Programming Assignment 4

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!:

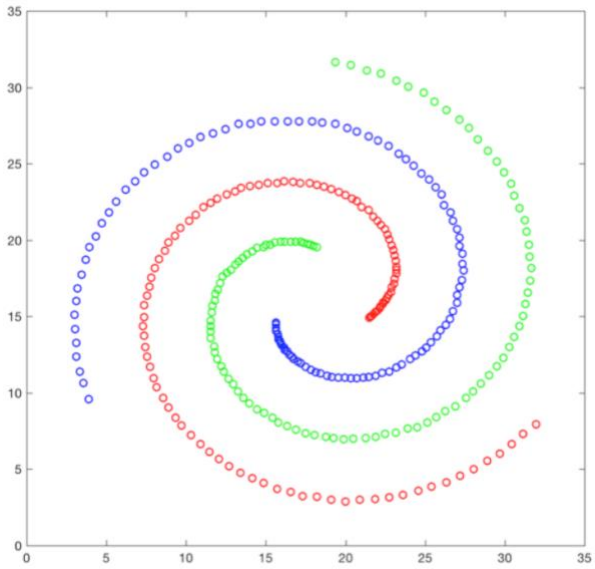

Figure 1: Plot of the given spiral dataset. ***Please note, there are 3 colors used in the figure above: green, blue and red (from outward to inward).***

  - 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. More on this is explained in a separate section below (page 3& 4). Please continue reading.***
  - 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 (for all)

1. (20 pts) Generate a figure from the given dataset that resembles Figure 1.
2. (40 pts) 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). For the definition of both, please continue reading.
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.

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.

Tasks only for MS students who are enrolled in CSCI-5930

MS1: (10 pts) Please draw the clustering results (like Figure 1) for all of task 2.
MS2: (10 pts) Please draw clustering results (like Figure 1) for all of task 3.
MS3: (10 pts) Please draw the dendrograms for each of the 4 hierarchical clustering results. *Hint: there are library functions to take care of the dendrogram plotting, and you are now allowed to leverage that.*
MS4: (10 pts) Consider the "cosine similarity" instead of "Euclidean distance" in solving for task 3.
MS5: (10 pts) Consider "L_3 distance" instead of "Euclidean distance (i.e., L_2 distance)" in solving for task 3.

Evaluating Clustering Results

Intrinsic Measure of Evaluation

- **Sum of Squared Error, SSE**

It is actually the sum of the squared distances between the given data samples with their corresponding centroids.
It is defined as:

$$SSE(C) = \sum_{i=1}^{k} \sum_{x \in C_i} dist(x, cen_i)^2$$

Here,

- C is a k clustering result representing the k partitions of the given m sample dataset. And obviously $\sum_{i=1}^{k} |C_i| = m$, meaning sum of the cardinalties of each of the k partitions equals total number data samples.
- *dist(a,b)* denotes the Euclidean (L_2) distance between two samples on an Euclidean space.
- The centroid (mean) of the i^{th} cluster is defined as: $cen_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$. To illustrate, the centroid of a cluster containing the three two-dimensional data samples: (1,1), (2,3) and (6,2) is: $(\frac{(1+2+6)}{3}, \frac{(1+3+2)}{3}) = (3,2)$.

Extrinsic Measure of Evaluation

- **Rand Index, RI**
  - This metric utilizes into the ground-true labels (when provided) of each of the data samples, that is quite unlike for unsupervised learning paradigm. But, here you were given a dataset with truth values associated to each sample, you could take advantage of that, right? RI metric is only applicable to this type of situation. Got it?
  - It is based on a series of decisions, one for each of the $\frac{m(m-1)}{2}$ pairs of the m data samples in the dataset. We want to assign two data samples (a, b) to the same cluster if and only if they are similar.
    - A true positive (TP) decision assigns two similar samples to the same cluster. *That is, if the data samples a and b belong to the same cluster according to the ground truth, as well as in predicted clustering, it counts as a TP assignment.*
    - A true negative (TN) decision assigns two dissimilar samples in to two different clusters. *That is, if the data samples a and b belong to two different clusters in the ground truth, as well as in predicted clustering, it counts as a TN assignment.*
    - The above two cases (i.e., TP and TN) are correctly clustering instances. However, there are two types of errors we can commit:
      - A False Positive (FP) assigns two dissimilar samples to the same cluster. *That is, if the samples a, and b belong to different clusters in the ground truth, but your predicted clustering puts them into the same cluster, you accrue one FP assignment.*
      - A False Negative (FN) assigns two similar samples into two different clusters. *That is, if the samples a, and b belong to the same cluster in the ground truth, but your predicted clustering puts them into two different clusters, you accrue one FN assignment.*

- Now, the Rand Index, RI measures the percentage of decisions that are correct. In simple term, it sounds a lot like accuracy, and we can treat it as the clustering accuracy:

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

- You can observe that cluster indices for each sample in a clustering result do not matter in computing the RI. You can literally use anything to mark the cluster memberships for each of the data samples, and you should get the same Rand Index value.
- For a complete workout example, please look at this webpage https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html
- And if you are looking for a sample Python code, here you go https://davetang.org/muse/2017/09/21/the-rand-index/