Machine Learning Unsupervised Model

20PAIE53

CT3 Mini Project

Inference:

Data Preprocessing:

- Number of rows and columns are 779 and 14 respectively.
- Obtained five-point summary of the attributes/features.
- Gathered information about the columns of the dataset.
- Checked for missing values and none were found.
- Splitted the features and Target variable. Performed standard scaling.

KMeans Clustering:

- The elbow plot suggests that the optimal number of clusters is around 6. This is where the WCSS starts to decrease at a slower rate.
- Applying Kmeans clustering with 6 clusters, the dataset was divided into distinct groups based on the molecular descriptors.
- The scatter plot of 'PiPC09' vs 'PCD' with hues representing the KMeans predicted labels shows the separation of data points into different clusters.

Agglomerative Clustering:

- Silhouette score is used to find the optimal number of clusters. The Silhouette score measures how similar an object is to its clusters.
- Agglomerative clustering with 3 clusters was selected based on silhouette scores.

- The scatter plot of 'PiPC09' vs 'PCD' with hues representing the predicted labels illustrates the clustering achieved.

DBSCAN Clustering:

- DBSCAN clustering was applied with specific parameters(epsilon = 1.5,min_samples = 15).
- This technique identified outliers as points labeled as -1 and grouped similar points into clusters.
- The scatter plot of 'piPC09' vs 'PCD' with hues representing the DBSCAN predicted labels illustrates the grouping of data points.

GMM:

- The GMM is initialized with 4 components.
- The model is fitted to the dataset and the cluster labels are predicted for each data point.
- Line plot is plotted between AIC and BIC. Lower the AIC and BIC values indicate better models.
- Scatter plot shows the structure of the data and how well the model seperated it into different groups.

FCM:

- Silhouette score of 0.123 shows that the clusters are overlapping and not well separated.
- Davies-Bouldin score of 2.737 indicates moderate separation between the clusters.
- Calinski-Harabasz score of 95.527 shows better defined clusters.

PCA:

- The Explained variance for each principal component is obtained and shows the variance captured by each component.

- The Cumulative variance ratio is plotted against the number of principal components to retain sufficient variance.
- Based on the cumulative variance plot, it is observed that around 85% is retained with 4 components and around 96% with 6 components.
- Using Hard Clustering(Agglomerative clustering) on the PCA transformed data with 6 components, a Silhouette score of 0.194 indicates a reasonable degree of separation between the clusters.
- Using Soft Clustering(GMM) in the PCA transformed data with 6 components yields a silhouette score of 0.194, indicating overlapping of clusters with less separation.

SVD:

- The Cumulative variance ratio is plotted against the number of principal components to retain sufficient variance.
- Based on the cumulative variance plot, it is observed that around 85% is retained with 4 components and around 96% with 6 components.
- Using Soft Clustering(GMM) on the SVD transformed data with 3 components, a Silhouette score of 0.194 indicates overlapping of clusters with less separation.
- Using Hard Clustering(Agglomerative) in the SVD transformed data with 3 clusters yields a silhouette score of 0.230, indicating a reasonable degree of separation between the clusters.