

K-means

Funcionamento:

Algoritmo de clustering que particiona os dados em K clusters.

Inicializa K centroides aleatoriamente (ou usando outras estratégias).

Itera alternando entre atribuir pontos ao centroide mais próximo e recalculando os centroides.

Estratégias de Inicialização:

Random: Escolhe K pontos aleatórios como centroides iniciais.

Forgy: Semelhante ao Random, mas escolhe K observações aleatórias como centroides.

Random Partition: Atribui aleatoriamente pontos a clusters e calcula os centroides.

K-means++: Escolhe o primeiro centroide aleatoriamente, e os subsequentes com probabilidade proporcional à distância do centroide mais próximo.

Prós:

Simple e fácil de implementar.

Eficiente em termos computacionais.

Contras:

Sensível à escolha inicial dos centroides.

Assume que os clusters são esféricos e de tamanhos semelhantes.

Necessidade de especificar o número de clusters antecipadamente.

Métodos Aglomerativos

Funcionamento:

Algoritmos de clustering hierárquico que começam com cada ponto como seu próprio cluster e fundem progressivamente os clusters mais próximos.

Estratégias:

Ward: Minimiza a soma dos quadrados dentro de cada cluster.

Complete (Ligação Completa): Usa o maior distanciamento entre pontos em dois clusters para a fusão.

Average (Ligação Média): Usa a distância média entre pontos de dois clusters.

Single (Ligação Simples): Usa o menor distanciamento entre pontos em dois clusters.

Prós:

Não requer a especificação do número de clusters.

Pode revelar estruturas interessantes nos dados.

Contras:

Computacionalmente intensivo para grandes datasets.

As fusões são irreversíveis, o que pode levar a decisões subótimas.

Gaussian Mixture Model (GMM)

Funcionamento:

Modelo baseado na suposição de que os dados são gerados a partir de uma mistura de várias distribuições gaussianas.

Utiliza o algoritmo Expectation-Maximization para estimar os parâmetros das gaussianas.

Prós:

Flexibilidade para modelar clusters com diferentes formas e tamanhos.

Fornece uma medida de probabilidade (soft clustering).

Contras:

Mais complexo e computacionalmente mais intensivo que K-means.

Sensível à inicialização e pode convergir para mínimos locais.

DBSCAN

Funcionamento:

Baseado na densidade espacial, forma clusters expandindo áreas de alta densidade.

Define 'core points' com muitos vizinhos próximos e expande clusters a partir deles.

Prós:

Não requer a especificação do número de clusters.

Pode encontrar clusters de formas arbitrárias.

Robusto contra outliers.

Contras:

Sensível aos parâmetros **eps** (distância) e **min_samples** (vizinhos mínimos).

Pode não funcionar bem com dados de densidade variável.

Cada um desses métodos tem suas próprias forças e fraquezas, e a escolha do método apropriado geralmente depende da natureza dos dados e do problema específico que está sendo abordado.