JOSÉ AHIRTON BATISTA LOPES FILHO

MACHINE LEARNING ENGINEERING

APRENDIZADO NÃO SUPERVISIONADO

AULA 01

SUMÁRIO

O QUE VEM POR AÍ?

Nesta disciplina exploraremos os princípios da análise de dados sem orientação externa, abordando clustering, redução de dimensionalidade e modelagem de tópicos. Estaremos atentos(as) aos avanços significativos nesses campos, capacitando vocês para compreenderem problemas complexos e técnicas-chave.

Além disso, discutiremos desafios como a interpretabilidade de modelos e a escalabilidade diante da crescente complexidade dos conjuntos de dados atuais. O desafio aqui é claro: oferecer uma compreensão sólida dos problemas abordados pela abordagem de Aprendizado Não Supervisionado e preparar os alunos e alunas para aplicar eficazmente essas técnicas em diferentes contextos.

HANDS ON

Neste Hands On, focaremos nos passos básicos do algoritmo K-Means em Aprendizado Não Supervisionado usando um dataset simples (penguim dataset). A implementação do K-Means do zero envolve compreender seus princípios e aplicá-los passo a passo. O K-Means ainda será revisitado por nós, mas, por ora, basta entender que ele é utilizado de modo a agrupar dados, em que cada cluster (ou grupo) é representado por um centroide e as observações são atribuídas ao centroide mais próximo.

O desafio é, então, encontrar os centroides ótimos de forma iterativa. Iniciamos com 'k' centroides aleatórios, atribuímos cada ponto ao cluster mais próximo, recalculamos os centroides e repetimos até a convergência. No início, selecionamos três observações aleatórias como centroides iniciais, refinando-os iterativamente até a convergência. Essa prática aprofunda a compreensão do K-Means e abre portas para aplicações avançadas em Aprendizado Não Supervisionado.

```
# Encontrando o centróide mais próximo de uma determinada
observação.

def get_nearest_centroid(obs):
    dists = np.sqrt(((obs - centroids) ** 2).sum(axis=1))
    return dists.idxmin()

get_nearest_centroid(X_train_scaled.loc[0])

# Aplicando a função a todo o conjunto de dados.

clusters = X_train_scaled.apply(get_nearest_centroid, axis=1)

# Plotando as atribuições dos clusters.

ax       =       X_train_scaled.plot.scatter(x="bill_depth_mm",
y="flipper_length_mm",
                                    c=clusters,        marker="x",
alpha=.5)
centroids.plot.scatter(x="bill_depth_mm",
y="flipper_length_mm",
                       c=centroids.index, ax=ax)

# Calculando o "tamanho" médio de cada cluster.

centroids = X_train_scaled.groupby(clusters).mean()
```

```
# Vamos plotar os novos centróides.

ax       =       X_train_scaled.plot.scatter(x="bill_depth_mm",
y="flipper_length_mm",
                                    c=clusters,        marker="x",
alpha=.5)
centroids.plot.scatter(x="bill_depth_mm",
y="flipper_length_mm",
                        c=centroids.index, ax=ax)

centroids

# Atribua pontos ao centróide mais próximo.

clusters = X_train_scaled.apply(get_nearest_centroid, axis=1)

Recalcular os centróides com base nos clusters.

centroids = X_train_scaled.groupby(clusters).mean()
```

Código-fonte 1 – Demonstração 1 – Criando um primeiro algoritmo de Aprendizado Não Supervisionado
Fonte: Elaborado pelo autor (2024)

SAIBA MAIS

O Aprendizado Não Supervisionado é um ramo da inteligência artificial e do aprendizado de máquina que lida com a análise de dados sem orientação externa. Mais especificamente, ao contrário do Aprendizado Supervisionado, na qual os algoritmos são treinados com exemplos rotulados, aqui os algoritmos são alimentados apenas com dados de entrada e encarregados de descobrir padrões, estruturas e relações por conta própria a partir de sua capacidade de interpretação. Esta abordagem oferece uma solução poderosa para lidar com a crescente disponibilidade de dados não rotulados e os desafios associados à análise deles.

A importância de aprender sobre esse paradigma em Aprendizado de Máquina se dá justamente pelo momento de inflexão em que vivemos na área de Inteligência Artificial atual junto ao processo de Transformação Digital vivido em muitos negócios. Passamos a testemunhar uma explosão sem precedentes na geração e armazenamento de dados.

Com o avanço da tecnologia e a proliferação de dispositivos conectados (parte da internet das coisas), uma quantidade massiva de informações é gerada a cada segundo em diferentes formatos, incluindo texto, imagens, áudio e vídeo. No entanto, uma parte significativa desses dados não vem acompanhada de rótulos ou categorias pré-definidas, tornando desafiador o processo de análise e extração de conhecimento útil.

