Relatório sobre a atividade "10 - Prática: Lidando com Dados do Mundo Real (II)"

Lucas Gabriel Arenhardt

Nesse módulo são apresentados diversos conceitos muito importantes em Machine Learning, além de suas aplicações em Python. Abaixo estão apresentados, de forma resumida, os conceitos aprendidos nas aulas:

K Vizinhos mais Próximos: é um algoritmo de aprendizado supervisionado que classifica ou prediz dados com base nos K pontos de dados mais próximos no conjunto de treinamento.

Dimensionality Reduction: esse método reduz o número de dimensões num conjunto de dados de forma a preservar o maior número possível de informações.

Principal Component Analysis: é uma técnica de redução de dimensionalidade. Funciona identificando as componentes principais nas quais os dados variam mais, projetando os dados nesses componentes e ordenando-os de forma que os primeiros componentes capturem a maior parte da variação nos dados, permitindo a redução do número de variáveis e mantendo a maior quantidade possível de informação original.

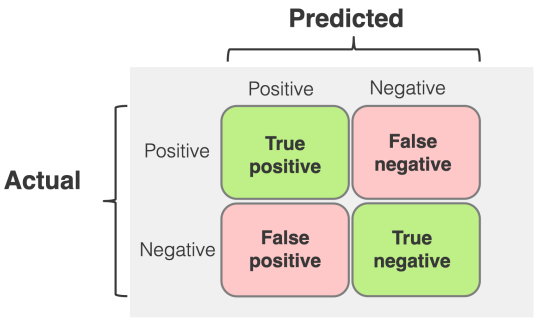
Data Warehouse: é um grande banco de dados que contém informação de várias fontes.

ETL: primeiramente, extrai-se os dados. Em seguida, as informações são transformadas na estrutura utilizada pelo Data Warehouse. Por último, é feito o upload no DW.

ELT: primeiramente, extrai-se os dados. Em seguida, é feito o upload diretamente no DW. Por último, é feita a transformação dos dados dentro do próprio DW (in-place).

Reinforcement Learning: é um tipo de aprendizado onde um agente aprende a tomar decisões com base nas interações com o ambiente. O agente executa ações e recebe recompensas ou punições, de acordo com o que foi executado. Assim, ele ajusta sua estratégia para maximizar as recompensas ao longo do tempo. Esse processo envolve explorar o conhecimento acumulado para melhorar o desempenho.

Confusion Matrix: é uma ferramenta usada para avaliar o desempenho de um modelo de aprendizado de máquina, especialmente em problemas de classificação. Um exemplo básico seria a seguinte matriz:



Recall: mede a capacidade do modelo de identificar corretamente todas as instâncias positivas dentro do conjunto de dados. Ele é dado pelo número de verdadeiros positivos dividido por (verdadeiros positivos + falsos negativos).

Precision: é uma métrica que avalia a qualidade das previsões positivas feitas por um modelo de classificação. É dado por: número de verdadeiros positivos dividido por (verdadeiros positivos + falsos positivos).

Specificity: métrica que mede a capacidade do modelo de identificar corretamente os casos negativos.

F1-Score: métrica que combina a precisão e o recall de um modelo de classificação em uma única medida.

ROC: a curva ROC é gerada ao plotar a Taxa de Verdadeiros Positivos x a Taxa de Falsos Positivos, para diferentes valores de corte. Quanto mais a linha estiver para o canto superior esquerdo, melhor.

AUC: é uma métrica que resume o desempenho do modelo ao longo de todos os pontos de corte. A AUC é o valor da área abaixo da curva ROC. Se o valor for 0.5, significa que o modelo é horrível. O valor 1 é perfeito.

Bias: é o quão longe a média das predições está do valor real.

Variance: é o quão dispersas estão as predições em relação ao valor real.

Erro: seu cálculo é dado por $Bias^2$ + Variance.

K-Fold Cross Validation: é uma técnica de validação cruzada usada para avaliar a performance de um modelo de aprendizado de máquina. Ela funciona dividindo o conjunto de dados em K subconjuntos de tamanhos aproximadamente iguais. O modelo, então, é treinado K vezes, cada vez usando K-1 dos subconjuntos para treinamento e o restante para validação.

Outliers: são valores muito discrepantes do conjunto e que podem ocasionar resultados indesejados.

Feature Engineering: é o processo de usar o conhecimento sobre os dados para criar novas variáveis que podem ser mais úteis do que as informações brutas. Um dos seus principais usos é para quando há muitas dimensões a serem trabalhadas (o que pode ocasionar um problema), então devem-se ser selecionadas as que mais impactam no resultado - ou até mesmo pode-se criar uma nova variável que relaciona outras duas ou mais.

Lidando com dados faltantes:

Substituição pela média: Substitui o valor nulo pela média dos valores da respectiva coluna.

Dropping: Exclui a linha que possui algum valor nulo (útil para quando já se tem uma boa quantidade de dados).

Machine Learning: Pode-se utilizar técnicas de Machine Learning para assumir os valores faltantes, por exemplo: KNN, Deep Learning e regressão

Lidando com dados desbalanceados:

Oversampling: Duplica amostras da classe minoritária.

Undersampling: Remove algumas amostras negativas.

SMOTE (Synthetic Minority Oversampling TEchnique): gera novas amostras da classe minoritária de forma artificial, utilizando o método do vizinho mais próximo.

Demais técnicas utilizadas em Feature Engineering:

Binning: agrupa valores contínuos em intervalos discretos.

Transforming: aplica funções matemáticas aos dados para permitir que eles se encaixem melhor ao treinamento.

Encoding: é o processo de transformar os dados em uma nova representação requerida pelo modelo.

Normalization: é o processo de normalizar os dados, ou seja, colocá-los na escala trabalhada pelo modelo.

Shuffling: é a ação de embaralhar os dados que o modelo irá trabalhar.

Conclusão: Após assistir às aulas foi possível adquirir um grande conhecimento sobre os conceitos apresentados e entender sua importância dentro do campo do Machine Learning. A limpeza dos dados é essencial para o bom funcionamento dos modelos, sendo uma das etapas mais cruciais durante o desenvolvimento.