Clusterização
ou
Agrupamento

IFSP Guarulhos

O que é Agrupamento

- Clusterização é um conjunto de técnicas usadas para particionar dados em grupos ou clusters.
- Os clusters são definidos como grupos de objetos de dados que são mais semelhantes uns aos outros dentro de cada agrupamento.

O que é Agrupamento

- Na prática, o agrupamento ajuda a identificar duas qualidades de dados:
  - Significância
  - Utilidade

Agrupamentos Significativos

- Permitem um melhor conhecimento de um domínio.
- Por exemplo:
  - Na medicina os pesquisadores aplicaram agrupamento em experimentos de expressão gênica.
  - Os resultados de agrupamento identificaram grupos de pacientes que respondem de forma diferente aos tratamentos médicos.

Agrupamentos Úteis

- Servem como uma etapa intermediária em um pipeline de dados.
- Exemplo:
  - A empresas usam agrupamentos para segmentação de clientes.
  - Os resultados do agrupamento segmentam os clientes em grupos com históricos de compra semelhantes, que as empresas podem usar para criar campanhas publicitárias direcionadas.

Visão geral das técnicas de agrupamento

- Você pode realizar o agrupamento usando muitas abordagens diferentes — tantas, na verdade, que existem categorias inteiras de algoritmos de agrupamento.
- Cada uma dessas categorias tem seus próprios pontos fortes e fracos.
- Isso significa que certos algoritmos de agrupamento resultarão em atribuições de agrupamento mais naturais, dependendo dos dados de entrada.

Visão geral das técnicas de agrupamento

- A seleção de um algoritmo de agrupamenot adequado para seu conjunto de dados geralmente é difícil devido ao número de opções disponíveis.
- Alguns fatores importantes que afetam essa decisão incluem as características dos agrupamentos, os recursos do conjunto de dados, o número de outliers e o número de objetos de dados.

Categorias de algoritmos de agrupamento

- Agrupamento em partições
- Agrupamento hierárquico
- Agrupamento baseado em densidade

Agrupamento em partições

- O agrupamento em partições divide os objetos de dados em grupos não sobrepostos.
- Dessa forma, nenhum objeto pode ser membro de mais de um agrupamento e cada agrupamento deve ter pelo menos um objeto.
- Essas técnicas exigem que o usuário especifique o número de agrupamentos, indicado pela variável k .

Pontos fortes do agrupamento em partições

- Os métodos de agrupamento em partições têm pontos fortes :

  - Funcionam bem quando os aglomerados têm uma forma esférica .
  - São escaláveis em relação à complexidade do algoritmo.

Pontos fracos do agrupamento em partições

- Os métodos de agrupamento em partições têm pontos fracos:

  - Não são adequados para agrupamentos com formas complexas e tamanhos diferentes.
  - Não funcionam quando usados com grupos de diferentes densidades .

Agrupamento em partições

- Muitos algoritmos de agrupamento em partições funcionam por meio de um processo iterativo para atribuir subconjuntos de pontos de dados em k agrupamentos.
- São algoritmos não determinísticos , o que significa que podem produzir resultados diferentes de duas execuções separadas, mesmo que as execuções sejam baseadas na mesma entrada.

Agrupamento hierárquico

- O agrupamento hierárquico determina as atribuições de agrupamento construindo uma hierarquia.
- Sendo implementado por uma abordagem de baixo para cima (por aglomeração) ou de cima para baixo (por divisão):
  - O **agrupamento por aglomeração** é a abordagem de baixo para cima, que mescla os dois pontos que são mais semelhantes até que todos os pontos tenham sido mesclados em um único agrupamento.
  - O **agrupamento por divisão** é a abordagem de cima para baixo, que começa com todos os pontos como um único agrupamento e divide em agrupamentos menos semelhantes em cada etapa até que restem apenas pontos de dados únicos.

Agrupamento hierárquico

- Esses métodos produzem uma hierarquia de pontos baseada em árvore chamada dendrograma .
- Semelhante ao agrupamento por partição, no agrupamento hierárquico o número de agrupamentos (k) é frequentemente predeterminado pelo usuário.
- Os agrupamentos são atribuídos cortando o dendrograma a uma profundidade especificada que resulta em k grupos de dendrogramas menores.

Pontos fortes do agrupamento hierárquico

- Pontos fortes dos métodos de agrupamento hierárquico:
  - Geralmente revelam os detalhes mais sutis sobre os relacionamentos entre os dados.
  - Fornecem um dendrograma interpretável.

