Agrupamento

Prof. Dr. Tiago Eugenio de Melo (EST/UEA)
tmelo@uea.edu.br

Aula
Clusterização

Referências Bibliográficas

- https://realpython.com/k-means-clustering-python
- https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering

Objetivos

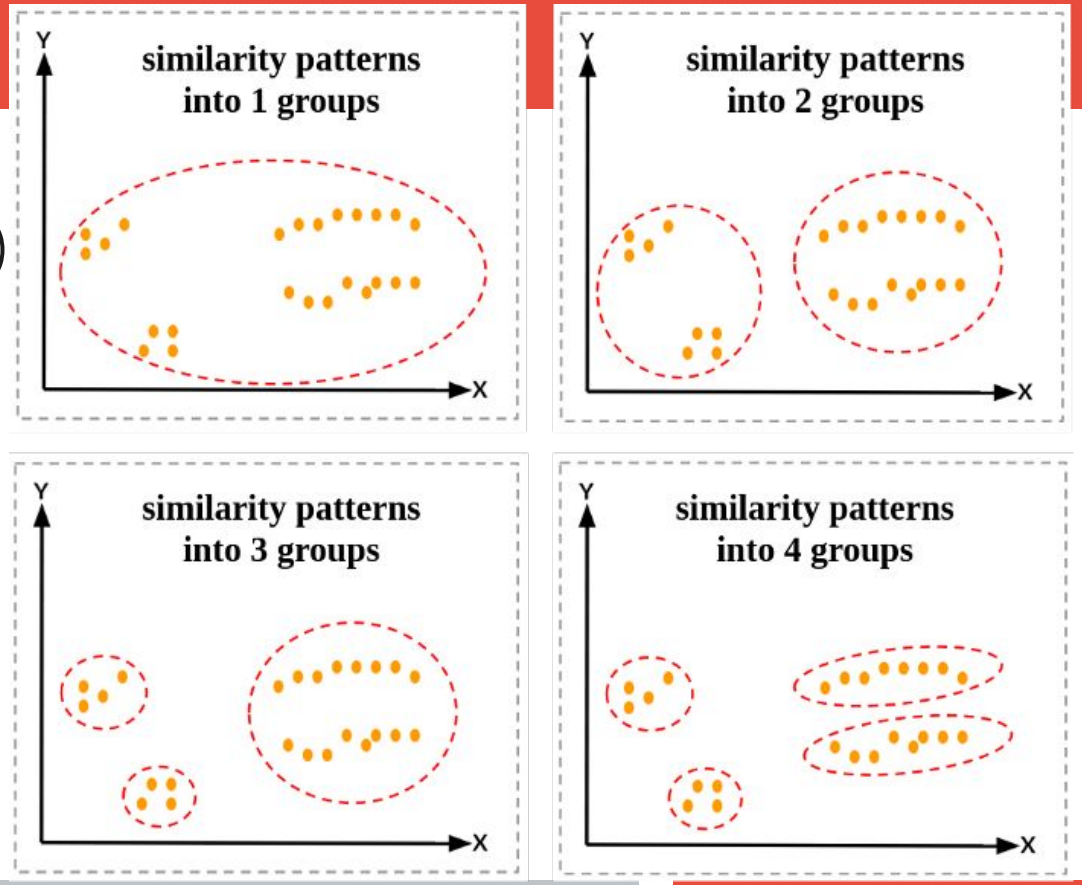
- Apresentar o conceito de clusterização.
- Apresentar o algoritmo de clusterização.

Introdução

- Clusterização (*clustering*)
  - É um conjunto de técnicas usadas para particionar objetos em grupos (*clusters*).

