ASSIGNMENT TWO

Semester 2 - 2021

PAPER NAME: Data Mining and Machine Learning

PAPER CODE: COMP809

DUE DATE: Sunday 24 Oct 2021 at midnight

TOTAL MARKS: 100

Student Name: .....

Student ID: .....

Note: This assignment must be complemented individually

Submission: A soft copy needs to be submitted through Turnitin (a link for this purpose will be set up in Blackboard) Include your actual code (no screenshot) in Appendix with appropriate comments for each task.

INSTRUCTIONS:

1. **ACADEMIC INTEGRITY GUIDELINES**

   The following actions may be deemed to constitute a breach of the General Academic Regulations Part 7: Academic Discipline, Section 2 Dishonesty During Assessment or Course of Study

   - 2.1.1 copies from, or inappropriately communicates with another person
   - 2.1.3 plagiarises the work of another person without indicating that the work is not the student's own – using the full work or partial work of another person without giving due credit to the original creator of that work
   - 2.1.4 Unauthorised collaboration in Assessment - collaborates with others in the preparation of material, except where this has been approved as an assessment requirement. This includes contract cheating where a student obtains services to produce or assist with an assessment
   - 2.1.5 resubmits previously submitted work without prior approval of the exam board
   - 2.1.6 Using any other unfair means

   Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your online submission on Blackboard immediately

Part A: Clustering Methods (40 marks)

For this question, you will explore the clustering methods you have learnt in this course. You have been given datasets from three very different application environments and you are required to explore three widely used clustering algorithms and deploy each of them on the different datasets.

The three algorithms that you have decided to explore are **1) K-Means 2) DBSCAN and 3) Agglomerative.**

The three datasets that you have been given are:

- Seoul Bike Sharing Demand
- *Sales Transactions*
- Water Treatment Plant

You need to complete three tasks as detailed below.

Task 1

For each activity in this task, ***you must explain each dataset and perform data exploration, data pre-processing and apply a suitable feature selection algorithm before*** **deploying** ***each clustering algorithm***. Your clustering results should include the following measures:

The time is taken, Sum of Squares Errors (SSE), Cluster Silhouette Measure (CSM). You may use Davis-Bouldin score as an alternative to SSE.

Submit Python code used for parts a) to c) below. You only need to submit the code for one of the 3 datasets.

a) Run the **K means algorithm** on each of the three datasets. Obtain the best value of K using either SSE and/or CSM. Tabulate your results in a 3 by 3 table, with each row corresponding to a dataset and each column corresponding to one of the three measures mentioned above. Display the CSM plot for the best value of the K parameter for each dataset. **[7 marks]**

b) Repeat the same activity for **DBSCAN algorithm** and tabulate your results once again, just as you did for part a). Display the CSM plot and the 3 by 3 table for each dataset. **[7 marks]**

c) Finally, use the **Agglomerative algorithm** and document your results as you did for parts a) and b). Display the CSM plot and the 3 by 3 table for each dataset. [**7 marks]**

Task 2

a) For each dataset identify which clustering algorithm performed best. Justify your answer. In the event that no single algorithm performs best on all three performance measures, you will need to carefully consider how you will rate each of the measures and then decide how you will produce an overall measure that will enable you to rank the algorithms. **[7 marks]**

b) For each winner algorithm and for each dataset explain why it produced the best value for the CSM measure. This explanation must refer directly to the conceptual design details of the algorithm. There is no need to produce any further experimental evidence for this part of the question. **[7 marks]**

c) Based on what you produced in a) above, which clustering algorithm would you consider being the overall winner (i.e., after taking into consideration performance across all three datasets). Justify your answer. **[5 marks]**

Part B: Predictions of $PM_{2.5}$ (60 marks)

Air pollution causing serious damage to public health and based on existing research, particulate matter (PM) smaller than $PM_{2.5}$ are currently considered to have the strongest correlation with effects of cardiovascular disease. Therefore, making accurate predictions of $PM_{2.5}$ is a crucial task. In this part, you are required to build prediction models based on multi-layer perceptron (MLP) and long short-term memory (LSTM).

Dataset: Penrose Air Quality Monitoring Station

The dataset for this experiment can be downloaded from the Environmental Auckland Data Portal. Your dataset includes $PM_{2.5}$ (output) and different predictors such as air pollution, Air Quality Index (AQI), and meteorological data collected on an hourly basis from the Penrose air quality monitoring station (ID:23). Two $PM_{2.5}$ lag measurements, lag_1 and lag_2, should be included in your dataset. Lag_1 is $PM_{2.5}$ measurements for the previous hour (*h-1*) and lag_2 is $PM_{2.5}$ concentration for *h-2*.