Logo, esse tipo de análise acaba sendo crucial em uma variedade de domínios, incluindo ciência de dados, mineração de dados, reconhecimento de padrões e muito mais. Com o Aprendizado Não Supervisionado, podemos segmentar clientes em grupos com base em seu comportamento de compra, identificar tópicos principais em grandes conjuntos de documentos de texto e detectar anomalias em dados de séries temporais, entre muitas outras aplicações.

Neste material de referência, o objetivo é fornecer uma visão abrangente dos fundamentos do Aprendizado Não Supervisionado, além de explorar suas aplicações em diferentes domínios, para além dos exemplos mais óbvios.

Ao longo das próximas páginas, abordaremos desde os conceitos básicos até técnicas avançadas, incluindo clustering, redução de dimensionalidade, modelagem

de tópicos e detecção de anomalias. Além disso, discutiremos os desafios enfrentados na aplicação de Aprendizado Não Supervisionada e as perspectivas futuras deste campo dinâmico na área de Inteligência Artificial.

A análise de dados sem orientação externa, característica do Aprendizado Não Supervisionado, enfrenta uma série de desafios que podem dificultar a identificação de padrões e a extração de insights significativos. Entre esses desafios, destacam-se:

- **Ausência de Rótulos ou Categorias Pré-Definidas:** ao contrário do Aprendizado Supervisionado, em que os dados são rotulados com informações conhecidas, no Aprendizado Não Supervisionado os dados não possuem rótulos prévios. Isso torna difícil identificar grupos ou classes naturais nos dados, já que não há orientação externa sobre como eles devem ser agrupados ou categorizados.
- **Complexidade e Dimensionalidade dos Dados:** muitos conjuntos de dados não supervisionados são altamente dimensionais, ou seja, contêm uma grande quantidade de variáveis ou atributos. Lidar com essa alta dimensionalidade pode tornar a análise mais complexa e aumentar o tempo de processamento dos algoritmos. Além disso, a presença de ruído ou dados não relevantes pode obscurecer os padrões verdadeiros nos dados.
- **Interpretabilidade dos Resultados:** outro desafio significativo na Análise Não Supervisionada é a interpretação dos resultados obtidos. Encontrar padrões nos dados é apenas o primeiro passo; compreender o significado desses padrões e sua relevância para o problema em questão pode ser uma tarefa complicada. Sem a orientação fornecida por rótulos ou categorias pré-definidas, a interpretação dos resultados pode ser subjetiva e exigir expertise adicional.
- **Escalabilidade e Eficiência:** à medida que os conjuntos de dados crescem em tamanho e complexidade, a escalabilidade e a eficiência dos algoritmos de Aprendizado Não Supervisionado tornam-se preocupações importantes. Estudando esse paradigma de aprendizado, passamos a perceber que certos algoritmos funcionam bem em conjuntos de dados pequenos mas podem não ser adequados para lidar com conjuntos de dados grandes em tempo hábil. Portanto, é crucial desenvolver técnicas que possam lidar

eficientemente com grandes volumes de dados sem comprometer a qualidade dos resultados.

Superar esses desafios requer uma abordagem cuidadosa e uma combinação de conhecimento teórico e prático. Ao reconhecer os desafios enfrentados na análise de dados sem orientação externa, podemos desenvolver estratégias e técnicas mais eficazes para extrair insights valiosos e tomar decisões informadas a partir de dados não rotulados. A figura 1 mostra uma visão geral do processo de treinamento de uma aplicação que se utiliza de Aprendizado Não Supervisionado:

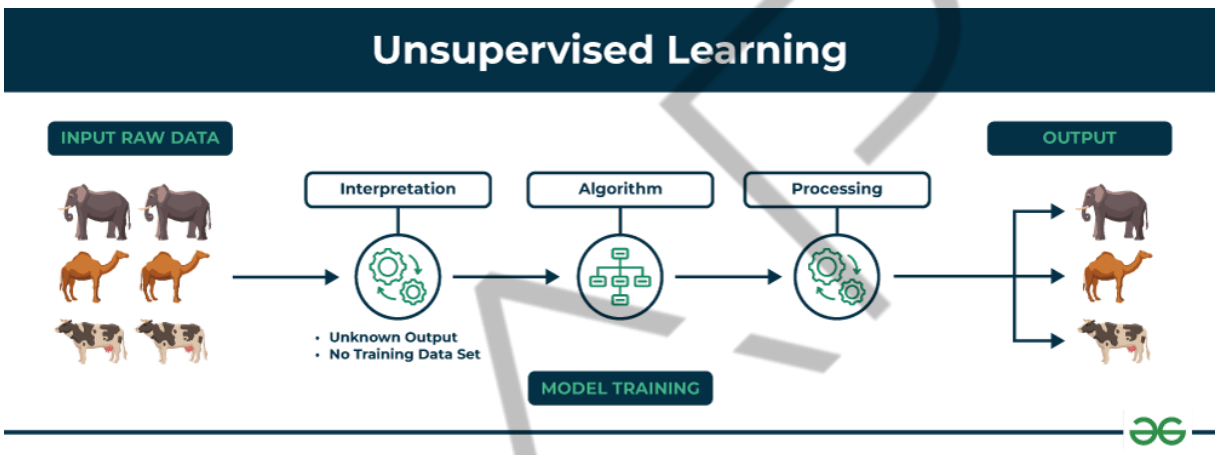

Figura 1 – Processo macro de treinamento de uma aplicação em Aprendizado Não Supervisionado
Fonte: GeeksforGeeks (2023)

