Pattern Recognition and Machine Learning

Assignment-9

Submission Deadline :

Guidelines for submission:

- Perform all tasks in a single collab file.
- Create a report regarding the steps followed while performing the given tasks.
- The report should not include excessive unscaled preprocessing plots.
- Try to modularize the code for readability wherever possible
- Submit the collab file [.ipynb] and report [.pdf] (separately) on the classroom assignment
- Submit the .py file on the floated form for the laboratory (it will be floated on _____) and follow the convention -> **Rollno_Name.py**(don't include any blank space)
- Plagiarism will not be tolerated.
- The report should be to the point and explanations for each task you performed should be included in the report.
- DO NOT copy 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 Wine Quality 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) Use any dimension reduction technique of your choice, visualize the data and by looking at the plot tell which value of k will be best suited for k-means clustering and why? (no need to use any method to find optimal k). [7]

b) Build a k-means clustering algorithm(can use sklearn library) and implement using the value of k what you have chosen above. Visualize part b by showing the clusters along with the centroids. [10]

c) Use different values of k and find the Silhouette Score and then tell which value of k will be optimal and why? [8]

d) 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]

Question 2: We will use the fashion-MNIST 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 784 dimensional data with 10 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) [45]

a) Implement a k-means clustering algorithm from scratch. [8]
b) Make sure that it should:
    i) Be a class which will be able to store the cluster centers. [2]
    ii) Take a value of k from users to give k clusters. [2]
    iii) Be able to take initial cluster center points from the user as its initialization. [2]
    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]
c) Train the k-means model on f-MNIST data with k = 10 and 10 random 784 dimensional points (in input range) as initializations. Report the number of points in each cluster. [8]
d) Visualize the cluster centers of each cluster as 2-d images of all clusters. [4]
e) Visualize 10 images corresponding to each cluster. [3]
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. [7]
g) Visualize 10 images corresponding to each cluster. [2]
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. [5]

Question 3: Hierarchical clustering is another unsupervised learning algorithm that is used to group together the unlabeled data points having similar characteristics.
Dataset: Brain Cancer Dataset : you can check the description about the dataset in the given link. [25]

A. Check out the dataset & normalize the data so that the scale of each variable will be the same. [5]
B. Use any dimension reduction technique and visualize the dataset & find out the number of communities available. [7]
C. Visualize the communities from part A. [5]
D. Apply Agglomerative hierarchical clustering (using sklearn). [5]
E. Apply K-means (sklearn) and make a comparison between these two approaches & justify your results. [8]