ELG 5225: Applied Machine Learning

Assignment 3

Due date posted in Bright Space

Submission

You must submit two documents. First, a report of the solutions including important code snippets as a PDF file. Second, the whole code should be in a separate python file (Notebooks are accepted). The file name must include your group number and assignment number, for example **Group1_HW2.pdf and Group1_HW2.py**.

Assignment must be submitted on-line with Bright Space. This is the only method by which we accept assignment submissions. We do not accept assignments sent via email, and we are not able to enter a mark if the assignment is not submitted on Bright Space! The deadline date is firm since you cannot submit an assignment passed the deadline. It is your responsibility to ensure that the assignment has been submitted properly.

Part 1: Calculations

1. Use the k-means algorithm and Euclidean distance to cluster the following 5 data points into 2 clusters: A1=(2,5), A2=(5,8), A3=(7,5), A4=(1,2), A5=(4,9). Suppose that the initial centroids (centers of each cluster) are A2 and A4. Using k-means, cluster the 5 points and show the followings for one iteration only:

   (a) Show step-by-step the performed calculations to cluster the 5 points. **(7 Marks)**

   (b) Draw a 10 by 10 space with all the clustered 5 points and the coordinates of the new centroids. **(4 Marks)**

   (c) Calculate the silhouette score and WSS score. **(5 Marks)**

Part 2: Programming

1. Use scikit-learn to implement Logistic Regression (LR) and K-Nearest Neighbor (K-NN) classifiers on the provided Diabetic dataset. The dataset has been standardized and split into training and testing. Through this assignment, **make sure to use the first 576 rows (75%) for training and the remaining 192 rows (25%) for testing.** Failing to do so might result in marks deduction. There are 2 classes in this dataset, and each sample in provided dataset has 8 features.

(a) Provide the accuracy of LR and K-NN classifier as baseline performances. **(6 Marks)**

(b) Provide 2D TSNE plots, one for the training set and one for the test set. **(4 Marks)**

2. Choose the best number of cluster for k-means clustering algorithm

(a) Plot the silhouette score vs the number of clusters. **(4 Marks)**

(b) Determine the optimal number of clusters for k-Means **(2 Marks)**

(c) Plot the clustered data with optimum number of clusters. **(4 Marks)**

3. Apply the following Dimensionality Reduction (DR) methods:
PCA(n_components=n, random_state=0)

(a) Find the best value of n_components, based on test accuracies, for both classifiers (LR and K-NN). **(3.5 Marks)**

(b) Plot the Number of Components-Accuracy graph with baseline performances for each classifier as shown below 3. The Graph should be plotted based on the test accuracy. (Use bar chart) **(4 Marks)**

(c) Provide 2D TSNE plots, one for the training set and one for the test set. **(2.5 Marks)**

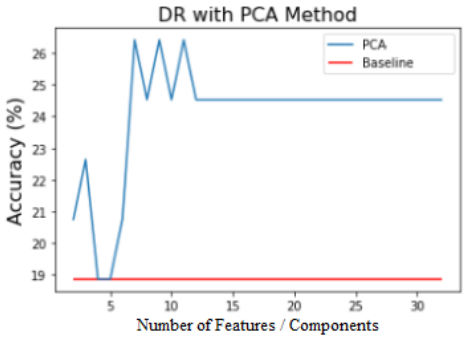

Figure 1: THIS FIGURE IS MERELY FOR ILLUSTRATION

4. Use the following Feature Selection methods (one for each method). Find the best number of features based on both, the LR and K-NN classifiers' test accuracies.

(a) Filter Methods (Information Gain, Variance Threshold etc.). Plot the number of features versus accuracy graph with the improved baseline performance as shown in Q3, using only the method that gives you the best test accuracy. **(4 Marks)**

(b) Wrapper Methods (Forward or Backward Feature Elimination, Recursive Feature Elimination etc.). Plot the number of features versus accuracy graph with the improved baseline performance as shown in Q3, using only the method that gives you the best test accuracy. **(4 Marks)**

(c) Provide 2D TSNE plots, one for the training set and one for the test set, using only the best method (either the filter or wrapper). **(2 Marks)**

5. Choose the best number of cluster for k-means clustering algorithm on the processed data, using the best features from Q3 or best number of dimensionality Q4.
   (a) Plot the silhouette score vs the number of clusters **(6 Marks)**
   (b) Determine the optimal number of clusters for k-means **(4 Marks)**

6. Choose the best number of neurons for SOM algorithm, using the best features from Q3 or best number of dimensionality Q4. You might find it easier if you use the MiniSom library.
   (a) Plot the silhouette score vs the number of neurons (max 30 neurons) **(5 Marks)**
   (b) Determine the optimal number of neurons for SOM **(5 Marks)**
   (c) Plot the initial and final Neuron positions **(5 Marks)**

7. Tune the epsilon (0.3-0.7) and minpoints (2-15) values to obtain the same number of clusters in Q6 by using DBSCAN. Provide two separate plots; where you show only the best 10 combinations of epsilon and minpoints that brings you closer to the desired cluster number with the highest silhouette score.
   (a) First plot is epsilon vs number of clusters. **(5 Marks)**
   (b) Second plot is minpoints vs number of clusters. **(5 Marks)**

8. In your report, make sure to include a conclusion section where you comment on the followings:
   (a) The results of K-means from Q2 and Q5 (i.e., before and after applying DR methods and FS). **(4 Marks)**
   (b) The results of TSNE plots from Q1, Q3 and Q4. **(5 Marks)**

Important Notes

- Report should include answers for all question briefly. All plots must have titles and proper axis labels. **Otherwise, you will lose one point for each missing item.** The code file is requested in case of need to verify.
- Make the following parameter's assumption whenever needed, random_state=0