Swinburne University of Technology Hawthorn Campus
Dept. of Computer Science and Software Engineering

COS10022 Introduction to Data Science

Assignment 3 - *Semester 1, 2021*

Assessment Title: Data Prediction and Visualisation

Assessment Weighting: 35%

Due Date: Saturday, 29th May 2021 at 23:59 (AEDT)

Assessable Item:

- A written report of **no more than** 12-page long with the Assignment Cover Sheet. The plagiarism check must be **lower than 20%** in the main content part. Otherwise, it is not considered a qualified submission and the submission will not be marked.

Purpose of Assignment

This assignment aims at evaluating students' achievement of the following unit learning outcomes:

1. **Assessing the skill of starting a data analytics process from A to Z.**
2. **Assessing the ability of data analytics and model creating.**
3. **Assessing the capacity of data visualisation and result in communication.**

This is an **individual** assignment. You need to start with the data cleaning process, make decisions on which model to use, implement the model, and get the outcome for analysis.

Refer to the Unit Outline for the late submission penalty.

"Data Prediction and Visualisation"

Key Lessons:

This is the last assignment from COS10022. This one is an individual task for all students to practice and experience the whole process of carrying out a data analytic project from A to Z. The goal of this assignment is to create a prediction model for identifying whether the received wine is white wine or red wine.

To achieve the goal, you will need to start from data preparation, decide which model to be used, develop the model, train and test the developed model and visualise the outcome for result communication.

Introduction

Using machine learning technique to perform the prediction tasks is nowadays quite common in our daily life. With the proper design, training and tuning, this kind of technique can be widely used in different domains such as disease detection, factory automation, quality check, etc. Because of the privacy and logistic issues, only physicochemical (inputs) and wine type (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). The given dataset contains 2,000 tuples with header for identifying different attributes. The number of instances from the red wine and the white wine is identical for both groups. It is a subset from the open dataset available online. For more details, consult: [Web Link] or the reference [Cortez et al., 2009].

Reference

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Assignment Goal

The goal of this assignment is to create a predictive model for predicting the wine type of the given input. The predictive model you choose should be capable of taking features of a tuple and predict the corresponding wine type. Visualisation is essential for you to communicate with the clients to present the results.

Assignment Task

Your task is to create a predictive model to match the requirement from the assignment goal. You are expected to start from the beginning of Data Science Lifecycle and goes to the very end steps to present a complete data science project. Don't forget to examine the given dataset to make sure the goal is achievable by processing the given attributes. Although the dataset doesn't contain missing data, the data was collected in the form that the same type of the wine is gathered together in a bunch. Thus, shuffling the data is essential before partitioning them into training and test sets. After getting the performance of your prediction result, visualisation technique is essential for you to communicate the result with the client in the report.

There are 100 marks in this assignment. Your report must address the following tasks:

1. Observing the given dataset and make a description of what you see and what you discovered. For example, you can talk about whether there are missing values in the collected data, whether this is a balanced dataset, how many usable attributes are there in the collected data, etc. **[20 marks]**

2. Make the decision on which model you are going to use and give the reason. Describe what input/output will be for your model. For example, you can talk about the input/output dimension. Describe the components inside your model. For example, if it is a neural network model, you can talk about the layers, the number of neurons, the connection types, etc. **[10 marks]**
3. Describe what treatments you have made to the dataset. For example, reveal whether you do any transfer/mapping, shuffling, cleaning, and how the training/test set is partitioned. **[10 marks]**
4. Describe the training process or how the model is built into a completed form. If your model requires the training process, reveal the outcome obtained in the training process. Reveal the test results with proper benchmarks. **[30 marks]**
5. Use visualisation elements (figures) to attract the reader's eye with sufficient information to enhance communication with the reader. **[20 marks]**
6. A comprehensive and well-written report. **[10 marks]**

There are already abundant works dedicated to studying the problem of predicting the outputs on different datasets. Similar works can be found online. You are encouraged to explore some of the existing literature and, where applicable, adapt their ideas into your work. When you do so, please include all the necessary **in-text citations** and the **end-of-report reference list**.

The Harvard Referencing format must be used when citing and referencing external information resources: http://www.swinburne.edu.au/library/referencing/harvard-style-guide/

Submission Requirement

To fulfil the requirement of this assignment, the following item must be submitted in **pdf** format, named ***COS10022_Assignment3_YourStudentID*** and submitted:

- A written report of **not more than** 12-page long with the Assignment Cover Sheet (digitally signed).

Failure to adhere to the submission requirements will immediately result in "N" grade for this assignment.

Rubric: Subtask 1

Marks	Data Observation (20 marks)
15.1 – 20.0	The observation on the dataset is comprehensive and some pilot studies on the subsets have been applied.
10.1 – 15.0	A general dataset observation is provided.
5.1 – 10.0	The observation on the dataset is rough.
0.1 – 5.0	The observation on the dataset contains some incorrect information.
0.0	No observation is produced.

Rubric: Subtask 2

Marks	Model Selection and Design (10 marks)
5.1 - 10.0	The model selection is reasonable and the proper design to adopt the selected input is feasible.
0.1 – 5.0	The reason of selecting a model is vague or the design of the model is partially unreasonable.
0.0	The model selection and design are not revealed in the report.

Rubric: Subtask 3

Marks	Data Treatment (10 marks)
5.1 - 10.0	The description of how the data is processed from the raw data to the usable dataset is comprehensive and easy to understand.
0.1 – 5.0	The workflow of how data treatment is processed is vague.
0.0	The data treatment is not revealed in the report.

Rubric: Subtask 4

Marks	Model Training/Creating and Test (30 marks)
20.1 – 30.0	The description of how the model is built/trained is comprehensive. The training and the test results are revealed with proper measurements.
10.1 – 20.0	The description of how the model is built/trained is reasonable. The training and the test results are revealed with reasonable measurements.
0.1 – 10.0	The description of how the model is built/trained is vague. The training and the test results are revealed but the measurements are incorrect, or the performance is not acceptable.
0.0	The model training/building process is not revealed, and the results are not mentioned. The test is not performed.

Rubric: Subtask 5

Marks	Visualisation (20 marks)
10.1 – 20.0	The visualisation used in the report is informative and in a professional way. The presented information via the visualisation is clear and accompanied by corresponding descriptions in the report.
0.1 – 10.0	Some visualisation is used but the information revealed by the visualisation is not strong or redundant information is contained in the figures.
0.0	No visualisation is used in the report.

Rubric: Subtask 6

Marks	Report Presentation (10 marks)
8.1 – 10.0	The report is well organised, and the presentation is professional with proper visualisation elements. The report contains a high volume of references in the related subject. The report is submitted with no issue. The report is within the length limitation. The report contains the required coversheet.
7.1 – 8.0	The report contains a high volume of references in the related subject. The report is submitted with no issue. The report is within the length limitation. The report contains the required coversheet.
6.1 – 7.0	The report is submitted with minor issues. The report is within the length limitation. The report contains the required coversheet.
5.1 – 6.0	The report is submitted with some issues. The report exceeds the length limitation. The report does not contain the required coversheet.
0.0 – 5.0	The report is not well organised or not submitted.