Algoritmos Não Supervisionados – K-Means

ROCKETSEAT

Plataforma completa de aprendizado contínuo em programação.

#BoostingPeople

rocketseat.com.br

Algoritmos Não Supervisionados

K-Means

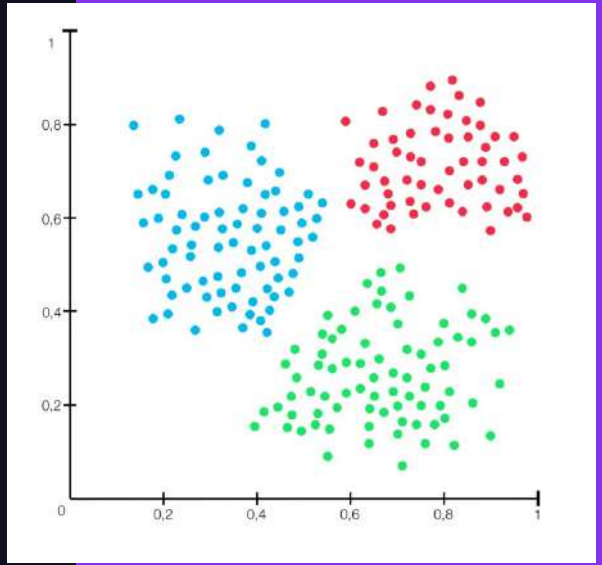
O objetivo deste módulo é apresentar conceitualmente os principais algoritmos de clusterização para que possamos desenvolver projetos de machine learning que realizem agrupamento de objetos. E faremos um projeto explorando o primeiro destes algoritmos, que é o K-Means, onde faremos o processo completo desde o EDA até a entrega do modelo através de uma aplicação para inferência batch.

Agenda

- O que é clusterização
- Um passeio pelos algoritmos de clusterização
- O que é o algoritmo K-Means
- Medidas de Distância
- Métricas de algoritmos de clusterização
- Projeto – K-Means

O que é clusterização

Os algoritmos de clusterização são técnicas de aprendizado de máquina e mineração de dados que agrupam um conjunto de dados em clusters ou grupos, com base na similaridade entre os itens. A ideia é agrupar objetos semelhantes em um mesmo grupo e objetos diferentes em grupos distintos. Existem diversos algoritmos de clusterização, cada um com suas próprias características e abordagens, mas todos compartilham o mesmo objetivo fundamental: identificar padrões nos dados e organizar esses dados em grupos significativos.



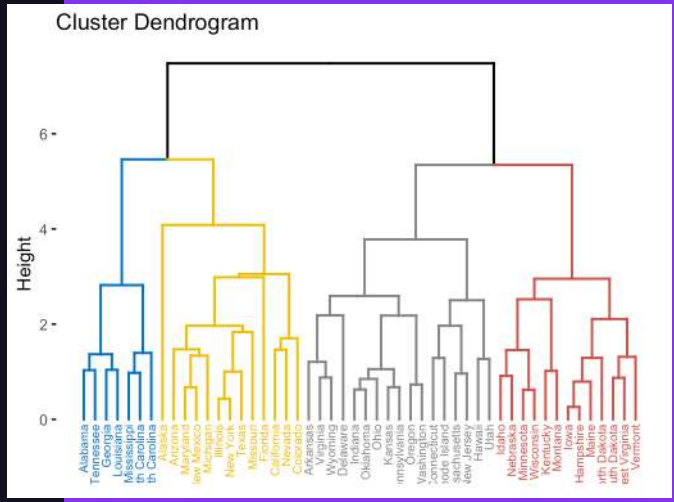
O que é clusterização

Alguns usos de algoritmos de clusterização:

