JOSÉ AHIRTON BATISTA LOPES FILHO

PÓS TECH

MACHINE LEARNING ENGINEERING

APRENDIZADO NÃO SUPERVISIONADO

AULA 02

SUMÁRIO

O QUE VEM POR AÍ?

Nesta aula sobre clustering, veremos como essa aplicação específica em Aprendizado Não Supervisionado tem emergido como uma ferramenta essencial na análise de dados, permitindo a identificação de padrões e estruturas subjacentes em conjuntos de dados complexos.

Ainda demonstraremos o funcionamento do algoritmo K-Means, um dos métodos mais populares de clustering, o qual destaca-se pela sua simplicidade e eficácia em dividir os dados em grupos distintos com base na similaridade. No entanto, uma variedade de outras técnicas, como o EM e o clustering hierárquico, também desempenham papéis importantes na resolução de problemas de clustering em diferentes domínios.

HANDS ON

No Hands On desta aula, entenderemos melhor como o K-Means pode ser utilizado em um fluxo de análise de negócio. No exemplo, utilizaremos o clustering via K-Means como uma ferramenta poderosa, de modo a explorar padrões e tendências no vasto universo cinematográfico. Utilizando o conjunto de dados IMDb 5000 Movie Dataset, (que contém uma riqueza de atributos tais como diretor, duração, orçamento, pontuação IMDb e mais) poderemos aplicar o algoritmo K-Means para identificar grupos semelhantes de filmes.

Por exemplo: podemos criar clusters com base em gênero, orçamento e popularidade. Isso nos permite descobrir padrões interessantes, como agrupamentos de filmes de baixo orçamento com altas pontuações IMDb em determinados gêneros. Com essas informações, os estúdios de cinema podem ajustar suas estratégias de produção e marketing para melhor atender às preferências do público. Em suma, o clustering via K-Means oferece uma abordagem analítica poderosa para desvendar insights valiosos no mundo do cinema e além.

```
import pandas as pd

import numpy as np

dataset = pd.read_csv('imdb_movie_dataset.csv', encoding='utf-8')

dataset.drop(['color','language'], axis= 1, inplace= True)


from sklearn.cluster import KMeans


# Excluindo os valores ausentes e selecionando apenas a pontuação GOB e IMDB

selected_dataset=dataset.loc[

(dataset['GOB']>0) &
dataset['imdb_score']>0][['imdb_score','GOB']]
```

```
# Agrupando o conjunto de dados usando o algoritmo K-Means
cls = KMeans(n_clusters=2)

# Ajustar o modelo ao algoritmo
cls.fit(selected_dataset)

# Trazendo os centróides e o rótulo de cada grupo
centroids=cls.cluster_centers_
labels = cls.labels_
```

Código-fonte 1 – Demonstração 2 – Criando uma aplicação de análise de tendências via Clustering K-Means
Fonte: Elaborado pelo autor (2024)

SAIBA MAIS

O Aprendizado de Máquina pode ser dividido em dois grandes paradigmas, Supervisionado e Não Supervisionado, e, embora ambas as abordagens tenham como objetivo extrair informações úteis dos dados, elas diferem fundamentalmente em suas metodologias e objetivos.

No Aprendizado Supervisionado, o algoritmo é treinado em um conjunto de dados rotulados, ou seja, cada exemplo de treinamento possui uma etiqueta associada, que indica a classe ou categoria à qual ele pertence. O objetivo é, então, aprender uma função que mapeia as entradas para as saídas com base nos exemplos de treinamento fornecidos. Isso permite que o algoritmo faça previsões precisas para novos dados não rotulados.

Por outro lado, em Aprendizado Não Supervisionado o algoritmo é treinado em um conjunto de dados não rotulado, em que não há informações sobre as classes ou categorias dos exemplos de treinamento. O objetivo é então descobrir a estrutura intrínseca nos dados, identificando padrões e relações ocultas entre as observações, com base na capacidade de interpretabilidade dos próprios algoritmos.

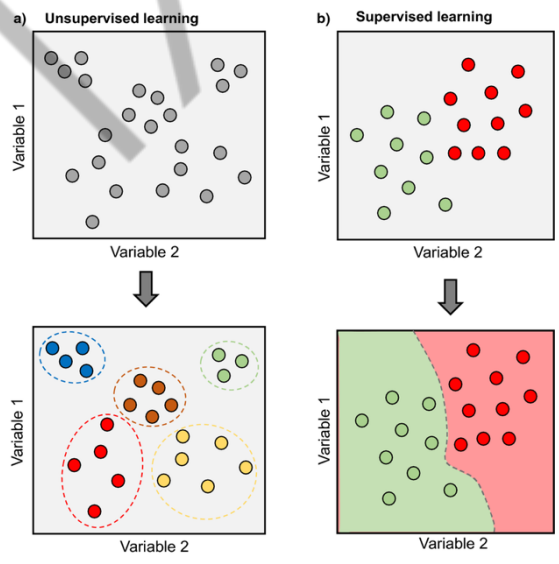

Figura 1 – Visão da diferença entre problemas resolvidos via Aprendizado Supervisionado e Não Supervisionado
Fonte: Morimoto; Ponton (2021)

Um dos principais tipos de tarefa no Aprendizado Supervisionado é a classificação, em que o objetivo principal é atribuir uma classe ou categoria a cada instância de dados. Por exemplo: em um conjunto de dados como o MNIST, que

consiste em imagens de dígitos escritos à mão, a tarefa de classificação pode ser identificar corretamente o dígito representado em cada imagem, atribuindo a ela o rótulo correspondente de '0' a '9'.

Já dois tipos comuns de tarefa em Aprendizado Não Supervisionado são a Estimação de Densidade e o Clustering, ou Agrupamento de Dados. A estimação de densidade e o clustering são duas técnicas fundamentais em Análise de Dados e Aprendizado de Máquina, frequentemente usadas para entender a estrutura subjacente dos dados e identificar padrões significativos.

