IML
Report
Lab -7

Question 1:

In this Question, One has to Classify the data given on Flower parts . There are 4 columns depicting the length of different part of flower namely,sepal length ,sepal width ,petal length and petal width and the target cl=olumn is the classification of flower in one of the 3 categories-{Iris Setosa,Iris versicolor,Iris Virginica}.
We are asked to first preprocess the data and split the dataset in a certain ratio. Then we have to apply a classification algorithm taught in the previous class other than PCA itself and find the accuracy we got predicting the accuracy of the classifier by putting x_test data in it. I have used logistic Regression as the classifier and got accuracy=`83.33333333333334%`*.*
Now we apply Principal component analysis Analysis to actually predict the accuracy .
About Principal component analysis-

- *The Principal Component Analysis is* ***a popular unsupervised learning technique for reducing the dimensionality of data****. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D.*

I have imported library -`from sklearn.decomposition import PCA`

Using the top few principal components to project the data to a certain chosen dimensions,we used the following line.

*.I have taken this line -`pca = PCA(0.95)` **to get 95 percent of essential features that are gonna contribute to predicting the accuracy.***

The best part of the PCA method is that in just one line it picks up components that are worth taking in a classification algorithm. The dimension of the training set is reduced to 2 from 4.

Let us walkthrough what these lines means-

- `pca.explained_variance_ratio_`*-The explained variance ratio is **the percentage of variance that is attributed by each of the selected components**. Ideally, you would choose the number of components to include in your model by adding the explained variance ratio of each component until you reach a total of around 0.8 or 80% to avoid overfitting. My variance ratio come out to be-*`[0.93042172, 0.04694468]`
- `Pca.n_components_-` *Here the pca algorithm already chose it as 2. It represent the dimension too which training set is reduced to.*
- `X_pca`*-This gives us the refined data set that the PCA algorithm has taken out from the original data set which will help in increasing the accuracy and reduce the time of execution.*
- `X_train_pca, X_test_pca, y_train, y_test = train_test_split(X_pca, y, test_size=0.2, random_state=30)`*-Splitting the dataset with X_pca as the data st chosen this time.*

- *Now again using the logistic regression classification method to predict the accuracy.*

```
model = LogisticRegression(max_iter=1000)
model.fit(X_train_pca, y_train)
model.score(X_test_pca, y_test)
```

- ★ *After applying the PCA algorithm the final accuracy comes out to be 0.9. (which is better than accuracy which we got before.)*

Then I have taken different dimensions and generated refined dataset X_pca several time to find the best accuracy that i can get.

For ,

- *N_component =3 ,Accuracy=*`0.8666666666666667`
- *N_component =1,Accuracy=*`0.9[Taken(pca = PCA(0.8)]`
- *N_component =4 ,Accuracy=*`0.9`

Google Collab Link-https://colab.research.google.com/drive/1NC1DeZWll6iUsTiyvywN42S84xMgkz74#scrollTo=FcAoy_TQ2xSG

—------------------------------------Thank You—---------------------------