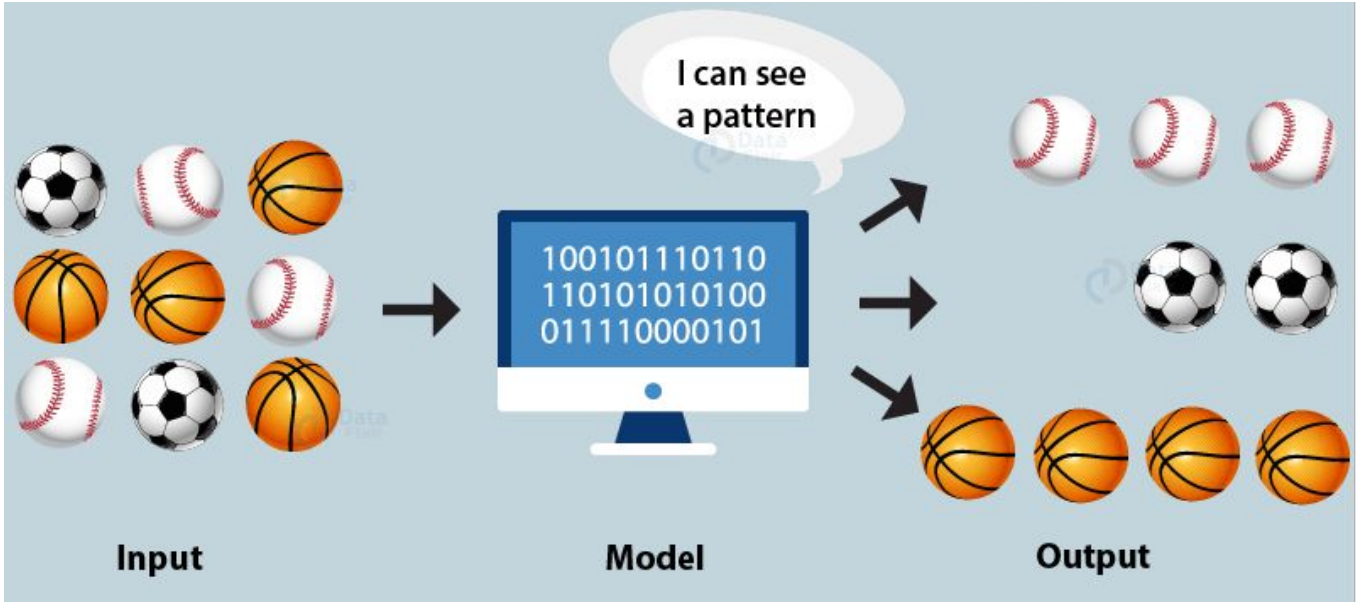
Introdução

- Clusterização (*clustering*)



Introdução

- Exemplo:



Aplicações

- Segmentação de clientes
- Agrupamento de documentos
- Segmentação de imagens
- Mecanismos de recomendação

Introdução

- Na prática, clusters ajudam a identificar duas qualidades dos dados:
  - Significado.
  - Utilidade.

Significado

- Clusters expandem o conhecimento em um determinado domínio.
- Por exemplo, no campo médico, é possível agrupar pessoas que tenham reações semelhantes a determinado medicamento.

Utilidade

- Clusters participam de um processo de análise de dados.
- Por exemplo, as empresas podem agrupar (segmentar) clientes pelo perfil de compras.

Abordagens

- Existem três principais abordagens:
  - Partição.
  - Hierárquico.
  - Baseado em densidade.

Partição

- Partição divide os objetos em grupos sem sobreposição.
- Cada objeto deve pertencer a um único grupo.
- Cada grupo deve ter, ao menos, um objeto.
- Exemplos de algoritmos:
  - K-means.
  - K-meoids.

Partição

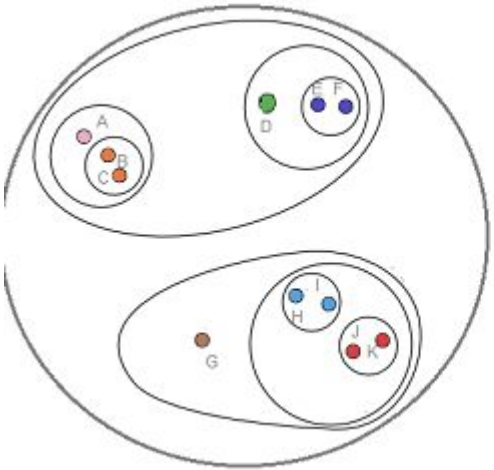
- Características:
  - Algoritmos são não-determinísticos.
  - Positivas:
    - Eles funcionam bem quando os clusters têm corpo esférico.
    - Eles são escaláveis com relação à complexidade dos algoritmos.
  - Fraquezas:
    - Não são apropriados para clusters com formatos diferentes e com diferentes tamanhos.