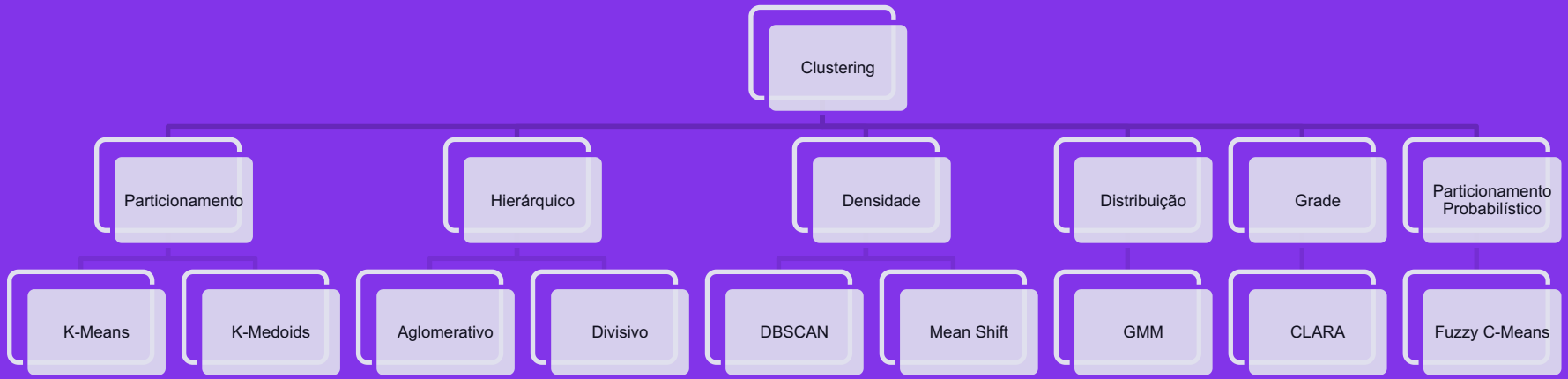
Segmentação de mercado: Empresas utilizam a clusterização para segmentar seus clientes com base em características demográficas, comportamentais ou de consumo. Isso permite direcionar estratégias de marketing mais eficazes e personalizadas para cada segmento de clientes.

Análise de redes sociais: Em redes sociais, os algoritmos de clusterização podem ser usados para identificar comunidades de usuários com interesses semelhantes, facilitando a personalização de conteúdo e recomendações.

Agrupamento de documentos: Na mineração de texto, os algoritmos de clusterização podem ser usados para agrupar documentos semelhantes, facilitando a organização e a recuperação de informações.



Um passeio pelos algoritmos de clusterização



Um passeio pelos algoritmos de clusterização

K-Means

Agrupa dados em k clusters, onde k é definido pelo usuário. Ele atribui pontos de dados aos clusters com base na proximidade dos centroides, recalculando-os iterativamente para minimizar a variância intra-cluster.

K-Medoids

Variação do K-Means que usa medoids (dados reais) como representantes de clusters. Ele atribui medoids inicialmente e, em seguida, reajusta iterativamente, minimizando uma função de dissimilaridade. É mais robusto a outliers

Hierárquico Aglomerativo

Começa considerando cada ponto de dado como um cluster separado e, em seguida, mescla iterativamente os clusters mais próximos, formando uma hierarquia de clusters até que todos os pontos estejam em um único cluster.

Hierárquico Divisivo

Começa considerando todos os pontos de dados em um único cluster e, em seguida, divide iterativamente o cluster em subclusters menores até que cada ponto de dado esteja em seu próprio cluster individual.

Um passeio pelos algoritmos de clusterização

DBSCAN

Agrupa pontos em regiões densas, identificando clusters de diferentes formas e tamanhos. Seus principais parâmetros são epsilon (distância máxima entre pontos vizinhos) e minPts (número mínimo de pontos para formar um cluster).

Mean Shift

Busca os máximos locais da função de densidade de probabilidade para encontrar os centros dos clusters. Ele ajusta iterativamente os centros dos clusters em direção às regiões de maior densidade de pontos, até convergir para os centros finais.

