# CS3481 Fundamentals of Data Science

# Assignment 3

## *Objective*

In this assignment, you will study the hierarchical clustering approach introduced in the class using Python.

## *Detailed Requirement*

We have introduced the hierarchical clustering approach in the class. In this assignment, you will apply this approach to the ***Wine*** data set from the UCI Machine Learning Repository. You may also import this data set directly through the `datasets` module from `scikit-learn`.

You can perform hierarchical clustering using the method `linkage` from the module `scipy.cluster.hierarchy`.

After performing hierarchical clustering, you could visualize the clustering result in the form of a dendrogram by using the method `dendrogram`. You could also study the clustering solution for a specific number of clusters in the hierarchy by using the method `fcluster`.

To perform clustering, you should only use the input attributes but not the class label. To improve the clustering results, you could consider removing outliers from the data set, and applying a suitable normalization operation to the input attributes.

## *Assignment Submission*

You should submit a report to summarize your work. The following tasks are to be performed:

1. Compare the hierarchical structures generated using single link, complete link and group average for the ***Wine*** data set. (30%)
2. For some of these hierarchical structures, observe the set of distance values at which cluster merge occurs, and identify possible patterns from these values. (20%)
3. Select different clustering solutions from the hierarchical structures, and compare the cluster groupings with the corresponding K-means clustering solutions (using the method `KMeans` from the module `sklearn.cluster`) based on, for example, the extent to which the clusters can capture the actual groupings of the data points according to their class labels. (30%)
4. Select different subsets of attributes from the data sets and re-perform hierarchical clustering. Compare the resulting hierarchical structures based on the selected attribute subsets with the original hierarchical structures. (20%)

Please provide a detailed description of the results of the above tasks in your report.

# Supplementary Instructions for Assignment 3

To perform hierarchical clustering and K-means clustering in Python, we need to include the following modules:

```
from matplotlib import pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
from sklearn.cluster import KMeans
```

Using the Iris data set as an example:

```
from sklearn import datasets
iris = datasets.load_iris()
X = iris.data
```

Perform hierarchical clustering using the complete link approach:

```
Z = linkage(X,'complete')
```

Visualize the hierarchical clustering result in the form of a dendrogram:

```
plt.figure(figsize=(25, 10))
dendrogram(Z)
plt.show()
```

Extract a clustering solution with a specific number of clusters in the hierarchy (The example below corresponds to the case of three clusters):

```
kclusters = fcluster(Z, 3, criterion='maxclust')
kclusters
```

Perform K-means clustering for a given number of clusters (The example below corresponds to the case of three clusters):

```
km = KMeans(n_clusters=3)
km.fit(X)
km.labels_
```