Um algoritmo de clusterização

É uma técnica de aprendizado de máquina e mineração de dados que agrupa um conjunto de dados em clusters ou grupos com base em suas similaridades. Esses algoritmos são usados para identificar padrões nos dados e organizar as informações em grupos significativos. Alguns exemplos de uso incluem segmentação de mercado, análise de redes sociais e agrupamento de documentos. Essas aplicações permitem direcionar estratégias de marketing, personalizar conteúdo e facilitar a organização e recuperação de informações.

Kmeans:

O Que É O Algoritmo K-Means: O algoritmo do K-Means é um método de clusterização amplamente utilizado em análise de dados e aprendizado de máquina. Ele agrupa dados em k-clusters, onde k é definido pelos cientistas de dados. O algoritmo começa definindo centroides aleatórios para cada cluster e, em seguida, atribui pontos a cada cluster com base em uma medida de distância, geralmente a distância euclidiana. Os centróides são recalculados como a média dos pontos atribuídos a cada cluster, e esse processo de atribuição e recálculo é repetido até que não haja mudanças significativas ou um número máximo de iterações seja alcançado. O resultado final pode variar dependendo da inicialização aleatória dos centroides e do número de clusters escolhidos.

Como Definir a Quantidade de Clusters: Existem diferentes métodos para definir o número ideal de clusters em um algoritmo de caminhos. O método do cotovelo envolve plotar a função de custo em relação ao número de clusters e observar quando ocorre uma mudança significativa na inclinação da curva. O método da silhueta avalia a qualidade dos clusters formados por diferentes valores de K e escolhe o valor que maximiza a média da silhueta. O método Gap Statistics compara a dispersão intra-cluster para diferentes valores de K e escolhe o número ideal de clusters que maximiza a lacuna estatística. A validação externa utiliza informações externas sobre os dados para determinar o número correto de clusters. O conhecimento de domínio também pode ser usado para determinar o número de clusters de forma mais precisa.

Medidas de Distância: Nesta aula, discutimos diferentes medidas de distância que nos permitem comparar objetos e itens em um plano multidimensional. A distância Euclidiana é a mais comum e simples, calculada como a raiz quadrada da soma dos quadrados das diferenças entre as coordenadas dos pontos em cada dimensão. A distância de Manhattan é a soma das diferenças absolutas entre as coordenadas dos pontos em cada dimensão. A distância de Minkowski é uma generalização das distâncias Euclidiana e Manhattan. Também exploramos a distância de Chebyshev, a distância de Cosseno e a distância de Hamming, cada uma com suas próprias características e aplicações.

Métricas de Algoritmos de Clusterização: As métricas dos algoritmos de clusterização são importantes para avaliar a qualidade dos resultados. O índice de silhueta mede a coesão intra-cluster e a separação inter-cluster, indicando o quão bem os pontos estão agrupados dentro de seus clusters e separados dos outros clusters. O índice Davies-Bouldin mede a dispersão dentro de cada cluster em relação à separação entre clusters. O índice Calinski-Harabasz calcula a relação entre a dispersão intra-cluster e a dispersão entre clusters, pontuando mais alto para clusters densos e bem separados. O índice Dunn mede a razão entre a menor distância inter-cluster e a maior distância intra-cluster, pontuando mais alto para clusters compactos e distantes uns dos outros. O índice Range Ajustado compara os rótulos atribuídos pelos algoritmos de classificação com

os rótulos verdadeiros, sendo útil quando se deseja confrontar informações supervisionadas com algoritmos não supervisionados. O índice de Validade Interna consiste em várias medidas internas, como compacidade e separação dos clusters, para avaliar a qualidade da clusterização sem rótulos verdadeiros.

Clusterização Hierárquica:

A **clusterização hierárquica** é uma técnica de agrupamento no aprendizado de máquina não supervisionado que se destaca pela estrutura de árvore dos clusters e pela flexibilidade no número de clusters. É útil para análises exploratórias, especialmente em situações onde a relação entre os dados é mais relevante do que a formação de grupos distintos. Não requer um número de clusters definido previamente e é sensível a mudanças nos dados. Comparado ao K-Means, pode identificar estruturas complexas, mas é menos eficiente computacionalmente.

