Clustering

Everaldo Neto

...nas últimas aulas

- Técnicas e modelos de aprendizagem supervisionada
  - classificação, regressão...

...nesta aula

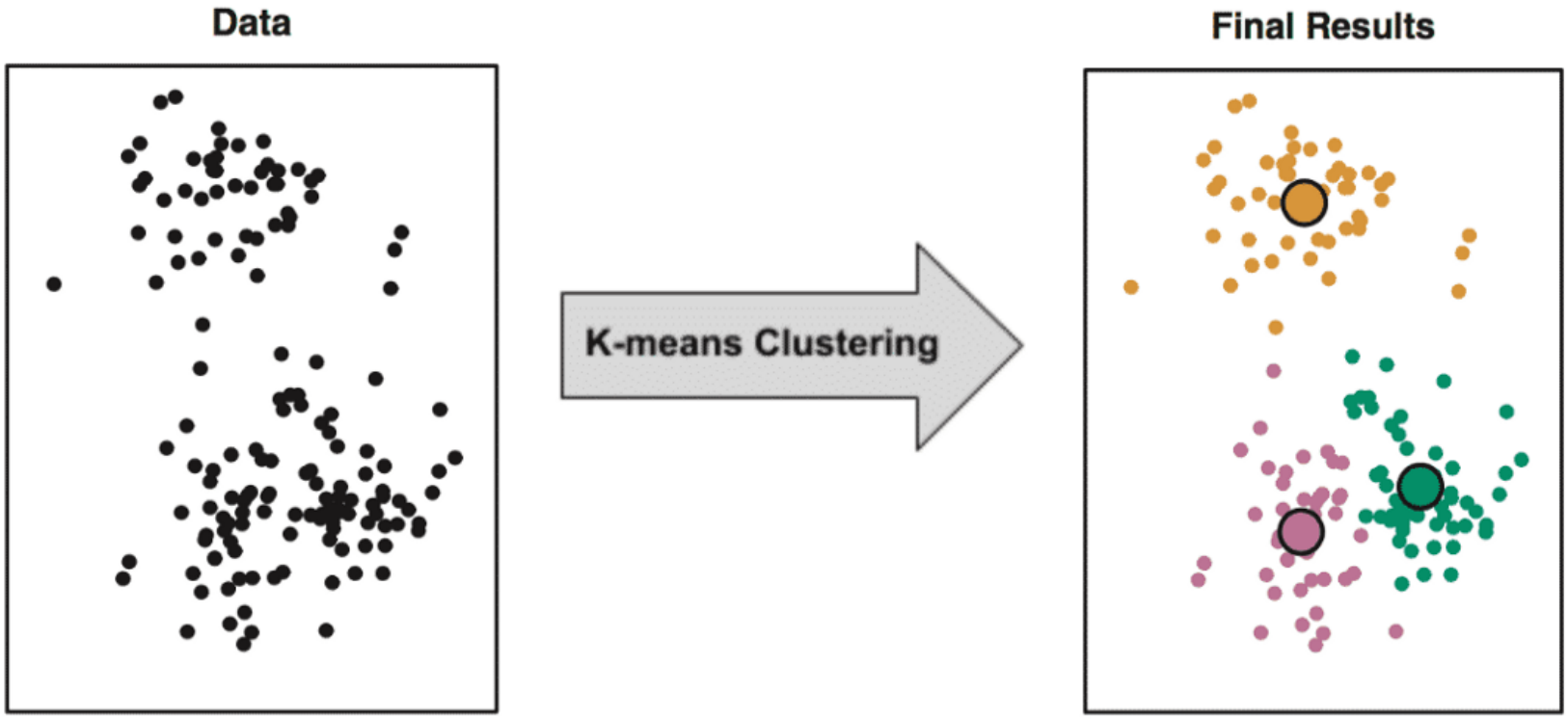
- Discutir técnicas e algoritmos de clustering
- Métricas de avaliação de clustering
- Resumo e códigos disponíveis em: https://github.com/everaldocsneto/aulas

