)
- Varia entre 0 e 1

Silhouete Score

- Quão similar está uma instância do seu cluster (coesão) comparada a os outros clusters (separação)
- Varia entre -1 e 1: quanto maior o valor mais separado o ponto está

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

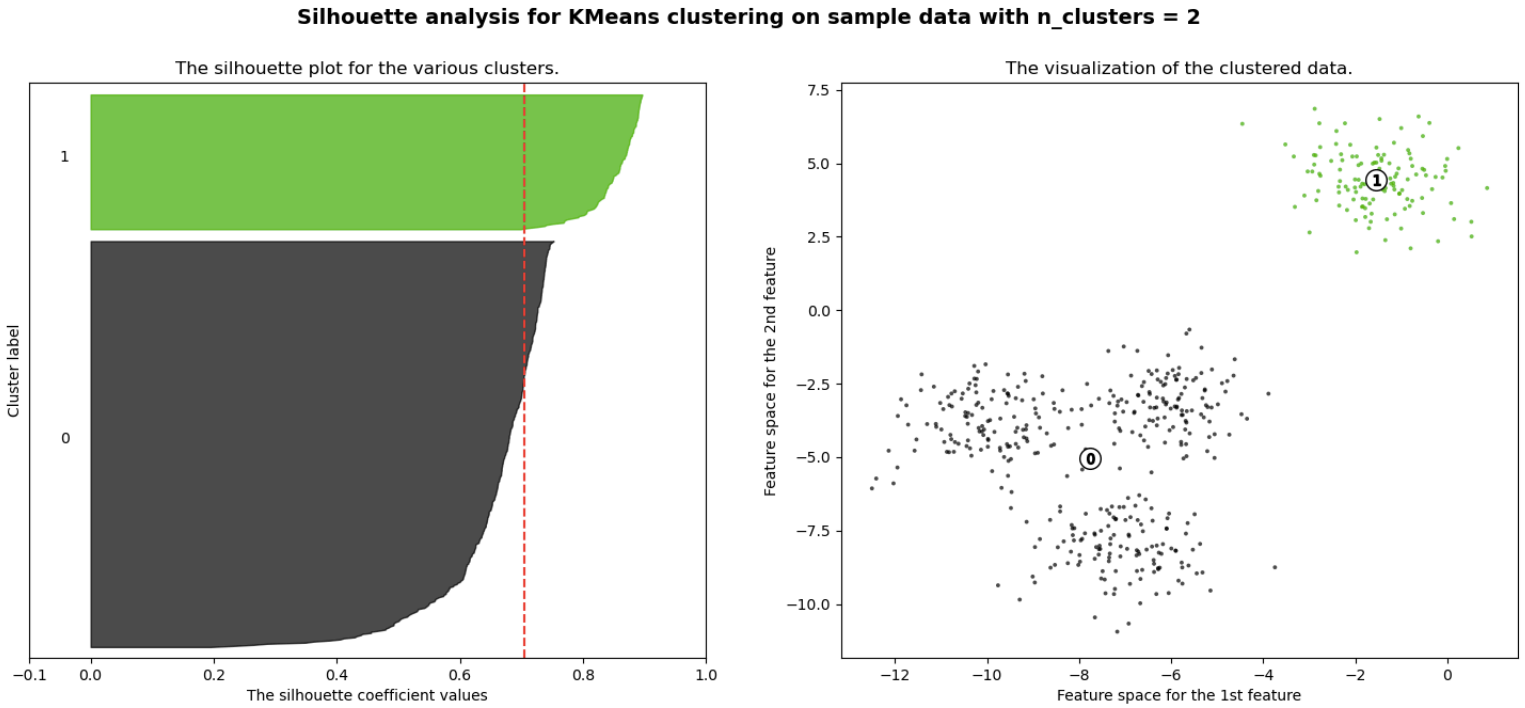
$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

