CO3093/CO7093 Resit - Classification & Clustering

School of Informatics
University of Leicester

Submission deadline

Assessment Number	Resit
Contribution to overall mark	100%
Submission Deadline	**Thursday 12 August at 5:00 pm**

Assessed Learning Outcomes

This Resit coursework assessment aims at testing your ability to

- carry out an exploratory data analysis
- build up a classification model and evaluate its performance
- cluster a dataset using the KMeans algorithm
- communicate your findings on the data

How to submit

For this assignment, you need to submit the followings:

1. A short report on your findings in exploring the given dataset, a description of your model and its evaluation, and your clustering and its justification.
2. The Python source code written in order to complete the tasks set in the paper. You should submit a single Python code file, say my_solution.py for all the questions you have answered.
3. A signed coursework cover, which can be found from Blackboard.

Please put your source codes, report and signed coursework cover into a zip file Resit_YouremailID.zip (e.g., Resit_emt12.zip) and then submit your assessment through the module's blackboard site by the deadline.

Problem Statement

Consider this credit score dataset (credit_scores.csv), which can be downloaded from Blackboard. The given dataset contains historical data on bank customers. The data contains 20 features such as credit history, job, employment, credit history, credit status, etc.

Objective: Your goal is to predict the creditworthiness of the customers and to propose a clustering for these customers.

Exploring the data

Even though the data is more or less cleaned, your first task is to prepare the data – load the dataset and carry out data cleansing bearing in mind the questions you would like to answer. Answer the following questions:

1. Load the data and conduct an exploratory analysis of the data trying to make sure values you think are numerical are being treated as such. Visualise the distributions for the variables representing features of interest. Comment on the distributions you have explored in particular that of the variable credit status.
2. Drop columns that may obviously not be relevant to the classification problem.
3. Through data visualisation, highlight if there might be features that are more relevant in predicting the credit worthiness of a given customer.
4. Check minimum and maximum values of numerical data. Normalise the data if need be.

Classification & Clustering

1. Construct a logistic regression model for the credit worthiness of customer with any predictors you feel are relevant. Justify why your model was appropriate to use.
2. Write down the mathematical equation of your classification model and evaluate your model. Make sure to withhold a subset of the data for testing. You should aim for a model with a higher accuracy. Give the ROC curve of your model and the value of the AUC. Comment on the performance of your model.
3. Use the K-Means algorithm to cluster your dataset. Justify your clustering using for example the Elbow method and visualise your data before and after clustering.

Marking Criteria

The following areas are assessed:

1. Exploring and understanding the data **[30 marks]**
2. Building up and evaluating the model **[30 marks]**
3. Clustering and justification **[20 marks]**
4. Report (up to 8 pages) interpreting the results. **[20 marks]**

Indicative weights on the assessed learning outcomes are given above. The following is a guide for the marking:

- **First (≥ 70 marks)**: As in **Second Upper** plus well-justified models by the data exploration and a concise and well-structured report containing any decisions that may be recommended.

- **Second Upper (60 to 69 marks)**: A good coverage of data cleansing techniques exploring the dataset, a good visualisation of the clusters, a predictive model with an appreciable accuracy with a rationale behind it, a working code and a well-structured report on the results obtained from the dataset.

- **Second Lower (50 to 59 marks)**: Some techniques used for data cleansing are overlooked, a predictive model partially justified with an appreciable accuracy, a working clustering, a partially commented code with very few functions, and a narrative of the findings about the dataset with few deficiencies.

- **Third (40 to 49 marks)**: Essential data cleansing techniques are covered, a predictive model is given with some justification, a working but basic block code with no clustering, and a written report describing some of the work done.

- **Fail (≤ 39 marks)**: Not satisfy the pass criteria and will still get some marks in most cases.

Last Updated 10 July 2021 by Emmanuel Tadjouddine