A estimação de densidade é o processo de inferir a distribuição de probabilidade subjacente dos dados observados. Em outras palavras, busca-se modelar a densidade de probabilidade dos dados em todo o espaço de características. Métodos comuns de estimação de densidade incluem o estimador de densidade de Kernel, histogramas e misturas de gaussianas. A estimação de densidade é útil para entender a distribuição dos dados, identificar outliers e anomalias e para gerar amostras de dados sintéticos.

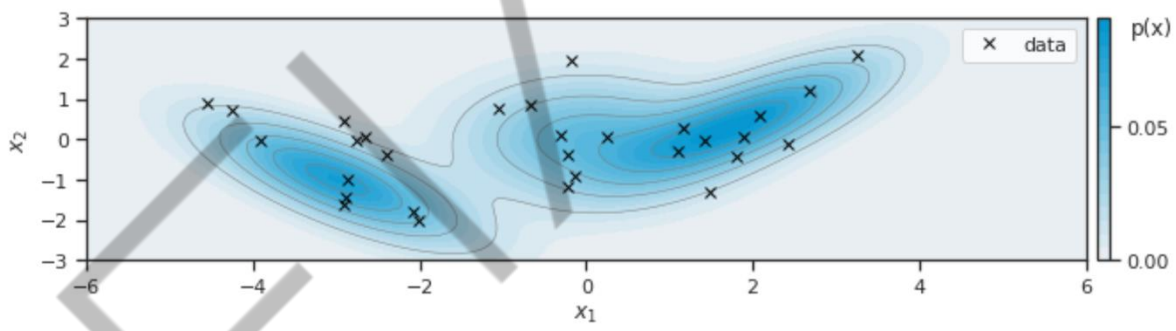

Figura 2 – Exemplo de problema de Estimação de Densidade
Fonte: Graves; Clancy (2019)

Já o clustering é uma técnica essencial em análise de dados, que visa encontrar estruturas intrínsecas em conjuntos de dados não rotulados. Em essência, o clustering busca agrupar objetos similares nos mesmos grupos, enquanto mantém objetos diferentes em clusters distintos. Existem diferentes algoritmos de clustering, como K-Means, Hierarchical Clustering e DBSCAN.

O clustering é amplamente utilizado em várias aplicações, como segmentação de clientes, agrupamento de documentos, análise de redes sociais e reconhecimento de padrões em imagens. Essa técnica tem uma longa história e é considerada uma das principais tarefas em Aprendizado Não Supervisionado.

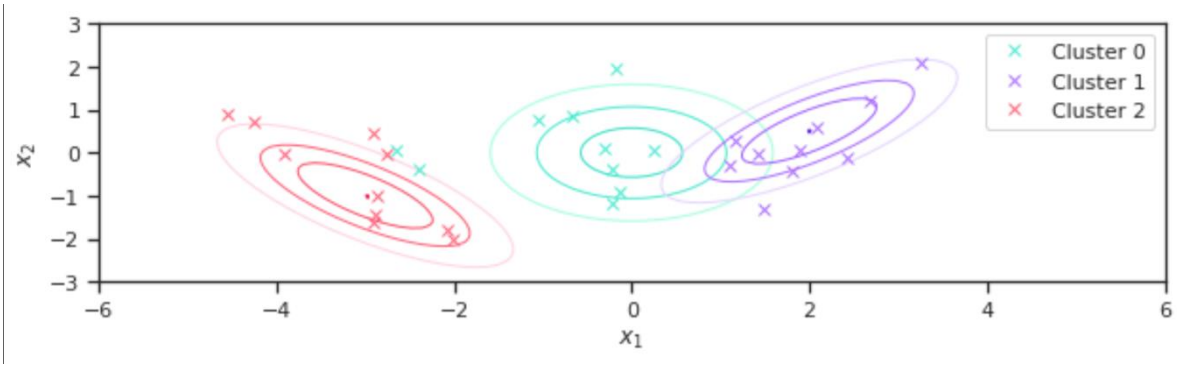

Figura 3 – Exemplo de Clustering
Fonte: Graves; Clancy (2019)

A ideia de clustering remonta a várias décadas atrás, com os primeiros trabalhos documentados datando dos anos 60. Um marco significativo foi o desenvolvimento do algoritmo K-Means por James MacQueen em 1967. Desde então, houve uma proliferação de algoritmos de clustering, cada um com suas próprias características e áreas de aplicação específicas.

O clustering é uma técnica versátil e é aplicada em uma variedade de campos. Na área de marketing, por exemplo, as empresas podem usar clustering para segmentar clientes com base em padrões de compra semelhantes, permitindo estratégias de marketing mais direcionadas e eficazes. Na biologia, o clustering é usado para agrupar genes com padrões de expressão semelhantes, facilitando a identificação de genes relacionados e a compreensão de redes de regulação genética.

A estimação de densidade e o clustering estão intimamente relacionados. De fato, muitos métodos de clustering dependem de estimativas de densidade para identificar a estrutura dos clusters. Por exemplo: o algoritmo DBSCAN usa a estimação de densidade para identificar regiões de alta densidade como clusters, enquanto o algoritmo Gaussian Mixture Model (GMM) combina estimativas de densidade gaussiana para modelar a distribuição dos dados e atribuir pontos a clusters.

No clustering, ao contrário da classificação, o agrupamento não implica necessariamente atribuir rótulos ou classes aos clusters. Em vez disso, visa encontrar estruturas de dados subjacentes que não foram previamente definidas. Por exemplo: em um conjunto de dados com as características altura e peso de pessoas adultas e crianças, a utilização de uma técnica de agrupamento vai identificar os grupos com as características de crianças e adultos, sem a necessidade dessas categorias estarem predefinidas.

