Assessment 2 – Data Classification (75 marks)

This assessment includes two parts and counts for 75% of the overall assessment for this module.

The first part is 20 marks and should be submitted by noon on 22/11/2019 and the second one is 55 marks and should be handed by 20/12/2019.

The programming language you should use to finish this assessment is Python (in version 3 and above). In particular, you can use functions from the following packages: Numpy, Pandas, Matplotlib, Seaborn and Sklearn.

All Python skills needed to do this assessment have been covered in the practical sessions –practical notes are available on Canvas.

The information of the dataset which you will work with can be viewed in the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

Note that you should work on the red wine dataset (winequality-red.csv) only, which can be downloaded from 'Data Folder' in the link given above.

**Part one (20 marks)**

**Task 1:** Data pre-processing and data exploration (15 marks)

a. Use Pandas to load data
b. Merge all the data with "quality" labels between 6-10 into Class 1 and similarly form Class 2 for the data with "quality" labels between 1-5.
c. Report the number of features and number of rows in each class
d. Choose an attribute and generate a boxplot for the two pre-defined classes.
e. Show one scatter plot, that is, one feature against another feature. It is your choice to show which two features you want to use.

**Task 2:** Computing probabilities using Python code for the given red wine dataset (5 marks)

f. Prior probability:
    i. What is the probability of a wine classified as Class 1 (P(Class 1))?
    ii. What is the probability of a wine classified as Class 2 (P(Class 2))?
g. Conditional probability:
    i. What is the probability of a wine having a pH value greater than 3.6 given it is classified as Class 1 (P(pH>3.6|Class 1))?
h. Posterior probability
    i. What is the probability of a wine classified as Class 1 when it has a pH value greater than 3.6?

**Task 3:** Writing a report to summarize what you have done. Explain figures you have put into your report clearly and report your findings and conclusions. The maximum number of pages is two and it should include less than 400 words. (This report counts for 0 marks: this is a chance for you to practice on how to write a report and to obtain feedback from a tutor.) Please use a single column format. The font size should be set to 11 or 12-point size.

**Hand in date**: Noon (12pm) on 22/Nov/2019 via Canvas.

**What to submit:**

Hand in two files:

1) A .ipynb file showing your completed programming code (worth 20 marks)
2) A report of maximum two pages in pdf format (worth 0 mark). The aim of this report is to give you a chance to practise how to write a report and a tutor will give you feedback during the demo.

**Demo**:

A demo will be taking place during the practical sessions on 25/11/2019 and 27/11/2019, separately, depending on which group you are timetabled for your practical session.

During the demo (8 minutes -10 minutes per student), you will be asked to run the code you have submitted in your .ipynb file and you also need to answer some questions to show that you understand the work you have done.

The final mark of your Part one will be given based on your performance during the demo.

| Demonstration Mark (out of 5) | Maximum Coursework Mark (out of 20) |
|---|---|
| 0-1: (almost) no understanding of the work | 0~9 |
| 2-3: some understanding of the work | 10~15 |
| 4-5: fully understands the work, maybe with a minor error | 16 ~20 |

**Part Two (55 marks)**

In this part, you will continue to work on the dataset you have used and modified in Part one, that is, the red-wine dataset with two classes, where Class One includes those red wine with a quality value in between 6-10 (inclusive) and Class Two includes those red-wine with a quality value in between 1-5 (inclusive).

**Task 1:** Divide the data set into a training set (I) and a test set. Usually, we use 20%-30% of the total data points as the test data. It is your choice on how to set the exact ratio. But you need make it clear in your report. You should further divide the training set (I) into a smaller training set (II)and a validation set using the same ratio. (**5 marks**)

**Task 2: PCA Analysis on the red-wine two classes dataset (5 marks)**

a) Perform a PCA analysis on the training data set (I) (2 marks)
b) Plot the training data in the PC1 and PC2 projection and label the data in the picture according to its class. (2 marks)
c) Report variances captured by each principal component (1 mark)

**Task 3: Do a classification using the logistic regression model (13 marks)**

a) In your report, describe the model you have used, including (6 marks):
   What is the cost function? You need to give a mathematical expression describing it.
   Which optimization algorithm has been used in your code?
   Did you use a regularisation term? If you used one, what is it?

b) Define your own function ([num1, num2]=misPatterns(predictions, labels)) using Python: the inputs of this function are predictions and labels; and the outputs of this function are the number (num1) of misclassified patterns whose label is 1 but prediction is 2, and the number (num2) of misclassified patterns whose label is 2 but prediction is 1. (4 marks)
c) Train the model on the training set and report the performance on the test set including accuracy rate and results obtained using the misPatterns function you have defined in b). (3 marks)

**Task 4: Investigate how the size of the training dataset affects the model performance on the test set (10 marks)**

a) Produce a learning curve of the size of training set (II) against the accuracy rate. The accuracy rate should be measured on both the training set and the validation set (5 marks).
b) Report what is the best training data size you would like to use for this work and explain why you chose it (2 marks).
c) Report the performance on the test set obtained using the model trained from the best size (3 marks).

**Task 5: Investigate how the number of features extracted from PCA affects the model performance on the test set (12 marks)**

a) Perform a PCA analysis on the training data set (II) and obtain the projected training set. (2 marks)
**b)** Producing a learning curve of the number of principal components against the accuracy rate. The accuracy rate should be measured on both the training set (II) and the validation set (5 marks).
**c)** Report what is the best number of principal components you would like to use for this dataset and explain why you chose it (2 marks).
**d)** Report the performance on the test set obtained using the model trained from the best number of principal components (3 marks).

**Task 6: Writing a report (10 marks)**

In this report, you need to summarize what you have done, which model you have used, what results you have obtained, and what are your findings and conclusions. The highest mark will give a report with outstanding presentation and clarity, no significant grammatical/ spelling or structural errors, and outstanding level of analysis with critical evaluation/reflection where it is required.

Hand in date: by noon on 20/12/2019 via Canvas.

What to submit:

Submit two files: identified by your student's ID number.

1) A .ipynb file to show your completed Python code.
2) A report with no more than 4 pages and less than 1200 words (please use a single column format. Font size should be set to 11 or 12 point) in the pdf format**.**

**Marking received on this piece of coursework may be subject to a viva (oral examination).**