SIT384 Cyber security analytics

Pass Task 7.1P: K-Means and Hierarchical Clustering

Task description:

In machine learning, clustering is used for analyzing and grouping data which does not include pre-labelled class or even a class attribute at all. K-Means clustering and hierarchical clustering are all unsupervised learning algorithms.

K- means is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. It is a division of objects into clusters such that each object is in exactly one cluster, not several.

In Hierarchical clustering, clusters have a tree like structure or a parent child relationship. Here, the two most similar clusters are combined together and continue to combine until all objects are in the same cluster.

In this task, you use K-Means and Agglomerative Hierarchical algorithms to cluster a synthetic dataset and compare their difference.

You are given:

- np.random.seed(0)
- make_blobs class with input:
    - n_samples: 200
    - centers: [3,2], [6, 4], [10, 5]
    - cluster_std: 0.9
- KMeans() function with setting: init = "k-means++", n_clusters = 3, n_init = 12
- AgglomerativeClustering() function with setting: n_clusters = 3, linkage = 'average'
- Other settings of your choice

You are asked to:

- plot your created dataset
- plot the two clustering models for your created dataset
- set the K-Mean plot with title "KMeans"
- set the Agglomerative Hierarchical plot with title "Agglomerative Hierarchical"
- calculate distance matrix for Agglomerative Clustering using the input feature matrix (linkage = complete)
- display dendrogram

Sample output as shown in the following figure is **for demonstration purposes only**. Yours might be different from the provided.

KMeans

Agglomerative Clustering
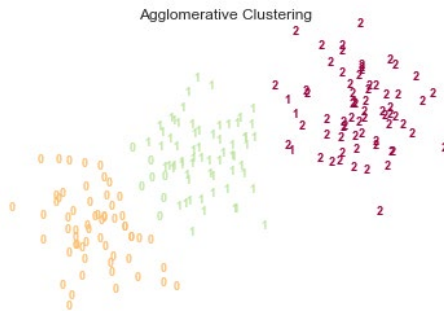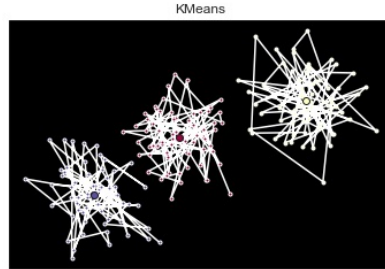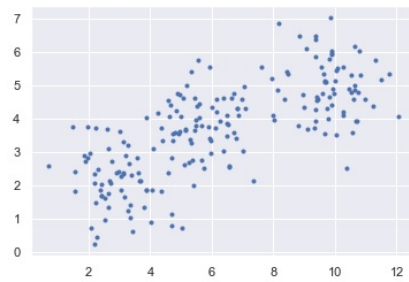

```
[[0.         0.68875871 0.29027041 ... 0.29329872 0.29428699 0.87806224]
 [0.68875871 0.         0.69237564 ... 0.39577438 0.68193381 0.25426008]
 [0.29027041 0.69237564 0.         ... 0.37736577 0.01220405 0.80511311]
 ...
 [0.29329872 0.39577438 0.37736577 ... 0.         0.3710466  0.59870543]
 [0.29428699 0.68193381 0.01220405 ... 0.3710466  0.         0.79324401]
 [0.87806224 0.25426008 0.80511311 ... 0.59870543 0.79324401 0.        ]]
```
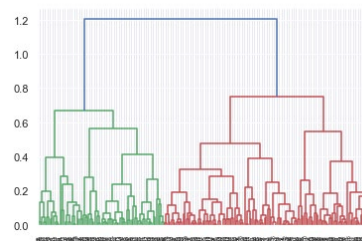

Submission:

Submit the following files to OnTrack:

1. Your program source code (e.g. task7_1.py)
2. A screen shot of your program running

Check the following things before submitting:

1. Add proper comments to your code