24-787 Homework 8

Due date: 04/10/2021 @ 11:59 pm EST

Note: In case a problem requires programming, it should be programmed in `Python`. In Programming, you should use plain `Python` language, unless otherwise stated. For example, if the intention of a Problem is familiarity with numpy library, it will be clearly noted in that problem to use numpy. Please submit your homework through Gradescope.

Submissions: There are two steps to submitting your assignment on Gradescope:

1. **HW08_Writeup**: Submit a combined pdf file containing the answers to theoretical questions as well as the pdf form of the FILE.ipynb notebooks.

- To produce a pdf of your notebooks, you can first convert each of the .ipynb files to HTML.
- To do this, simply run: `ipython nbconvert -to html FILE.ipynb` for each of the notebooks, where FILE.ipynb is the notebook you want to convert. Then you can convert the HTML files to PDFs with your favorite web browser.
- If an assignment has theoretical and mathematical derivation, scan your handwritten solution and make a PDF file.
- Then concatenate them all together in your favorite PDF viewer/editor. The file name (FILE) for naming should be saved as HW-assignmentnumber-andrew-ID.pdf. For example for assignment 1, my FILE = HW-1-lkara.pdf
- Submit this final PDF on Gradescope, and **make sure to tag the questions correctly!**

2. **HW08_Code**: Submit a ZIP folder containing the FILE.ipynb notebooks for each of the programming questions. The ZIP folder containing your iPython notebook solutions should be named as HW-assignmentnumber-andrew-ID.zip

Q1: Kmeans algorithm in Python from scratch (50 pts)

For this question you are given a dataset named `alpha_shape.csv` of 3-dimension (x, y, z). You will use this dataset to create cluster using kmeans algorithm.

(a) 5pts Initialize 3 centroids randomly within the bounds of data (i.e. the x, y, z of each centroid must be within min and max of x, y, z of the data). Visualize these centroids and individual datapoints using matplotlib. You will need to output a 3d scatter plot with dataset and centroids clearly distinguishable. You can change size,color or shape for centroids.

(b) 20pts Now using 3 centroids (i.e. k=3) initialized earlier implement k-means algorithm from scratch to cluster the data into 3 clusters. This means that you can not use sklearn modules, you can use numpy or pandas to handle arrays and process database. Output centroids and labels (cluster 1, cluster 2, cluster 3) for each datapoint in your notebook. (Note: You can use the recitation code utilised on April 2nd, 2021)

(c) 10pts For this question you are required to output a total of 4 plots:

1) Visualize clusters (use differentiated color schemes for different clusters) and centroids using a 3-d scatter plot.

Visualize decision boundaries for clusters in the following planes (your output would be 2-d plots):

2) $x = \mu(x_i)$, ($\mu(x_i)$ is the mean of x for the dataset)

3) $y = \mu(y_i)$, ($\mu(y_i)$ is the mean of y for the dataset)

4) $z = \mu(z_i)$, ($\mu(z_i)$ is the mean of z for the dataset)

Try to show the decision boundaries for all 3 clusters in 2),3),4). This is not always true though, some centroids may be farther from other centroid for every point in a plane. You may need to change the plot bounds to visualize sectors corresponding to all 3 clusters. (Hint: you can use meshgrid command to create test points to visualize boundary condition. The output plot for 1) would be 3-d scatter plot but for 2),3) 4) would be a 2-d plot)

(d) 5pts Using **sklearn kmeans** module fit the data into 3 clusters and visualize the scatter plot of these clusters similar to 1) in question (c). Output centroids and label for each datapoint. Also Visualize scatter plot similar to 1) in (c)

(e) 10pts In this question you are required to use the Elbow method to determine the ideal cluster numbers for the database. You can use sklearn for this question. Vary your k value (of clusters) from 1 to 20. Use negative of `kmeans.score` from sklearn to evaluate each of your models. Finally, Plot a Score vs K Clusters graph and qualitatively decide what the best k should be. Also, comment why should you decide this value. You should submit a plot of "Score vs K clusters". Also mention a couple of lines outlining your argument about the K value decision.

Q2: Clustering Algorithms (50 pts)

For each of the datasets shown in Figure1 (**provided as text files with this assignment**), you will implement three clustering methods; you may assume that the number of clusters is known (k=5). **You may use inbuilt scikit functions to implement your code.** Plot the results, using a different color for each cluster.

You will use the following three methods: spectral clustering, k-means, hierarchical single link. Feel free to play with the hyper parameters to obtain good results.