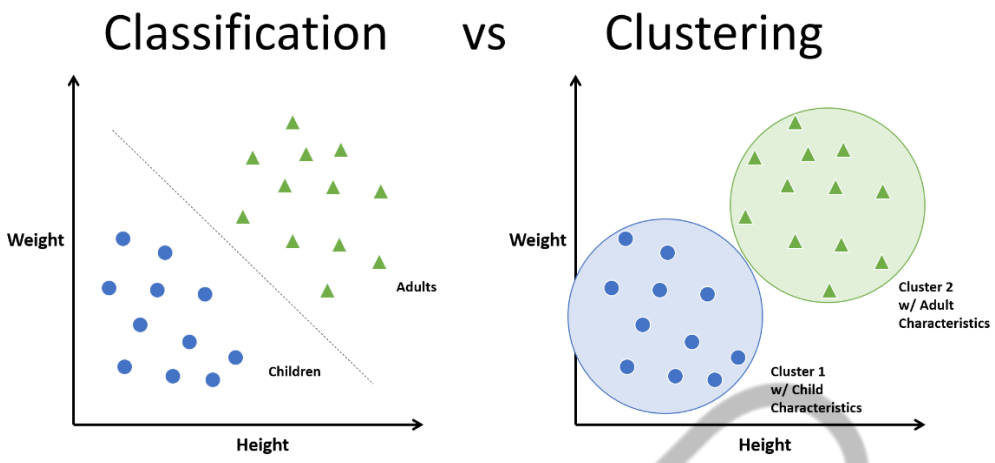

Figura 4 – Exemplo de Clustering
Fonte: Keerthana (2024)

O objetivo principal do clustering é então encontrar estruturas intrínsecas nos dados, agrupando objetos similares e separando objetos dissimilares. Isso pode ser útil para descobrir padrões ocultos nos dados, identificar segmentos de mercado em análises de marketing e agrupar genes relacionados em biologia molecular, entre muitas outras aplicações. Por exemplo: em análises de mercado, empresas podem usar clustering de modo a organizar seus clientes em grupos com características semelhantes, permitindo direcionar campanhas de marketing de forma mais eficaz.

Figura 5 – Exemplo de Clustering
Fonte: ECLOUDVALLEY (2021)

O primeiro conceito chave no clustering é então o de similaridade (ou dissimilaridade) entre os objetos. A similaridade é uma medida que quantifica os quão próximos dois objetos estão um do outro, enquanto a dissimilaridade é o oposto, quantificando o quão diferentes eles são ou estão. Essas medidas podem ser calculadas de várias maneiras, dependendo do tipo de dados e do domínio do problema.

Outro conceito fundamental é o de centroides. Em muitos algoritmos de clustering, como o K-Means, cada cluster é representado por um ponto central, chamado de centroide. O centroide é uma representação média dos pontos no cluster e é usado para calcular a proximidade dos outros pontos ao cluster. Além disso, é importante entender o conceito de espaço de características. Cada objeto em um conjunto de dados é representado por um vetor de características em um espaço de características multidimensional. O clustering é realizado neste espaço de características, em que a proximidade entre os objetos é calculada com base nas características que eles compartilham.

Fora esses, outro aspecto crucial do clustering é a definição do número de clusters. Este é um parâmetro importante que precisa ser especificado antes da aplicação do algoritmo de clustering. O número ideal de clusters pode variar dependendo do domínio do problema e dos objetivos da análise. Por exemplo: o funcionamento básico do algoritmo K-Means envolve a definição de um número pré-determinado de clusters, denotado por 'K'.

Inicialmente esse algoritmo seleciona aleatoriamente 'K' pontos como centroides iniciais, que servirão como representantes de cada cluster. Em seguida, cada ponto de dados é atribuído ao cluster cujo centroide é o mais próximo, com base em uma medida de distância, geralmente a distância euclidiana.

Uma vez que todos os pontos de dados tenham sido atribuídos a clusters, os centroides são recalculados como a média de todos os pontos atribuídos ao cluster. Este processo de atribuição e atualização dos centroides é repetido iterativamente até que a convergência seja alcançada, ou seja, até que não ocorram mais alterações significativas nos clusters ou nos centroides (também conhecido como critério de parada do algoritmo).

O algoritmo K-Means é conhecido por sua simplicidade e eficiência computacional, o que o torna uma escolha popular para muitas aplicações de clustering. No entanto, é importante destacar que o sucesso do K-Means depende fortemente da escolha adequada do número de clusters 'K' e da inicialização dos centroides.

Nesta aula, também aprenderemos mais sobre o Método do Cotovelo (Elbow Method), uma abordagem comum tendo em vista determinar o número ideal de

clusters, o qual envolve plotar a variabilidade explicada em função do número de clusters e identificar o ponto no qual a curva começa a se nivelar, indicando o número ideal de clusters. Clustering é então uma técnica poderosa de análise de dados com uma ampla gama de aplicações em diversos domínios.

Hoje em dia o clustering pode ser encontrado em aplicações avançadas nos mais diversos problemas, como segmentação de clientes, detecção de anomalias e processamento de imagens. Como comentado no início dessa seção, uma das aplicações mais comuns de clustering é a segmentação de clientes em marketing.

Ao se agrupar clientes com base em comportamentos de compra semelhantes, as empresas podem personalizar suas estratégias de marketing e oferecer produtos e serviços mais relevantes para cada perfil de consumidor. O clustering também pode ser usado para identificar grupos de clientes com maior propensão a saírem de uma base de dados ou deixarem de consumir (churn) ou para recomendar produtos com base nos padrões de compra.

Outra aplicação importante é a detecção de anomalias em conjuntos de dados. Ao identificar padrões incomuns, ou outliers, o clustering pode ajudar a detectar fraudes em transações financeiras, falhas em sistemas de monitoramento de saúde ou comportamentos suspeitos em redes de computadores. Algoritmos como o DBSCAN são especialmente úteis nesse contexto, pois podem identificar grupos de pontos de dados densos, enquanto os outliers são isolados em clusters menores.