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

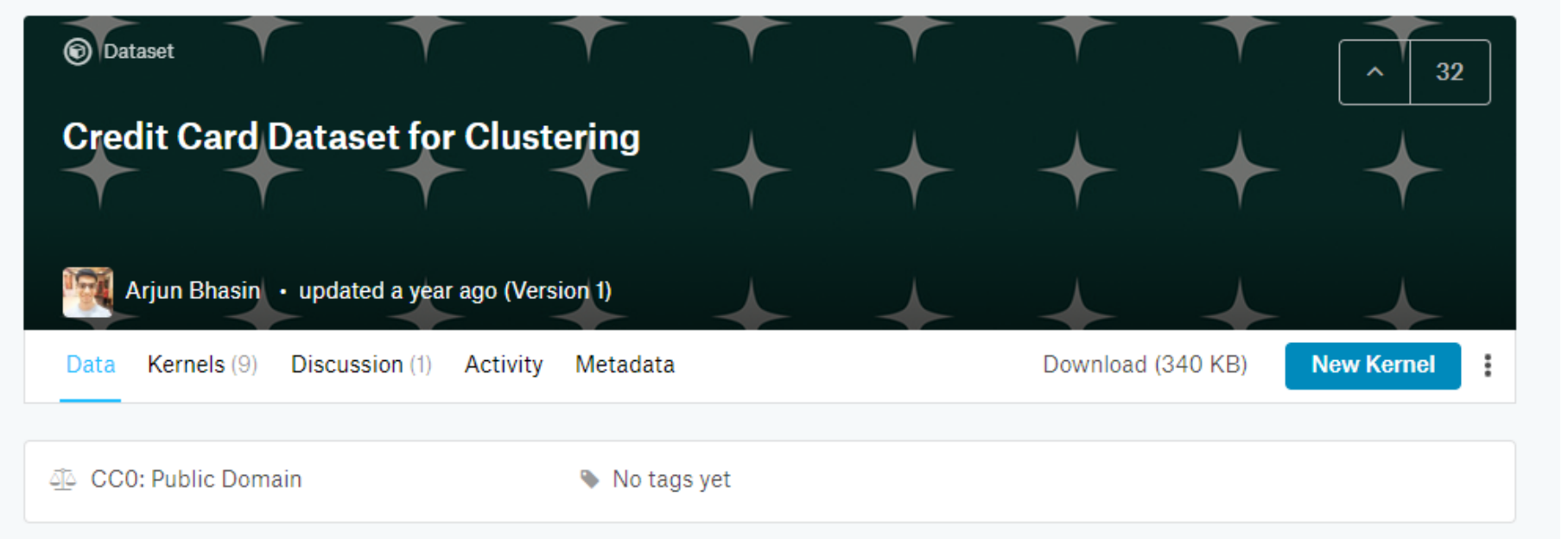


Aplicações

- 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 (...)

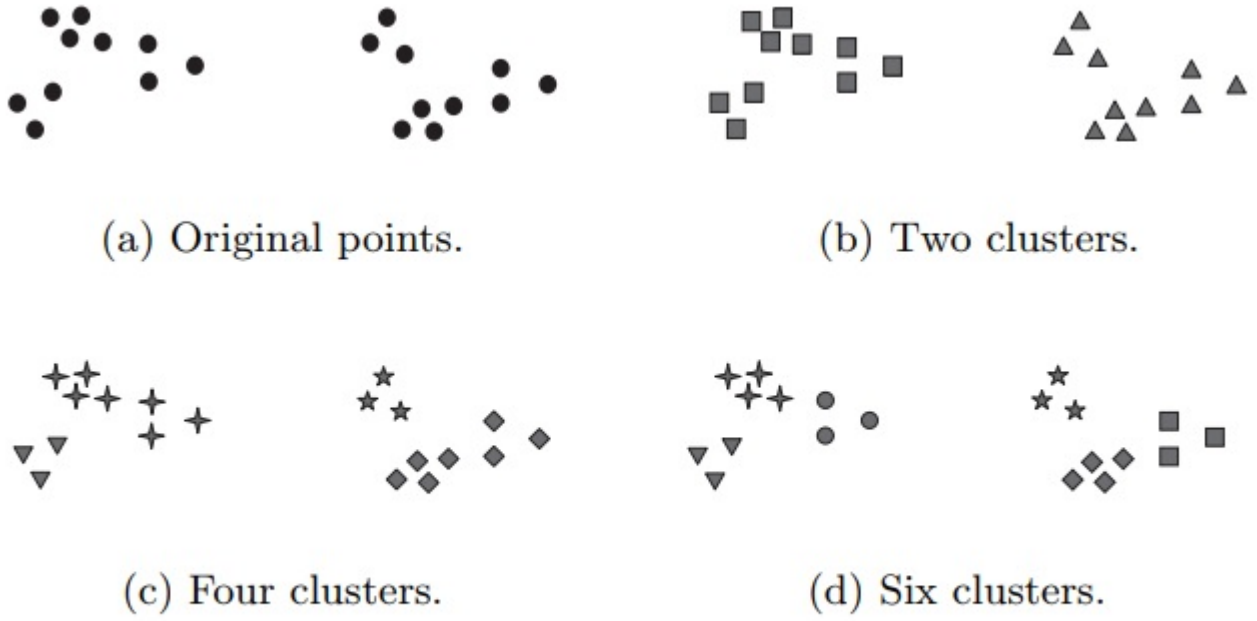


...resumindo!!!