Gaussian Mixture Models

Assume que os pontos de dados são gerados a partir de uma mistura de várias distribuições gaussianas. Ele estima os parâmetros dessas distribuições (como médias, covariâncias e pesos) para modelar os clusters nos dados.

CLARA

É uma adaptação do K-Medoids para grandes datasets. Usa amostragem para criar um subconjunto, aplica o K-Medoids nesse subconjunto e ajusta os clusters para os dados completos, mantendo apenas os medoids ótimos, tornando o processo eficiente para grandes datasets.

O que é o algoritmo K-Means

O K-Means é um algoritmo de clusterização amplamente utilizado em análise de dados e aprendizado de máquina. Ele agrupa dados em k clusters. O algoritmo segue estes passos:

1) Inicialização dos Centroides: Começa selecionando aleatoriamente k pontos como centroides iniciais. Esses centroides são representantes dos clusters.

2) Atribuição de Pontos aos Clusters: Cada ponto de dado é atribuído ao cluster cujo centróide está mais próximo, com base em uma medida de distância, geralmente a distância euclidiana.

3) Atualização dos Centroides: Os centroides dos clusters são recalculados como a média de todos os pontos atribuídos a esse cluster.

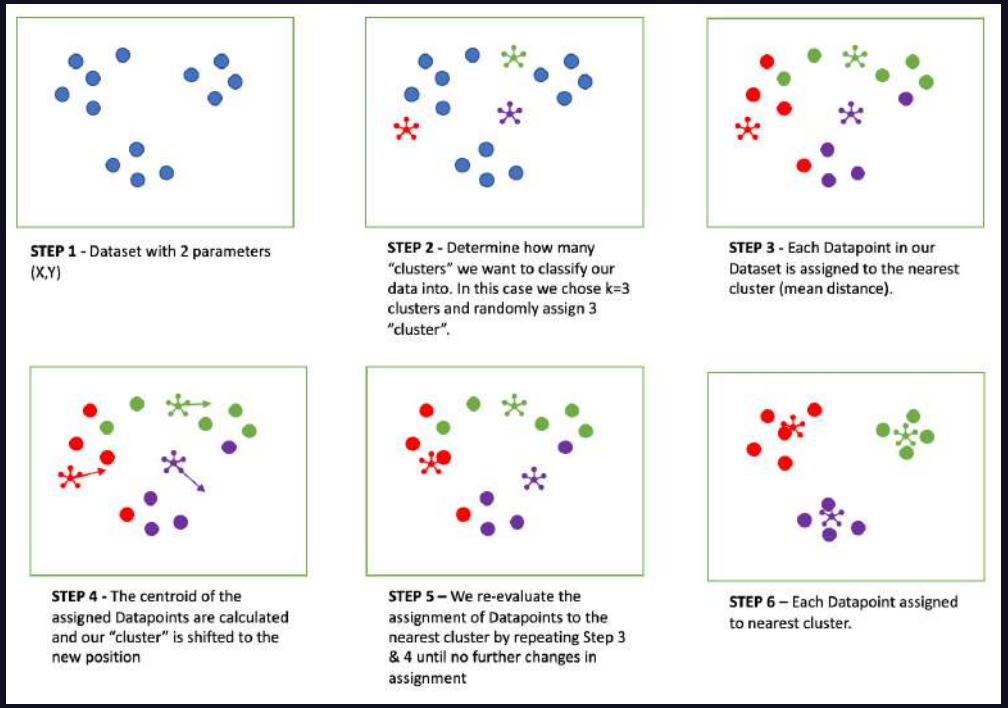
O que é o algoritmo K-Means

4) Reatribuição de Pontos: Os pontos são reatribuídos aos clusters com base nos novos centroides.