Distância média para as instâncias do mesmo cluster

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

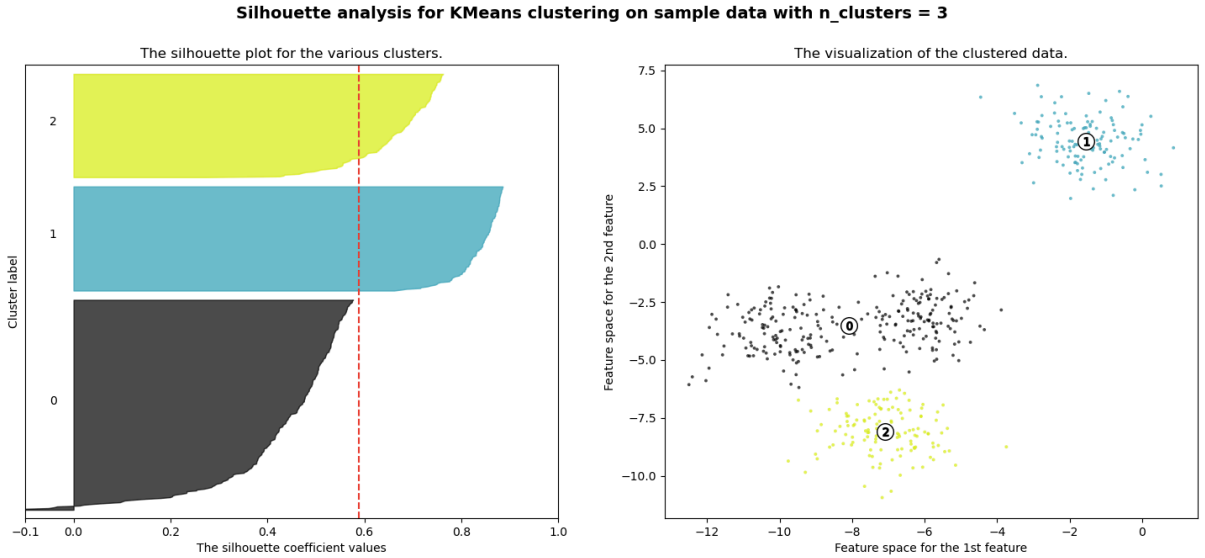
Menor distância média para todas as instâncias nos outros clusters