- Clustering pode ser vista como a tarefa de separar objetos em grupo
  - baseia-se nas *características* que estes objetos possuem
  - o agrupamento dos objetos é feito de acordo com algum *critério pré-determinado*

Desafios

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



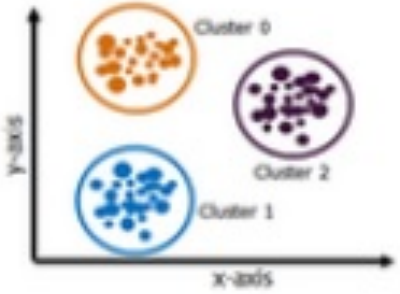
(a) Original points. (b) Two clusters.

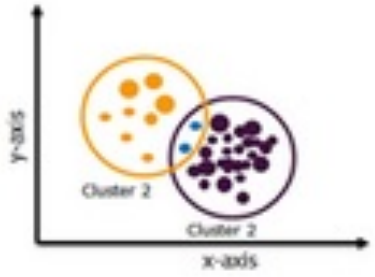
(c) Four clusters. (d) Six clusters.

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

Tipos de clusters

- Exclusivos
- Sobrepostos





Abordagens

- Existem diferentes abordagens para agrupar objetos a partir de um conjunto de dados
- Estudar algumas dessas abordagens e seus principais algoritmos
  - Partição (K-means)
  - Hierárquica (HAC)
  - Densidade (DBscan)

Métricas de avaliação

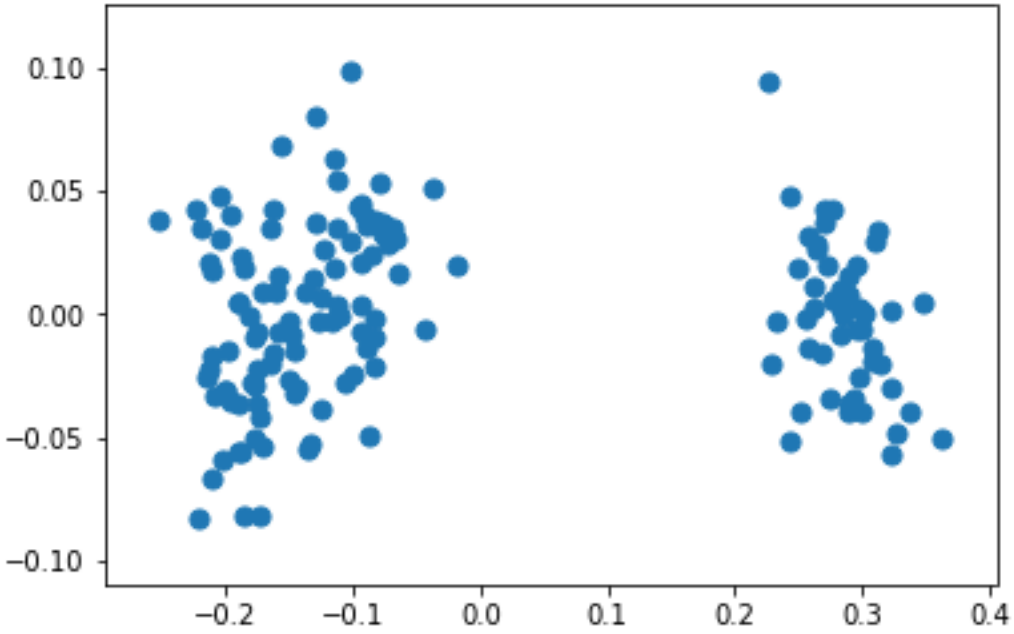
- É esperado que os clusters gerados sejam homogêneo e isolados
  - Silhouete Score
  - Rand Index
  - V-score

Preliminares

- Algoritmos de clustering utilizam alguma função de si/dissimilaridade para agrupar objetos
- Diferentes métricas
  - Euclidiana, Cosseno, Manhattan, Jaccard, ...
- Distâr

$$d(x, z) = ||x - z||_2 = \sqrt{\sum_{j}^{p} (x_j - z_j)^2}$$

Dataset



K-means (+ detalhes no notebook)

