Relatório LAMIA 10
Lidando com Dados do Mundo Real (II)

Kaique Medeiros Lima

Introdução

Esse card é dividido entre duas seções do curso Machine Learning, Data Science and Generative AI with Python, More Data Mining and Machine Learning Techniques e Dealing with Real World Data. Esse relatório tem o intuito de representar os conteúdos abordados.

Descrição da atividade

More Data Mining and Machine Learning Techniques

K-N-N Concepts: K-Nearest-Neighbour é um algoritmo de classificação/predição de dados com base nos pontos K de dados do conjunto de treinamento.

Dimensionality Reduction: Método de redução de dimensões do conjunto de dados visando a preservação máxima do número de informações.

Principal Component Analysis (PCA): Uma técnica para reduzir a dimensionalidade. Ela identifica as principais direções nas quais os dados variam mais, projeta os dados nessas direções, e os organiza de maneira que as primeiras direções capturem a maior parte da variação nos dados. Isso reduz o número de variáveis enquanto preserva o máximo da informação original.

Data Warehouse: Banco de dados, contém diversas informações de diversos autores.

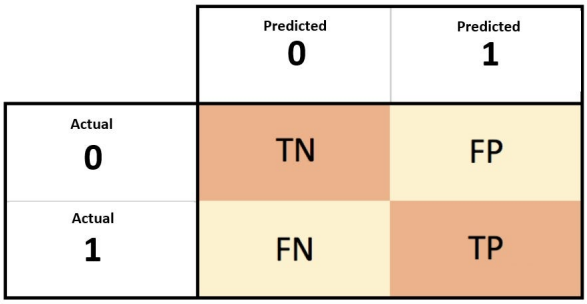
ETL: Primeiro, extrai-se os dados, depois, as informações são manipuladas, e por último, upload para a Data Warehouse.

ELT: Os dados são extraídos e depois são upados diretamente ao Data Warehouse, sendo manipulados lá, in-place.

Reinforcement Learning: É um método de aprendizado onde um agente desenvolve suas habilidades de decisão ao interagir com o ambiente. A cada ação executada, o agente recebe uma recompensa ou punição, dependendo do resultado. Com o tempo, ele adapta sua estratégia para aumentar as recompensas recebidas. O processo requer que o agente equilibre a exploração de novas ações com a utilização do conhecimento já adquirido para otimizar seu desempenho.

Confusion Matrix: Tipo de gráfico para a interpretação de performance da máquina. Em problemas de classificação, a máquina pode dizer que algo simplesmente não é 99.9% das vezes, mas sem embasamento. Por isso, a Confusion Matrix reporta verdadeiros positivos/negativos.

Exemplo de Confusion Matrix:



Dealing with Real World Data

Recall:

$$\frac{VERDADEIROS\ POSITIVOS}{VERDADEIROS\ POSITIVOS + FALSOS\ NEGATIVOS}$$

Também conhecido como sensibilidade, taxa de verdadeiros positivos, ou completude. É o percentual de positivos corretamente previstos. É uma boa métrica quando você se preocupa muito com falsos negativos, como em detecção de fraudes, onde não pode deixar os casos de fraudes passarem.

Precision:

$$\frac{VERDADEIROS\ POSITIVOS}{VERDADEIROS\ POSITIVOS + FALSOS\ POSITIVOS}$$

Também conhecido como positivos corretos. Representa o percentual de resultados relevantes. É uma boa métrica quando você se preocupa com falsos positivos, como em triagens médicas e testes de drogas, onde erros podem afetar seriamente a vida das pessoas.

Specificity:

$$\frac{VERDADEIROS\ NEGATIVOS}{VERDADEIROS\ NEGATIVOS + FALSOS\ POSITIVOS}$$

Taxa de verdadeiros negativos. Mede a proporção de verdadeiros negativos entre o total de negativos reais.

F1-Score:

$$2 \times \frac{PRECISION * RECALL}{PRECISION + RECALL}$$

A média harmônica de precisão e sensibilidade. É usada quando é importante equilibrar precisão e recall.

RMSE: Root Mean Squared Error, uma medida de precisão que avalia a magnitude dos erros. É utilizado na diferença entre valores previstos e reais, sem levar em conta o tipo de erro específico (certo ou errado).

ROC Curve: Receiver Operating Characteristic Curve (ROC) é um gráfico que representa a taxa de verdadeiros positivos (recall) contra a taxa de falsos positivos. Pontos acima da diagonal indicam um bom desempenho, superior ao aleatório. Quanto mais a curva se inclina para o canto superior esquerdo, melhor é o desempenho do classificador.

AUC: Area Under Curve (AUC). Representa a probabilidade de que um classificador classifique uma instância positiva aleatoriamente escolhida mais alta do que uma instância negativa aleatoriamente escolhida. Um AUC ROC de 0.5 indica um desempenho despresável ao caso, enquanto 1.0 indica um desempenho perfeito.

Conclusão

Após esse card, aprendi sobre diversos conceitos de relacionados a manipulação e limpeza de dados e a interpretação dos resultados. Aprendi sobre K-N-N, redução de dimensionalidade, Data Warehouse, ETL, ELT, Reinforcement Learning, Confusion Matrix, Recall, Precision, Specificity, F1-Score, RMSE, ROC Curve e AUC. Esses conceitos são essenciais para a compreensão de como as máquinas aprendem e como podemos interpretar seus resultados, visto que a limpeza dos dados é extremamente importante para a qualidade dos resultados e é um dos maiores desafios e diferenciais de um bom cientista de dados.