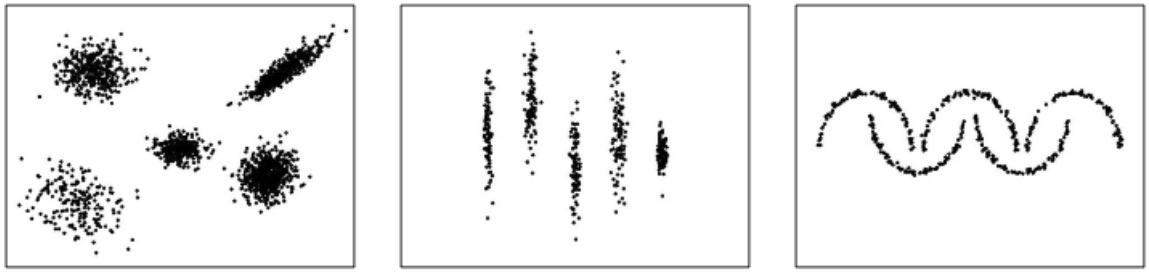

Figure 1: Clustering datasets

(a) Spectral Clustering from Scratch (10 pts) This problem is open ended in that you will have to study spectral clustering on your own.

1. Use the *normalized symmetric laplacian matrix*. You may find the following sites useful:

   https://en.wikipedia.org/wiki/Spectral_clustering
   https://charlesmartin14.wordpress.com/2012/10/09/spectral-clustering/

2. For 2 data points x_i, x_j you can construct the affinity matrix W (or, could be denoted as A) as follows:

   $$W_{i,j} = exp\Big(- \Big(\frac{(x_i - x_j)^2}{\sigma^2} \Big) \Big) \tag{1}$$

   from which you will also compute the D matrix. You will have to choose an appropriate value for the standard deviation in the distance function (through trial and error). With this, the normalized symmetric laplacian will be:

   $$L = I - D^{-0.5} W D^{-0.5} \tag{2}$$

3. Use scikit's kmeans function to cluster the points obtained from the set of eigenvectors (v) of L (where, w,v = np.linal.eig(L)). You can explicitly specify the number of clusters to be 5 for each dataset. Note that if done correctly, your first eigenvalue(w) should be zero (rigid body translation), so you should start from the second eigenvector.

4. You will have to decide the appropriate number of eigenvectors to use.

   Show your clustering results using the following number of eigenvectors: 1,2,5,8.

   Rate the quality of the results via visual inspection.

   For instance:

   Quality of Spectral Clustering as a function of the number of eigenvectors:

   Data Set (a): $5 > 2 = 1 > 8$

   Which means, for data Set (a), the clustering you obtain with 5 eigenvectors is better than the results for 2 eigenvectors, which is equal to the results with 1 eigenvector etc.

5. Produce 12 color plots for spectral clustering (3 datasets X 4 cases per dataset).

(b) Spectral Clustering using scikit (10 pts)

1. Cluster each of the three datasets using Spectral Clustering. Assume the number of clusters is 5 in all cases.

   (*Refer:* `https://scikit-learn.org/stable/modules/clustering.html#spectral-clustering`)

2. You will have to decide the appropriate number of eigenvectors to use.

   Show your clustering results using the following number of eigenvectors: 1,2,5,8.

   Rate the quality of the results via visual inspection.

(c) Kmeans Clustering using scikit (10 pts)

1. Cluster each of the three datasets using kmeans. Assume the number of clusters is 5 in all cases. You can use the Euclidian distance as your distance measure.

   (*Refer:* `https://scikit-learn.org/stable/modules/clustering.html#k-means`)

2. Produce a total of 3 color plots showing the clustering results for the three data sets (one plot per data set).

(d) Hierarchical Single Link using scikit (10 pts)

1. Cluster each of the three datasets using hierarchical single link. Again, assume the number of clusters is 5 in all cases. You can use the Euclidian distance as your distance measure.

   (*Refer:* `https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering`)

2. Produce a total of 3 color plots showing the clustering results for the three data sets (one plot per data set).

(e) Discussion (10 pts)

1. For each dataset, rank the three clustering methods based on their performance. You may use the best performing eigenvector setting for spectral clustering to compare it against kmeans and hierarchical clustering. Is one method consistently better than the others? In 4-5 sentences summarize your observations. In particular, try to characterize which method is best for which kind of dataset and why.

2. If labels were provided, could you use support vector machines to accurately classify these datasets? Explain your reasoning.