Programming Assignment 4: Clustering Analysis

Shen-Shyang Ho (Dr.)

November 14, 2022

- In this assignment, you will be using the dataset assigned to you in Assignment 1.
- We will reuse the histogram representation for images you have created in Assignment 2 for clustering. Use only the images/histogram for the four weed classes and ignore the negative class images.
- The labels will be used as ground truths for performance evaluation when we use external performance measure.
- You will use the following clustering methods: **K-means, Spectral Clustering, Hierarchical Clustering, DBSCAN, Bisecting K-means**
- Scikit-learn (https://scikit-learn.org/stable/user_guide.html) will be used in this assignment.
- In particular, most important coding information should be available in https://scikit-learn.org/stable/modules/clustering.html

1. Convert the images from the four weed classes (ignoring the negative class images) to grayscale pixel intensity histograms. (You should have done this in Assignment 2) and normalize the histogram dataset.
2. **[K-means Clustering and its variants]** Using $K = 4$, we will investigate the behavior of the K-means, bisecting K-means, and spectral clustering algorithm.
   (a) Perform 50 trials for K-mean clustering (Use KMeans with init = 'Random') using different random initial centroids on your assigned dataset.
   (b) Perform clustering performance evaluation using Fowlkes-Mallows index (sklearn.metrics.fowlkes_mallows_score). Compute the Fowlkes-Mallows index for each trial. Plot a histogram with 10 equal bins (for range of [0,1]) showing the distribution of the Fowlkes-Mallows index for the 50 trials.
   (c) Perform clustering performance evaluation using Silhouette Coefficient (sklearn.metrics.silhouette_score). Compute the Silhouette Coefficient for each trial. Plot a histogram with 10 equal bins (for range of [-1,1]) showing the distribution of the Silhouette Coefficient for the 50 trials.
   (d) Repeat Step (a)-(c) for (1) KMeans with init='k-means++', (2) bisecting K-means (sklearn.cluster.BisectingKMeans with init = 'Random' and all other parameters to be default), and (3) spectral clustering (sklearn.cluster.SpectralClustering with default parameters) (**4 points - 1 point for each method**)
   (e) What is the Fowlkes-Mallows index? What is the main difference between Fowlkes-Mallows index and Silhouette Coefficient? (**0.5 point**)
   (f) Computer the average Fowlkes-Mallows index using the values you obtained in (b) for each method. Which one performs the best? (**0.5 point**)

(g) Computer the average Silhouette Coefficient using the values you obtained in (c) for each method. Which one performs the best? Is it consistent with the result you obtain in (f)? Explain your observation. (**0.5 point**)

3. **[Model Selection Process for Clustering Task]** Perform model selection for the k-mean clustering algorithm (Use KMeans with init = 'Random') using $K = 2, 3, 4, 5, 6, 7, 8$ and using the average Silhouette Coefficient (similar steps as Question 2(c) and 2(g)) as clustering performance evaluation. Again, to obtain the average Silhouette Coefficient, perform 50 trials for each K. (**1 point**)

   (a) Plot a graph with K as the x-axis and average Silhouette Coefficient as the y-axis. (**0.25 point**)

   (b) What is the best K to use for your dataset based on the plotted graph? (**0.25 point**)

4. **[Density-based Clustering]** Perform dimension reduction on your histogram dataset to reduce the dimension to 2 (similar to Assignment 1 Question 2(e)). Perform DBSCAN on the 2D dataset to obtain 4 to 10 clusters. (**1 point**)

   (a) How many clusters did you obtain? What is the Silhouette Coefficient for the clustering you obtain? What are the eps and min_samples parameter values you used to get your clustering? (**0.25 point**)

   (b) Plot your clustering result similar to the one shown in https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html #sphx-glr-auto-examples-cluster-plot-dbscan-py (**0.25 point**)

5. **[Hierarchical Clustering]** We will investigate the four different strategies: Single (MIN), Complete (MAX), Average, and Ward, for agglomerative clustering (i.e., hierarchical clustering) we learned in the lecture using sklearn.cluster.AgglomerativeClustering with number of clusters set to 4 using the 2D dataset constructed in Question 4.

   (a) Use the four linkage values 'ward', 'complete', 'average', 'single' for sklearn.cluster.AgglomerativeClustering to construct 4 clusterings. (**1 point**)

   (b) What are the Silhouette Coefficients for each clustering in (a)? Based on the Silhouette Coefficients, which strategy performs the best? (**0.25 point**)

   (c) Plot the four clustering results (separately and label them clearly in your submission) similar to the one in 4(b) (but no noise point). Based on your observation of the plots, which strategy performs the best? Is it consistent with the result you obtain in (b)? (**0.25 point**)