Pontos fracos do agrupamento hierárquico

- Pontos fracos dos métodos de agrupamento hierárquico:
  - Apresentam alto custo computacional devido a complexidade de seu algoritmo.
  - São sensíveis à ruídos e discrepâncias.

Agrupamento Baseado em Densidade

- O agrupamento baseado em densidade realiza as atribuições de agrupamento baseado na densidade de pontos de dados em uma região.
- Os agrupamentos são atribuídos onde há altas densidades de pontos de dados separados por regiões de baixa densidade.

Agrupamento Baseado em Densidade

- Ao contrário das outras categorias de agrupamentos, essa abordagem não exige que o usuário especifique o número de agrupamentos.
- Existe um parâmetro baseado na distância que atua como um limite ajustável. Esse limite determina quão próximos os pontos devem ser para serem considerados um membro do agrupamento.
- Exemplos de algoritmos de agrupamento baseados em densidade incluem agrupamento espacial baseado em densidade de aplicativos com ruído (DBSCAN), e pontos de ordenação para identificar a estrutura de agrupamento (OPTICS) .

Pontos Fortes do Agrupamento Baseado em Densidade

- Os pontos fortes dos métodos de agrupamento baseados em densidade incluem:
  - Destacam-se na identificação de grupos de formas não esféricas .
  - São resistentes a outliers .

Pontos Fracos do Agrupamento Baseado em Densidade

- Os pontos fracos dos métodos de agrupamento baseados em densidade incluem:
  - Não são adequados para agrupamento em espaços de grande dimensão.
  - Têm dificuldade em identificar aglomerados de densidades variadas.

O algoritmo K-Means – K médias

- O algoritmo k -means convencional requer apenas alguns passos.
- O primeiro passo é selecionar aleatoriamente a quantidade de k centroides, onde k é igual ao número de agrupamentos escolhidos.
- Centroides são pontos de dados que representam o centro de um agrupamento.

O algoritmo K-Means – K médias

- O elemento principal do algoritmo funciona por meio de um processo de duas etapas chamados de maximização e de expectativa .
- Na etapa de expectativa é atribuído para cada ponto de dados a sua centroide mais próxima.
- Em seguida, a etapa de maximização calcula a média de todos os pontos para cada agrupamento e define o novo centroide.

O algoritmo K-Means – K médias

- Versão convencional do algoritmo k -means:

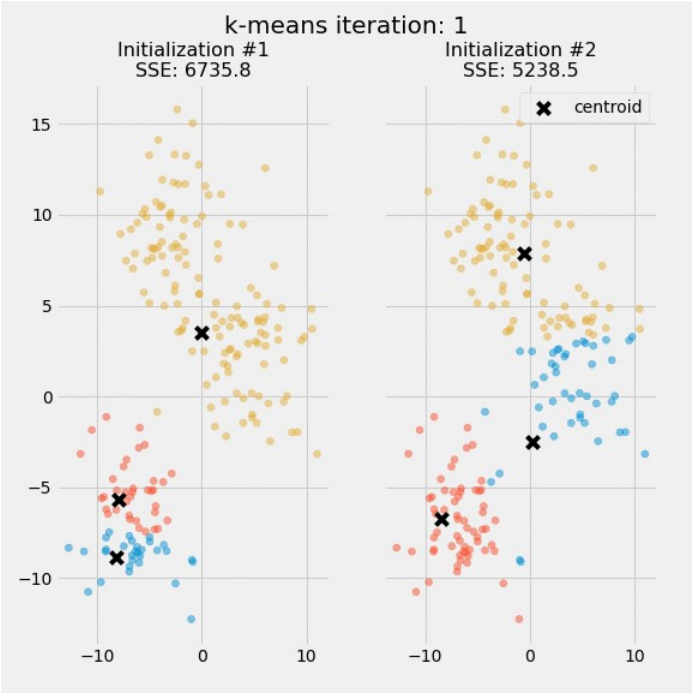
Algorithm 1 k-means algorithm

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.

O algoritmo K-Means – K médias

- A qualidade do número de agrupamentos é determinada pelo cálculo da soma do erro quadrático (SER) após os centroides convergirem.
- O SER é definido como a soma das distâncias euclidianas ao quadrado de cada ponto ao seu centróide mais próximo.
- Como esta é uma medida de erro, o objetivo do k - means é tentar minimizar esse valor.

Inicialização dos centroides



- a inicialização dos centróides é um passo importante.
- A etapa de inicialização aleatória faz com que o algoritmo k-means seja não determinístico, ou seja, a atribuição dos agrupamentos será variável se executarmos o mesmo algoritmo duas vezes.