

Aprendizado de Máquina

Aula 7.4 - Algoritmos de agrupamento

Adriano Rivolli

rivolli@utfpr.edu.br

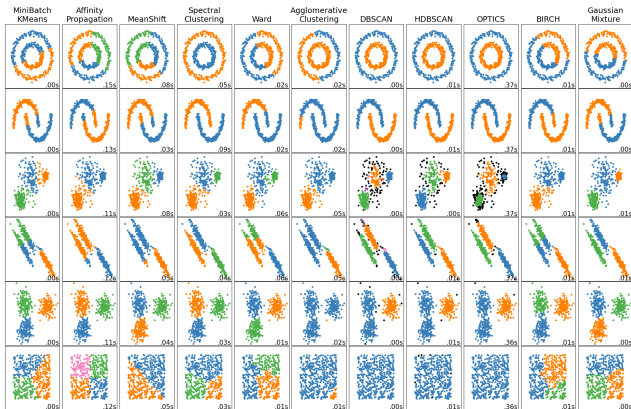
Especialização em Inteligência Artificial

Universidade Tecnológica Federal do Paraná (UTFPR)
Câmpus Cornélio Procopio
Departamento de Computação

Conteúdo

- 1 K-means
- 2 Hierárquico
- 3 DBSCAN

Visão geral dos métodos



Fonte: <https://scikit-learn.org/stable/modules/clustering.html>

K-means

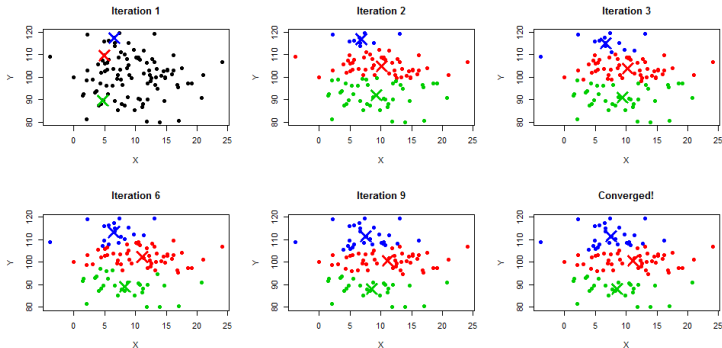
Visão geral

- Algoritmo mais popular e amplamente utilizado
- Baseado em centroide
- O usuário define o número de *clusters*
- É um algoritmo iterativo que gera grupos sem sobreposição
- Características principais:
 - ▶ Simplicidade
 - ▶ Velocidade
 - ▶ Escalabilidade

Etapas do algoritmo

- 1 Inicialização
 - ▶ Número de grupos
 - ▶ Inicializa os centroides de modo aleatório
- 2 Atribuição
 - ▶ Calcula a distância de todos os pontos para o centroide
 - ▶ Atribui cada instância para o centroide mais próximo
- 3 Alteração
 - ▶ Recalcula os centroides conforme as instâncias de cada grupo
 - ▶ Repete a Etapa 2 e 3 até a convergência
- 4 Convergência
 - ▶ Não há mais alterações dos centroides
 - ▶ Número máximo de iterações

Visualização



Fonte:

[method-k-means-steps-example.png](https://www.learnbymarketing.com/wp-content/uploads/2015/01/method-k-means-steps-example.png)

<https://www.learnbymarketing.com/wp-content/uploads/2015/01/>

Exemplos

■ Ferramenta de visualização:

- ▶ <https://www.naftaliharris.com/blog/visualizing-k-means-clustering/>

■ Agrupamento do dataset Iris:

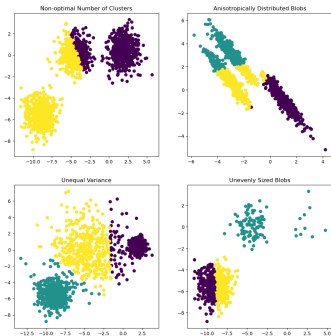
- ▶ https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_iris.html

■ Agrupamento do dataset Dígitos:

- ▶ https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html

Possíveis problemas


Unexpected KMeans clusters



Soluções

- Redução do espaço das características
- Seleção do número de grupos adequadamente
- Usar diferentes inicializações

Como escolher o número correto de k ?

