Relatório <10> - <Lidando com Dados do Mundo Real (II)>

<Breno Augusto Oliveira Abrantes>

Descrição da atividade

Nessa atividade vamos explorar vários conceitos ligados ao Machine Learning, indo do inicialmente apresentado KNN até mesmo o Q-learning, durante o caminho são apresentados inúmeros conceitos que nos ajudarão no futuro a criar bons modelos com base na concretização desse conhecimento. São apresentados conceitos totalmente novos como os citados acima, mas alguns já vistos durante o curso voltam para serem retomados, como o caso dos outliers. Assim, são realizados vários testes e novos métodos para amplificar nosso conteúdo.

K-Nearest Neighbors: Conceitos

O K-Nearest Neighbors (KNN) é uma técnica de aprendizado supervisionado bastante simples. Quando recebemos um novo ponto de dados, os vizinhos mais próximos no gráfico votam para decidir sua classificação. Usamos uma métrica de distância, como popularidade, e o "K" em KNN indica quantos vizinhos vamos considerar. Escolher o valor de K é muito importante: ele precisa ser grande o suficiente para garantir que tenhamos uma amostra significativa, mas pequeno o bastante para não incluir pontos irrelevantes. Por exemplo, ao recomendar filmes, o KNN pode prever o gênero ou a classificação de um filme com base nos filmes semelhantes que estão mais próximos, utilizando metadados como gêneros e classificações.

Redução de Dimensionalidade: Análise de Componentes Principais (PCA)

A redução da dimensionalidade é útil porque facilita a visualização e a compressão de dados, mantendo a variação mais importante. A técnica PCA, junto com a Decomposição de Valor Singular (SVD), projeta dados em hiperplanos de menor dimensão, o que ajuda a simplificar a análise de conjuntos de dados complexos. Exemplos práticos dessa técnica incluem compressão de imagens e reconhecimento facial. Um caso interessante é o conjunto de dados "Iris", que permite visualizar características de flores em duas dimensões ao invés de quatro.

Aprendizado por Reforço

O aprendizado por reforço envolve treinar um agente, como o Pac-Man, para explorar um ambiente (o labirinto) e aprender com as consequências de suas ações. O Pac-Man toma decisões baseadas nas mudanças de estado, influenciadas por recompensas ou penalidades, como comer uma pílula (positivo) ou ser pego por um fantasma (negativo). O Q-learning é uma forma popular de aprendizado por reforço, onde cada ação recebe um valor Q que é ajustado conforme o agente explora o ambiente. À medida que o Pac-Man aprende, ele aprimora suas escolhas para maximizar as recompensas, baseando-se em valores Q armazenados. Além disso, o texto menciona processos de decisão de Markov e programação dinâmica, que são conceitos matemáticos que ajudam a otimizar o aprendizado, dividindo o problema em partes menores e reutilizando soluções já calculadas. Essa técnica pode ser aplicada não apenas em jogos, mas também em diversas situações que envolvem prever comportamentos com base em condições e possíveis ações.

Entendendo a Matriz de Confusão

Na matriz de confusão, aprende-se que ela é uma ferramenta essencial para avaliar a performance de um modelo de ML (Machine Learning), especialmente quando a precisão sozinha não é suficiente. Ela ajuda a identificar os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, permitindo uma compreensão mais profunda da eficácia do modelo. Por exemplo, um modelo que prevê uma doença rara pode ter uma aparência de alta precisão, mas se sempre prever que ninguém está doente (falsos negativos), sua utilidade é bastante limitada. A matriz de confusão organiza esses resultados visualmente em uma tabela, comparando os valores reais e previstos. O objetivo é ter a maioria dos resultados na diagonal da matriz, o que indica previsões corretas. Também é importante prestar atenção aos rótulos, pois a organização pode variar. Em problemas de classificação com múltiplas categorias, a matriz pode ser ampliada, e o uso de mapas de calor pode ajudar a visualizar a precisão de forma mais clara.

Métricas para Avaliação de Classificadores (Precisão, Recall, F1, ROC, AUC)

A matriz de confusão permite calcular várias métricas importantes, como sensibilidade e precisão. A sensibilidade mede a proporção de verdadeiros positivos em relação à soma de verdadeiros positivos e falsos negativos. Essa métrica é especialmente relevante quando os falsos negativos são críticos, como na detecção de fraudes. Já a precisão refere-se à proporção de verdadeiros positivos em relação à soma de verdadeiros positivos e falsos positivos, sendo útil em uma situação como em triagens médicas. Já curva ROC plota a taxa de verdadeiros positivos em comparação com a taxa de falsos positivos em diferentes limiares; quanto mais a curva se aproxima do canto superior esquerdo, melhor. A AUC fornece uma medida numérica dessa performance, variando de 0.5 (aleatório) a 1.0 (perfeito).

Bias / Variance Tradeoff

O trade-off entre viés e variância é um conceito importantíssimo na modelagem preditiva, para transformar e ajustar um modelo a chegar na sua melhor performance. O viés se refere ao erro sistemático que ocorre quando um modelo não consegue capturar a verdadeira relação nos dados, resultando em previsões imprecisas. Por outro lado, a variância indica a sensibilidade do modelo às flutuações nos dados de treinamento. Modelos com baixo viés e alta variância podem prever com precisão em média, mas suas previsões podem ser bastante dispersas. Ao contrário, modelos com alto viés e baixa variância tendem a ser mais consistentes, embora ainda aconteçam erros ocasionais em suas previsões. Modelos complexos tendem a ter alta variância, capturando até os ruídos do conjunto de dados, enquanto modelos simples geralmente apresentam baixo viés, mas não conseguem ter a complexidade abrangida pelo modelo retrocitado. O cálculo do erro total considera tanto o viés quanto a variância, enfatizando a importância de minimizar o erro total para alcançar um desempenho ideal do modelo.

