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_HW3.pdf and Group1_HW3.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=(3,6), A2=(6,3), A3=(8,6), A4=(2,1), A5=(5,9). 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 Naive Bayes Classifier (NB) and K-Nearest Neighbor (KNN) classifiers on the provided Mobile Crowd Sensing (MCS) dataset. The dataset consists of some features such as Latitude, Longitude, Day, Hour, Minute, Duration, RemainingTime, Resources, Coverage, OnPeakHours, GridNumber, Ligitimacy. Legitimacy will be used as an output and according to day feature you need to generate training and test datasets. The values 0, 1 and 2 in column day should be used for training dataset and only the value 3 in column day should be used for test dataset.

ID indicates the index and it cannot be used as the feature. Feature day is different for training and test datasets so it cannot be used as feature in datasets. Please use color code and different line style in your figures.

(a) Create training and test datasets for remaining parts according to day feature in the dataset (column: "day") **(6 Marks)**

(b) Provide confusion matrixes and F1 scores of NB and KNN classifier as baseline performances. **(6 Marks)**

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

2. Apply the following Dimensionality Reduction (DR) methods:PCA(n_components=n, random_state=0) and Auto Encoder (AE). AE structure should be from higher to lower number of neurons, from outer to center.

   (a) To find the best reduced dimensions of PCA and AE based on f1 score of test dataset using both classifiers (NB and KNN), plot the number of components (dimension) vs f1 score together with baseline performances for each classifier. The Graph should be plotted based on the f1 score of test dataset. The total number of figure will be 4 in this part. **(16 Marks)**

   (b) Provide 2D TSNE plots for the best performance in previous part (The best dimensionality reduction performance using one of the NB and KNN classifiers) one for the training set and one for the test set. **(4 Marks)**

3. Use the following Feature Selection methods (one for each method). Find the best number of features based on both, the NB and KNN classifiers f1 scores

   (a) Filter Methods (Information Gain, Variance Threshold etc.). Plot the number of features vs f1 score with the improved baseline performance. **(8 Marks)**

   (b) Wrapper Methods (Forward or Backward Feature Elimination, Recursive Feature Elimination etc.). Plot the number of features versus accuracy graph with the baseline performance. **(8 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). **(4 Marks)**

4. Latitude and longitude features should be considered for clustering based methods. Choose the best number of cluster among 8, 12, 16, 20 and 32 clusters. Legitimate only clusters should be found and the maximum number of legitimate only members in legitimate only clusters should be reported.

   (a) Apply K-means algorithm to plot the number of clusters (8,12,16,20 and 32) vs the total number of legitimate only members inside the legitimate only clusters. **(8 Marks)**

   (b) Apply SOFM algorithm to plot the number of clusters (8,12,16,20 and 32) vs the total number of legitimate only members inside the legitimate only clusters. **(8 Marks)**

(c) Apply DBSCAN algorithm to plot the number of clusters (8,12,16,20 and 32) vs the total number of legitimate only members inside the legitimate only clusters. You need to try different midPoint and epsilon parameters to obtain the 5 different cluster numbers. If you cannot obtain specific numbers you can report approximate numbers to 8,12,16,20 and 32. **(8 Marks)**

5. Please create the conclusion part in your report

(a) Please list your conclusions for questions 1,2,3 and 4. Minimum one conclusion for each question should be given. **(4 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