Machine Learning
Lab : 7
Agglomerative Clustering

January 2025

Perform Date: January 20-24, 2025

1 Objective

1.1 Implementing Agglomerative Clustering on the given dataset

2 Description

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.

The Agglomerative Clustering performs a hierarchical clustering using a bottom up approach: each observation starts in its own cluster, and clusters are successively merged together. The linkage criteria determines the metric used for the merge strategy.

Following are the different linkage criteria:

1. **Ward** minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and in this sense is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.
2. **Maximum or complete linkage** minimizes the maximum distance between observations of pairs of clusters.
3. **Average linkage** minimizes the average of the distances between all observations of pairs of clusters.
4. **Single linkage** minimizes the distance between the closest observations of pairs of clusters.

3 Implementation Guidelines

3.1 Part A

[ ]:
```
import pandas as pd
import numpy as np
import seaborn as sns
```

[ ]:
```
from scipy import ndimage
from scipy.cluster import hierarchy
from scipy.spatial import distance_matrix
from matplotlib import pyplot as plt
from sklearn import manifold, datasets
from sklearn.datasets import make_blobs
```

[ ]:
```
#Generate random dataset using make_blobs function
# Input n_samples, centers, cluster_std parameters in  the make_blobs
```

[ ]:
```
#Plot the scatter plot of the randomly generated data
```

[ ]:
```
#use sklearn to perform Agglomerative Clustering
from sklearn.cluster import AgglomerativeClustering
```

[ ]:
```
#Use different linkage method such as single and complete
```

[ ]:
```
#Get the cluster labels
```

[ ]:
```
#Plot the clusters generated

# Create a figure of size 6 inches by 4 inches.
plt.figure(figsize=(6,4))

# These two lines of code are used to scale the data points down,
# Or else the data points will be scattered very far apart.

# Create a minimum and maximum range of X1.
x_min, x_max = np.min(X1, axis=0), np.max(X1, axis=0)

# Get the average distance for X1.
X1 = (X1 - x_min) / (x_max - x_min)

# This loop displays all of the datapoints.
for i in range(X1.shape[0]):
    # Replace the data points with their respective cluster value
    # (ex. 0) and is color coded with a colormap (plt.cm.spectral)
    plt.text(X1[i, 0], X1[i, 1], 'c',
    color=plt.cm.nipy_spectral(agglom.labels_[i] / 10.),
    fontdict={'weight': 'bold', 'size': 15})
```

```
# Remove the x ticks, y ticks, x and y axis
plt.xticks([])
plt.yticks([])
#plt.axis('off')



# Display the plot of the original data before clustering
plt.scatter(X1[:, 0], X1[:, 1], marker='.')
# Display the plot
plt.show()
```

[ ]:
```
#Plot the dendrogram for the Agglomerative clustering
#use spicy library
```

3.2 Part B

[ ]:
```
#Clustering on Iris Dataset
```

[ ]:
```
#load the dataset
```

[ ]:
```
#Plot the dataset using scatterplot
```

[ ]:
```
# Cluster the datset using Agglomerative clustering
from sklearn.cluster import AgglomerativeClustering
```

[ ]:
```
#Identify cluster labels
```

[ ]:
```
#Plot the clusters
```

[ ]:
```
#Plot the dendrogram for the Agglomerative clustering
#use spicy library
```

4 Exercise

1. What is the difference between Hierarchical Clustering and K-Means Clustering ?
2. Describe the dataset used in this lab exercise.
3. Give insights into the model trained for Iris dataset.
4. Use the algorithm in python to cluster the following 8 examples into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).
   (a) Calculate the distance matrix for the given datapoints using 'Scipy' library.
   (b) Fit the model on distance matrix. Use linkage method as 'single' and 'complete'. Show the cluster labels assigned for each linkage method. Take values of 'distance_threshold'

as 0,1,2 and 3 and identify the change in the number of clusters.

(c) Plot the dendrogram for the data points using linkage method as 'single' and 'complete'.

5 Reference

1. https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py
2. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
3. https://www.youtube.com/watch?v=RdT7bhm1M3E
4. https://medium.com/@MaheshGadakari/hierarchical-agglomerative-clustering-hac-with-single-li
5. http://www.econ.upf.edu/~michael/stanford/maeb7.pdf