Universidade do Minho
Licenciatura em Engenharia Informática
Mestrado integrado em Engenharia Informática

Aprendizagem e Decisão Inteligentes
3°/4° Ano, 2° Semestre
Ano letivo 2021/2022

Enunciado Prático n° 4
Março, 2022

Tema

Regressão Linear

Enunciado

Regressão linear é uma técnica de *machine learning* usada para estimar o objetivo de um caso de estudo, sendo dado um conjunto dados que caracterizam o comportamento de uma série de casos passados.

Tarefas

Numa primeira fase devem descarregar o *dataset* [lego_sets] disponível na plataforma de *e-learning* da UMinho, secção [Conteúdo]. LEGO é uma marca de blocos de construção e são, normalmente, vendidos em conjuntos para construir um objeto específico. Este *dataset* contém dados sobre esses conjuntos, incluindo o número de peças de cada conjunto, preço de venda, idade sugerida, entre outros. O objetivo deste exercício foca-se no desenvolvimento de um modelo de regressão linear capaz de prever o preço do *set* (i.e., 'list_price'), dado um conjunto de *features* disponibilizados no *dataset*.

T1. Carregar, no *Knime*, o *dataset* [lego_sets] e aplicar nodos de exploração de dados como forma de permitir a análise dos dados;

T2. Proceder ao tratamento e limpeza dos dados:

a. Excluir o conjunto de *features* que considere irrelevantes para o desenvolvimento e validação do modelo de aprendizagem;
b. Aplicar técnicas de *feature engineering* e *data encoding* nas respetivas *features* do *dataset*, incluindo:
    i. Técnicas de *encoding*;
    ii. Tratamento de valores em falta;
    iii. Tratamento de *outliers*;
    iv. Tratamento de casos repetidos.

T3. Analise a correlação das *features* através de uma matriz de correlação. Dada esta informação, que *features* sugeria remover?

T4. Particionar os dados utilizando 80% para treino e 20% para teste;

T5. Aplicar um nodo *Linear Regression Learner* para treinar um modelo de regressão e um nodo *Linear Regression Predictor* para obter previsões utilizando o modelo treinado;

T6. Avaliar o desempenho da previsão do modelo de regressão utilizando o nodo *Numeric Scorer*;

T7. Criar um *scatter plot* como forma de observar os valores previstos *versus* valores reais;

T8. Calcular o erro residual e analisar a sua distribuição;

T9. Atendendo aos resultados e conhecimentos adquiridos ao longo da ficha prática, explore a utilização de outras técnicas de tratamento e limpeza dos dados, de modo a minimizar o erro do modelo de aprendizagem.