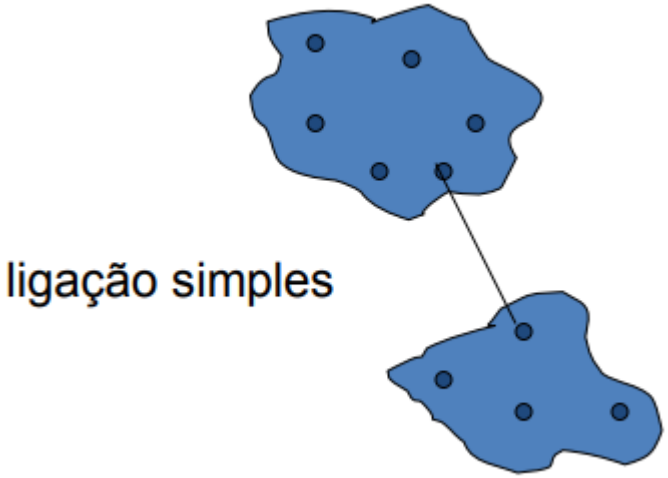
- Divide o conjunto de dados em k partições
- Utiliza o conceito de *centróides*
- Uso do algoritmo K-means++ para atribuir o melhor centróide
- Uso da métrica silhouete para encontrar o melhor k

HAC (+ detalhes no notebook)

- Algoritmo aglomerativo (bottom-up)
- Converge quando agrupa todos os objetos em um único cluster
- Não requer um número de cluster, essa quantidade pode ser inferida
- Gera um dendograma, que permite analisar proximidade entre os objetos
- Diferentes métodos para agrupar os objetos

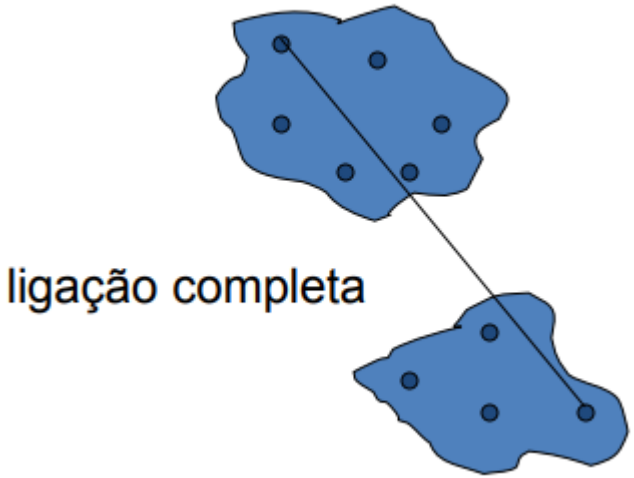
Métodos de agrupamento

- Método de ligação simples (Single linkage)
  - Medida de similaridade entre dois clusters é definida pela menor distância de qualquer ponto do 1º cluster para qualquer ponto do 2º cluster.



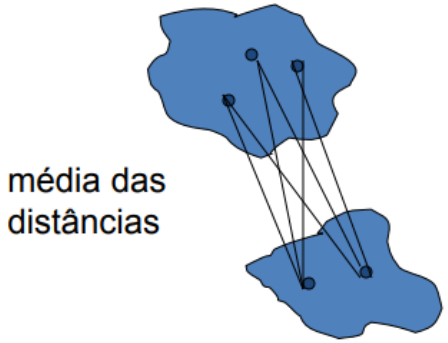
Métodos de agrupamento

- Método de ligação completa (Complete linkage)
  - Medida de similaridade entre dois clusters é definida pela maior distância de qualquer ponto do 1º cluster para qualquer ponto do 2º cluster.



