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

Dados e Aprendizagem Automática
1°/4° Ano, 1° Semestre
Ano letivo 2024/2025

Practical Exercise no. 4

Theme

Linear and Logistic Regression

Exercise

Linear regression and logistic regression are two supervised learning techniques applied in the field of machine learning used to estimate the value/class of a case study, given a set of characteristics and the statistical patterns analysed in a series of past case studies. Linear regression aims to estimate a certain numerical value given a set of variables (regression algorithm). Logistic regression focuses on estimating the class of a given case study (classification algorithm).

Tasks

I – Linear Regression with Ecommerce Customers Dataset

An online clothing retailer intends to invest in improving one of its online sales platforms, given the income each provides. The respective platforms available are (1) mobile application and (2) web platform. Given the problem, a linear regression model was proposed as a way of estimating the performance of each option and thus evaluating the best decision. To this end, the company provided a dataset (available at https://bit.ly/3mGDpu0) containing the sales history of its customers and their respective information (e.g. email, address, time on the mobile platform, time on the web platform, total revenue acquired, among others).

After downloading the dataset, it is intended to:

T1. Load the dataset using the *pandas.read_csv(...)* function;

T2. Apply methods for data exploration and visualisation;

T3. Define the set of input and output variables of the model (i.e., input = [«Avg. Session Length», «Time on App», «Time on Website», «Length of Membership»]; output = [«Yearly Amount Spent»]);

T4. Prepare and organise the set of case studies from the dataset in training and test data, using the *sklearn.model_selection.train_test_split (..., test_size = 0.3)* function;

T5. Train the linear regression model (*sklearn.linear_model.LinearRegression*) using the training data;

T6. Analyse the converged coefficients of the linear regression model and identify their significance in the context of the problem at hand;

T7. Evaluate the *Mean Absolute Error, Mean Squared Error* and *Root Mean Squared Error* the model developed in the *'Yearly Amount Spent'* prediction (use the functions available in the *sklearn.metrics* library) and carry out the respective critical analysis).

II – Logistic Regression with Advertising Dataset

The aim of this exercise is to estimate whether or not a particular Internet user has clicked on an advert, using a logistic regression classification model. To develop this model, a dataset is provided (available at https://bit.ly/3CMO63B) showing the habits of various Internet users, illustrating a set of characteristics about each user and their decision-making.

After downloading the dataset, it is intended to:

T1. Load the dataset using the *pandas.read_csv(...)* function;

T2. Apply methods for data exploration and visualisation;

T3. Define the set of input and output variables of the model (i.e., input = [«Daily Time Spent on Site», «Age», «Area Income», «Daily Internet Usage», «Male»]; output = [«Clicked on Ad»]);

T4. Prepare and organise the set of case studies from the dataset in training and test data, using the *sklearn.model_selection.train_test_split (..., test_size = 0.3)* function;

T5. Training various logistic regression models (*sklearn.linear_model.LogisticRegression* using the training data. Train with 3 different classifiers (*solver=...*);

T6. Evaluate the classification performance of the models by creating confusion matrices *sklearn.metrics.confusion_matrix(...)* and classification reports *sklearn.metrics.classification_report(...)*;

T7. Given the results observed in **T6**, what conclusions did you draw? In which situations does the model succeed/fail? How can the proposed learning model be improved? What is the best model (set of parameters)?