Programming Assignment 3

Dataset & Assumptions & Restrictions to follow

- **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!:

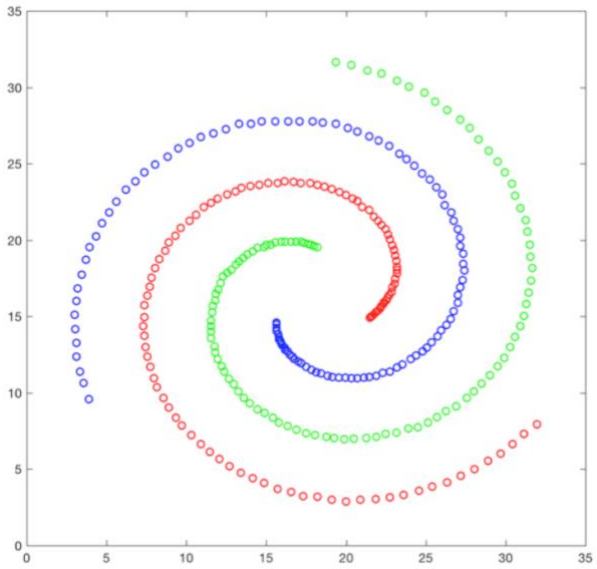

Figure 1: Plot of the given spiral dataset. ***Please note, there are 3 colors used in the figure above: green, blue and red (from outward to inward).***

  - 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, we can leverage this extra information to evaluate the clustering results externally, a metric affectionately called the RandIndex (an extrinsic metric for evaluation), besides measuring the sum-of-squared-error (intrinsic metric) which you can find in my lecture note***
  - It should be noted that this type of dataset is difficult to cluster! But, I have trust in you; you are clever enough to employ the appropriate clustering algorithm to properly cluster the dataset. You need to explore most of the clustering approaches you learned in the class.
  - **You may assume that there is no need to normalize the dataset.**
  - **Use the Euclidean distance measure for all the distance calculations.**
  - **NO LIBRARY FUNCTIONS OF k-means and hierarchical clustering WILL BE ALLOWED.**

Tasks

1. Generate a figure from the given dataset that resembles Figure 1.
2. Implement the k-means clustering algorithm. And do the following:
   - 2.a) Run your k-means algorithm on the given dataset setting the value k=3 (because visually we only have 3 clusters to worry about). And do not forget to randomly initialize the 3 centroids.
   - 2.b) Once your k-means algorithm has converged above, stop and from your clustering result compute the intrinsic performance metric: **Sum of Squared Error, SSE** (smaller the better), and the extrinsic performance metric: **Rand-Index, RI** (higher the better).
   - 2.c) Repeat Task (2.a) & (2.b) another 9 (nine) times randomizing again the initial centroids, and report out of the 10 runs of k-means what is the best SSE & RI you could get.
   - 2.d) Please draw the clustering results (like Figure 1).
3. (40 pts) Implement the Hierarchical clustering algorithm. And do the following:
   - 3.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 and RI.
   - 3.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 and RI.
   - 3.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 and RI.
   - 3.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 and RI.
   - 3.e) Please comment, out of the 4 clustering results (3.a), (3.b), (3.c) and (3.d) which method gets you the best SSE as well as RI.
   - 3.f) Please draw the clustering results (like Figure 1).