No campo do processamento de imagens, o clustering é amplamente utilizado para segmentar e categorizar imagens com base em características visuais. Por exemplo, ele pode ser usado para agrupar pixels de uma imagem em regiões semelhantes, facilitando a identificação de objetos e padrões. Algoritmos de clustering como Spectral Clustering são frequentemente empregados nesse contexto, pois podem lidar com a alta dimensionalidade dos dados de imagem e capturar relações espaciais complexas.

O uso de técnicas de clustering em projetos do mundo real vem com uma série de desafios e considerações práticas que profissionais de dados devem enfrentar. Um dos desafios mais importantes é a escolha adequada de métricas de distância para medir a similaridade entre os pontos de dados. Outro desafio comum é lidar com dados ausentes durante o processo de clustering.

A presença de valores ausentes pode distorcer a estrutura de agrupamento e levar a resultados não representativos. É importante implementar estratégias robustas para lidar com dados ausentes, como imputação de valores, exclusão de observações ou utilização de algoritmos de clustering que sejam robustos à presença de dados ausentes.

Um passo crucial na implementação bem-sucedida de clustering é o entendimento do domínio do problema. Isso inclui compreender as características dos dados, os padrões de comportamento esperados e as necessidades específicas de stakeholders. Uma compreensão sólida do domínio ajuda na seleção adequada de algoritmos de clustering e na interpretação dos resultados.

A avaliação adequada do desempenho do algoritmo de clustering é essencial para garantir resultados confiáveis e úteis. Métodos de validação interna, como o método do cotovelo e o índice de silhueta, podem ser utilizados para determinar o número ideal de clusters e avaliar a qualidade dos agrupamentos. Além disso, é importante validar os resultados do clustering com base em conhecimento especializado e interpretação visual dos dados.

Em projetos do mundo real, a escalabilidade e a eficiência computacional do algoritmo de clustering são aspectos críticos a serem considerados, especialmente ao lidar com grandes volumes de dados. A seleção de algoritmos e implementações eficientes pode reduzir significativamente o tempo de processamento e os recursos computacionais necessários para executar o clustering em grandes conjuntos de dados.

Em resumo, ao aplicar técnicas de clustering em projetos do mundo real, é importante estar ciente dos desafios comuns enfrentados e considerar cuidadosamente as implicações práticas de cada decisão tomada durante o processo de implementação. Com uma abordagem cautelosa e orientada pelo domínio, é possível obter insights valiosos e significativos por meio do clustering em uma variedade de contextos e aplicações. Ao longo desta aula vamos explorar os principais conceitos, técnicas e aplicações do clustering, examinando também os desafios comuns enfrentados ao aplicar técnicas de clustering e considerações práticas para uma implementação bem-sucedida em projetos do mundo real.

Livros

"Pattern Recognition and Machine Learning", de Christopher M. Bishop: este livro aborda uma ampla gama de tópicos em Reconhecimento de Padrões e Aprendizado de Máquina, incluindo técnicas de Aprendizado Não Supervisionado, como clustering e redução de dimensionalidade.

"Introduction to Data Mining", de Pang-Ning Tan, Michael Steinbach e Vipin Kumar: esta obra fornece uma introdução abrangente à Mineração de Dados, cobrindo tanto aspectos supervisionados quanto não supervisionados, com ênfase em técnicas como clustering e criação de regras de associação.

Artigos Científicos

"Some Methods for classification and Analysis of Multivariate Observations", de James MacQueen: este artigo apresenta métodos fundamentais para a análise de observações multivariadas, sendo parte do que veio a ser a abordagem pioneira para clustering conhecida como algoritmo K-means. MacQueen propõe uma técnica iterativa para particionar um conjunto de dados em clusters que se tornou uma base importante para muitos algoritmos de clustering subsequentes.

"Algorithm AS 136: A K-Means Clustering Algorithm**",** de John A. Hartigan e Manchek A. Wong: neste artigo, Hartigan e Wong descrevem o algoritmo K-means, que é amplamente utilizado para clustering em conjuntos de dados não rotulados. O artigo fornece uma descrição detalhada do algoritmo e sua implementação eficiente, tornando-se uma referência fundamental na área de clustering.

"Finding Groups in Data: An Introduction to Cluster Analysis" de Leonard Kaufman e Peter J. Rousseeuw: este livro oferece uma introdução abrangente à análise de cluster, abordando diferentes métodos, técnicas de avaliação e aplicações práticas. Kaufman e Rousseeuw exploram uma variedade de algoritmos de clustering, fornecendo insights valiosos sobre como identificar estruturas de grupo em conjuntos de dados complexos.

Estas obras demonstram um pouco do pensamento evolutivo e incremental quando o assunto é clustering. A identificação de padrões em grupos é algo que já fazemos há bastante tempo como humanidade, aprendendo a sermos mais eficazes e lidando com conjuntos de dados e características de mundo real cada vez mais

complexas e detalhadas. Vamos fazer dos algoritmos de clustering parte do seu arsenal de ferramentas em ciência de dados?

Algoritmo Expectation-Maximization (EM)

A partir dos notebooks que trouxemos como exemplo nessa aula, você já teve contato com o algoritmo Expectation-Maximization (EM), uma técnica avançada de clustering usada em cenários em que os dados têm distribuições complexas ou desconhecidas. Ele é especialmente eficaz quando os clusters não têm formas simples, como em distribuições multimodais ou não gaussianas. O EM é uma abordagem iterativa, que estima os parâmetros de um modelo de mistura de Gaussianas a partir dos dados observados.

Em essência, o algoritmo EM é um método iterativo para estimar os parâmetros de um modelo probabilístico quando há variáveis latentes envolvidas. Ele consiste em duas etapas principais: a etapa de Expectativa (E-step) e a etapa de Maximização (M-step). Na etapa de Expectativa (E-step), são calculadas as probabilidades das variáveis latentes, dadas as observações e os parâmetros atuais do modelo. Essas probabilidades são chamadas de probabilidades de responsabilidade. A fórmula para calcular as probabilidades de responsabilidade $\gamma(\mathrm{znk})$ para cada ponto de dados x_n e cluster k é dada por:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)}$$

Em que:

- π_k^{new} é a nova proporção de pontos de dados no cluster k.
- μ_k^{new} é o novo vetor médio do cluster k.
- Σ_k^{new} é a nova matriz de covariância do cluster k.
- N é o número total de pontos de dados.
- $\gamma(z_{nk})$ são as probabilidades de responsabilidade calculadas na etapa de Expectativa.

