Relatório- Card- 10-Prática: Lidando com Dados do Mundo Real (II)

Jefferson korte junior

Descrição da atividade:

Inicialmente fomos introduzidos ao conceito de **K-Nearest-Neighbors: Concepts:** O algoritmo K-Nearest Neighbors (KNN), uma técnica simples de aprendizado supervisionado e mineração de dados. KNN é utilizado para classificar novos pontos de dados com base nos pontos de dados já classificados. A ideia central é que, ao receber um novo dado, o algoritmo verifica os K vizinhos mais próximos, medidos por uma métrica de distância (como em um gráfico de dispersão), e utiliza o "voto" desses vizinhos para determinar a classificação do novo ponto.
Por exemplo, se tivermos dados de filmes, onde quadrados azuis representam filmes de ficção científica e triângulos vermelhos representam dramas, e precisarmos classificar um novo filme, o algoritmo verifica os K vizinhos mais próximos e faz a decisão com base nas classificações dos vizinhos. Se K = 3, por exemplo, e houver 2 filmes dramáticos e 1 de ficção científica, o novo filme seria classificado como drama.
O valor de K é importante: ele deve ser pequeno o suficiente para evitar incluir vizinhos irrelevantes, mas grande o suficiente para garantir uma amostra representativa. O texto enfatiza que a escolha de K muitas vezes requer experimentação.
No exemplo prático, a ideia é aplicar KNN para encontrar filmes semelhantes entre si, usando metadados como classificações e gêneros. Com isso, seria possível criar recomendações de filmes, como os "clientes que também assistiram" da Amazon, e prever a classificação de novos filmes com base nos vizinhos mais próximos.

Redução de dimensionalidade é uma técnica em ciência de dados e machine learning essencial para lidar com dados complexos e de alta dimensionalidade, usada para diminuir o número de variáveis (dimensões) em um conjunto de dados, enquanto se tenta preservar a maior quantidade possível de informações relevantes.Tudo isso é capaz de fazer usando métodos como Análise de Componentes Principais(PCA) e a Decomposição de Valor Singular (SVD), Isso é útil por várias razões, incluindo simplificação de modelos, melhoria da visualização de dados e aceleração de algoritmos de processamento de dados.

O exemplo da flor Iris é utilizado para ilustrar a aplicação prática dessa técnica, onde quatro características (dimensões) são resumidas em duas, sem sacrificar as nuances necessárias para classificação.

ETL e ELT:

ETL é um processo essencial especialmente em ambientes de data warehousing. Este processo é fundamental para coletar, transformar e carregar dados de diferentes fontes para

um local centralizado, onde podem ser analisados e utilizados para tomada de decisões. Incluindo três passos:

Extração: Coleta de dados brutos de várias fontes, como bancos de dados, arquivos, aplicativos SaaS, sensores IoT.

Transformação: Processamento dos dados brutos para converter em um formato adequado para análise. Isso pode incluir a remoção de dados inconsistentes, conversão de tipos de dados, remoção de duplicatas

Carregar: A etapa final do ETL é carregar os dados transformados em um sistema de destino, como um data warehouse, banco de dados ou sistema de BI..

ELT: é um método de integração de dados onde os dados são primeiro extraídos, depois carregados diretamente no sistema de destino, e então transformados dentro do sistema de destino

Extrair: A primeira etapa do ELT, assim como no ETL, é a extração de dados de várias fontes. Essas fontes podem incluir bancos de dados, arquivos, APIs e muito mais.

Carregar: A diferença principal no ELT é que os dados extraídos são carregados diretamente no sistema de destino, como um data lake ou data warehouse.

Transformar: : A transformação dos dados ocorre após o carregamento, dentro do próprio sistema de destino. Utiliza-se o poder de processamento do data warehouse para transformar os dados.