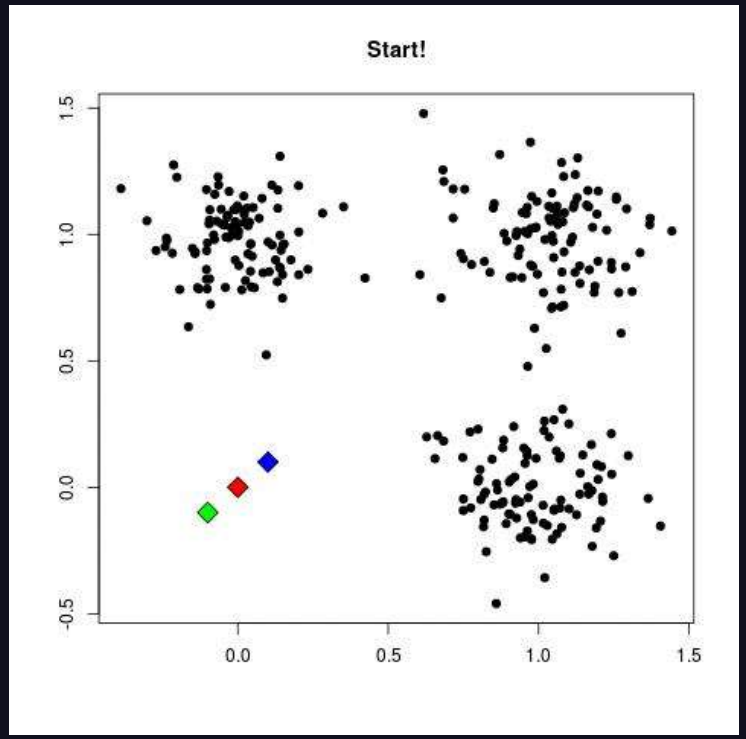
5) Iteração: Os passos 3 e 4 são repetidos até que não haja mudanças significativas na atribuição dos pontos aos clusters ou um número máximo de iterações seja alcançado.

O algoritmo converge quando os centroides não mudam significativamente entre as iterações ou quando atinge a quantidade máxima de iterações. No entanto, o resultado final pode variar dependendo da inicialização aleatória dos centroides e do número de clusters escolhido.

O que é o algoritmo K-Means



O que é o algoritmo K-Means



O que é o algoritmo K-Means

Como definir o parâmetro k (quantidade de clusters) ?

Método do Cotovelo (Elbow Method): Este método envolve plotar o valor da função de custo (por exemplo, a soma dos quadrados das distâncias intra-cluster) em relação ao número de clusters (k) e observar onde ocorre uma mudança significativa na inclinação da curva. O ponto onde a curva começa a se nivelar é frequentemente escolhido como o número ideal de clusters.

Método da Silhueta (Silhouette Method): Este método avalia a qualidade dos clusters formados por diferentes valores de k. Calcula a silhueta média de todos os pontos de dados em cada cluster para diferentes valores de k e escolhe o valor de k que maximiza a média da silhueta.

O que é o algoritmo K-Means

Como definir o parâmetro k (quantidade de clusters) ?

Método Gap Statistics: Este método compara a dispersão intra-cluster para diferentes valores de k com uma dispersão esperada sob um modelo de referência nulo (por exemplo, dados aleatórios). O número ideal de clusters é aquele que maximiza a lacuna estatística entre as duas dispersões.

Validação Externa: Em alguns casos, pode-se ter acesso a informações externas sobre os dados que indicam o número correto de clusters. Nesses casos, métodos de validação externa, como índices de validação externa ou comparação com rótulos conhecidos, podem ser usados para determinar o número ideal de clusters.

O que é o algoritmo K-Means

Como definir o parâmetro k (quantidade de clusters) ?

Conhecimento de Domínio: Às vezes, o conhecimento prévio do domínio pode ajudar a determinar o número de clusters de forma mais precisa. Por exemplo, em ciências sociais, pode-se ter conhecimento sobre o número de grupos ou categorias existentes.

