Universidade do Minho
Departamento de Informática
Mestrado [integrado] em Engenharia Informática

Dados e Aprendizagem Automática
1°/4° Ano, 1° Semestre
Ano letivo 2023/2024

Enunciado Prático n° 4
19 de outubro de 2023

Tema

Regressão Linear e Logística

Enunciado

Regressão linear e regressão logística são duas técnicas de aprendizagem supervisionada, aplicadas nas áreas de *machine learning* usadas no sentido de estimar o valor/classe de um caso de estudo, dado um conjunto de características e os padrões estatísticos analisados numa série de casos de estudo passados. A regressão linear tem como objetivo estimar um determinado valor numérico dado um conjunto de variáveis (algoritmo de regressão). A regressão logística foca-se em estimar a classe de um determinado caso de estudo (algoritmo de classificação).

Tarefas

Exercício 1 - Regressão Linear:

Uma companhia de comércio *online* de vestuário tenciona investir em melhorar uma das suas plataformas de venda *online*, atendendo ao rendimento que cada uma proporciona. As respetivas plataformas disponíveis são: (1) aplicação móvel; (2) plataforma *web*. Atendendo ao problema, foi proposto o desenvolvimento de um modelo de regressão linear, como forma de se estimar o rendimento de cada opção e com isto avaliar a melhor decisão. Para o efeito, a empresa disponibiliza um *dataset* (disponível em https://bit.ly/3mGDpu0) contendo o histórico de venda dos seus clientes e respetivas informações (p. ex., *email*, endereço, tempo na plataforma móvel, tempo na plataforma *web*, rendimento total adquirido, entre outros).

Após descarregar o *dataset*, pretende-se:

T1. Carregar o *dataset*, utilizando a função *pandas.read_csv(...)*;

T2. Aplicar métodos para exploração e visualização de dados;

T3. Definir o conjunto de variáveis de entrada e saída do modelo (i.e., entrada = [«Avg. Session Length», «Time on App», «Time on Website», «Length of Membership»]; saída = [«Yearly Amount Spent»]);

T4. Preparar e organizar os conjuntos de casos de estudo do *dataset* em dados de treino e teste, utilizando a função *sklearn.model_selection.train_test_split (..., test_size = 0.3)*;

T5. Treinar o modelo de regressão linear (*sklearn.linear_model.LinearRegression*), usando o conjunto de dados de treino;

T6. Analisar os coeficientes convergidos do modelo de regressão linear e identificar o seu significado no contexto do problema em causa;

T7. Avaliar a *Mean Absolute Error, Mean Squared Error* e *Root Mean Squared Error* do modelo desenvolvido na previsão de *'Yearly Amount Spent'* (utilize as funções disponíveis na biblioteca *sklearn.metrics*) e efetuar a respetiva análise critica.

Exercício 2 - Regressão Logística:

Neste exercício pretende-se estimar se um determinado utilizador de internet clicou, ou não, num anúncio publicitário, através do uso de um modelo de classificação de regressão logística. Para desenvolver este modelo, é disponibilizado um *dataset* (disponível em https://bit.ly/3CMO63B) apresentando os hábitos de vários utilizadores de internet, ilustrando um conjunto de características acerca de cada utilizador e da sua tomada de decisão.

Após descarregar o *dataset*, pretende-se:

T1. Carregar o *dataset*, utilizando a função *pandas.read_csv(...)*;

T2. Aplicar métodos para exploração e visualização de dados;

T3. Definir o conjunto de variáveis de entrada e saída do modelo (i.e., entrada = [«Daily Time Spent on Site», «Age», «Area Income», «Daily Internet Usage», «Male»]; saída = [«Clicked on Ad»]);

T4. Preparar e organizar os conjuntos de casos de estudo do *dataset* em dados de treino e teste, utilizando a função *sklearn.model_selection.train_test_split (..., test_size = 0.3)*;

T5. Treinar vários modelo de regressão logística (*sklearn.linear_model.LogisticRegression*), usando o conjunto de dados de treino. Treinar com 3 classificadores diferentes (*solver=...*);

T6. Avaliar a performance de classificação dos modelos, através da criação de matrizes de confusão *sklearn.metrics.confusion_matrix(...)* e relatórios de classificação *sklearn.metrics.classification_report(...)*;

T7. Atendendo aos resultados observados na **T6**, quais as conclusões adquiridas? Em que situações o modelo acerta/falha? Como melhorar o modelo de aprendizagem proposto? Qual o melhor modelo (conjunto de parâmetros)?