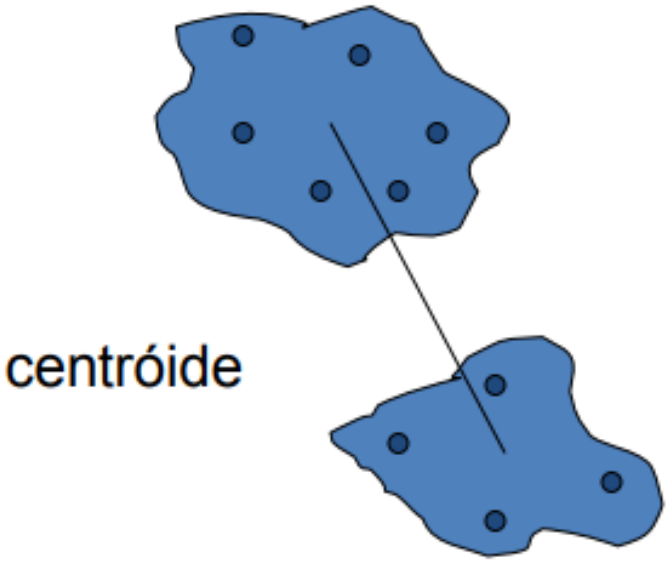
Métodos de agrupamento

- Método da média das distâncias (Average linkage)
  - Medida de similaridade entre dois clusters é definida pela média das distâncias de todos os pontos do 1º cluster em relação aos pontos do 2º cluster.

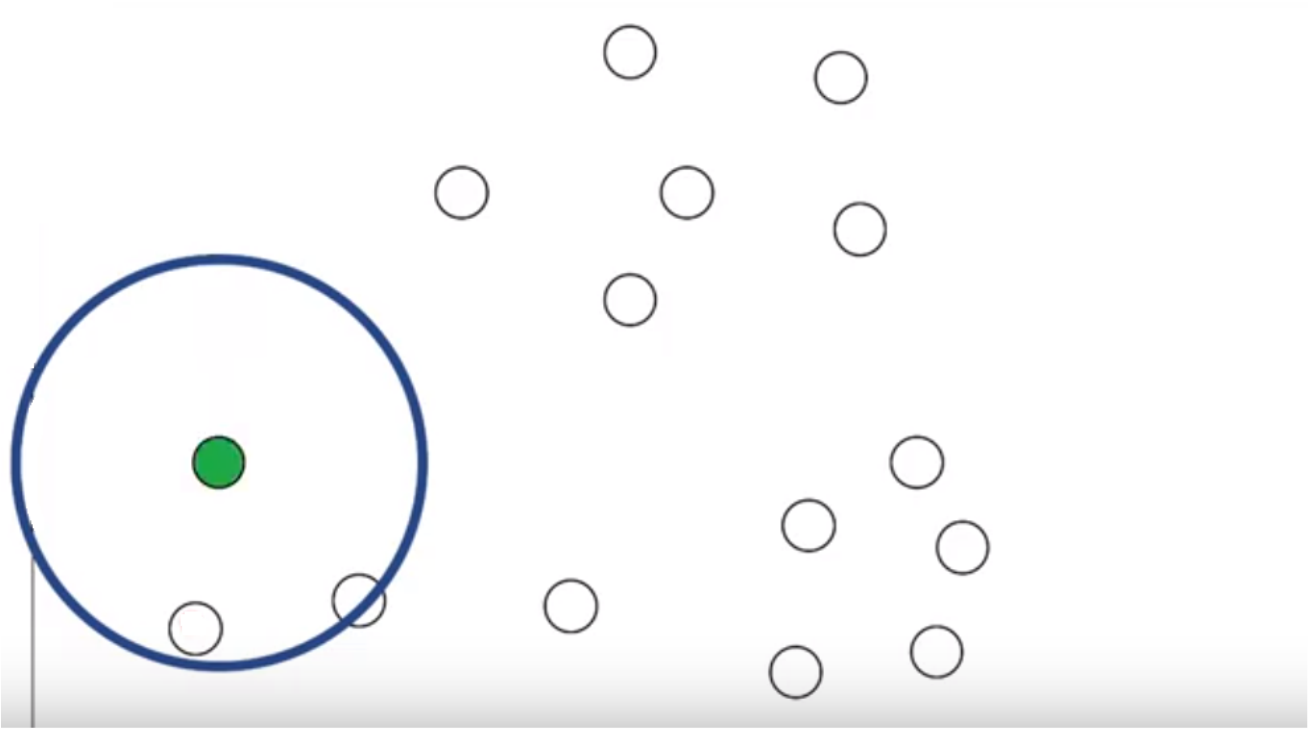


Métodos de agrupamento

- Método do centróide (Centroid method)
  - Medida de similaridade entre dois clusters é definida pela distância entre os pontos médios do 1º e 2º clusters.



DBScan (+ no notebook)



Métricas de avaliação (+ no notebook)

- Considera o rótulo dos dados
  - V-score
  - ARI
- S/ considerar o rótulo dos dados
  - Silhouette index

Considerações

- Aprendizado não supervisionado
- Clustering
- Aplicações
- Técnicas e Algoritmos de clustering
- Validação de clustering