ASSIGNMENT – 2

Total Marks: 150

Instructions:

1) Use Python programming language and **submit working code,** failing which will fetch zero marks.
2) You are allowed to submit .ipynb or .py working files
3) Use Regression.py for Q3. You are allowed to save model files, so that on execution it should give same and efficient result.
4) You need to submit the README.pdf
5) Mention methodology, assumptions, plots, results, and analysis you may have in README.pdf.
6) You need to submit a single zip file with README.pdf, code files, model files
7) Submission format: A2_ML_ <Register_Number>_<Name>.zip

Question 1: Exploratory Data Analysis (EDA): Load Iris Dataset provided with assignment (Iris.data) into a Pandas Dataframe **(35 Marks)**

1. Observe the dataframe and answer the following question. [3]
   (a) Print the first five rows of the dataframe.
   (b) How many samples are available in the dataframe?
   (c) What is the total number of attributes in the dataframe? [5]
2. Convert the Dataframe into a NumPy array.
   (a) What is the shape of the NumPy array?
   (b) Split the NumPy array into two NumPy arrays and report shape of each array:
      i. X ("features") containing first 4-columns of the dataset.
      ii. Y ("labels") containing last column of the dataset.
   (c) Print the unique elements of the labels array.
   (d) Create a dictionary to map unique labels into numerical labels.
   (e) Convert the labels array into an integer-valued array using the dictionary-map.
3. Usually to explore the data complexity, we visualize the scatter plot of the data. In the scatter plot, the samples of all the classes are visualized simultaneously, which provides information about the class separation. We can visualize only the 2D or 3D data using the scatter plots. [16]
   (a) Visualize the dataset using scatter plots amongst all possible pairs of attributes in the form of scatter plots. Which attribute pairs are best in terms of class separation?

(b) For features with dimensions higher than three, we may use T-distributed Stochastic Neighbor Embedding (t-SNE) to reduce the number of features. Use the t-SNE to reduce the feature-set to 2 dimensions and visualize the scatter plot. What is your inference regarding the class separation?

(c) Another popular way to reduce the feature-set dimensions is Principal Component Analysis (PCA). Now, use PCA to project features into 2D-space. Visualize the scatter plot of the obtained principal components. Note: PCA is affected by scale so you need to scale the features in your data before applying PCA. Use StandardScaler to help you standardize the dataset's features.

(d) Write your inference for both the techniques used in (b) and (c).

4. We can also create a histogram of each attribute to get an idea of the distribution. Plot histogram for each feature using matplotlib. Can you comment upon the distributions observed? [8]

5. Use NumPy to compute the following statistics for the features inside the dataset: [3]

   (a) Max. Value

   (b) Average value

   (c) Variance

Question 2: Clustering using K-Means Algorithm (You need to implement it from scratch, usage of any library for model is not allowed). In this question, cluster words belonging to four categories: animals, countries, fruits and veggies. Words are arranged in four different files: first entry in each line is a word followed by 300 features (word embedding) describing the meaning of that word. [Dataset is attached in the assignment] **(85 Marks)**

Hint: Combine the data from four files to get one dataset and shuffle the data before performing clustering.

Train a k-means model to cluster instances into k clusters with

(a) Euclidean Distance

(b) Cosine Similarity

For both the metrics;

I. vary the value of k from 1 to 10

II. compute Precision, Recall and F-score for each set of clusters.
(To calculate Precision and Recall in clustering, Refer: Link)

III. Plot k in horizontal axis and precision, recall and f-score in the vertical axis in same plot

Compare both the set of clusters you obtained and discuss what is the best setting for k-means clustering for the given dataset.

Question 3: Linear Regression:

For this question you have to use ‘regression.data’ attached with the assignment. The dataset consists of nine columns, and the last column represents the target variable. The remaining columns denote the features. **(30 Marks)**

(a) A file named ‘Regression.py’ containing a ‘Regression’ class is attached with the assignment. You need to fill the suitable code in this class. In this class you can use ‘.fit()’ of ‘LinearRegression’ from the sklearn. However, you have to write ‘.predict()’ from scratch using the outcomes of ‘.fit()’. [5]

(b) Split your data using 5-fold splits: one-fold used as validation set and remaining four folds as training sets (k-cross validation). Use the ‘Regression’ class to prepare a table containing training and validation Mean Squared Error (MSE) for each fold. Implement your own MSE function. [10]

(c) Now use ‘LinearRegression’ from the sklearn to make the predictions and prepare similar table as in (b). [10]

(d) Write your inferences from both the approaches in (b) and (c) [5]