- 
- Método da Silhouette
 - ▶ Calcular a medida para diferentes valores de k
 - ▶ Usar o k que resulta no mais alto valor
 - Se os resultados forem ruins, considerar usar outro método

Vantagens e desvantagens

- Simplicidade
- Obtém grupos bem separados
- Eficiente mesmo para grandes datasets
- Sensível à inicialização (não determinístico)
- Dificuldade com grupos não esféricos
- Grupos desbalanceados
- Por se tratar de um algoritmo guloso pode convergir a um mínimo local

Variações

■ K-means++

- ▶ Inicializa os centroides de tal maneira que eles fiquem distantes uns dos outros

■ K-Medoides


- ▶ Ao invés de usar um centroide usa um medoide
- ▶ Um medoide é um ponto real existente no dataset

■ Mini-batch K-means

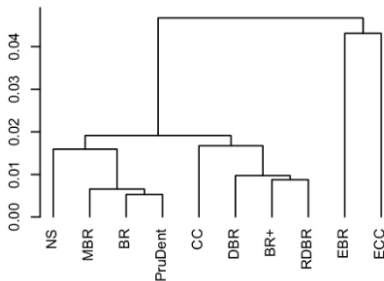
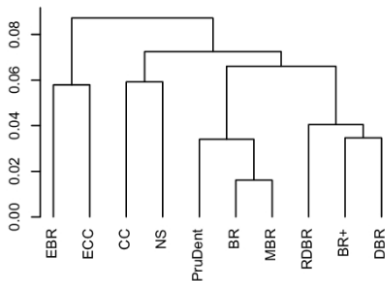
- ▶ Utiliza subconjuntos dos dados para calcular os grupos
- ▶ As instâncias são amostradas aleatoriamente a cada iteração

Hierárquico

Introdução

- 
- Uma família de métodos que cria grupos aninhados
 - O agrupamento pode ser visualizado no formato de uma árvore (dendrograma)
 - A raiz da árvore corresponde a um grupo com todas as instâncias
 - As folhas correspondem as instâncias individuais

Dendrograma



Fonte: Autoria própria

Abordagens

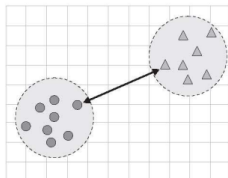
■ Aglomerativa

- ▶ Cada instância é considerada um grupo
- ▶ A cada etapa une os dois grupos mais próximos

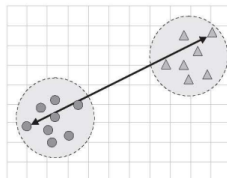
■ Divisiva

- ▶ Todo o conjunto é considerado um grupo
- ▶ A cada etapa divide-se o grupo em duas partes

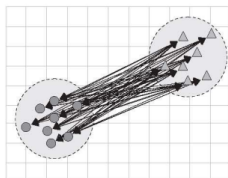
Calcular a distância entre grupos



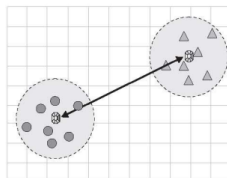
(a) Ligação mínima



(b) Ligação máxima



(c) Ligação média



(d) Centroide

Ligações

■ Ligação mínima (*single linkage*)

- ▶ Menor distância entre os membros de 2 grupos
- ▶ Produz grupos conectados (bom para formatos não elípticos)
- ▶ Suscetível a ruídos e *outliers*

■ Ligação máxima (*complete linkage*)

- ▶ Maior distância entre os membros de 2 grupos
- ▶ Produz grupos mais compactos, porém desbalanceados
- ▶ Menos suscetível a ruídos e *outliers*

■ Ligação média (*average linkage*)

- ▶ Distância média entre os membros de 2 grupos
- ▶ Meio-termo entre as 2 abordagens anteriores

Ligações

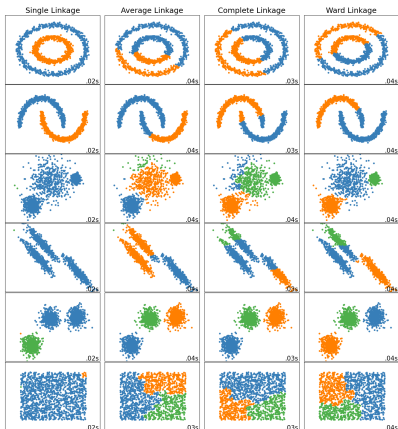
■ Ligação Centroide (*centroid linkage*)