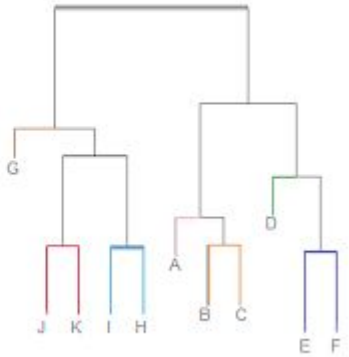
Hierárquico

- Determina uma construção hierárquica dos dados.
- Gera um dendograma.
- O número K de clusters é pré-determinado pelo usuário.
- Os algoritmos são determinísticos.

Hierárquico

- Exemplo:





Hierárquico

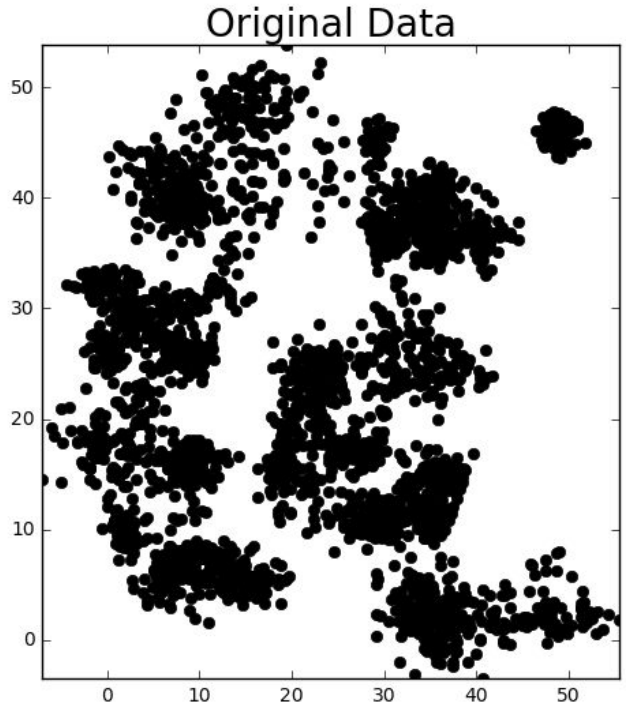
- Características:
  - Positivo:
    - Revela uma relação mais detalhada entre os objetos.
  - Fraqueza:
    - São computacionalmente "pesados".
    - Eles são sensíveis a ruídos (*noise*) e exceções (*outliers*).

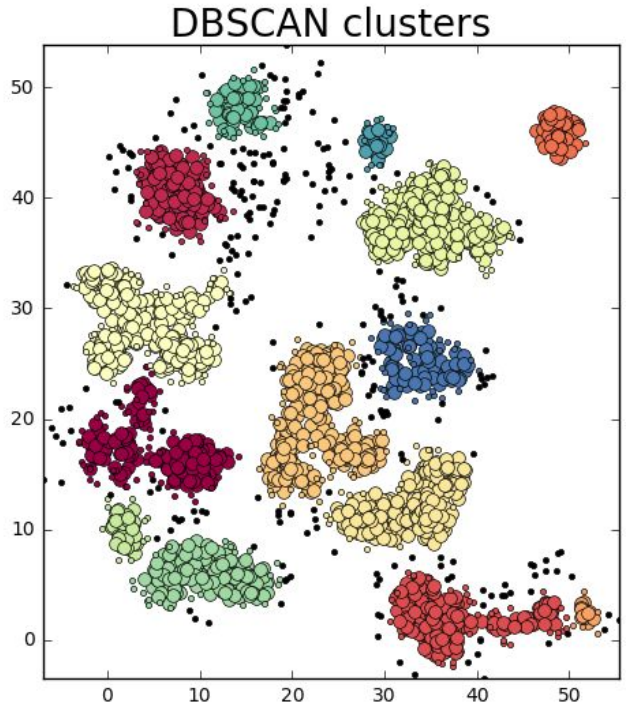
Densidade

- A determinação dos clusters é baseada na densidade dos pontos.
- Exemplos de algoritmos:
  - DBSCAN.
  - OPTICS.

