Spring 2023 4710 Machine Learning: Assignment 5

Programming elements:

Clustering & Dimensionality reduction

In class programming:

1. Principal Component Analysis

a. Apply PCA on CC dataset.

Table

Description automatically generated

**Explanation:** On applying PCA to the preprocessed data X using PCA() method from the sklearn.decomposition module and Using n\_components parameter to set the components to 2 we create a new dataframe finalDf with 2 components principal component 1 and 2 to store the principal component values.

b. Apply k-means algorithm on the PCA result and report your observation if the silhouette score has improved or not?

Chart, line chart

Description automatically generated

Silhouette Score: 

**Explanation:** Using K-Means algorithm from the sklearn.cluster module to cluster the preprocessed data X**.** Using plt.plot() to plot the clusters. Using the K-Means algorithm from the sklearn.cluster module to cluster the final data finalDf with n\_clusters=3. Using silhouette\_score() function from the sklearn.metrics to calculate the silhouette score for the clustering.

The silhouette score ranges from -1 to 1, where a score closer to 1 indicates that the clusters are well-separated and distinct, while a score closer to -1 indicates that the clusters are overlapping or poorly defined. The calculated silhouette score is 0.572, which is closer to 1, therefore, it can be concluded that clustering is performing well.

c. Perform Scaling+PCA+K-Means and report performance.

Table

Description automatically generated

**Explanation:** Performing PCA on scaled data X\_Scale using PCA(). Using n\_components parameter to set the components to 2 and storing the principal components in new dataframe principalDf1. Using .concat() to concatenate the principalDf1 with the 'TENURE' and displaying the ‘finalDF2’ dataframe.

Chart, line chart

Description automatically generated

Silhouette Score: 

**Explanation:** Using K-Means algorithm from the sklearn.cluster module to cluster the preprocessed data finalDf2 into a range of 1 to 10 clusters. Using plt.plot() to plot the clusters. Using the K-Means algorithm from the sklearn.cluster module to cluster the final data finalDf2 with n\_clusters=3. Using silhouette\_score() function from the sklearn.metrics to calculate the silhouette score for the clustering.

Compared to the previous silhouette score of 0.572, the new score of 0.384 is lower. This suggests that the performance of the clustering algorithm has decreased after performing feature scaling.

2. Use pd\_speech\_features.csv

a. Perform Scaling

scaler = StandardScaler() #performing feature selection # feature scaling to have zero mean and unit variance.

X\_Scale = scaler.fit\_transform(X)

b. Apply PCA (k=3)

Table

Description automatically generated

**Explanation:** Performing PCA on scaled data X\_Scale using PCA(). Using n\_components parameter to set the components to 3 and storing the principal components in new dataframe principalDf2. Using .concat() to concatenate the principalDf2 with the 'CLASS' and displaying the ‘finalDF3’ dataframe.

c. Use SVM to report performance

Table

Description automatically generated

**Explanation:** Training and predicting SVM model on our data set using the SVC() method and then and then predicting the target values on the test set. Then printing the classification report, confusion matrix, and the accuracy score of the model.

The SVM model achieved an overall accuracy of 0.7489. However, the precision, recall and F1-score for class 0 is 0.00, which means the model is not able to predict class 0 correctly. This could be due to class imbalance in the dataset, or the model may not be well-suited for this particular problem.

3. Apply Linear Discriminant Analysis (LDA) on Iris.csv dataset to reduce dimensionality of data tok=2.

Graphical user interface, text, table

Description automatically generated

**Explanation:** Performing LDA on the datataset with n\_components= 2 to create two new features LD1 and LD2. Using .concat() to combine the new features LD1 and LD2 with ‘Species’ into a single dataset.

4. Briefly identify the difference between PCA and LDA

Ans: The main difference between PCA and LDA is that PCA is an unsupervised technique that reduces the dimensionality of the dataset by transforming the original variables into a new set of uncorrelated variables, called principal components. The objective of PCA is to maximize the variance of the data along the principal components. On the other hand, LDA is a supervised technique that finds a linear combination of features that maximizes the separation between classes. The objective of LDA is to maximize the between-class variance and minimize the within-class variance.