Diante dos desafios mencionados na análise de dados sem orientação externa, o Aprendizado Não Supervisionado surge como uma abordagem fundamental e poderosa para lidar com essas questões de forma eficaz. Algumas das razões que fazem esse paradigma ser essencial em problemas para os quais não se tem acesso a rótulos prévios são as seguintes:

- **Exploração da Estrutura Intrínseca dos Dados:** o Aprendizado Não Supervisionado permite que os algoritmos explorem a estrutura intrínseca dos dados, identificando padrões e relações subjacentes sem depender de rótulos ou categorias pré-definidas. Isso é crucial quando lidamos com conjuntos de dados complexos e não estruturados, nos quais os padrões podem não ser facilmente discerníveis a olho nu.
- **Flexibilidade e Adaptabilidade:** os algoritmos de Aprendizado Não Supervisionado são conhecidos por serem altamente flexíveis e adaptáveis, o que significa que podem se ajustar dinamicamente aos dados sem a

necessidade de intervenção humana. Isso é especialmente útil em cenários nos quais os padrões nos dados podem mudar ao longo do tempo (como na leitura de dados de um maquinário em uma série temporal) ou diferentes aspectos dos dados podem ser mais relevantes em momentos diferentes.

- **Descoberta de Insights Ocultos:** ao explorar os dados sem orientação externa, o Aprendizado Não Supervisionado tem o potencial de descobrir insights valiosos e padrões ocultos que podem não ser evidentes à primeira vista. Isso pode levar a descobertas significativas e inesperadas, ajudando a gerar novas hipóteses e direcionar a investigação em novas direções;
- **Redução da Dependência de Especialistas Humanos:** ao contrário de abordagens que requerem a intervenção humana para rotular os dados, o Aprendizado não Supervisionado reduz a dependência de especialistas humanos e permite uma análise mais automatizada e escalável. Isso é especialmente vantajoso em cenários em que os recursos humanos são limitados ou a rotulação manual dos dados é impraticável devido ao tamanho ou à complexidade dos dados.
- **Aplicabilidade em uma Variedade de Domínios:** o Aprendizado Não Supervisionado é amplamente aplicável em uma variedade de domínios, incluindo ciência de dados, medicina, finanças, biologia e muito mais. Sua capacidade de lidar com dados não rotulados e extrair insights valiosos torna-o uma ferramenta poderosa e versátil para análise de dados em diferentes contextos.

Em resumo, o Aprendizado Não Supervisionado desempenha um papel crucial na análise de dados sem orientação externa, fornecendo uma abordagem flexível, adaptável e automatizada para lidar com os desafios associados à análise de dados não rotulados. Ao explorar a estrutura intrínseca dos dados e descobrir insights ocultos, o Aprendizado Não Supervisionado oferece uma maneira poderosa de extrair valor dos dados e obter uma compreensão mais profunda do mundo ao nosso redor.

Logo, o Aprendizado Não Supervisionado desempenha um papel fundamental em uma variedade de aplicações práticas, permitindo a descoberta de padrões e estruturas nos dados. Três das técnicas mais amplamente utilizadas em Aprendizado Não Supervisionado incluem construção de regras de associação, clustering e

redução de dimensionalidade. Para iniciar, vamos explorar como cada uma dessas técnicas é aplicada em diferentes domínios dentro desse contexto.

Construção de Regras de Associação

A construção de regras de associação é uma técnica usada para descobrir relações entre diferentes itens em um conjunto de dados. Um exemplo comum dessa técnica é a análise de cestas de compras em supermercados, em que o objetivo é identificar quais produtos tendem a ser comprados juntos. Essa informação é valiosa para estratégias de marketing e planejamento de estoque. Além disso, a construção de regras de associação é aplicada em sistemas de recomendação, onde sugere itens com base nas preferências do usuário e padrões de compra anteriores.

Clustering

O clustering é uma técnica usada para agrupar itens semelhantes em conjuntos, ou clusters, com base em suas características compartilhadas. No domínio do marketing, por exemplo, o clustering pode ser usado para segmentar clientes em grupos com base em seus comportamentos de compra, permitindo estratégias de marketing mais direcionadas e personalizadas. Em ciência da saúde, o clustering pode ser aplicado para identificar subgrupos de pacientes com características semelhantes, auxiliando no diagnóstico e tratamento de doenças. Na figura 2 há um exemplo de processo de clustering.

