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

Dados e Aprendizagem Automática
1º Ano, 1º Semestre
Ano letivo 2021/2022

Enunciado Prático nº 4
4 de novembro de 2021

Tema

Regressão Linear e Logística

Enunciado

Regressão linear e regressão logística são duas técnicas supervisionadas aplicadas nas áreas de *machine learning* usadas no âmbito 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. Respectivamente, a regressão linear tem como objectivo estimar um determinado valor numérico dado um conjunto de variáveis (algoritmo de regressão), enquanto 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 comercio online de vestuário tenciona investir os seus esforços em melhorar uma das suas plataformas de venda online, atendendo ao rendimento que cada uma proporciona. As respectivas plataformas disponíveis são: (1) aplicação móvel; (2) plataforma website. Atendendo ao problema, foi proposto o desenvolvimento de um modelo de regressão linear, como forma de estimarmos o rendimento de cada opção, e com isto avaliarmos a melhor decisão. Para isso, a empresa disponibiliza um *dataset* (disponível em https://bit.ly/3mGDpu0), contendo o histórico de venda dos seus clientes e respectivas informações (e.g., email, endereço, tempo na plataforma móvel, tempo na plataforma website, rendimento total adquirido, etc.).

Após descarregarem o *dataset*, devem de seguida:

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 efectuar a respectiva análise critica.

Exercício 2 - Regressão Logística:

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

Após descarregarem o *dataset*, devem de seguida:

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 o modelo de regressão logística (*sklearn.linear_model.LogisticRegression*), usando o conjunto de dados de treino;

T6. Avaliar a performance de classificação do modelo, através da criação de uma matriz de confusão *sklearn.metrics.confusion_matrix(…)* e um relatório 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?