Data Science (Eng) 774/874
Post-block Assignment 2

Total Marks (MEng): 75
Total Marks (PGDip): 45
09 March 2021

Kindly complete the assignment and submit your assignment report as an electronic submission in PDF format on SUNLearn by 16 April 2021 at 23:50. The submission only needs to include the assignment report.

The datasets that should be used for this assignment are available on SUNLearn under the *Datasets* section.

Question 1 [25]

A medical research lab is conducting research on breast cancer and contracted you as a Data Scientist to help them respond to some pertinent questions with regards to the early detection of the disease. In order to achieve your goal, you are provided with the *BreastCancer* dataset.
As a Data Scientist, you are required to conduct an Exploratory Data Analysis (EDA) of the dataset.

The following are some of the tasks that you are required to perform:

1. Draw a scatter plot from the dataset using the *radius_mean* and the *perimeter_mean* features [**6**].
2. Based on the plot obtained in question 1.1, what can you say about the *radius_mean* and the *perimeter_mean* attributes? [**4**]
3. Compute the Pearson correlation coefficient between the *radius_mean* and the *perimeter_mean*. Does your score confirm the observation made in question 1.2? [**5**]

   Hint: The Pearson Correlation Coefficient, p, is a linear coefficient (score) that determines the level of correlation between two random attributes, *A*, *B* within a dataset. The coefficient p varies between 0 and 1. The closer the score is to 1, the more the features are correlated.

4. **The following sub question is only applicable to the MEng module level, Data Science (Eng) 874 (not 774), students:**

   Implement a k-Nearest Neighbors (kNN) algorithm on the dataset for the following values of $k \in \{3, 7, 15, 31, 61\}$. What is the classification accuracy (accuracy score) for each value of k? (Use 70% training subset size for each value of k)
   As a data scientist, which value of k would you select for your production model and why? [**10**]

Question 2 [20]

Consider the *WineComposition* dataset which includes different categories of wines based on their chemical composition. As a Data Scientist, you are given a task to find out if there exist some useful patterns in the attribute space. The following are your tasks:

1. Use any software tool to plot the data distribution of each feature within the dataset. Explain why such a plot is necessary in your exploratory data analysis [**10**].

2. Use a software tool of your preference and implement the k-Means clustering algorithm using only the two features: *Alcohol* and *Magnesium*. Compare the results when varying the k-value between 2 and 7. Which k-value yields the best looking clusters? Plot the cluster diagrams and discuss what you see [**6**].

   Notes: The *WineComposition* dataset has multiple features. When implementing the k-Means algorithm, you may consider the *Alcohol* and *Magnesium* features. Regardless of the tool that you use, the *init* parameter must be set to *random* and the *radnom_state* parameter must be set to 40 (The *init* parameter is a method for initialization, the *random_state* parameter is used to make the randomness deterministic - These parameters might not exist in some tools)

3. During the clustering process, determining the best value for k is generally conducted after fitting several k-Means models [As done in question 2.2.]. However, the Elbow Method can help you to determine which value of k is suitable for a specific problem.

   Based on the above definition, plot an elbow graph that can prove that the value of k found in question 2.2. is an optimal one [**4**].
   Hint: Research about the Elbow Method for k-Means clustering.

Question 3 [10]

This question is for the MEng module level, Data Science (Eng) 874, students only.
Consider a university admission dataset, *AdmissionDataset*, which contains 400 records of university students' admission probabilities. Apply a linear regression model to predict the chance of admission (probability) of a student based on the following features (present in the dataset): *GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research.* Use 80% of the dataset for training and 20% for testing.

1. Predict the probability of admission for the following records: $record_1 = [322, 109, 5, 4.5, 3.5, 8.80, 0]$, $record_2 = [307, 52, 5, 4.4, 3.5, 8.20, 2]$ [**6**]

2. Explain the steps you (or your software tool) followed to arrive at your predictions [**4**].

Question 4 [20]

Consider the Iris dataset, *iris.* Each row in the dataset represents an iris flower, including its species and dimensions of its botanical parts, petal and sepal, in centimetres.

Perform the following operations on the dataset:

1. Implement the following models to classify the type of flowers (species) within the dataset: k-Nearest Neighbor (kNN), Naive Bayes (NB) and Decision Trees (DT).In your evaluation process, consider the following performance metrics: Accuracy, Precision and Recall. How can you justify the lower performance of the NB method in comparison to kNN and DT ?

   Note: Use 80% training set and 20% testing set for all the models. For the kNN, use $k = 3$ and for the NB method, use the Gaussian Naive Bayes [**10**].

2. **This sub-question is for the MEng (874) students only.**
   Discuss the importance of computing a confusion matrix when dealing with the multiclass classification scheme. Furthermore, compute the confusion matrix for the NB model implemented in question 4.1. and discuss the results. [**10**].