Densidade

- Exemplo:





Densidade

- Características:
  - Positivas:
    - Algoritmos resistentes a *outliers*.
  - Fraquezas:
    - Eles não são apropriados para dados multi-dimensionais.
    - Eles têm dificuldades em clusters com densidades variadas.

K-Means

K-Means

- K-means é executado em poucos passos.
- O primeiro passo é escolher randomicamente k centróides.
- K representa o número de clusters.
- Centróides são os pontos (objetos) que representam o centro de um grupo.

K-means

Algorithm 1 k-means algorithm

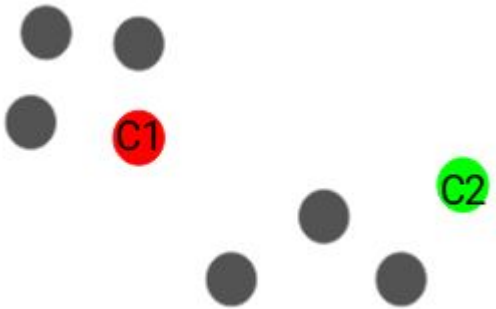
1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

K-Means

- Passo 1
  - O primeiro passo é escolher o valor de K.

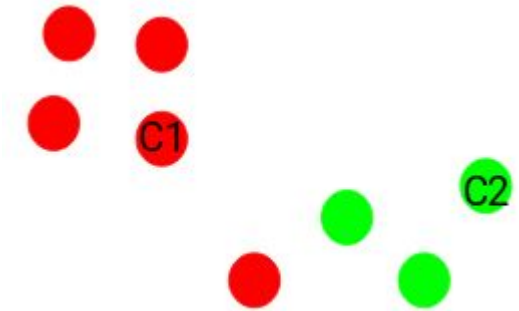
K-Means

- Passo 2:
  - Selecione K pontos como centróides.



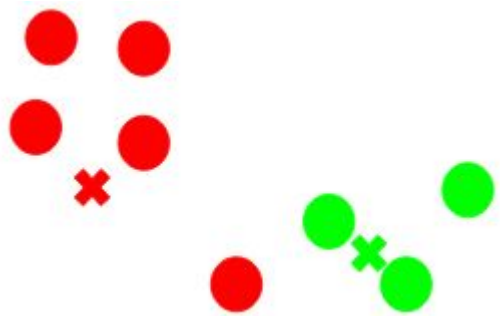
K-Means

- Passo 3:
  - Defina os grupos por proximidade com os centróides.



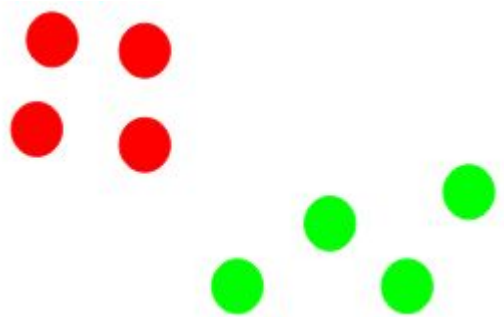
K-Means

- Passo 4:
  - Recalcule os novos centróides.



K-Means

- Passo 5:
  - Repita os passos 3 e 4.



K-Means

- Exemplo:
  - http://shabal.in/visuals/kmeans/1.html

Critérios de Parada

- Existem três critérios de parada:
  - Novos centróides são os mesmos.
  - Os pontos permanecem nos mesmos clusters.
  - Máximo número de iterações é alcançado.

Trabalho Prático III

- Etapas:
  - Definição das equipes.
  - Criação de notebook com a implementação do algoritmo K-Means.
  - Avaliação dos resultados considerando as métricas de precisão, revocação e F1.
  - Apresentação e discussão dos resultados.
  - Exibição [1] dos grafos corretos e os gerados.

[1] https://networkx.org