Para realizar um algoritmo de classificação hierárquica, é necessário definir a distância entre os objetos, construir uma matriz de distância, criar um dendrograma e escolher o número de clusters cortando o dendrograma em uma certa altura. A métrica de similaridade é baseada na distância escolhida, como a euclidiana ou manhattan. O dendrograma pode ser construído de forma aglomerativa, unindo clusters próximos, ou divisiva, dividindo um único cluster em dois até que todos os objetos tenham seu próprio cluster.

O algoritmo de classificação hierárquica divisivo trabalha de cima para baixo, dividindo um cluster em sub-clusters com base em um ponto de corte na matriz de distância. Esse processo é repetido até que cada ponto seja seu próprio cluster ou um número desejado de clusters seja atingido. Pode-se usar algoritmos como K-Means ou DBSCAN para essa divisão. Um dendrograma pode ser construído para ilustrar a progressão da divisão dos clusters.

O dendrograma é uma representação visual de agrupamentos hierárquicos de dados. Ele é composto por eixos x e y, nós que dividem clusters e ramos que subdividem os nós. Cortes no dendrograma determinam o número de clusters. O dendrograma é útil para análise exploratória, identificação de outliers e determinação do número de clusters em um processo de agrupamento. A estrutura do dendrograma ajuda a entender a estrutura dos dados e a visualizar a formação dos grupos.

Redução de Dimensionalidade

Os algoritmos de redução de dimensionalidade constituem uma classe de técnicas matemáticas aplicadas na análise de dados, cujo propósito fundamental é a simplificação dos dados ao reduzir o número de variáveis envolvidas. Essa redução é especialmente valiosa em contextos onde os dados apresentam alta complexidade dimensional, o que não apenas dificulta a análise visual e estatística, mas também aumenta o custo computacional e pode degradar o desempenho de algoritmos de aprendizado de máquina devido ao fenômeno conhecido como "maldição da dimensionalidade". Na prática, esses algoritmos trabalham transformando um grande conjunto de variáveis em um menor, preservando tanto quanto possível as informações essenciais. Este processo é realizado através da identificação de padrões, correlações e estruturas fundamentais nos dados originais.
Principais objetivos e benefícios:

- Redução de Complexidade: Dados de alta dimensão podem ser complexos de analisar e visualizar. A redução de dimensionalidade ajuda a simplificar esses dados para facilitar a interpretação.
- Eliminação de Ruído: Ao focar em componentes principais ou características principais, a redução de dimensionalidade pode ajudar a eliminar o ruído, destacando apenas as características mais significativas dos dados.
- Eficiência Computacional: Dados com menos dimensões requerem menos recursos computacionais para processamento. Isso é crucial em casos de aprendizado de máquina e análise de grandes volumes de dados, onde o tempo de processamento e a capacidade de memória podem ser limitantes
- Melhoria no Desempenho de Algoritmos: Muitos algoritmos de aprendizado de máquina têm seu desempenho prejudicado pela "maldição da dimensionalidade", que é quando o aumento no número de dimensões leva a um espaçamento maior entre os pontos de dados.
- Reduzindo a dimensionalidade, pode-se melhorar a acurácia e eficácia desses algoritmos.
- Visualização de Dados: É difícil visualizar dados com muitas dimensões diretamente. A redução para duas ou três dimensões permite o uso de gráficos bidimensionais ou tridimensionais para explorar e comunicar os dados de forma efetiva.
- Descoberta de Estruturas Subjacentes: A redução de dimensionalidade pode revelar estruturas ocultas nos dados, que não são aparentes em uma análise de alta dimensão. Isso pode incluir agrupamentos ou padrões correlacionados

As métricas para avaliar algoritmos de redução de dimensionalidade incluem erro de reconstrução, coeficiente de correlação de distâncias, taxa de compressão de dados e precisão na classificação. O erro de reconstrução compara os dados originais com os reconstruídos. O coeficiente de correlação de distâncias avalia a preservação das relações de distância. A taxa de compressão indica a eficiência na redução de dimensionalidade, mas não garante qualidade na reconstrução. A precisão na classificação é útil em problemas de aprendizado supervisionado para comparar dados originais e reduzidos.

