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

Willian Augusto Soder de Souza

O objetivo deste relatório é citar os principais conhecimentos adquiridos com as duas seções indicadas do curso "Machine Learning, Data Science and Generative AI with Python". Essas seções apresentam diversos conceitos extremamente importantes dentro do Machine Learning, além de fornecer uma introdução ao Python utilizando algumas bibliotecas, como a Scikit-learn. Abaixo segue uma lista dos principais conceitos ensinados no vídeo, acompanhada de uma breve explicação sobre cada um deles:

- **K-Nearest Neighbor (KNN):** é um algoritmo de aprendizado de máquina usado para classificação e regressão. Para classificar um novo dado, o algoritmo KNN encontra os 'K' dados mais próximos (vizinhos) no conjunto de treinamento. A classe mais frequente entre esses vizinhos é atribuída ao novo dado e, para prever um valor numérico, o KNN calcula a média dos valores dos 'K' vizinhos mais próximos.
- **Dimensionality Reduction:** é uma técnica usada para reduzir o número de variáveis (ou dimensões) em um conjunto de dados, mantendo o máximo de informação possível. Isso ajuda a simplificar o modelo, diminuir o tempo de processamento e visualizar dados complexos.
- **Principal Component Analysis (PCA):** é uma técnica de redução de dimensionalidade que transforma um conjunto de dados com muitas variáveis em um novo conjunto de variáveis chamadas componentes principais. Essas componentes são combinações lineares das variáveis originais e são ordenadas de forma que a primeira componente capture a maior parte da variância dos dados, a segunda captura a segunda maior parte, e assim por diante.
- **Data Warehousing:** é um processo de coleta, armazenamento e gerenciamento de grandes volumes de dados para análise e tomada de decisões.
- **ETL (Extract, Transform, Load):** é um processo onde os dados são extraídos de várias fontes, transformados (limpeza, agregação, formatação) para se ajustarem ao formato desejado, e então carregados em um armazém de dados (data warehouse).
- **ELT (Extract, Load, Transform):** é um processo onde os dados são extraídos e carregados diretamente no armazém de dados primeiro. A transformação dos dados ocorre depois, dentro do próprio armazém.
- **Reinforcement Learning:** é uma técnica de aprendizado de máquina onde um agente aprende a tomar decisões através da interação com um ambiente. O agente executa ações e recebe recompensas ou penalidades com base nas suas ações. O objetivo é aprender uma política (ou estratégia) que maximize a soma total das recompensas recebidas ao longo do tempo. O agente ajusta seu comportamento com base no feedback do ambiente para melhorar seu desempenho e alcançar objetivos específicos.
- **Precision:** é a proporção de verdadeiros positivos (casos corretamente identificados como positivos) em relação ao total de casos identificados como positivos (verdadeiros positivos + falsos positivos). Mede a exatidão do classificador ao identificar a classe positiva.
- **Recall:** é a proporção de verdadeiros positivos em relação ao total de casos realmente positivos (verdadeiros positivos + falsos negativos). Mede a capacidade do classificador de identificar todos os casos positivos.
- **F1-Score:** é a média harmônica entre precision e recall. É uma métrica que combina ambas para fornecer uma visão equilibrada do desempenho do classificador, especialmente quando há um desbalanceamento entre as classes.

- **ROC Curve (Receiver Operating Characteristic):** é um gráfico que mostra a relação entre a taxa de verdadeiros positivos (recall) e a taxa de falsos positivos para diferentes limiares de decisão. A curva ROC ajuda a visualizar a capacidade do classificador de distinguir entre classes.
- **AUC (Area Under the Curve):** é a área sob a curva ROC. Mede a habilidade geral do classificador em distinguir entre classes. Um AUC de 1 indica um classificador perfeito, enquanto um AUC de 0,5 indica um classificador que não é melhor que o acaso.
- **Bias/Variance:** Bias refere-se ao erro sistemático introduzido por um modelo ao assumir uma simplificação excessiva dos dados, enquanto a Variance refere-se à sensibilidade do modelo às flutuações nos dados de treinamento.
- **K-Fold Cross-Validation:** é uma técnica para avaliar a performance de um modelo de aprendizado de máquina.
- **Data Cleaning:** refere-se ao processo de identificar e corrigir ou remover erros e inconsistências nos dados, como valores ausentes, duplicados ou incorretos. O objetivo é garantir que os dados sejam precisos, consistentes e úteis para análise.
- **Normalization:** é o processo de ajustar os dados para que estejam em uma escala comum, a fim de melhorar a eficácia dos algoritmos de aprendizado de máquina.
- **Outlier:** é um dado que se desvia significativamente dos outros valores em um conjunto de dados. Pode ser muito maior ou muito menor do que a maioria dos dados.
- **Feature Engineering:** é o processo de criar, transformar ou selecionar variáveis (ou atributos) a partir dos dados brutos para melhorar o desempenho de modelos de aprendizado de máquina.
- **The Curse of Dimensionality:** refere-se aos problemas que surgem quando o número de características (ou dimensões) em um conjunto de dados aumenta significativamente.
- **Imputing Missing Data (Mean Replacement):** é uma técnica para lidar com dados ausentes substituindo os valores faltantes pela média dos valores existentes em uma variável.
- **Imputing Missing Data (Dropping):** é uma técnica para lidar com dados ausentes que consiste em remover as linhas ou colunas com valores faltantes.
- **Imputing Missing Data (Machine Learning):** é uma técnica que usa algoritmos de aprendizado de máquina (KNN, Deep Learning, Regression, etc.) para prever e preencher valores ausentes com base em padrões encontrados nos dados.
- **Unbalanced Data:** ocorre quando as classes em um conjunto de dados têm distribuições muito diferentes (desequilibradas).
- **Oversampling:** é uma técnica usada para lidar com dados desequilibrados, onde se aumenta a quantidade de exemplos da classe minoritária.
- **Undersampling:** é uma técnica usada para lidar com dados desequilibrados, onde se reduz a quantidade de exemplos da classe majoritária.
- **SMOTE:** é uma técnica de oversampling usada para lidar com dados desequilibrados. Em vez de simplesmente duplicar exemplos da classe minoritária, o SMOTE gera novas amostras sintéticas.
- **Binning:** é uma técnica de pré-processamento de dados que agrupa valores contínuos em intervalos discretos, ou "bins".
- **Transforming:** é o processo de aplicar funções matemáticas ou operações aos dados para alterar sua escala, distribuição ou formato.
- **Encoding:** é o processo de transformar dados categóricos em uma representação numérica que seja aceita pelo modelo.
- **Scaling:** é o processo de ajustar os valores das variáveis para que estejam dentro de uma mesma escala.

- **Shuffling:** é o processo de reorganizar aleatoriamente as amostras em um conjunto de dados.

Em resumo, esses são os conceitos apresentados no curso. Vale ressaltar que foi ensinada a prática desses conceitos no Python, e suas aplicações encontram-se nos arquivos de código. Neste relatório, tentei fazer um resumo dos conceitos teóricos.

CONCLUSÃO

Concluindo, a manipulação de dados e a metrificação de resultados são fundamentais para o sucesso na aplicação de modelos de machine learning. A capacidade de limpar, transformar e preparar os dados garante que os modelos recebam informações precisas e relevantes, melhorando sua eficiência e precisão. Além disso, a avaliação contínua dos modelos através de métricas apropriadas permite ajustar e aprimorar os algoritmos, garantindo resultados robustos e confiáveis. Portanto, dominar essas habilidades é essencial para qualquer profissional que deseja aproveitar o potencial do machine learning em suas análises e soluções.