O algoritmo EM é então iterado até que os parâmetros convirjam para um máximo local da verossimilhança dos dados. Essa é uma ferramenta poderosa para modelagem de dados complexos e amplamente aplicável em áreas como reconhecimento de padrões, bioinformática e processamento de linguagem natural.

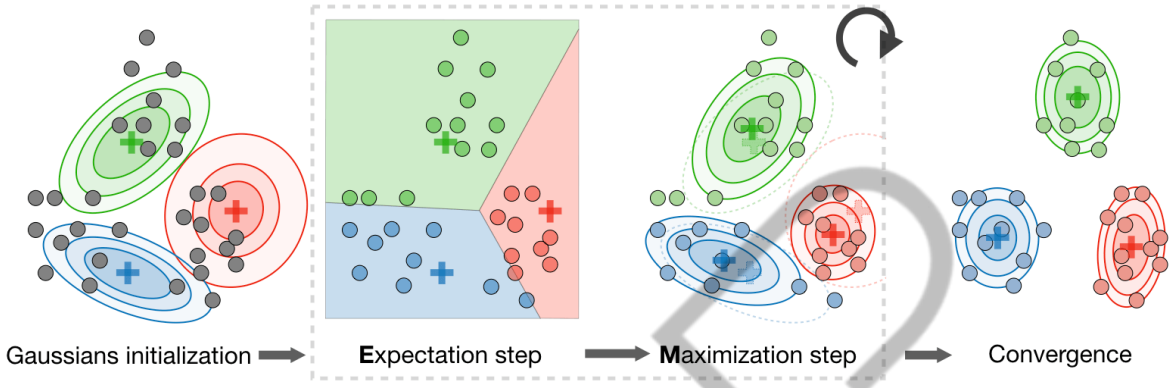

Figura 6 – Exemplo de utilização de algoritmo EM
Fonte: Amidi; Amidi (2018)

K-Means

O algoritmo K-Means é uma técnica de clustering amplamente utilizada na análise e mineração de dados. Neste tópico, exploraremos sua explicação matemática detalhada, incluindo a função objetivo de minimização da inércia, a derivação das equações de atualização dos centroides e a análise da complexidade computacional do K-Means.

A função objetivo do algoritmo K-Means é minimizar a inércia, também conhecida como a soma dos quadrados das distâncias de cada ponto ao seu centroide mais próximo. Formalmente, a função objetivo é dada por:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$$

Em que:

- N é o número total de pontos de dado.
- K é o número de clusters.
- x_n é o n-ésimo ponto de dado.
- μ_k é o centróide do k-ésimo cluster.

- r_{nk} é uma variável indicadora que é 1 se o ponto de dado x_n pertence ao cluster k e 0 caso contrário.

As equações de atualização dos centroides são derivadas da minimização da função objetivo. Para minimizar a inércia, os centroides são atualizados iterativamente por meio das seguintes equações:

$$\mu_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n$$

Essas equações calculam o centroide de cada cluster como a média de todos os pontos atribuídos ao cluster. A atribuição dos pontos aos clusters é feita com base na distância euclidiana entre cada ponto e os centroides dos clusters. Cada ponto é atribuído ao cluster cujo centroide é o mais próximo, de acordo com a seguinte regra:

$$C(x) = \arg\min_i d(x, y_i)$$

Em que:

- $C(x)$ é o cluster ao qual o ponto x será atribuído.
- $d(x,yi)$ é a distância entre o ponto x e o centróide yi do i-ésimo cluster.
- arg min$_i$ denota a operação que retorna o índice do cluster cujo centroide tem a menor distância ao ponto x.

Essa equação indica que atribuímos o ponto x ao cluster cujo centroide yi minimiza a distância $d(x,yi)$.

Em resumo, temos o seguinte processo (vide figura 7):

- **Inicialização dos centroides**: o algoritmo K-Means começa selecionando aleatoriamente ‘K’ pontos de dados como centroides iniciais. Isso é conhecido como inicialização gaussiana, em que os centroides são escolhidos a partir de uma distribuição normal.
- **Expectativa (E-step)**: na etapa de expectativa, cada ponto de dado é atribuído ao centroide mais próximo com base em alguma medida de distância, geralmente a distância euclidiana. Isso é feito calculando a

distância entre cada ponto de dado e todos os centroides e atribuindo o ponto ao cluster cujo centroide está mais próximo.

- **Maximização (M-step)**: na etapa de maximização, os centroides são recalculados para minimizar a inércia dos clusters. Cada novo centroide é definido como o centroide médio de todos os pontos de dado atribuídos ao cluster correspondente na etapa de expectativa.
- **Convergência**: os passos de expectativa e maximização são repetidos iterativamente até que ocorra a convergência. A convergência é alcançada quando os centroides não mudam significativamente entre iterações ou quando um critério de parada pré-definido é satisfeito, como um número máximo de iterações ou uma pequena mudança na inércia.

