CSE 590-04 **Introduction to Machine Learning** Spring 2022

Homework No. 5
Due April 6 (11:59pm), 2022

Objectives

1. *Apply K-Means and Agglomerative clustering algorithms to real data*
2. *Analyze and optimize the parameters of each clustering method*
3. *Analyze and compare the clustering results*

Problem 1 (60 points)

a) Use the *K-Means* algorithm to cluster the provided data. Vary the number of clusters from 2 to 20 and select the optimal number. Justify your choice based on the **SSE vs. No. clusters plot**.
b) Using the number of clusters selected in (a), generate the silhouette plot.
c) Using the silhouette coefficients, identify 5 samples that are at the core of each cluster and 2 samples that are at the boundary of any two clusters (if they exist). Display the original images associated with these samples and comment on the results.

Problem 2 (75 points)

a) Use the hierarchical *agglomerative* algorithm, with the **Ward's method** to compute the distance between two clusters, to cluster the provided data. Generate the dendrogram and use that to identify the optimal number of clusters. Justify your choice.
b) Using the number of clusters selected in (a), generate the silhouette plot.
c) Repeat (a) and (b) using **single-link** and **complete-link**. Compare the silhouette plots of the 3 methods and identify the best distance for this data. Justify your choice.
d) Using the silhouette coefficients of the best method identified in (c), identify 5 samples that are at the core of each cluster and 2 samples that are at the boundary of any two clusters (if they exist). Display the original images associated with these samples and comment on the results.

(15 points)

For each clustering method (K-Means, Agglomerative), compute the adjusted rand index by comparing the generated clusters to the provided ground truth (**this should be the only time you use the ground truth**). Using these ARI's and the visualizations generated for each problem, identify the best clustering method for this application. Justify your choice.

What to submit?

- A report that
  - **Describes** your experiments, the parameters considered for each method, etc.
  - **Summarizes**, **explains** (using concepts covered in lectures) and **compares** the results (using plots, tables, figures)
- Do not submit your source code
- Your report needs to be a single file (MS Word or PDF)
- Your report cannot exceed 10 pages using a font of 12
- Assign numbers to all your figures/tables/plots and use these numbers to reference them in your discussion