Medidas de Distância

Distância Euclidiana: É a distância mais comum e simples. Calculada como a raiz quadrada da soma dos quadrados das diferenças entre as coordenadas dos pontos em cada dimensão.

Distância Manhattan: Também conhecida como distância de cidade, é a soma das diferenças absolutas entre as coordenadas dos pontos em cada dimensão.

Distância de Minkowski: É uma generalização das distâncias Euclidiana e Manhattan. Dependendo do valor de um parâmetro p, pode reduzir-se à distância Euclidiana (quando p = 2) ou à distância Manhattan (quando p = 1).

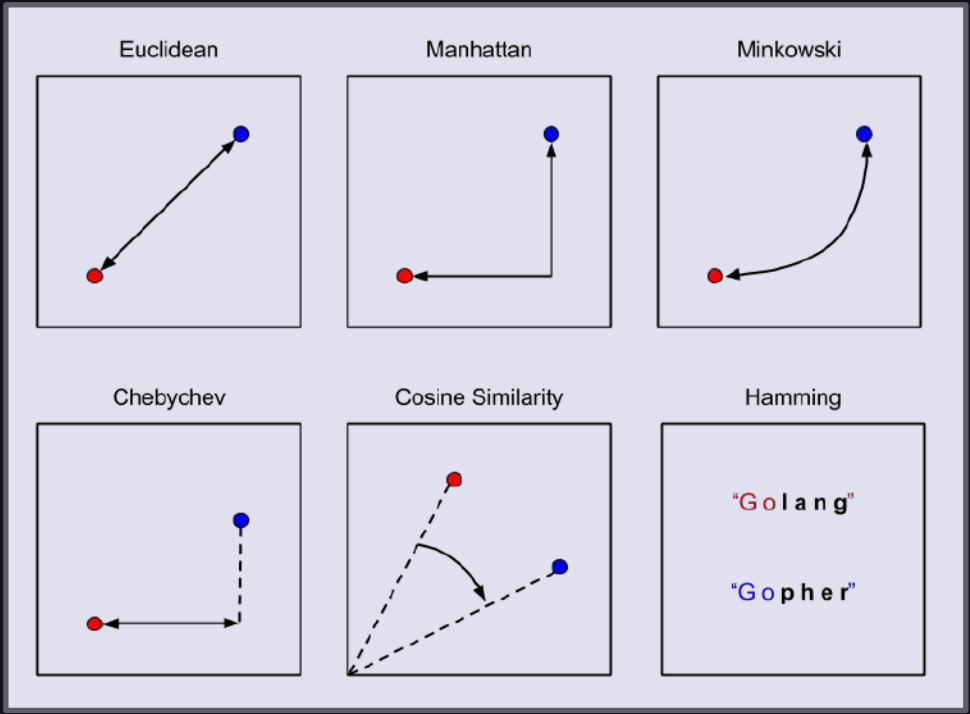
Medidas de Distância

Distância de Chebyshev: É a maior diferença absoluta entre as coordenadas dos pontos em qualquer dimensão.

Distância de Coseno: É uma medida da similaridade entre dois vetores, definida como o cosseno do ângulo entre eles. É útil para dados em espaços de alta dimensão.

Distância de Hamming: É usada para dados categóricos e conta o número de dimensões nas quais os pontos diferem.

Medidas de Distância



Métricas de Algoritmos de Clusterização

Índice de Silhueta: Avalia a coesão intra-cluster e a separação inter-cluster. Uma pontuação alta indica que os pontos estão bem agrupados dentro de seus clusters e mal conectados aos clusters vizinhos.

Índice Davies-Bouldin (DB): Mede a dispersão dentro de cada cluster em relação à separação entre clusters. Quanto menor o valor, melhor a separação entre os clusters.

Índice de Calinski-Harabasz (CH): Calcula a relação entre a dispersão intra-cluster e a dispersão entre clusters. Pontua mais alto para clusters densos e bem separados.

