Assignment 3, Hierarchical Clustering, Density-based Clustering, Bayesian and Nearest Neighbor Classification: 50 points

Due October 18, 2024, 11:59 PM CST

Instructions

Your homework submission must cite any references used (including articles, books, code, websites, and personal communications). All solutions must be written in your own words, and you must program the algorithms yourself. Submit your solutions as a PDF to Canvas: `https://canvas.tamu.edu/`.

Your programs must be written in Python. The relevant code to the problem should be submitted as a separate file (*e.g.*, Python file (.py), Jupter/Google Colab Notebook (.ipynb)). Please **do not** include screenshots of your code in your submission. If a problem involves programming, then the code should be shown as part of the solution to that problem. If you solve any problems by hand just digitize that page and submit it (make sure the problem is clearly labeled and legible). **Clearly** label your figures and tables.

If you have any questions, please reach out to Dr. Peeples.

1 Hierarchical Clustering- 20 points

Given the following dataset shown in Figure 1, please answer the following questions. Please answer the following questions. You may either solve by hand, typeset, or creating python script:

1. Show the dendrogram resulting from the single-link hierarchical agglomerative clustering approach using the $L_1 - norm$ as the distance between two points.

2. Show the merge order tree and table, stopping when you have $k = 4$ (hint: Slide 11, Lecture 10). Merge the clusters by taking the minimum feature values of data points.

3. Cluster the dataset using bisecting k-Means with $k = 4$ and plot the dataset with the corresponding cluster assignments.

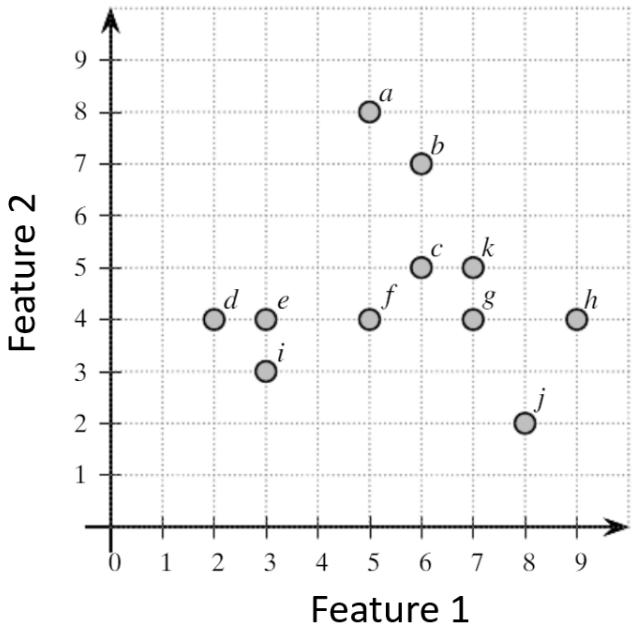

Figure 1: Dataset for Problem 1 and 2

4. How do the results of the hierarchical agglomerative clustering compare with the clustering of bisecting k-Means for this dataset? Do the two clustering approaches achieve the same results? Clearly discuss the results for each algorithm and include quantitative/qualitative assessments.

2 DBSCAN Clustering- 10 points

Consider the points in Figure 1. We have the two following distance functions:

$$L_\infty(\mathbf{x}, \mathbf{y}) = \max_{i=1}^{d} \{|x_i - y_i|\}$$

$$L_{min}(\mathbf{x}, \mathbf{y}) = \min_{i=1}^{d} \{|x_i - y_i|\}$$

Answer the following questions. Please complete by **hand** or **typeset**. Clearly show and compute distance matrix.

1. Using $\epsilon = 2$, $minpts = 5$, and L_∞, find all core, border, and noise points.
2. Using $\epsilon = 1$, $minpts = 6$, and L_{min}, find all core, border, and noise points.

3 Mean Shift Clustering - 10 points

Please access the Iris Flower Dataset via Sklearn and select the following two attributes: petal length and petal width.

1. Use Mean Shift to cluster the data with bandwidth = 1 (use the default settings of the algorithm in Sklearn).
   (a) Plot the two features with the associated class labels for each data point.
   (b) Plot the two features with the associated cluster labels for each data point.
   (c) Report the Silhouette index for the resulting clustering assignments.
2. Compare the initial bandwidth clustering results with the clustering results of the bandwidth estimation function is Sklearn.
   (a) Report the estimated bandwidth value.
   (b) Plot the two features with the associated cluster labels for each data point using the bandwidth from the estimation function.
   (c) Which bandwidth leads to the "optimal" clustering for the dataset? Justify your answer.

4 Bayesian and Nearest Neighbor Classification- 10 points

You will be implementing two different types of classifiers to distinguish between species of rock crabs of genus Leptograpsus. Please download the dataset provided along with this assignment: *Assignment_3_Dataset.txt*. The dataset is composed of 200 samples of different crab specimens. There are five features that were captured for each specimen. These features include anatomical properties: the front lip, rear width, length, width, and depth of the crab. Your goal is to discriminate between the two species of crab using these provided features (total of 5). The species of crab are given as binary labels in the first column of the provided dataset. The remaining five columns of data are the real-valued features. You will be considering two classifiers: Gaussian Naive Bayes and k-Nearest Neighbors (KNN).

Please answer the following questions by creating python script to implement the classifiers:

1. Break apart the observations into training and testing sets. Use the first 70% of the data for training (first 140 samples) and the remaining 30% of the data for testing (remaining 60 samples). Use the default setting for each classifier and provide the following:
   (a) Training set confusion matrix

(b) Testing set confusion matrix

(c) Test accuracy, precision, recall, f1 score

2. Compare the performance of the two classifiers for the crab dataset. What are the advantages and disadvantages of each classifier?