PCA:

Transforma os dados para um novo sistema de coordenadas, reduzindo a dimensionalidade ao escolher os primeiros componentes principais que capturam a maior variância possível. Antes de aplicar o PCA, é essencial normalizar os dados. A matriz de covariância é calculada a partir dos dados normalizados. Os autovetores representam as direções dos novos eixos com maior variância. Os componentes principais são selecionados com base nos autovetores com os maiores autovalores. Por fim, os dados originais são projetados nos autovetores selecionados, resultando em um novo espaço com menos dimensões.

O SVD (Singular Value Decomposition) está intimamente ligado ao PCA (Principal Component Analysis), pois o SVD fornece uma abordagem computacional para realizar o PCA. O SVD ajuda a desmontar uma pilha de livros (dados) em três pilhas menores, mostrando direções, importância e relação dos padrões. O PCA encontra os melhores ângulos para capturar informações dos dados, sendo esses ângulos os componentes principais. Ao aplicar o SVD aos dados centrados, obtemos os mesmos componentes principais buscados pelo PCA.

T-SNE:

O t-SNE é um algoritmo poderoso de redução de dimensionalidade, amplamente usado para visualizar conjuntos de dados complexos em 2D ou 3D. Desenvolvido por Lawrence Van Den Maarten e Geoffrey Hinton em 2008, ele captura a estrutura local dos dados, revelando agrupamentos em espaços de baixa dimensão. O algoritmo transforma distâncias entre pontos em probabilidades condicionais, facilitando a identificação de padrões em dados genéticos

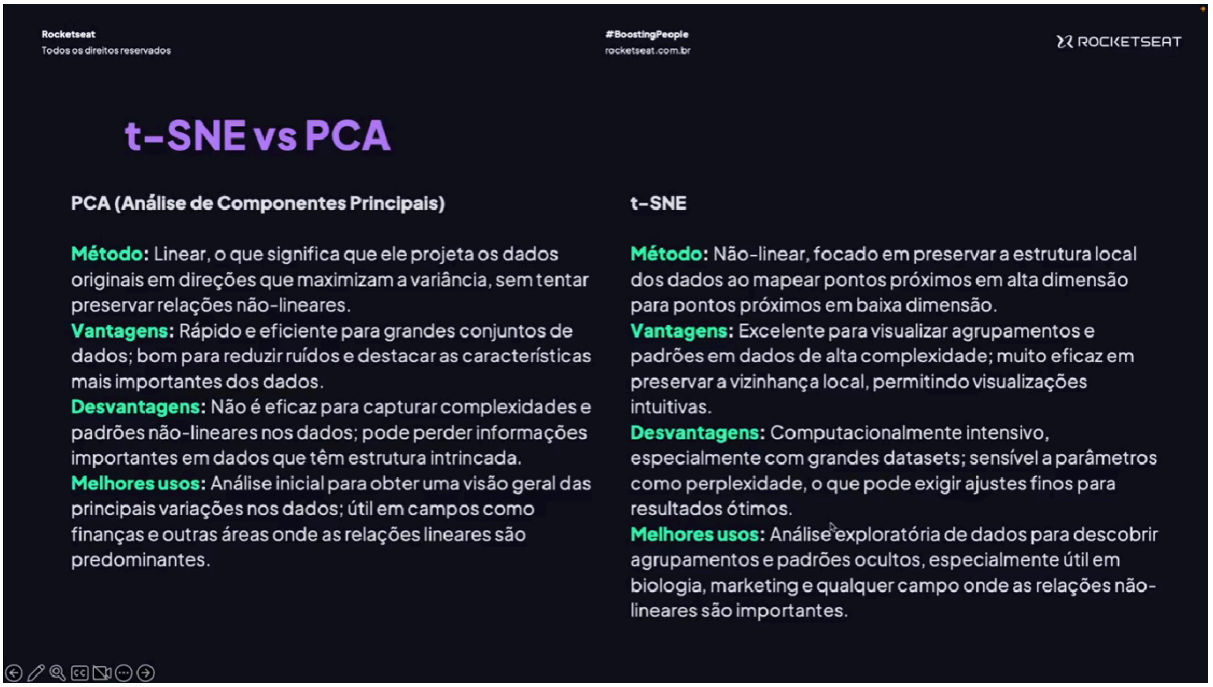
O t-sne começa com a compreensão de que em um espaço de alta dimensão, cada ponto de dados pode ser visto como um ponto em um espaço multidimensional. O algoritmo transforma as distâncias entre os pontos em probabilidades condicionais que representam similaridade. A ideia é que pontos que são próximos uns dos outros têm uma alta probabilidade de serem ‘vizinhos’ e aqueles que estão distantes têm uma probabilidade baixa.

