Programming Assignment 4: Clustering Analysis

Shen-Shyang Ho (Dr.)

November 14, 2023

- In this assignment, you will be using the 4-class dog dataset assigned to you in Assignment 1.
- We will use the features extracted from the last convolution layer of a "ResNet18" deep learning architecture as image representation for this clustering assignment using "forward_hook" in PyTorch (https://pytorch.org/docs/stable/generated/torch.nn.modules.module.register_module_forward_hook.html) OR "feature_extraction" in TorchVision (https://pytorch.org/vision/stable/feature_extraction.html). There are other approaches that work. **You can also use those approaches**. **You must do proper reference to the website(s) you seek help from.**
- 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. (**Feature Extraction**)
   - Resize each cropped image to a 224×224 pixel image. (Similar to Assignment 1 Question 2(a))
   - Normalize the resized image dataset.
   - Extract features for each image from the last convolution layer of "ResNet18" (You can follow https://kozodoi.me/blog/20210527/extracting-features. But you must reference this website in your solution) **(2.5 points)**
2. (**Dimension Reduction**) Perform dimension reduction on your new dog image representation dataset to reduce the dimension to 2 (similar to Assignment 1 Question 2(f)). **(0.5 points)**
3. (**Clustering Algorithm**) 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. (**Clustering Evaluations**) 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) Rank the methods from the best to the worst for our dataset based on Fowlkes-Mallows index. (**0.5 point**)

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