Indian Institute of Technology Dharwad

CS214: Artificial Intelligence Laboratory

Lab 12: Hierarchical and DBSCAN clustering

You are given with Iris flower dataset file. The file Iris.csv consists of 150, 4-dimensional data which includes 50 samples from each of the three species of Iris (Iris setose, Iris virginica and Iris versicolor). Column 1 to 4 of the given file are the four features (attributes) that were measured from each sample: the length and the width of the sepals and petals (in centimetres) respectively. Column 5 is the class label (species name) associated with each of the samples of Iris flower. Task here is to reduce the data into 2-dimensional data using PCA and then partition (cluster) the reduced dimensional data using different clustering techniques. While performing the PCA, ignore the fifth column of the data. Use the class information in fifth column only for computing purity score.

1. Data Preprocessing and Dimensionality Reduction.
   (a) Load the Iris dataset from the CSV file.
   (b) Apply PCA on the 4-dimensional data and select the first two directions (eigenvectors corresponding to the two leading eigenvalues) to convert the data into 2-dimensional form. **Exclude** the attribute `Species` when performing PCA.
   (c) Save this 2d dataset as `iris_dataset_2D`. And visualize it.
2. Hierarchical Clustering and Comparison.
   (a) Load the `iris_dataset_2D` reduced datasets.
   (b) Apply Agglomerative Hierarchical Clustering with `n_clusters = 3` and `linkage ="ward"` on the reduced dataset. Assign cluster labels to each data point and visualize the clusters using different colors.
   (c) Compute the purity score of this clustering(used linkage =”ward”).
   (d) Apply Agglomerative Hierarchical Clustering with `n_clusters = 3` and `linkage ="complete"` on the reduced dataset. Assign cluster labels to each data point and visualize the clusters using different colors.
   (e) Compute the purity score of this clustering(used linkage =”complete”).
   (f) Apply Agglomerative Hierarchical Clustering with `n_clusters = 3` and `linkage ="average"` on the reduced dataset. Assign cluster labels to each data point and visualize the clusters using different colors.
   (g) Compute the purity score of this clustering(used linkage =”average”).
   (h) Apply Agglomerative Hierarchical Clustering with `n_clusters = 3` and `linkage ="single"` on the reduced dataset. Assign cluster labels to each data point and visualize the clusters using different colors.
   (i) Compute the purity score of this clustering(used linkage =”single”).
   (j) Plot the dendrogram to visualize the hierarchical clustering process, use the linkage method that yields the highest purity score, apply a threshold of 6, and determine the optimal number of clusters.

   Use `scipy.cluster.hierarchy` functions such as `linkage` and `dendrogram`.

(k) Apply Agglomerative Hierarchical Clustering with `n_clusters = optimal number of clusters` on the reduced dataset.

(l) Assign cluster labels to each data point and visualize the clusters using different colors.

(m) Compute the purity score after clustering.

(n) Compare this purity score with the purity score of obtained by K-means, k-medoids cluster for optimal clusters (performed in previous lab).

3. DBSCAN Clustering.

   (a) Load the `iris_dataset_2D` dataset again.

   (b) Apply DBSCAN clustering on a given dataset using default parameters (`eps=0.5`, `min_samples=5`) and compare its performance with K-Means and Agglomerative Clustering.

      i. Plot the data points with different colours for each cluster for default `eps` and `min_samples`.

      ii. Compute the purity score after examples are assigned to clusters obtained for default `eps` and `min_samples`.

   (c) Apply the DBSCAN clustering using different values for `eps` and `min_samples`. Consider `eps` = 1 and 5, and `min_samples` = 4 ,10. For each combination of `eps` and `min_samples`, observe the number of clusters formed. Here, `eps` (Epsilon) denotes the radius of the boundary from every example, and `min_samples` is the minimum number of examples present inside the boundary with radius of `eps` for an example.

      i. Plot the data points with different colours for each cluster for each combination of `eps` and `min_samples`.

      ii. Compute the purity score after examples are assigned to clusters obtained for each combination of `eps` and `min_samples`.

      iii. Find and display the frequent Species in optimal clusters.

Functions/Code Snippets:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import AgglomerativeClustering
from scipy.cluster.hierarchy import dendrogram, linkage

clustering= AgglomerativeClustering(n_clusters=2, *,
metric='euclidean', memory=None, connectivity=None,
compute_full_tree='auto', linkage='ward',
distance_threshold=None, compute_distances=False).fit(X)

labels = clustering.labels_
# DBSCAN clustering
from sklearn.cluster import DBSCAN
dbscan_model=DBSCAN(eps=1, min_samples=10).fit(train_data)
DBSCAN_predictions = dbscan_model.labels_
```

```python
# Optional: Dendrogram
linked = linkage(X, method='ward')
plt.figure(figsize=(8, 4))
dendrogram(linked)
plt.title("Hierarchical Clustering Dendrogram")
plt.show()

from sklearn.cluster import DBSCAN
from sklearn.datasets import load_iris
from sklearn.preprocessing import StandardScaler

# DBSCAN clustering
db = DBSCAN(eps=0.6, min_samples=5)
labels = db.fit_predict(X_scaled)
```