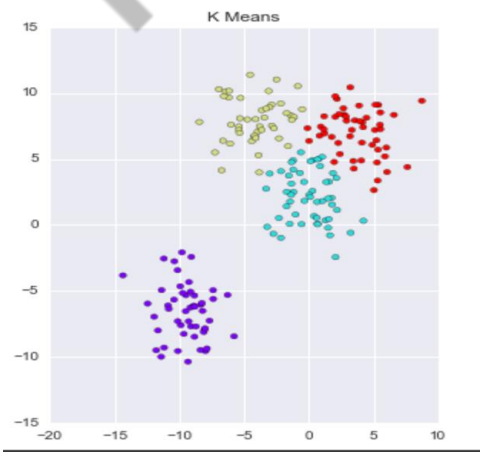

Figura 2 – Visualização 2D de um processo de clustering via Algoritmo K-Means
Fonte: Elaborado pelo autor (2024)

Redução de Dimensionalidade

A redução de dimensionalidade é uma técnica usada para reduzir a quantidade de variáveis em um conjunto de dados, preservando ao mesmo tempo o máximo de informação possível. Isso é útil quando lidamos com conjuntos de dados altamente dimensionais, em que o número de variáveis é grande em relação ao número de observações. Uma aplicação comum da redução de dimensionalidade é a visualização de dados, em que os dados são reduzidos a um número menor de dimensões para facilitar a interpretação e a análise visual.

Essas são apenas algumas das muitas aplicações do Aprendizado Não Supervisionado na prática e, à medida que os dados continuam a desempenhar um papel central em uma variedade de domínios, a utilização de Aprendizado Não Supervisionado na descoberta de padrões e na extração de conhecimento só tende a crescer.

Para aqueles(as) que desejam se aprofundar no campo do Aprendizado Não Supervisionado (para além desse material de referência), existem várias fontes de estudo canônicas e artigos base que servem como pilares fundamentais na compreensão dos conceitos e técnicas envolvidas.

A seguir, algumas das referências mais influentes na área.

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

"A Survey of Clustering Data Mining Techniques", de Pavel Berkhin: este artigo oferece uma visão geral abrangente dos métodos de clustering, incluindo uma análise detalhada de algoritmos populares e suas aplicações.

"Principal Component Analysis" de I. T. Jolliffe: este artigo clássico apresenta uma explicação detalhada do método de redução de dimensionalidade conhecido como Análise de Componentes Principais (PCA), um dos principais algoritmos em Aprendizado Não Supervisionado.

"Data Clustering: 50 Years Beyond K-Means", de Anil K. Jain: este artigo examina os avanços em técnicas de clustering, indo além do algoritmo K-Means e explorando outras abordagens e desafios associados à clusterização de dados.

Revistas Científicas

IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI): esta revista é uma fonte valiosa de artigos de pesquisa sobre reconhecimento de padrões, aprendizado de máquina e análise de dados, incluindo muitos trabalhos relevantes na área de Aprendizado Não Supervisionado.

Journal of Machine Learning Research (JMLR): esta revista publica pesquisas de alta qualidade em todas as áreas de aprendizado de máquina, incluindo Aprendizado Não Supervisionado. Muitos dos artigos aqui oferecem insights inovadores e avanços na teoria e prática do Aprendizado Não Supervisionado. Essas fontes fornecem uma base sólida para aqueles(as) que desejarem explorar mais a fundo os conceitos e técnicas em Aprendizado Não Supervisionado.

Indo além das videoaulas e deste material de referência, pesquisadores(as) e estudantes interessados(as) com certeza expandirão seu conhecimento e contribuirão para o desenvolvimento contínuo deste campo dinâmico e em constante evolução!

Como visto anteriormente, o Aprendizado Não Supervisionado é um paradigma dentro do Aprendizado de Máquina, que visa descobrir padrões intrínsecos e estruturas nos dados sem a necessidade de rótulos de saída conhecidos. Diferentemente do que acontece no Aprendizado Supervisionado, em que os modelos são treinados com dados rotulados, o Aprendizado Não Supervisionado lida com dados não rotulados, explorando a estrutura subjacente dos dados para encontrar informações úteis.

Uma das técnicas mais comuns em Aprendizado Não Supervisionado é o clustering, o qual conheceremos em mais detalhes posteriormente, em que os dados são agrupados em clusters com base em métricas de similaridade entre os dados, como visto em nosso Hands On.

Algoritmos como K-Means, DBSCAN e de Clustering Hierárquico são amplamente utilizados para essa finalidade, visando identificar grupos naturais nos dados.