O algoritmo t-SNE consiste em cinco etapas. Primeiramente, mede-se as similaridades entre os pontos de dados em um espaço de alta dimensão. Em seguida, cria-se um mapa em um espaço menor, mantendo as amizades o mais fiel possível. Posteriormente, compara-se as similaridades nos espaços original e reduzido, ajustando as posições para preservar as amizades. Por fim, otimiza-se as posições até obter uma configuração onde as relações sejam bem representadas. Essa técnica facilita a visualização de padrões em dados de alta dimensão.

O t-SNE é uma técnica que converte distâncias euclidianas em probabilidades para preservar a estrutura local dos dados. Ele usa distribuições gaussianas e t de Student para mapear de alta para baixa dimensão, lidando com a maldição da dimensionalidade. A otimização por gradiente descendente minimiza a diferença entre as probabilidades das duas dimensões. A divergência de Kullback-Leibler mede o quão bem o espaço de baixa dimensão representa o de alta dimensão.

O t-SNE apresenta desafios e limitações, como a sensibilidade aos hiperparâmetros, como perplexity, que afeta a aparência dos gráficos. A interpretação de clusters pode ser limitada, pois o t-SNE pode criar clusters de forma exagerada. Além disso, o custo computacional pode ser alto, especialmente para conjuntos de dados grandes, devido à comparação de todos os pares de pontos. É importante ter cautela ao interpretar os resultados do t-SNE.

O hiperparâmetro perplexity no t-SNE é comparado a decidir com quantas pessoas conversar em uma festa, influenciando quantos vizinhos cada ponto considera. Uma perplexidade alta pode perder detalhes, enquanto uma baixa pode criar grupos isolados. Não há um valor ideal universal, sendo necessário experimentar diferentes valores para obter a melhor representação dos dados.



Regras de Associação

Regras de associação são uma técnica de mineração de dados utilizada para descobrirrelações significativas, frequentes e úteis entre conjuntos de itens em grandes bases de dados transacionais. Elas são especialmente úteis para identificar padrões ocultos que podem ajudar na tomada de decisões em várias áreas, como marketing, varejo, saúde e finanças.

As regras de associação são expressas na forma de implicações, onde a presença de um conjunto de itens em uma transação implica a presença de outro conjunto de itens. Formalmente, uma regra de associação é representada com:

(A ->B), onde (A) e (B) são conjuntos de itens, e a regra indica que as transações que contêm (A) tendem a também conter (B).

Antecedente (A): O conjunto de itens que aparece antes da implicação. Também chamado de premissa ou corpo da regra.

Consequente (B): O conjunto de itens que aparece após a implicação. Também chamado de conclusão ou cabeça da regra.

Fundamentos Item: Um objeto ou entidade de interesse. Por exemplo, em uma base de dados de supermercado, um item pode ser “leite” ou “pão”.

Itemset: Um conjunto de itens. Por exemplo, {leite, pão} é um itemset.

Transações:Uma transação é uma coleção de itens comprados ou ocorridos juntos. Cada transação é representada por um itemset. Exemplo: Uma transação pode ser {leite, pão, manteiga}.

Banco de Dados Transacional: Uma coleção de transações. Cada linha na base de dados representa uma transação.

Principais Objetivos e Benefícios Identificação de Relacionamentos Significativos:

Descobrir relações úteis entre itens que podem não ser evidentes à primeira vista. Isso ajuda as empresas a entender melhor os hábitos de compra dos clientes e a desenvolver estratégias mais eficazes.