Aprendizado por Reforço (RL): É uma área do aprendizado de máquina onde um agente aprende a tomar decisões em um ambiente para maximizar uma recompensa cumulativa. Diferente do aprendizado supervisionado (onde há um conjunto de dados rotulados) e do aprendizado não supervisionado (onde o objetivo é encontrar padrões ocultos), no RL, o agente aprende com as consequências de suas ações.

Matriz de confusão: Uma matriz de confusão é uma ferramenta poderosa utilizada principalmente na avaliação do desempenho de algoritmos de classificação em aprendizado de máquina. Ela é uma tabela que permite a visualização dos resultados de um modelo de classificação, mostrando como as previsões do modelo se comparam aos resultados reais.

Componentes da Matriz de Confusão:

Verdadeiro Positivo (TP): Quando o modelo prevê "sim" e o valor real também é "sim"

Falso Positivo (FP): Quando o modelo prevê "sim", mas o valor real é"não" (também chamado de erro tipo I):

Falso Negativo (FN): Quando o modelo prevê "não", mas o valor real é"sim" (também chamado de erro tipo II).

Verdadeiro Negativo (TN): Quando o modelo prevê "não" e o valor real também é "não"

Exemplo:

Imaginemos um modelo que classifica imagens como contendo um gato ou não.

há um gato (50 vezes).

TP (Verdadeiro Positivo): O modelo prevê que há um gato, e realmente

FP (Falso Positivo): O modelo prevê que há um gato, mas na verdade não há (5 vezes).

FN (Falso Negativo): O modelo prevê que não há um gato, mas na verdade há (10 vezes).

TN (Verdadeiro Negativo): O modelo prevê que não há um gato, e de fato não há (100 vezes).

Métricas derivadas da matriz de confusão:

1: Precisão: A precisão avalia a proporção de instâncias corretamente classificadas como positivas em relação ao total de classificações positivas, e é mais relevante quando se busca minimizar falsos positivos. Um exemplo clássico seria um teste de drogas, onde classificar alguém erroneamente como positivo pode ter consequências graves.

2: Recall (Sensibilidade): Mede a capacidade do modelo de identificar corretamente instâncias positivas, calculado como razão entre verdadeiros positivos e a soma dos verdadeiros positivos com os falsos negativos. É especialmente útil quando o foco está em minimizar erros falsos negativos, como em cenários de detecção de fraudes ou diagnóstico médico.

3: A pontuação **F1** é a média harmônica entre precisão e recall, útil quando é necessário equilibrar ambos, sem priorizar um sobre o outro. Ela oferece uma visão geral do desempenho do modelo quando tanto a precisão quanto o recall são importantes. A especificidade, ou taxa de verdadeiros negativos, é outra métrica derivada da matriz de confusão que se concentra em quantos dos negativos foram corretamente identificados, sendo relevante em situações onde falsos positivos devem ser minimizados.

A curva ROC (Receiver Operating Characteristic) e a AUC (Area Sob a Curva) medem o desempenho do modelo em diferentes limiares de decisão. A curva ROC traça o recall versus a taxa de falsos positivos, e a AUC oferece uma métrica numérica que varia entre 0.5 (classificação aleatória) e 1 (classificação perfeita). Juntas, essas métricas fornecem uma visão

abrangente do desempenho de classificadores e ajudam na comparação entre diferentes modelos.

VIES E VARIANCIA

O vies refere-se ao quanto as previsões de um modelo se afastam dos valores corretos, ou seja, se o modelo consistemente erra em uma direção específica

Variância é a medida de quanto as suas previsões mudam quando você usa diferentes conjuntos de dados de treinamento. Voltando ao exemplo dos dardos, imagine que agora você joga de várias maneiras diferentes, acertando dardos em várias partes do alvo, sem consistência. Isso é alta variância, onde seu modelo é muito sensível aos detalhes específicos dos dados de treinamento, capturando até mesmo os erros (ou "ruídos").