Clustering é uma técnica fundamental em Aprendizado Não Supervisionado, sendo que o objetivo é agrupar dados semelhantes em conjuntos distintos, chamados clusters. Esses clusters são definidos de forma que os pontos dentro de um mesmo cluster sejam mais semelhantes entre si do que com os pontos de outros clusters. O processo de clustering é essencialmente uma tarefa de agrupamento automático de dados com base em sua similaridade.

Existem várias abordagens e algoritmos de clustering, cada um com suas próprias características e pressupostos. O algoritmo K-Means, por exemplo, é amplamente utilizado devido à sua eficiência e simplicidade. Ele divide os dados em 'k' clusters, em que cada cluster é representado por um centroide e os pontos são atribuídos ao cluster mais próximo de seu centroide.

Outro algoritmo popular é o DBSCAN, que identifica clusters baseados na densidade dos pontos no espaço de características. Ele é capaz de encontrar clusters de formas arbitrárias e é robusto a outliers e ruídos nos dados. Na figura 3 é apresentada comparação do funcionamento das duas técnicas, DBSCAN e K-Means.

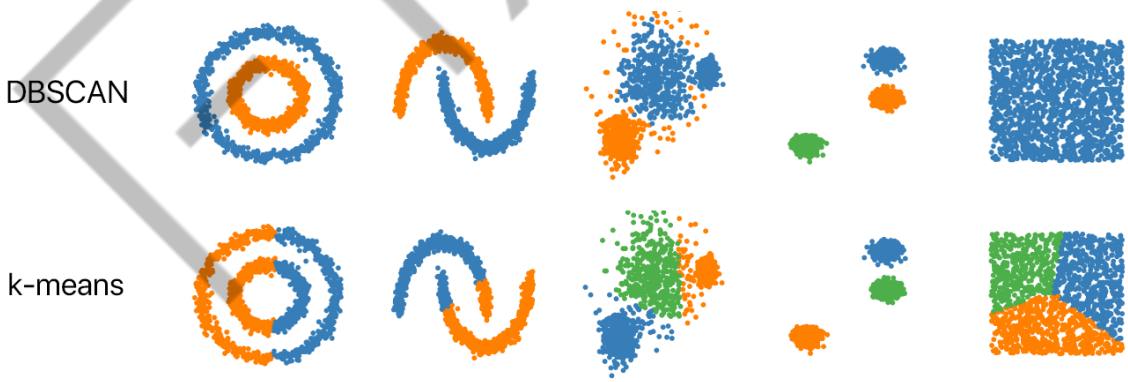

Figura 3 – Visualização 2D de processos de clustering via Algoritmos DBSCAN e K-Means
Fonte: Mattt (2021)

Além disso, o clustering hierárquico é uma técnica que constrói uma hierarquia de clusters, em que os clusters podem ser visualizados como uma árvore (dendrograma, como na figura 4), facilitando a interpretação da estrutura dos dados. O clustering tem diversas aplicações em diferentes áreas, como segmentação de

mercado, análise de redes sociais, biologia computacional e reconhecimento de padrões em imagens. Em resumo, o clustering desempenha um papel crucial na organização e interpretação de conjuntos de dados não rotulados, permitindo a identificação de padrões e estruturas úteis nos dados.

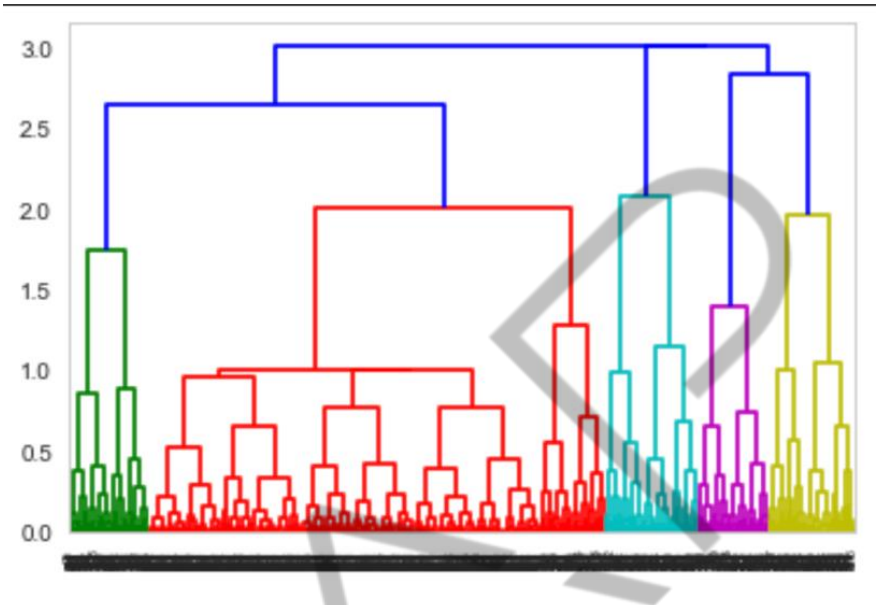

Figura 4 – Visualização de dendrograma por meio de clustering hierárquico
Fonte: Elaborado pelo autor (2024)

Outra técnica importante que discutiremos é a Redução de Dimensionalidade, a qual envolve a projeção dos dados de alta dimensionalidade em um espaço de menor dimensionalidade, preservando o máximo possível de informação. Métodos como Análise de Componentes Principais (PCA) e T-distributed Stochastic Neighbor Embedding (t-SNE) são comumente empregados para reduzir a complexidade dos dados e facilitar sua interpretação.