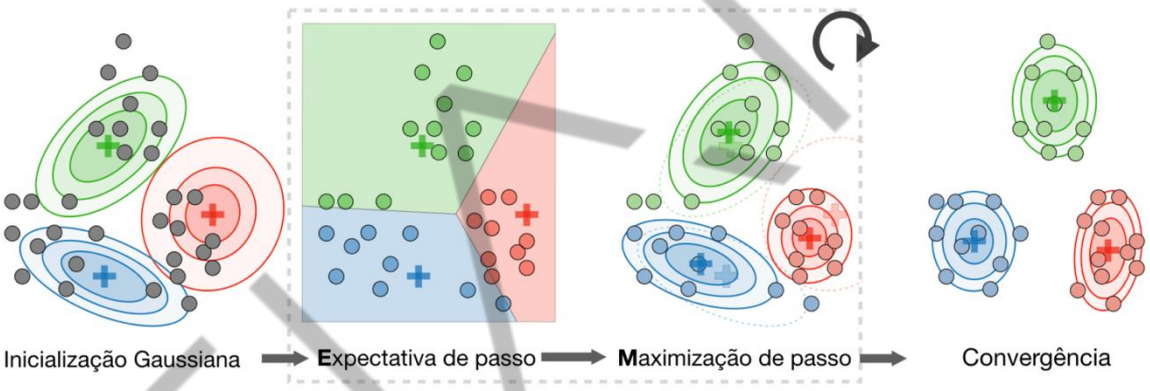

Figura 7 – Exemplo de utilização de algoritmo K-Means
Fonte: Amidi; Amidi (2018)

A complexidade computacional do K-Means é O (n·K·I·d)

Em que:

- n é o número de pontos de dados.
- K é o número de clusters.
- I é o número de iterações.
- d é a dimensionalidade dos dados.

Nos últimos anos, esforços têm se concentrado para lidar com grandes conjuntos de dados; estratégias como inicialização inteligente dos centroides,

paralelização e utilização de técnicas de redução de dimensionalidade podem ser empregadas para melhorar a eficiência do algoritmo.

Método do Cotovelo - Obtendo o valor de 'K'

O Método do Cotovelo, ou Elbow Method, é uma técnica comum usada para determinar o número ideal de clusters em um conjunto de dados durante a aplicação do algoritmo de clustering, como o K-Means. Neste tópico, exploraremos a introdução ao Método do Cotovelo, sua demonstração prática com exemplos de código em Python e as limitações associadas, além de alternativas para avaliação do número de clusters.

O Método do Cotovelo é baseado na ideia de que a variabilidade explicada pelo modelo aumenta à medida que o número de clusters aumenta, mas a uma taxa decrescente. O ponto em que a adição de um cluster extra não fornece uma melhoria significativa na variabilidade explicada é conhecido como o "cotovelo" do gráfico.

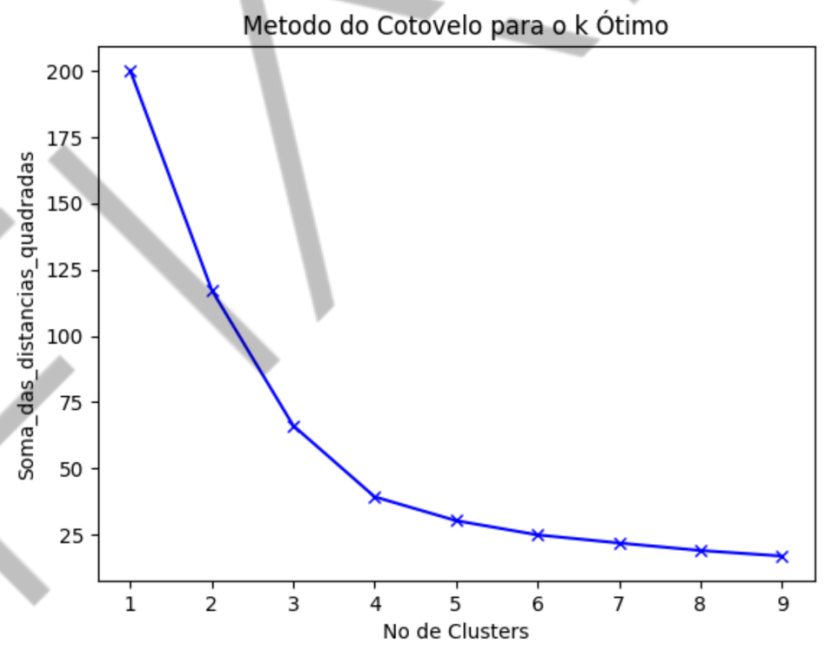

Figura 8 – Exemplo de utilização do Método do Cotovelo
Fonte: Elaborado pelo autor (2024)

Na prática, o Método do Cotovelo envolve plotar a variabilidade explicada em função do número de clusters e auxiliar na identificação do ponto em que a curva começa a se nivelar. Vamos ilustrar isso com exemplos de código em Python, utilizando bibliotecas como NumPy, scikit-learn e Matplotlib para calcular a inércia para diferentes valores de 'K' e plotar o gráfico do Método do Cotovelo.

```
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt


# Calcular inércia para diferentes valores de K
inertias = []
for k in range(1, 10):
        kmeans = KMeans(n_clusters=k)
        kmeans.fit(data)
        inertias.append(kmeans.inertia_)


# Plotar o gráfico do Método do Cotovelo
plt.plot(range(1, 10), inertias, marker='o')
plt.xlabel('Número de Clusters (K)')
plt.ylabel('Inércia')
plt.title('Método do Cotovelo')
plt.show()
```

Código-fonte 2 – Demonstração 3 – Exemplo de aplicação do Método do Cotovelo
Fonte: Elaborado pelo autor (2024)

Embora o Método do Cotovelo seja uma técnica útil para se determinar o número ideal de clusters, ele apresenta algumas limitações. Por exemplo, o método pode ser inconclusivo em conjuntos de dados complexos ou com distribuições não convencionais. Alternativas incluem o Método da Silhueta (Silhouette Method), que considera a coesão e a separação dos clusters, e métodos baseados em validação externa, como o Índice Rand (Rand Index) e o Índice de Davies-Bouldin (Davies-Bouldin Index).

Em resumo, o Método do Cotovelo é uma técnica valiosa para determinar o número ideal de clusters em algoritmos de clustering, como o K-Means. No entanto, é importante estar ciente de suas limitações e considerar alternativas quando necessário, a fim de tomar decisões mais informadas sobre a estrutura de clusters em conjuntos de dados complexos.

EM vs. K-Means, quando usar?

