Practical Session 10 : recap 2

Exercise 1 We shall work on synthetic data

1. Creation of a synthtic dataset
   - (a) Simulate synthetic data as follows
     ```
     np.random.seed(123)
     X = np.random.randn(50,2)
     X[0:25, 0] = X[0:25, 0] + 3
     X[0:25, 1] = X[0:25, 1] - 4
     ```
   - (b) Visualize this synthetic dataset
2. K-means clustering with two clusters
   - (a) Perform K-means clustering with K = 2
   - (b) How can you have access to the cluster assignments of the 50 observations ?
   - (c) Plot the data, with each observation colored according to its cluster assignment
3. K-means clustering with three clusters
   - (a) Perform k-means with 3 clusters
   - (b) Try with multiple random initial assignements giving different values to the variable `n_init`. Calculate each time the `inertia_`
4. Hierarchical clustering
   - (a) Define the linkage matrix using different options : "complete", "average" or "single".
   - (b) Plot the corresponding dendrograms obtained using the usual `dendrogram()` function
   - (c) Determine the cluster labels for each observation associated with a given cut of the dendrogram, using the `cut_tree()` function

Exercise 2

In this exercise we shall perform hierarchical and K-means clustering compare on the NCI60 cancer cell line microarray data, which consists of 6,830 gene expression measurements on 64 cancer cell lines. Each cell line is labeled with a cancer type. We shall ignore the cancer types in performing clustering, as these are unsupervised techniques. After performing clustering, we shall use this column to see the extent to which these cancer types agree with the results of these unsupervised techniques.

1. Import the two datasets `NCI60_labs.csv` and `NCI60_data.csv`
2. Shape of the data. Take a look at the cancer types for the cell lines
3. We shall now perform hierarchical clustering
   - (a) Perform hierarchical clustering of the observations using complete, single, and average linkage. Use standard Euclidean distance as the dissimilarity measure
   - (b) Cut the dendrogram at the height that will yield four clusters
4. Perform K-means clustering with four clusters
5. Use a confusion matrix to compare the differences in how the two methods assigned observations to clusters