A redução de dimensionalidade é uma técnica essencial em Aprendizado Não Supervisionado que visa reduzir a complexidade dos dados, projetando-os de um espaço de alta dimensionalidade para um espaço de menor dimensionalidade enquanto tenta manter o máximo de informação relevante possível. O principal objetivo é, então, lidar com problemas de "maldição da dimensionalidade", em que conjuntos de dados com muitas variáveis podem levar a um aumento na complexidade computacional e a problemas de sobreajuste (overfitting).

Um dos algoritmos mais amplamente utilizados para a redução de dimensionalidade é a Análise de Componentes Principais (PCA). O PCA é uma técnica estatística que busca identificar as direções de máxima variância nos dados e

projetar os pontos de dados ao longo dessas direções. Isso é alcançado calculando os autovetores e autovalores da matriz de covariância dos dados originais.

Durante o processo, o PCA seleciona as primeiras 'k' componentes principais que capturam a maior parte da variância nos dados. Essas componentes principais são uma combinação linear das variáveis originais e podem ser interpretadas como os eixos principais de variação nos dados.

Ao projetar os dados nesses 'k' componentes principais, podemos reduzir a dimensionalidade do conjunto de dados original, mantendo a maior parte da estrutura e das informações relevantes. Essa abordagem é valiosa para visualização de dados, redução de ruído, compressão de dados e aceleração de algoritmos de aprendizado de máquina. Na figura 5, é demonstrado um caso clássico de uso do PCA para análise dos componentes principais em um problema da área do marketing.

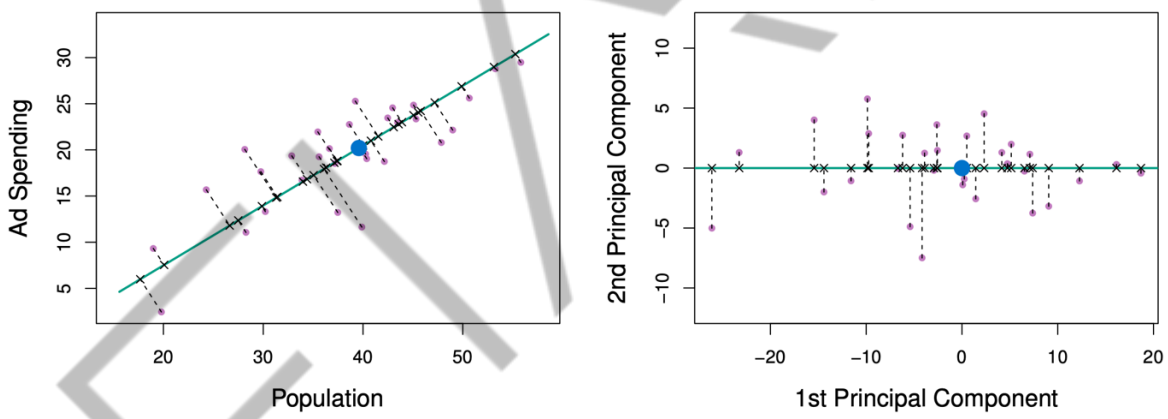

Figura 5 – Exemplo de Análise de Componentes Principais (PCA) para um problema clássico de investimento em mídia para marketing.
Fonte: James et al. (2023)

Além disso, o Aprendizado Não Supervisionado também inclui técnicas de detecção de anomalias, em que o objetivo é identificar padrões incomuns ou outliers nos dados, o que pode ser útil em uma variedade de aplicações, como detecção de fraudes, monitoramento de sistemas e diagnóstico de falhas em maquinários.

A detecção de anomalias é uma das aplicações mais interessantes e importantes do Aprendizado Não Supervisionado. Neste contexto, o objetivo é identificar padrões ou instâncias incomuns nos dados, que podem indicar eventos raros, falhas em sistemas, fraudes ou comportamentos anômalos. Essa capacidade

de detectar anomalias tem aplicações em uma ampla gama de campos, incluindo segurança cibernética, detecção de fraudes financeiras, monitoramento de saúde e manutenção preditiva.

