Homework 4 (Spring'25): Machine Learning with Dr. B

Dataset & Assumptions & Restrictions to follow

- **Back to the dataset from Homework 2:** `spiral-dataset.csv` (Courtesy to *H. Chang and D.Y. Yeung, Robust path-based spectral clustering. Pattern Recognition, 2008. 41(1): p. 191-203.*)
  - The spiral dataset represents three intertwined spirals, each with approximately 100 two-dimensional data points. Please see a plot of all the points below. The three spirals are intentionally given colors (blue, red and green) to emphasize the obvious 3-clusterings as you can see below. I believe you can appreciate how human eyes/head/brain can distinguish the three clusters quite easily!:

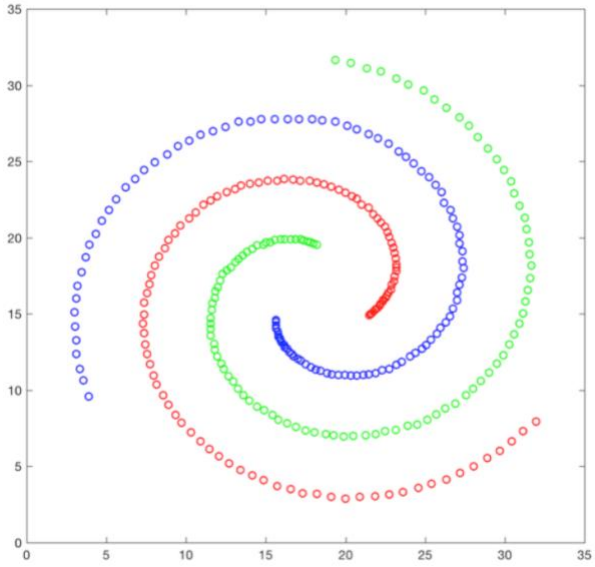

Figure 1: Plot of the given spiral dataset. ***Please note, there are 3 colors used in the figure above: green, blue and red (from outward to inward).***

  - The spiral dataset is available in the **spiral-dataset.csv** file. The file contains three columns, corresponding to the X and Y coordinates in the Cartesian plane, as well as the cluster number in the third column of the csv file are to denote only the membership of each data point to one of the three clusters. *Please note that the cluster numbers are irrelevant in clustering as it is an unsupervised learning algorithm.* ***However, as we happen to have the true clustering results here (indicated by the 3rd column in the dataset), we can leverage this extra bit of information to evaluate the clustering results externally, a metric known as the RandIndex (an extrinsic metric for clustering evaluation), besides measuring the sum-of-squared-error (intrinsic metric) which you can find in my lecture note. More on this is explained in a separate section below (page 3& 4). Please continue reading.***
  - As you saw in Homework 2, this type of dataset is difficult to cluster with a partitional clustering algorithm, such as k-means! Now that you know hierarchical clustering, you can do better.
  - **You may assume that there is no need to normalize the dataset.**
  - **NO LIBRARY FUNCTIONS OF clustering WILL BE ALLOWED.**

Tasks (for all)

1. Implement the Hierarchical clustering algorithm. And do the following:
   - 1.a) Using the "**single linkage**" method, run the hierarchical clustering algorithm on the dataset, and get a 3-cluster result (by cutting the dendrogram at a certain height), and report SSE, RI, Cophenetic Correlation Coefficient, Silhouette Score.
   - 1.b) Using the "**complete linkage**" method, run the hierarchical clustering algorithm on the dataset, and get a 3-cluster result (by cutting the dendrogram at a certain height), and report SSE, RI, Cophenetic Correlation Coefficient, Silhouette Score
   - 1.c) Using the "**average linkage**" method, run the hierarchical clustering algorithm on the dataset, and get a 3-cluster result (by cutting the dendrogram at a certain height), and report SSE, RI, Cophenetic Correlation Coefficient, Silhouette Score
   - 1.d) Using the "**centroid linkage**" method, run the hierarchical clustering algorithm on the dataset, and get a 3-cluster result (by cutting the dendrogram at a certain height), and report SSE, RI, Cophenetic Correlation Coefficient, Silhouette Score
   - 1.e) Please comment, out of the 4 clustering results (1.a), (1.b), (1.c) and (1.d) which method gets you the best SSE, RI, Cophenetic Correlation Coefficient, Silhouette Score.
   - 1.f) Please draw the clustering results obtained in (1.a), (1.b), (1.c) and (1.d) (like Figure 1).

Graduate Students Only

2. Could you please draw dendrograms for each hierarchical clusterings in (1.a), (1.b), (1.c) and (1.d)