Aumento das Vendas e Receita: Utilizar as associações descobertas para criar promoções cruzadas e pacotes de produtos que incentivem os clientes a comprar mais itens juntos, aumentando assim as vendas e a receita.

Otimização do Layout das Lojas: Analisar padrões de compra para otimizar a disposição dos produtos nas lojas físicas, facilitando a navegação dos clientes e aumentando a probabilidade de compras impulsivasPersonalização de Ofertas e

Campanhas: Segmentar os clientes com base nos padrões de compra e personalizar ofertas e campanhas de marketing, aumentando a taxa de conversão e a satisfação do cliente.

Detecção de Fraudes: Identificar padrões anômalos que possam indicar atividades fraudulentas, ajudando as instituições financeiras a prevenir e mitigar fraudes.

Melhoria na Gestão de Estoques: Prever a demanda por produtos com base nos padrões de compra, permitindo uma melhor gestão de estoques e redução de custos com armazenamento e desperdício

Métricas de Algoritmos de Regras de Associação

Suporte: Mede a frequência que um itemset aparece na base de dados

$$Suporte X = N\acute{u}mero de Transa\c{c}\tilde{o}es que tenham X / N\acute{u}mero Total de Transa\c{c}\tilde{o}es$$

Importância: Identifica itemsets frequentes que são relevantes para a geração de regras de associação. Um suporte baixo pode indicar que o padrão é raro e, portanto, menos interessante.

Lift: Mede a importância da regra de associação, comparando a confiança da regra com a frequência esperada dos itens se fossem independentes.

$$Lift A \Rightarrow B = Confian\c{c}a(A \Rightarrow B) \,/\, Suporte(B)$$

Maior que 1: Indica uma associação positiva, (os itens ocorrem juntos mais frequentemente do que o esperado) Igual a 1: Indica que os itens são independentes. Menor que 1: indica uma associação negativa (os itens ocorrem juntos menos frequentemente do que o esperado)

Convicção: Avalia a relação entre a confiança da regra e a expectativa de não ocorrência do consequente, considerando o suporte.

$$Convic\tilde{c}ao A \Rightarrow B = 1 - Suporte(B) \,/\, 1 - Confian\c{c}a(A \Rightarrow B)$$

Maior que 1: Indica uma dependência positiva, ou seja, B é mais provável de ocorrer com A do que por acaso. Igual a 1: Indica que A e B são independentes. Menor que 1: Indica uma dependência negativa, ou seja, B é menos provável de ocorrer com A do que por acaso.

Leverage: Mede a diferença entre a frequência observada de (A) e (B) ocorrendo juntos e a frequência esperada se (A) e (B) fossem independentes

$$Leverage A \Rightarrow B = Suporte A \cap B - (Suporte A \times Suporte B)$$

Positiva: A probabilidade de compra de A aumenta a probabilidade de compra de B, mais do que seria esperado por acaso Zero: Eles ocorrem juntos com a mesma frequência que ocorreria por acaso. Negativa: A probabilidade de compra de A diminui a probabilidade de compra de B, mais do que seria esperado por acaso

Use Lift para uma comparação relativa que mede a força da associação em relação à independência dos itens. Ideal para avaliar a relevância prática das regras.

Use Convicção Para medir a dependência entre itens, especialmente quando lida com itens raros e deseja uma medida da certeza da ocorrência do consequente. Ideal para regras com alta confiança.

Use Leverage para uma medida direta da diferença entre a ocorrência observada e a esperada. Ideal para dados balanceados.

Apriori:

Identifica itemsets frequentes gerando candidatos e utilizando a propriedade anti monotônica, que afirma que se um itemseté frequente, todos os seus subconjuntos também são frequentes

A motivação para desenvolver o Apriorie outros algoritmos de regras de associação foi a necessidade de encontrar padrões frequentes em bases de dados transacionais. O exemplo clássico usado para ilustrar esses algoritmos é a análise de cestas de compras (marketbasketanalysis), onde o objetivo é descobrir associações entre produtos comprados juntos pelos clientes

Como funciona o algoritmo Apriori Reunir Dados Iniciais:

Imagine que temos uma lista de compras de vários clientes. Cada lista de compras é chamada de “transação”. O primeiro passo é olhar para todos os itens únicos comprados.