- ▶ Calcula a distância entre os dois centroides

■ Ligação Ward (*Ward linkage*)

- ▶ Minimiza a soma das distâncias entre os pontos e o centroide
- ▶ Mede o quanto se aumenta de variabilidade ao unir os grupos
- ▶ Essa é a mesma abordagem implícita utilizada pelo K-means

Ligações (comparação)



Determinando o número de grupos


■ Inspeção visual

- ▶ Altura das junções
- ▶ Forma do dendrograma

■ Critério de qualidade

- ▶ Otimização de uma medida de validação interna/externa

Vantagens e desvantagens

- 
- Visualização intuitiva
 - Escolhe o número de grupos depois
 - Flexível para diferentes formas
 - Útil para análise de dados
 - Computacionalmente mais caro
 - Sensível a ruídos e *outliers*
 - Depende da escolha da medida de ligação

Variações

■ Adicionar restrições nas ligações

- ▶ Definir uma matriz com as instâncias que podem ser conectadas
- ▶ Definir outras regras de ligações

■ Bisecting K-means

- ▶ Uma variação que roda k-means combinado com o agrupamento hierárquico

DBSCAN

Introdução

- **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise
- Encontra grupos com alta densidade de amostras separados por regiões de baixa densidade
- Encontra qualquer forma no espaço
- Robusto a ruídos e *outliers*
 - ▶ A abordagem permite que instâncias não sejam agrupadas

Core point

■ Alta densidade:

- ▶ O raio de um **ponto central** contém um número mínimo de amostras

■ Um grupo é definido por:

- ▶ Um conjunto de pontos centrais
- ▶ Um conjunto de pontos que estão no raio de um ponto central (*border point*)

Ruídos

- Pontos que não **ponto central** e nem **ponto de borda**
- Pontos que não fazem parte do raio de nenhum **ponto central**
- Estes pontos não fazem parte de nenhum grupo
- São considerados *outliers*

Hiperparâmetros

■ Eps (*epsilon*)

- ▶ A distância máxima entre duas amostras para que uma seja considerada vizinha da outra

■ Min samples

- ▶ O número de amostras (ou peso total) em uma vizinhança para que um ponto seja considerado como ponto central.


Epsilon

- A análise do valor de k obtido pode indicar uma boa escolha deste hiperparâmetro
- Valores pequenos podem fazer com que a maioria dos pontos de dados sejam classificados como ruído
- Valores grandes podem fazer com que *clusters* diferentes se fundam, reduzindo a capacidade de detectar estruturas de cluster mais refinadas

Minímo de instâncias



- Controla o quão tolerante o algoritmo é em relação ao ruído
- Valores pequenos podem levar à identificação de mais clusters, aumentando a sensibilidade ao ruído
- Valores grandes tornam o algoritmo mais robusto ao ruído, mas podem falhar na identificação de clusters menores
- Um ponto de partida é a dimensionalidade do conjunto de dados mais um ou dois

Etapas do algoritmo

- 
- 1 Identificar os pontos centrais, pontos de bordas e ruídos
 - 2 Formar grupos ao redor dos pontos centrais
 - 3 Associar os pontos de bordas aos seus respectivos grupos
 - 4 Marcar os pontos de ruídos

Simulador: <https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/>

Vantagens e desvantagens

- 
- Ótimo para grupos de alta *versus* baixa densidade
 - Grupos em formato arbitrário
 - Robustez para *outliers* e ruído
 - Não requer a escolha de k
 - Usado para detecção de anomalias
 - Determinístico
- 
- Dificuldade em encontrar clusters de densidades variadas
 - Sensibilidade à escolha dos hiperparâmetros
 - Desafios com dados de alta dimensão

Variações

■ HDBSCAN

- ▶ Explora diferentes escalas de densidade enquanto o DBSCAN é globalmente homogêneo

■ OPTICS

- ▶ É considerado uma generalização do DBSCAN, usando uma faixa de valores eps