Programming Assignment 4:
Clustering Analysis

Shen-Shyang Ho (Dr.)

April 22, 2023

- 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. Perform dimension reduction on your histogram dataset to reduce the dimension to 2 (similar to Assignment 1 Question 2(e)).
3. Perform clustering using the following approaches on the 2D dataset you preprocessed in Item 2:
   - K-mean clustering and its variants for $K = 4$:
     (a) K-means clustering: (Use KMeans with init = 'Random') (**0.5 point**)
     (b) KMeans with init='k-means++' (**0.5 point**)
     (c) Bisecting K-means (sklearn.cluster.BisectingKMeans with init = 'Random') (**0.5 point**)
     (d) spectral clustering (sklearn.cluster.SpectralClustering with default parameters) (**0.5 point**)
   - DBSCAN (**0.5 point**)
     - What are the eps and min_samples parameter values you used to get 4 clusters? (**0.5 point**)
   - Agglomerative clustering (i.e., hierarchical clustering) - use sklearn.cluster.AgglomerativeClustering with number of clusters set to 4
     (a) Single link (MIN), (**0.5 point**)
     (b) Complete link (MAX), (**0.5 point**)
     (c) Group Average, and (**0.5 point**)
     (d) Ward's method (**0.5 point**)

Use the four linkage values 'ward', 'complete', 'average', 'single' for sklearn.cluster. AgglomerativeClustering

4. For all the methods in Item 3:

    (a) Perform clustering performance evaluation using Fowlkes-Mallows index (sklearn.metrics.fowlkes_mallows_score). Compute the Fowlkes-Mallows index for each method on the 2D dataset. (**0.5 point**)

    (b) Perform clustering performance evaluation using Silhouette Coefficient (sklearn.metrics.silhouette_score). Compute the Silhouette Coefficient for each method. (**0.5 point**)

    (c) What is the main difference between Fowlkes-Mallows index and Silhouette Coefficient? (**0.5 point**)

    (d) Rank the methods from the best to the worst for our dataset based on Fowlkes-Mallows index. (**0.5 point**)

    (e) Rank the methods from the best to the worst for our dataset based on Silhouette Coefficient. (**0.5 point**)

    (f) Which one do you think is better for our dataset? Fowlkes-Mallows index or Silhouette Coefficient? Why? (**0.5 point**)

5. Perform K-mean clustering and its variants for $K = 4$ in Item 3 (i.e., 4 methods need to be performed) on the processed dataset in Item 1 (i.e., perform clustering on the dataset without dimensionality reduction). (**0.5 point**)

    (a) Do 4(a) (**0.25 point**)

    (b) Do 4(b) (**0.25 point**)

    (c) Do 4(d) (**0.25 point**)

    (d) Do 4(e) (**0.25 point**)

    (e) Are the methods performing better with the original dataset or the one with dimension reduced? Explain your observation. (**0.5 point**)