Um dos principais desafios na detecção de anomalias é a natureza não rotulada dos dados, o que torna difícil definir o que constitui uma anomalia. Além disso, as anomalias podem assumir várias formas e manifestações, tornando-as difíceis de serem identificadas por métodos tradicionais. É aqui que o Aprendizado Não Supervisionado desempenha um papel crucial, fornecendo técnicas e algoritmos poderosos para identificar padrões incomuns e anomalias nos dados.

Ao aprender sobre Aprendizado Não Supervisionado, profissionais de ciência de dados e aprendizado de máquina são capacitados a explorar e a entender a estrutura intrínseca dos dados, sem depender de rótulos de saída conhecidos. Isso é especialmente importante em cenários nos quais os dados são escassos ou caros de obter, como em ambientes industriais ou de IoT (internet das coisas).

Além disso, compreender técnicas como clustering, redução de dimensionalidade e detecção de anomalias abre portas para uma série de aplicações práticas. Por exemplo: ao identificar padrões de comportamento suspeitos em transações financeiras, é possível detectar fraudes em tempo real e evitar perdas significativas. Da mesma forma, na área de saúde, a detecção precoce de anomalias em dados de pacientes pode levar a diagnósticos mais precisos e intervenções médicas mais eficazes.

Em resumo, aprender sobre Aprendizado Não Supervisionado não só nos permite explorar e entender os dados de uma maneira mais profunda, mas também nos capacita a desenvolver soluções mais robustas e eficazes para uma variedade de problemas do mundo real. Ao dominar as técnicas de Aprendizado Não Supervisionado, podemos desbloquear todo o potencial dos dados e impulsionar a inovação em uma ampla gama de campos e setores.

Ao longo das últimas décadas, o campo do Aprendizado Não Supervisionado tem passado por uma evolução significativa, impulsionada pelo avanço da tecnologia, o aumento na disponibilidade de dados e a demanda por soluções mais sofisticadas para análise de dados não rotulados. Inicialmente, as técnicas de clustering e redução

de dimensionalidade dominavam o cenário, oferecendo ferramentas poderosas para explorar a estrutura dos dados e extrair insights valiosos.

No entanto, à medida que os conjuntos de dados cresceram em escala e complexidade, surgiram novos desafios e oportunidades. Uma tendência crescente é o uso de técnicas de modelagem de tópicos, que visam descobrir padrões latentes nos dados e identificar temas ou tópicos subjacentes. A modelagem de tópicos é especialmente relevante em campos como processamento de linguagem natural (PLN), em que é essencial extrair informações semânticas e identificar padrões de coocorrência em grandes volumes de texto.

Um dos algoritmos mais proeminentes em modelagem de tópicos é o Latent Dirichlet Allocation (LDA), que assume que os documentos são uma mistura de tópicos e que cada palavra em um documento é atribuída a um tópico específico. O LDA e outras técnicas de modelagem de tópicos permitem a descoberta automática de temas subjacentes em grandes coleções de documentos, facilitando a organização e a compreensão de grandes volumes de dados textuais. Na figura 6 pode-se observar melhor o uso do LDA para um problema de Modelagem de Tópico.

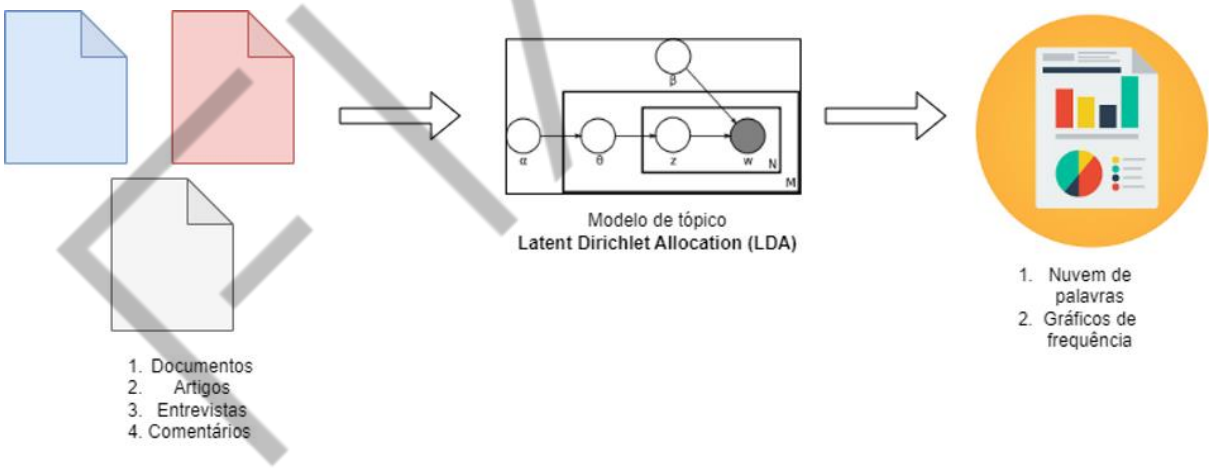

Figura 6 – Exemplo de Modelagem de Tópico se utilizando de LDA.
Fonte: Castro (2020)

Além da modelagem de tópicos, à medida que continuamos a avançar no campo do Aprendizado Não Supervisionado, podemos esperar ver uma maior integração de diferentes técnicas e uma ênfase crescente em interpretabilidade, escalabilidade e eficiência computacional. Em suma, o campo do Aprendizado Não Supervisionado está em constante evolução, impulsionado por novas tecnologias, demandas do mercado e desafios complexos de análise de dados.