LIMPANDO DADOS E NORMALIZAÇÃO:A limpeza de dados é uma das etapas mais importantes e desafiadoras no processo de ciência de dados, frequentemente demandando mais tempo do que a análise em si. Dados brutos são muitas vezes sujos e poluídos, o que pode distorcer significativamente os resultados de um modelo. Isso torna a limpeza e a preparação dos dados essenciais para garantir a qualidade dos resultados.

Existem diversos problemas que podem ocorrer com os dados, como outliers (valores fora do padrão esperado), dados ausentes, dados maliciosos (exemplos incluem ataques de sistemas automatizados), dados errôneos (causados por bugs ou falhas de software), dados irrelevantes, dados inconsistentes (onde informações são representadas de diferentes maneiras), e formatação inadequada (como diferenças regionais em datas e números de telefone). Cada um desses problemas precisa ser resolvido para evitar que interfiram no desempenho do modelo e conduzam a decisões erradas.

Um modelo bem ajustado com dados limpos pode muitas vezes superar modelos mais complexos que utilizam dados sujos. Além disso, é crucial questionar constantemente os resultados, mesmo que pareçam corretos, para evitar viéses não intencionais e garantir a confiabilidade das análises. Portanto, a qualidade e a quantidade dos dados têm um impacto direto na eficácia dos algoritmos e nos insights gerados.

NORMALIZANDO DADOS NUMÉRICOS:

A normalização e a escalabilidade dos dados são fundamentais ao preparar entradas para algoritmos de aprendizado de máquina. É crucial garantir que diferentes atributos numéricos estejam na mesma escala e sejam comparáveis, ajudando a evitar que atributos com valores mais altos dominem o modelo e introduzam vieses.

Ao trabalhar com atributos de diferentes escalas, como idade variando de 0 a 100 anos e renda de 0 a bilhões, a normalização se torna essencial, especialmente em modelos que não lidam bem com escalas variadas, como aqueles que utilizam distâncias, como KNN ou SVM. Ferramentas como Scikit-Learn oferecem métodos automáticos para normalizar e escalar dados.

Por exemplo, o PCA (Análise de Componentes Principais) possui opções para normalizar os dados automaticamente.

Além disso, dados categóricos, como respostas sim/não, devem ser convertidos em formatos numéricos, como 0 e 1, para serem utilizados em modelos. Após a normalização, é importante reverter o processo antes de apresentar os resultados, garantido que sejam interpretáveis na escala original

TECNICAS DE IMPUTAÇÃO PARA DADOS AUSENTES:

A engenharia de características é o processo de seleção, transformação e criação de atributos a partir dos dados de treinamento, visando melhorar a eficácia de um modelo. A escolha e a transformação adequadas de características são cruciais para prever resultados, e esse processo pode ajudar a reduzir a maldição da dimensionalidade, que torna a análise mais complexa quando há muitos atributos.

No que diz respeito à imputação de dados ausentes, é comum encontrar valores faltantes nos dados do mundo real. Uma abordagem simples e a substituição média, onde os valores ausentes são substituídos pela média da coluna correspondente. Essa técnica é rápida, mas pode ser insatisfatória em presença de outliers ou correlações importantes entre características. Uma alternativa é a substituição pela mediana, que pode ser mais adequada em conjuntos de dados com valores extremos.

Outra técnica é o KNN (M Nearest Neighbors), que consiste em encontrar as linhas mais semelhantes que possuem dados completos e usar a média desses valores para imputar os ausentes. Modelos de aprendizado de máquina também podem ser utilizados para prever os valores ausentes com base em outras características, utilizando redes neurais ou regressões. MICE, ou Imputação Múltipla por Equações Encadeadas, é uma técnica avançada que considera a incerteza em imputações.

No entanto, a melhor solução para lidar com dados ausentes muitas vezes é simplesmente obter mais dados. Coletar informações adicionais pode ser mais eficaz do que confiar apenas em técnicas de imputação, desde que se evite viés ao remover linhas com dados faltantes. Em suma, um entendimento sólido e prático dessas técnicas é fundamental para aumentar a qualidade e a precisão dos modelos de aprendizado de máquinas.