Em comparação com o K-Means, o EM é mais flexível e pode modelar clusters com formas mais complexas e distribuições não gaussianas. No entanto, como visto anteriormente, o EM é mais computacionalmente intensivo, pois envolve a estimativa de parâmetros de distribuição contínua, enquanto o K-Means lida com atributos discretos e assume clusters esféricos com a mesma variância.

Outra diferença importante é que o EM fornece probabilidades suaves de pertencimento a cada cluster, enquanto o K-Means produz atribuições rígidas de clusters. Isso pode ser útil em aplicações em que a incerteza na atribuição do cluster é importante, como na classificação de documentos ou na segmentação de imagens médicas.

Clustering Hierárquico

O clustering hierárquico é uma técnica de clustering que agrupa os dados em uma estrutura hierárquica de clusters. Ele é especialmente útil quando a estrutura de agrupamento dos dados não é conhecida e quando se deseja explorar diferentes níveis de granularidade nos clusters.

Neste tópico, abordaremos uma visão geral do clustering hierárquico e seus dois principais métodos: aglomerativo e divisivo. Também discutiremos como o clustering hierárquico cria uma árvore de clusters que pode ser visualizada como um dendrograma e forneceremos exemplos de aplicação em diferentes domínios.

O clustering hierárquico pode ser realizado de duas maneiras: aglomerativo e divisivo. No método aglomerativo, cada ponto de dados começa como um cluster separado e, em cada iteração, os clusters mais semelhantes são mesclados até que todos os pontos de dados estejam em um único cluster. Por outro lado, no método divisivo, todos os pontos de dados começam em um único cluster e, em cada iteração, o cluster é dividido em subgrupos menores até que cada ponto de dados esteja em seu próprio cluster.

Uma característica distintiva do clustering hierárquico é a criação de uma árvore de clusters, conhecida como dendrograma. Um dendrograma é uma representação visual da estrutura hierárquica dos clusters, em que os pontos de dados são agrupados em diferentes níveis de similaridade. A altura das ligações no dendrograma representa a dissimilaridade entre os clusters: quanto maior a altura, menos similares são os clusters.

O clustering hierárquico é amplamente utilizado em uma variedade de domínios, incluindo biologia, medicina e finanças. Por exemplo, no campo da biologia, o clustering hierárquico é usado para agrupar genes ou proteínas com base em sua expressão gênica ou similaridade estrutural. Em finanças, pode ser aplicado para agrupar ativos com comportamentos de mercado semelhantes.

Clustering e Alta Dimensionalidade

Na análise de dados moderna, é comum lidar com conjuntos de dados que possuem um grande número de atributos ou características, resultando em dados de alta dimensionalidade. Esses conjuntos de dados podem surgir em uma variedade de domínios, como genômica, processamento de linguagem natural, visão computacional e análise de redes sociais. No entanto, a alta dimensionalidade apresenta desafios únicos para o clustering.

Quando se trabalha com dados de alta dimensionalidade, várias questões surgem durante o processo de clustering. Primeiro, a visualização dos dados torna-se difícil, pois é impossível representar eficientemente espaços de alta dimensão em gráficos bidimensionais ou tridimensionais. Além disso, a distância entre pontos tende a perder significado em espaços de alta dimensão, pois muitos atributos podem ter pouco impacto na similaridade entre instâncias.

Para lidar com clustering de dados de alta dimensionalidade, são necessárias técnicas específicas que considerem as características únicas desses conjuntos de dados. Uma abordagem comum é a redução de dimensionalidade, que envolve a projeção dos dados em um espaço de menor dimensão, preservando ao máximo a estrutura dos dados originais. Técnicas como PCA (Análise de Componentes Principais) e t-SNE (t-Distributed Stochastic Neighbor Embedding) são amplamente utilizadas para esse fim.

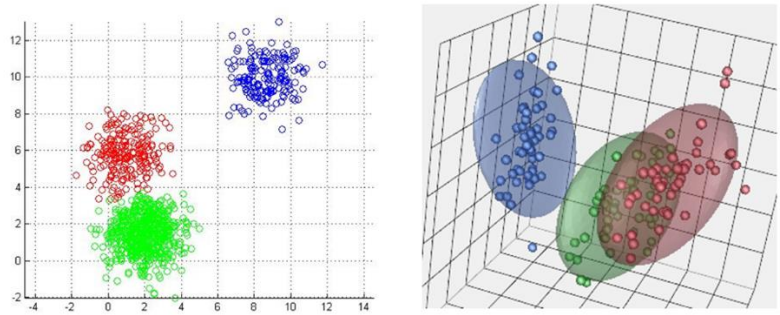

Figura 9 – Exemplo de diferentes análises dimensionais de dados alvos de processo de clustering.
Fonte: NPM (2015).

Além da redução de dimensionalidade, é importante selecionar cuidadosamente os atributos relevantes para o clustering e realizar o pré-processamento adequado dos dados para remover ruídos e redundâncias. Algoritmos de clustering robustos e escaláveis também são essenciais para lidar com conjuntos de dados de alta dimensionalidade de forma eficaz.

O clustering de dados de alta dimensionalidade tem aplicações em uma variedade de domínios, incluindo bioinformática, reconhecimento de padrões em imagens médicas, análise de texto e detecção de anomalias em grandes conjuntos de dados. Ao agrupar automaticamente instâncias de dados semelhantes em clusters, é possível extrair insights valiosos e identificar padrões complexos que podem não ser aparentes em uma inspeção superficial dos dados.

O uso de técnicas de aprendizado profundo, como Redes Neurais Convolucionais (CNNs) e Redes Neurais Recorrentes (RNNs), oferece novas oportunidades para o clustering em dados de alta dimensão e complexidade. Com o aumento da demanda por análise de dados em tempo real, há uma crescente necessidade de algoritmos de clustering capazes de lidar com grandes volumes de dados em tempo real, como o streaming de dados.