Ao passo que exploramos novas fronteiras e desenvolvemos soluções mais avançadas, o Aprendizado Não Supervisionado continuará desempenhando um papel fundamental na extração de conhecimento e insights a partir de dados não rotulados, abrindo novas possibilidades para a inovação e a descoberta. Vamos juntos?

MERCADO, CASES E TENDÊNCIAS

Matéria PT-BR – Provedor de Soluções de Nuvem (IBM)

Para mais detalhes sobre a visão de um player do mercado de nuvem sobre como funciona o Aprendizado Não Supervisionado e como ele pode ser usado a fim de se descobrir e agrupar dados, clique aqui.

Matéria EN-US – Provedor de Soluções de Nuvem (Google Cloud)

O Google é a empresa por trás de um dos frameworks mais utilizados para implementação de algoritmos de Machine e Deep Learning, o TensorFlow. Além disso, ele ainda oferece serviços dedicados para hospedagem de algoritmos de Machine Learning em nuvem. Neste link, é detalhada a visão Google Cloud sobre as tendências em Aprendizado Não Supervisionado.

Matéria PT-BR – Provedor de Soluções de Nuvem (AWS)

Mais detalhes sobre a visão de uma grande player do mercado de nuvem sobre as diferenças fundamentais e usos mais comuns de Aprendizado Supervisionado versus Aprendizado Não Supervisionado podem ser encontradas aqui.

O QUE VOCÊ VIU NESTA AULA?

Nesta aula, mergulhamos em uma experiência prática emocionante ao criar nosso próprio algoritmo K-Means em Python usando o popular dataset penguins. Então, aprendemos como inicializar os centroides aleatoriamente, a atribuir pontos de dados aos clusters mais próximos e a refinar os centroides iterativamente até a convergência. Essa jornada nos proporcionou uma compreensão profunda de como o K-Means funciona e como ele pode ser aplicado em problemas do mundo real.

Além disso, exploramos os fundamentos do Aprendizado Não Supervisionado, incluindo clustering, redução de dimensionalidade e detecção de anomalias. Descobrimos como o clustering é usado de forma a agrupar dados semelhantes e simplificar conjuntos de dados complexos, enquanto a redução de dimensionalidade nos ajuda a visualizar e interpretar dados de alta dimensionalidade de forma mais eficaz. Também entendemos a importância da detecção de anomalias na identificação de padrões incomuns nos dados.

Ao final da aula, além de criarmos nosso próprio algoritmo K-Means, saímos com uma compreensão mais ampla e prática das técnicas e conceitos essenciais em Aprendizado Não Supervisionado. Esta experiência prática nos preparou para explorar novos desafios, além de podermos aplicar nosso conhecimento em projetos futuros de aprendizado de máquina.

REFERÊNCIAS

BERKHIN, P. **Survey of clustering Data Mining Techniques**. Springer-Verlag: Berlin, 2006. Disponível em: <https://faculty.cc.gatech.edu/~isbell/classes/reading/papers/berkhin02survey.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.

CASTRO, G. **Análise de Texto Usando Modelo de Tópico**.2020. Disponível em: <https://www.dataside.com.br/dataside-community/a-i-e-machine-learning/analise-de-texto-usando-modelo-de-topico>. Acesso em: 12 jul. 2024.

GEEKS FOR GEEKS. **Unsupervised Learning**. 2023. Disponível em: <https://www.geeksforgeeks.org/ml-types-learning-part-2/>. Acesso em: 12 jul. 2024.

IEEE. **IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)**. 2024. Disponível em: <https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=34>. Acesso em: 12 jul. 2024.

JAIN, A. K. **Data Clustering**: 50 Years Beyond K-Means. Springer-Verlag: Berlin, 2009. Disponível em: <https://www.cs.odu.edu/~sampath/courses/w17/cs599/papers/reading7/45.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.

JMLR. **Journal of Machine Learning Research (JMLR)**. 2024. Disponível em: <https://www.jmlr.org/>. Acesso em: 12 jul. 2024.

JOLLIFFE, I. T. **Principal Component Analysis**. New York: Springer, 2002. Disponível em: <http://cda.psych.uiuc.edu/statistical_learning_course/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf>. Acesso em: 12 jul. 2024.

MATTT. **DBSCAN**. 2021. Disponível em: <https://github.com/NSHipster/DBSCAN>. 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.