Encontrar Itens Populares: Primeiro, contamos quantas vezes cada item individual aparece nas listas de compras. Por exemplo, se leite aparece em 3 listas de 5, ele é comprado em 60% das vezes.

Filtrar Itens Frequentes: Definimos um limite mínimo de popularidade, chamado “suporte mínimo”. Só consideramos os itens que atingem ou superam esse limite. Por exemplo, se decidirmos que um item precisa estar em pelo menos 50% das listas, mantemos apenas os itens que atendem a esse critério.

Criar Combinações de Itens: Agora, olhamos para todas as combinações de dois itens a partir dos itens populares. Por exemplo, combinamos “leite e pão”, “leite e manteiga”, e assim por diante.

Contar Combinações Populares: Contamos quantas vezes cada combinação de dois itens aparece nas listas de compras. Por exemplo, se “leite e pão” aparecem juntos em 40% das listas, essa combinação é mantida.

Repetir com Combinações Maiores: Repetimos o processo, agora olhando para combinações de três itens a partir das combinações populares de dois itens. Continuamos aumentando o número de itens na combinação e contando quantas vezes aparecem juntos, até que não haja mais combinações populares.

Criar Regras de Associação: Com as combinações populares, criamos regras. Por exemplo, se “leite e pão” são populares juntos, podemos criar a regra: “Se alguém compra leite, também compra pão”. Calculamos a “confiança” dessas regras, que é a frequência com que a combinação ocorre em relação a um dos itens. Por exemplo, se toda vez que alguém compra leite, eles também compram pão, a confiança é alta

Tópicos Complementares

Os Gaussian Mixed Models (GMM) são modelos probabilísticos usados para agrupar dados, assumindo uma mistura de distribuições gaussianas. Cada cluster é uma distribuição gaussiana e o GMM é uma combinação ponderada delas. Comparado ao algoritmo K-means, o GMM é uma versão probabilística que não assume clusters esféricos de tamanho similar, como o K-means. Enquanto o K-means é um método de clusterização rígido, onde cada ponto de dados pertence a apenas um cluster.

O GMM é utilizado em diversos casos, como agrupamento de dados, modelagem de distribuições complexas, segmentação de imagens, reconhecimento de padrões e fala, e detecção de anomalias. Ele agrupa dados em subconjuntos homogêneos modelados por distribuições gaussianas diferentes, sendo útil para representar distribuições multimodais ou complexas. Na segmentação de imagens, os GMMs ajudam a identificar regiões com características semelhantes. No reconhecimento de fala, modelam características acústicas de fonemas, e na detecção de anomalias, identificam comportamentos raros nos dados.
O algoritmo Gaussian Mixture Model (GMM) é explicado passo a passo. Inicialmente, é necessário definir a quantidade de clusters, podendo ser baseado no conhecimento do problema ou utilizando métricas como o Bayesian Information Criteria (BIC). Em seguida, são inicializadas as médias, covariâncias e pesos de forma aleatória. O algoritmo Expectation Maximization (EM) é utilizado para ajustar os parâmetros das distribuições gaussianas, visando melhor modelar os dados observados. A convergência do modelo é alcançada quando os parâmetros não variam significativamente entre iterações. O GMM oferece flexibilidade ao fornecer não apenas os clusters, mas também a probabilidade de pertencimento a cada um deles.

Detecção de Anomalias:

Processo de identificar padrões ou valores que se desviam significativamente do comportamento normal ou esperado em um conjunto de dados. Tal processo é essencial para garantir a confiabilidade e a qualidade dos dados.

Algoritmo LOF (Local Outlier Factor): Técnica de aprendizado de máquina não supervisionado usada para detectar anomalias em conjuntos de dados. Compara a densidade de cada ponto dado com a densidade dos seus vizinhos mais próximos. Se um ponto se encontra em uma região com uma densidade significativamente menor que a dos seus vizinhos, é considerado uma anomalia.