Download Penrose **$PM_{2.5}$** concentration, air pollution data (**SO_2, NO, NO_2**), and meteorological data **Solar Radiation** (W/m^2), **Air Temperature** (°C), **Relative Humidity** (%), **Wind Direction** (°) and **Wind Speed** (m/s)). The dataset should be hourly measurement starting from January 2016 to December 2020 (4 years).

Note: Unit of measurement for $PM_{2.5}$ and air pollution data should be ($\mu g/m^3$).

Data Pre-processing **[5 marks]**

Make sure your dataset all has the same temporal resolution (i.e. hourly measurement). Perform data exploration and identify missing data and outliers (data that are out of the expected range). For example, unusual measurements of air temperature 40(°C), Relative Humidity measurements above 100, negative or unexplained high concentrations are outliers.

- Provide attribute-specific information about outliers and missing data. How can these affect dataset quality?
- Based on this analysis, decide, and justify your approach for data cleaning. Once your dataset is cleaned move to the next step for feature selection.

Feature Selection **[5 marks]**

Choose **five attributes** of your dataset that has the highest correlation with $PM_{2.5}$ concentration using Pearson Correlation or any other feature selection method of your choice with justification.

- Provide the correlation plot (or results of any other feature selection method of your choice) and elaborate on the rationale for your selection.
- Describe your chosen attributes and their influence on $PM_{2.5}$ concentration.
- Provide graphical visualisation of variation of $PM_{2.5}$ variation.
- Provide summary statistics of the $PM_{2.5}$ concentration.
- Provide summary statistic of predictors of your choice that has the highest correlation in tabular format.

Experimental Methods

Use 70% of the data for training and the rest for testing of the MLP and LSTM models. Use a Workflow diagram to illustrate the process of predicting $PM_{2.5}$ concentrations using the MLP and LSTM models. **[5 marks]**

For both models, provide root mean square error (RMSE), Mean Absolute Error (MAE) and correlation coefficient (R^2) to quantify the prediction performance of each model.

Multilayer Perceptron (MLP)

1) In your own words, describe multilayer perceptron (MLP). You may use one diagram in your explanation (one page). **[5 marks]**

2) Use the sklearn.MLPClassifier with default values for parameters and **a single hidden** layer with k= 25 neurons. Use default values for all parameters other than the number of iterations. Determine the best number for iteration that gives the highest performance on the testing dataset. Use this as a baseline for comparison in later parts of this question. **[5 marks]**

3) Experiment with **two hidden layers** and experimentally determine the split of the number of neurons across each of the two layers that gives the highest classification accuracy. In part 2, we had all k neurons in a single layer, in this part we will transfer neurons from the first hidden layer to the second iteratively in step size of 1. Thus, for example in the first iteration, the first hidden layer will have k-1 neurons whilst the second layer will have 1, in the second iteration k-2 neurons will be in the first layer with 2 in the second and so on. **[5 marks]**

4) From the results in part 3 of this question, you will observe a variation in the obtained performance metrics with the split of neurons across the two layers. Give explanations for some possible reasons for this variation and which architecture gives the best performance? **[5 marks]**

Long Short-Term Memory (LSTM)

1) Describe LSTM architecture including the gates and state functions. How does LSTM differ from MLP? Discuss how does the number of neurons and batch size affect the performance of the network? **[5 marks]**

2) To create the LSTM Model, apply Adaptive Moment Estimation (ADAM) to train the networks. Identify appropriate cost function to measure model performance based on training samples and the related prediction outputs. To find the best epoch, based on your cost function results, complete 30 runs keeping the learning rate and the number of batch size constant at 0.01 and 4 respectively. Provide a line plot of the test and train cost function scores for each epoch. Report the summary statistics (Mean, Standard Deviation, Minimum and Maximum) of cost function as well as the run time for each epoch. Choose the best epoch with justification. **[5 marks]**

3) Investigate the impact of differing the number of batch size, complete 30 runs keeping the learning rate constant at 0.01 and use the best number of epochs obtained in previous step 2. Report the summary statistics (Mean, Standard Deviation, Minimum and Maximum) of cost function as well as the run time for each batch size. Choose the best batch size with justification.. **[5 marks]**

4) Investigate the impact of differing the number of neurons in the hidden layer while keeping the epoch (step 2) and Batch size (step 3) constant for 30 runs. Report the summary statistics (Mean, Standard Deviation, Minimum and Maximum) of cost function as well as the run time. Discuss how does the number of neurons affect performance and what is the optimal number of neurons in your experiment? **[5 marks]**

Model Comparison

1) Plot model-specific actual and predicted $PM_{2.5}$ to visually compare the model performance. What is your observation? **[5 marks]**

2) Compare the performance of both MLP and LSTM using RMSE. Which model performed better? Justify your finding. **[5 marks]**