Aumentar a interpretabilidade e a explicabilidade dos resultados do clustering continua a ser um desafio importante, especialmente em domínios nos quais a transparência e a justificativa das decisões tomadas são críticas.

Ademais, à medida que os dados e os problemas os quais resolvemos se tornam mais complexos, é importante escolher algoritmos de clustering adequados e compreender suas características e limitações.

Algoritmos como o EM e o K-Means oferecem abordagens eficazes para se agrupar dados, mas é crucial considerar a dimensionalidade dos dados, a distribuição dos clusters e a escalabilidade do algoritmo. Além disso, explorar técnicas de pré-processamento de dados e avaliação de clusters pode ajudar a obter insights significativos desses ambientes de dados complexos.

MERCADO, CASES E TENDÊNCIAS

Matéria EN-US – Empresa de Biotecnologia (Tissue Dynamics)

O vice-diretor editorial global da WIRED, Greg Williams, conversou com o professor Yaakov Nahmias, fundador da Tissue Dynamics, sobre o futuro do desenvolvimento de medicamentos e como IA e sensores podem auxiliar no aumento das taxas de sucesso clínico nesse campo. O vídeo e artigos destacam como as empresas farmacêuticas estão utilizando machine learning de modo a identificar padrões em grandes conjuntos de dados genômicos e mais.

Matéria EN-US – Instituição Financeira (JPMorgan Chase)

O JPMorgan Chase tem utilizado Inteligência Artificial (incluindo técnicas de clustering), de modo a melhorar o atendimento ao cliente e reduzir fraudes financeiras. Eles empregam algoritmos de machine learning para analisar grandes volumes de dados e identificar padrões de comportamento suspeitos entre clientes, ajudando a prevenir fraudes e garantir uma melhor experiência para eles. Saiba mais aqui.

O QUE VOCÊ VIU NESTA AULA?

Nesta aula sobre clustering, exploramos mais uma vez o fascinante mundo da análise de dados não supervisionada, agora conhecendo em mais detalhes o algoritmo K-Means. A partir da utilização do IMDb 5000 Movie Dataset como nosso conjunto de dados de exemplo, mergulhamos na aplicação prática do K-Means de modo a identificar padrões e agrupamentos dentro dos filmes.

Por meio de uma série de etapas, desde o pré-processamento dos dados até a análise dos resultados, fomos capazes de visualizar como os filmes podem ser agrupados com base em suas características, como gênero, orçamento e popularidade.

Além disso, demos um passo adiante na compreensão dos clusters ao empregar o método do cotovelo para determinar o número ideal de clusters. Essa técnica revela-se crucial para, por exemplo, uma segmentação eficaz dos filmes, permitindo-nos identificar um ponto de inflexão em que a adição de mais clusters não contribuía significativamente para a explicação da variância nos dados.

No final, esta aula não apenas forneceu uma introdução prática ao clustering com o K-Means, mas também lhe equipou com ferramentas valiosas para explorar e compreender conjuntos de dados complexos de forma mais significativa.

REFERÊNCIAS

AMIDI, A.; AMIDI, S. **VIP Cheatsheet**: Unsupervised Learning – Stanford CS 229 – Machine Learning, 2018. Cheatsheet, 2018. Disponível em: <https://github.com/afshinea/stanford-cs-229-machine-learning/blob/master/en/cheatsheet-unsupervised-learning.pdf>. Acesso em: 12 jul. 2024.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006. Disponível em: <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>. Acesso em: 12 jul. 2024.

ECLOUDVALLEY. **Introduction to Machine Learning: Is AutoML Replacing Data Scientists?** 2021. Disponível em: <https://www.ecloudvalley.com/en/p/ai-machine-learning>. Acesso em: 12 jul. 2024.

GRAVES, A.; CLANCY, K.. **Unsupervised Learning**: The Curious Pupil. 2019. Disponível em: <https://deepmind.google/discover/blog/unsupervised-learning-the-curious-pupil/>. Acesso em: 12 jul. 2024.

HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. **Journal of the Royal Statistical Society**. Series C (Applied Statistics), v. 28, n. 1, pp. 100-108, 1979. Disponível em <https://www.stat.cmu.edu/~rnugent/PCMI2016/papers/HartiganKMeans.pdf>. Acesso em: 12 jul. 2024.

JAMES, G. et al. **An Introduction to Statistical Learning with Applications in Python**. [s.l.]: Springer, 2023. Disponível em: <https://www.statlearning.com/>. Acesso em: 12 jul. 2024.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data**: An Introduction to Cluster Analysis. [s.l.]: Wiley, 1990. Disponível em: <https://doi.org/10.1002/9780470316801>. Acesso em: 12 jul. 2024.

KEERTHANA, V. **Spectral Clustering**: A Comprehensive Guide for Beginners. 2024. Disponível em: <https://www.analyticsvidhya.com/blog/2021/05/what-why-and-how-of-spectral-clustering/>. Acesso em: 12 jul. 2024.

MACQUEEN, J. **Some Methods for classification and Analysis of Multivariate Observations**. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967. Disponível em: <https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s5_v1_article-17.pdf>. Acesso em: 12 jul. 2024.

MORIMOTO, J.; PONTON, F. Virtual reality in biology: could we become virtual naturalists? **Evo Edu Outreach**, v. 14, n. 7, 2021. Disponível em: <https://doi.org/10.1186/s12052-021-00147-x>. Acesso em: 12 jul. 2024.

NPM. **Clusters**. 2015. Disponível em: <https://www.npmjs.com/package/clusters>. Acesso em: 12 jul. 2024.

TAN, P. et al. **Introduction to Data Mining**. Essex: Pearson, 2014. Disponível em: <https://www.ceom.ou.edu/media/docs/upload/Pang-Ning_Tan_Michael_Steinbach_Vipin_Kumar_-_Introduction_to_Data_Mining-Pe_NRDK4fi.pdf>. Acesso em: 12 jul. 2024.

PALAVRAS-CHAVE

Inteligência Artificial. Aprendizado de Máquina. Aprendizado Não Supervisionado.