Pattern Recognition and Machine Learning

Lab Assignment-6

Early Bird Submission Deadline*: Tuesday Batch: 6 Mar, 23:59*

Thursday Batch: 8 Mar, 23:59

Late Submission Deadline*: Tuesday Batch: 7 Mar, 2023, 23:59*

Thursday Batch:Mar 9, 2023, 23:59

Final deadline*: Tuesday Batch: Mar 8, 2023, 23:59*

Thursday Batch: Mar 10, 2023, 23:59

Guidelines for submission:

1. Perform all tasks in a single colab file.
2. Create a report regarding the steps followed while performing the given tasks. The report should not include excessive unscaled preprocessing plots.
3. Try to modularize the code for readability wherever possible
4. Link for In-Lab Submission: Link
5. Submit the colab[.ipynb], python[.py] and report[.pdf] files here : Link
6. Plagiarism will not be tolerated

Guidelines for Report:

1. The report should be to the point. Justify the space you use!
2. Explanations for each task should be included in the report. You should know 'why' behind whatever you do.
3. Do not paste code snippets in the report.

Question 1: K-Means clustering is an unsupervised learning algorithm which groups the unlabeled dataset into different clusters. [30]

You have been given the Glass Classification Dataset and the details about the dataset is given in the link.
Do the pre-processing of the data before performing the following tasks

a) Build a k-means clustering algorithm(can use sklearn library) and implement using the value of k which you find suitable. Visualize this part by showing the clusters along with the centroids. **[10 Marks]**

b) Use different values of k and find the Silhouette score and then tell which value of k will be optimal and why? **[8 Marks]**

c) There are few methods to find the optimal k value for k-means algorithm like the Elbow Method . Use the above method to find the optimal value of k. **[5 Marks]**

d) Apply bagging with the KNN classifier as the base model. Show results with different values of K(=1,2,3). Comment on whether the accuracy changes or not after bagging with KNN along with the proper reason in terms of variance and bias.**[7 Marks]**

Question 2: We will use the Olivetti dataset for this question (you can download it from any other source also including libraries). Flatten and preprocess the data (if required) before starting the tasks. It will become a *4096* dimensional data with *40* classes, more details are available in the link. Inbuilt functions of sklearn can not be used for this question (except for functions for auxiliary tasks) **[40 Marks]**

a) Implement a k-means clustering algorithm from scratch. **[8 Marks]**

b) Make sure that it should:

i) Be a class which will be able to store the cluster centers. **[1 Marks]**

ii) Take a value of k from users to give k clusters. **[2 Marks]**

iii) Be able to take initial cluster center points from the user as its initialization. **[1 Marks]**

iv) Stop iterating when it converges (cluster centers are not changing anymore) or, a maximum iteration (given as max_iter by user) is reached. **[2 Marks]**

c) Train the k-means model on Olivetti data with k = 40 and 10 random 4096 dimensional points (in input range) as initializations. Report the number of points in each cluster. **[8 Marks]**

d) Visualize the cluster centers of each cluster as 2-d images of all clusters. **[4 Marks]**

e) Visualize 10 images corresponding to each cluster. **[3 Marks]**

f) Train another k-means model with 10 images from each class as initializations , report the number of points in each cluster and visualize the cluster centers. **[5 Marks]**

g) Visualize 10 images corresponding to each cluster. **[2 Marks]**

h) Evaluate Clusters of part c and part f with Sum of Squared Error (SSE) method. Report the scores and comment on which case is a better clustering. **[4 Marks]**

Question 3: DBSCAN is another unsupervised learning algorithm that is used to group together the unlabeled data points having similar characteristics. The dataset used in this question is the Wholesale Customers Dataset. and *make_moons* dataset from sklearn **[30 Marks]**

A. Check out the dataset & preprocess the data so that the scale of each variable will be the same. **[5 Marks]**

B. Find out the covariance between the pair of features with which you can best visualize the outliers. Also, visualize the same set of features. **[7 Marks]**

C. Apply DBSCAN to cluster the data points and visualize the same. **[5 Marks]**

D. Apply KNN on the same dataset and compare the visualization with DBSCAN. Comment on what you observe with reason in the report. **[5 Marks]**

E. Use the make_moons function of sklearn to create a datasat of 2000 points. Add some noise to the plot, i.e., randomly add data points to the plot with a 20% probability. Apply DBSCAN and KNN to cluster them and finally compare the plots and comment on which one is better.**[8 Marks]**