Métricas de Algoritmos de Clusterização

Índice Dunn: É a razão entre a menor distância inter-cluster e a maior distância intra-cluster. Pontua mais alto para clusters mais compactos e mais distantes uns dos outros.

Índice Rand Adjustado (ARI): Comparação entre os rótulos atribuídos pelos algoritmos de clusterização e os rótulos verdadeiros, quando disponíveis.

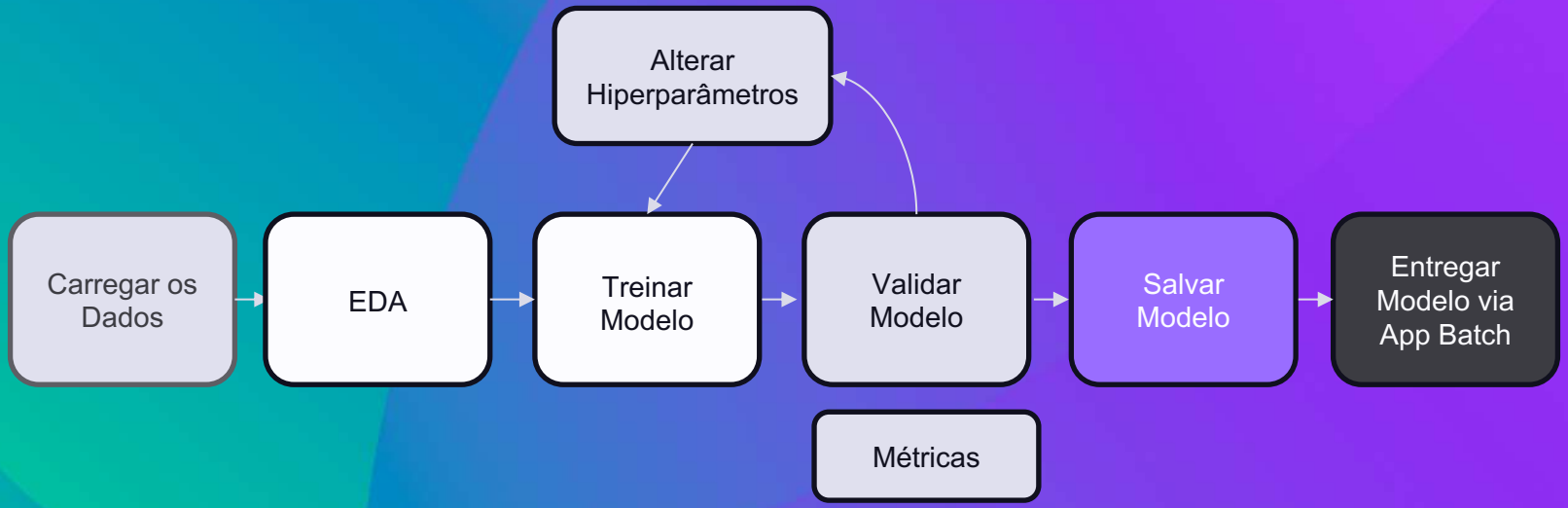
Índice de Validade Interna (IV): Consiste em várias medidas internas, como compacidade e separação dos clusters, para avaliar a qualidade da clusterização sem rótulos verdadeiros.

Projeto – K-Means

Uma empresa de concessão de crédito para pequenas empresas possui um catálogo de seus clientes (PJ) com informações como idade da empresa, faturamento mensal, nível de inovação, entre outras. E para que esta empresa de concessão possa dar um atendimento mais apropriado para cada tipo de cliente, eles desejam agrupar estes clientes conforme estas características.

Desta forma, para que seja possível classificar novos clientes, iremos construir um algoritmo de clusterização que agrupe os clientes em segmentos, com base nas informações disponíveis sobre o mesmo.

Estrutura do Projeto



Code Time ...

rocketseat.com.br