CS 6220 Data Mining — Assignment 5

Clustering: Agglomerative and K-Means Clustering

This assignment will require you to apply and interpret some of the techniques related to Agglomerative clustering that were introduced in class. Keep in mind that the main objective of this assignment is to highlight the insights that we can derive from applying these techniques—the coding aspect is secondary. Accordingly, you are welcome to consult any online documentation and/or code that has been posted to the course website, so long as all references and sources are properly cited. You are also encouraged to use code libraries, but be sure to acknowledge any source code that was not written by you by mentioning the original author(s) directly in your submission (comment or header).

Objectives:

1. Apply clustering to a dataset
2. Evaluate a clustering method

Submission:

Through the assignment submission portal on Canvas, submit your ipynb with a pdf of your assignment solution; no need to zip the files.

Grading Criteria:
Follow the instructions in the pdf, and complete each task. You will be graded on the application of the modules' topics, the completeness of your answers to the questions in the assignment notebook, and the clarity of your writing and code.

Assignment Description

The Data The iris dataset contains measurements of sepal length, sepal width, petal length, and petal width of 150 iris flowers, with 50 samples each of three species: setosa, versicolor, and virginica. For the clustering we will work only with Sepal Length and Sepal Width of each flower. As we are working with only two features, you can plot this using scatter plot to visualize the data.

The Idea: Hierarchical Clustering
Your objective here is to assess the performance of hierarchical clustering on the provided iris dataset. For this assignment, we will use agglomerative clustering with euclidean distance using ward and complete linkage as discussed in the weekly lessons. For this, we will plot the dendrogram of the iris hierarchy, and then select the number of clusters from the dendrogram diagram. If you draw a horizontal line in the dendrogram diagram, the number of clusters is defined by the number of intersections with the dendrogram vertical lines with the horizontal line.

The Idea: Choosing k for k-means
Your objective here will be to assess the performance of k-means clustering on the provided iris dataset. Recall that the number of clusters, k, is an input parameter to the k-means algorithm. A variety of measurements are available for estimating the optimal value of k. In this assignment, we will look at the 2d plot of the data (as the data has 2 features), and choose the appropriate number of clustering.

What to Do
Load the dataset using sklearn's `load_iris()` function, and remove features other than Sepal Length and Sepal Width from the dataset.

You can use the following code snippet to see the dendrogram of the cluster hierarchy using the `ward` method.

```
import scipy.cluster.hierarchy as shc
plt.figure(figsize=(10, 7))
plt.title("Iris Dendrograms")
dend = shc.dendrogram(shc.linkage(data, method='ward'))
```

Use Agglomerative Clustering from `sklearn` with ward linkage. Also, use complete linkage to compare the clusters with the ward method. Using the dendrogram, determine a few number of clusters (k). Show the clusters in 2d scatter plot for different k values. Show plots for at least 4 different k values.

Implement k-means clustering using the same number of clusters, and show the clusters in 2d scatter plot.

What to Provide
Your output should contain the following:

- Show dendrogram of the iris data using hierarchical clustering using the `ward` method.
- Scatter plots of the data in 2d showing the clusters in different colors using Agglomerative Clustering for different k values. Show the plots side by side for *ward* and *complete* linkage.
- Scatter plot of the data in 2d showing the clusters in different colors using K-Means clustering for different k values. Also show the cluster centers in the plot.

Given this output, respond to the following questions:

1. Based on the scatter plot of the clustered data, which makes the most sense? Give logical interpretation from the clusters.
2. Compare the plots and clusters found by euclidean distance and complete linkage.
3. Compare the scatter plots from Agglomerative and K-Means clustering.