Data Cleaning and Normalization

A limpeza de dados é uma etapa crítica em projetos de ciência de dados, influenciando diretamente a qualidade e a confiabilidade dos resultados. Dados brutos frequentemente contêm erros e inconsistências, podendo comprometer a validade das análises. Problemas comuns incluem outliers, dados ausentes e informações maliciosas. O tratamento de outliers é fundamental, pois eles podem distorcer as análises. Dados ausentes exigem

decisões cuidadosas, como substituições ou remoções, enquanto dados maliciosos precisam de verificação rigorosa para garantir a integridade dos resultados. A normalização dos dados, que envolve uniformizar formatos, é crucial para evitar erros de interpretação. Esse processo, embora muitas vezes considerado menos glamoroso, é essencial para garantir que as conclusões tiradas sejam robustas e precisas. Ao questionar constantemente a qualidade dos dados, é possível evitar vieses e garantir que a modelagem seja feita com informações confiáveis.

Normalizing Numerical Data

Normalizar dados é vital ao preparar informações para algoritmos de machine learning. A normalização garante que os atributos estejam em uma escala comparável, o que é essencial para a precisão do modelo. A falta de normalização pode levar a um viés na análise, principalmente quando atributos variam em escalas muito diferentes. Embora nem todos os modelos exijam normalização, aqueles que trabalham com múltiplas variáveis se beneficiam enormemente. É importante converter dados textuais e categóricos em formatos numéricos, facilitando a interpretação dos resultados. Lembrar-se de reverter a escala dos dados após a modelagem também é fundamental para a clareza na apresentação dos resultados. Esse cuidado com a normalização e a clareza é muitas vezes negligenciado, mas é essencial para garantir a confiabilidade das análises.

Feature Engineering and the Curse of Dimensionality

Engenharia de características é um componente crucial na construção de modelos de machine learning, envolvendo a seleção e transformação de atributos a partir dos dados disponíveis. O sucesso do modelo depende da qualidade dessas características. O processo inclui identificar atributos relevantes, normalizar dados e criar novas características que ajudem a capturar tendências. No entanto, a "maldição da dimensionalidade" representa um desafio, pois adicionar muitas características pode tornar os dados esparsos, dificultando a identificação de padrões e a eficiência do treinamento do modelo. Portanto, a seleção cuidadosa das características é fundamental, garantindo que apenas as mais relevantes sejam utilizadas para alcançar resultados efetivos.

Imputation Techniques for Missing Data

Imputar dados ausentes é uma prática essencial já que muitos conjuntos de dados contêm elementos faltantes que podem comprometer ou até mesmo destruir a qualidade das previsões. A substituição média é uma técnica comum, mas pode ser problemático em presença de outliers. A mediana, menos influenciada por valores extremos, muitas vezes oferece uma alternativa mais robusta. No entanto, essa abordagem pode ignorar correlações importantes entre características. Descartar linhas com dados ausentes pode ser viável, mas deve ser feito com cautela para não afetar a representatividade do conjunto de dados. A escolha da técnica de imputação deve sempre considerar o contexto dos dados para garantir a precisão das análises.

Binning, Transforming, Encoding, Scaling, and Shuffling

A "fiação" permite agrupar dados numéricos em categorias, facilitando a modelagem em cenários onde as medições não sabem se estão completamente corretas, acarretando incertezas. O "quantile binning" garante que cada faixa contenha um número igual de amostras, melhorando a representação dos dados. Transformações matemáticas, como logaritmos, podem ajudar a linearizar relações não lineares. A codificação de dados, especialmente a "one-hot encoding", transforma categorias em formatos numéricos adequados para modelos de aprendizado profundo. A normalização e o escalonamento garantem que todos os atributos tenham uma distribuição próxima, o que é fundamental para o desempenho do modelo. Utilizar ferramentas apropriadas facilita esse processo e melhora a eficácia da modelagem.

Conclusões

A aplicação de boas práticas em ciência de dados é fundamental para garantir o sucesso de projetos de machine learning. Compreender conceitos como viés e variância ajuda a encontrar um equilíbrio que previne problemas de ajuste, como o overfitting e o underfitting. A limpeza e normalização dos dados são passos essenciais, pois dados bem tratados influenciam diretamente a precisão das análises e previsões. Ao aprender a

importância da engenharia de características, é possível identificar e transformar os atributos mais relevantes, otimizando o desempenho do modelo. Bibliotecas como Scikit facilitam a implementação dessas práticas, permitindo se possa modelos úteis e robustos. Portanto, cada técnica aprendida forma uma grande base para um cientista de dados ter em seu leque, ainda que não sejam sempre utilizadas, ter elas ao seu alcance ajudam a sempre buscar novos meios de melhorar seu projeto.

https://github.com/brenooabrantes/BT-Machine-Learning-Lamia/tree/main/CARD10

Referências

https://inst.eecs.berkeley.edu/~cs188/fa24/projects/

https://pymdptoolbox.readthedocs.io/en/latest/

https://scikit-learn.org/stable/

https://realpython.com/knn-python/

https://docs.python.org/3/library/re.html