MANIPULAÇÃO DE DADOS DESEQUILIBRADOS:

O desafio do manuseio de dados desequilibrados na engenharia de características, especialmente em contextos como a detecção de fraudes, onde a maioria dos dados são negativos (não fraudulentos) e apenas uma pequena fração representa casos positivos (fraudulentos). Essa discrepância pode levar um modelo a prever sempre a classe majoritária, resultando em alta precisão aparente, mas sem eficácia real na detecção dos casos positivos.

Uma forma de abordar esse problema é por meio da sobreamostragem, que envolve a duplicação de amostras da classe minoritária para aumentar sua representação no conjunto de dados. Embora essa abordagem possa ser útil, ela também tem suas limitações e pode não ser a melhor solução, especialmente se resultar em perda de informações valiosas.

Outra alternativa é a subamostragem, que consiste em remover casosda classe majoritária. No entanto, essa abordagem geralmente não é recomendada, a menos que o volume de dados seja tão grande que justifique a exclusão de informações. Em vez disso, buscar mais poder computacional é uma solução mais eficaz.

Uma técnica superior a ambas é o SMOTE (Synthetic MinorityOver-sampling Technique), que gera amostras artificiais da classe minoritária utilizando algoritmos de vizinhos mais próximos (KNN). Essa abordagem cria novos pontos de dados a partir de médias de amostras existentes, o que pode melhorar a performance do modelo ao preservar a estrutura dos dados.

Binning, Transforming, Encoding, Scaling, and Shuffling:

A fiação, ou binning, consiste em transformar dados numéricos em categóricos, agrupando valores em faixas. Por exemplo, as idades podem ser agrupadas em intervalos como 20-29 ou 30-39. Essa abordagem ajuda a lidar com incertezas nas medições e simplifica a modelagem ao reduzir a precisão dos dados. O quantile binning é uma forma de fiação que assegura que cada categoria tenha o mesmo número de amostras, melhorando a distribuição dos dados em cada grupo.

A codificação, especialmente a codificação one-hot, cria variáveis indicadoras para representar categorias. No caso de números de 0 a 9, por exemplo, são criadas 10 variáveis, onde apenas uma será "1" e as outras serão "O". Essa técnica é essencial para transformar dados categóricos em um formato que algoritmos de aprendizado de máquina, especialmente redes neurais, consigam processar.

escalonamento e normalização são práticas que ajustam as magnitudes dos dados, assegurando que todas as características tenham pesos comparáveis. Por exemplo, normalizar idades e rendas evita que características de maior magnitude tenham um impacto desproporcional nas previsões, melhorando a desempenho do modelo.

embaralhamento envolve aleatorizar a ordem dos dados de treinamento para evitar padrões indesejados que influenciem o aprendizado. Isso ajuda a garantir que o modelo aprenda de forma generalizada, em vez de se adaptar a uma ordem especifica dos dados. Essas técnicas são fundamentais para otimizar como os dados são utilizados no treinamento de modelos de aprendizado de máquina, maximizando a eficácia das previsões.

Conclusão: Este relatório aborda diversos conceitos e técnicas importantes no campo da ciência de dados e aprendizado de máquina. Desde o entendimento do algoritmo K-Nearest-Neighbors (KNN) e a importância da redução de dimensionalidade até os processos de ETL e ELT, cada seção destaca métodos essenciais para a preparação e análise de dados. O aprendizado por reforço tambem é explorado. A matriz de confusão é apresentada como uma ferramenta vital para a avaliação de modelos de classificação, e sobre métricas derivadas como precisão, recall e a pontuação F1. Além disso, a importância de balancear o viés e a variância para evitar modelos ineficazes, e a necessidade crítica da limpeza e normalização de dados para assegurar a qualidade das análises.

Referências:

Machine Learning, Data Science and Deep Learning with Python:

6: More Data Mining and Machine Learning Techniques

7:Dealing with Real-World Data