Silhouete Analysis

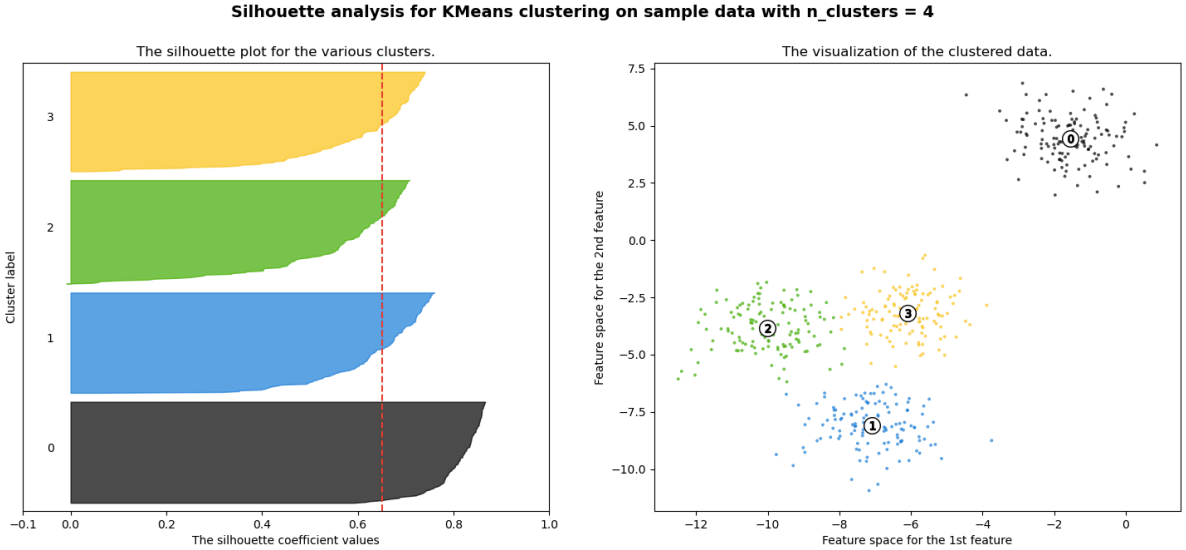


Score = 0.7

Silhouete Analysis

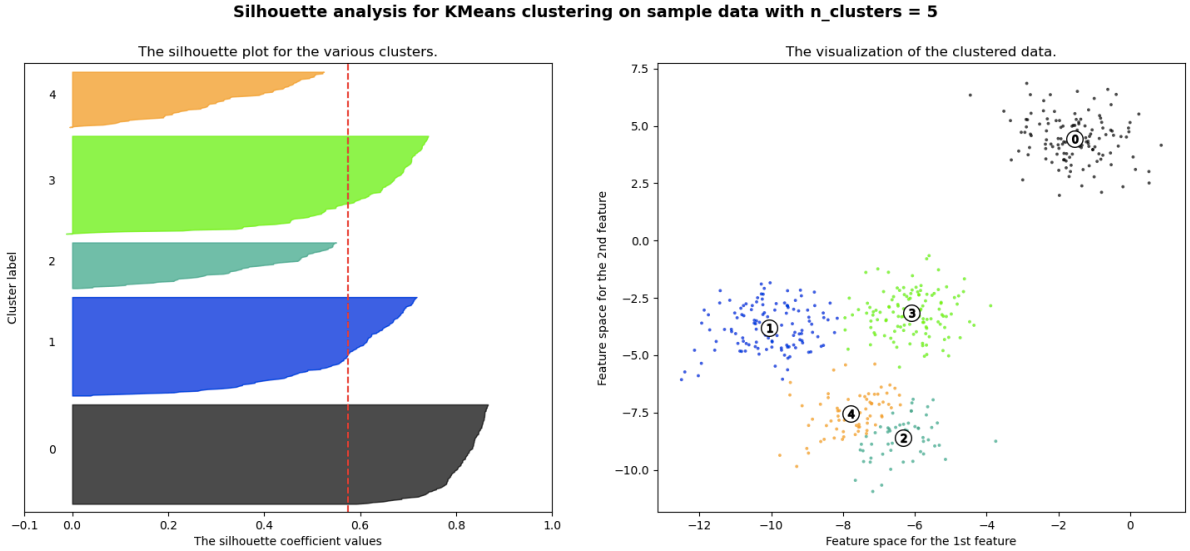


Score = 0.588

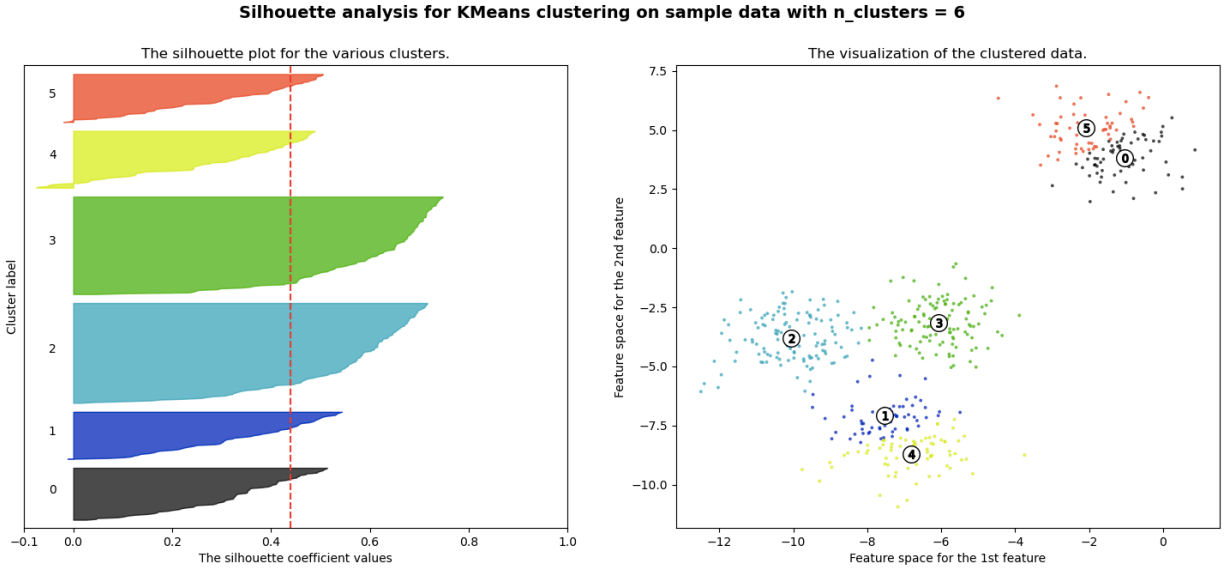


Score = 0.65

Silhouete Analysis



Score = 0.57



Score = 0.43