91.350.202/580.202 HW #4 Vishal Gadiraju 01672049

Principal Component Analysis (PCA):

PCA finds the directions of maximum variance in a high dimensional data and projects it onto a smaller dimensional subspace while retaining most of the information.

The IRIS dataset contains measurements for 150 iris flowers from three different species.

The three classes in the dataset are:

1. Iris-setosa (n=50)
2. Iris-versicolor (n=50)
3. Iris-virginica (n=50)

And the four features of the Iris dataset are:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

For Principal Components Analysis (PCA),we require eigenvalue analysis of the covariance matrix in order to find the linear combinations of the data variables with greatest variance. These linear combinations are nothing but Principal Components.

We have performed PCA on the Iris dataset and projected it into 2 components.

The eigenvector corresponding to the largest eigenvalue of the covariance matrix is the first-Principal Component (PC) whereas the next largest eigenvalue of covariance matrix corresponds to the second-Principal Component (PC).

RESULTS/ANALYSIS:

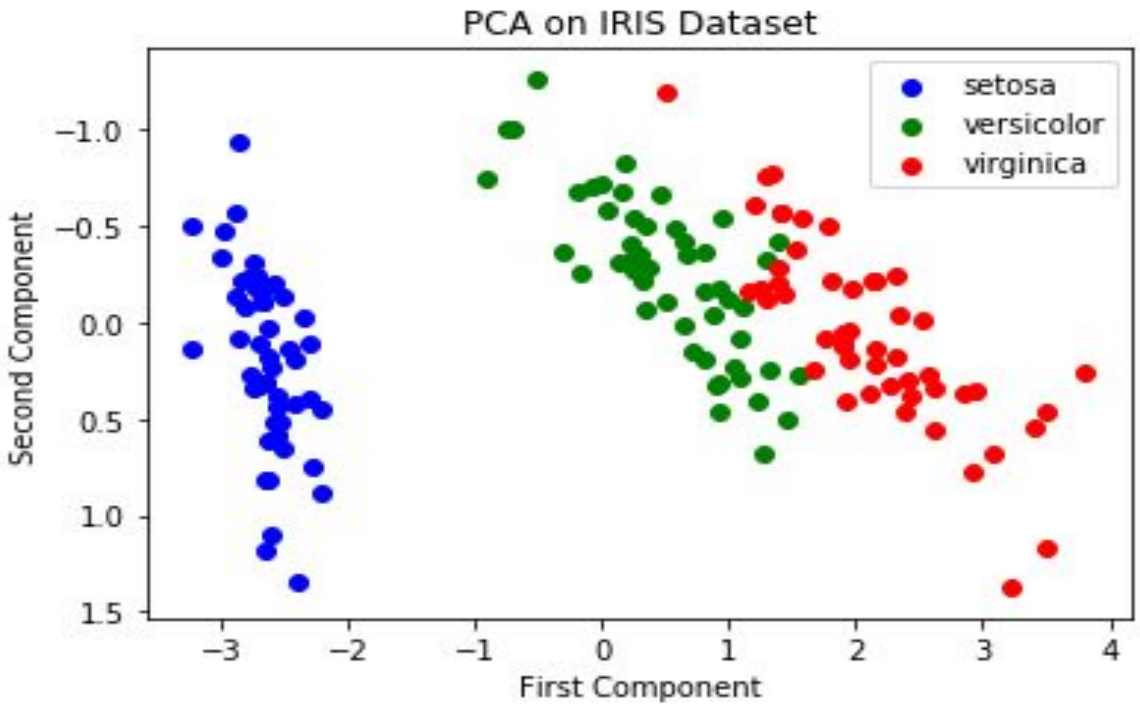
	First Component	Second Component
Explained Variance Ratio	0.92461621	0.05301557

Explained Variance	4.19667516	0.24062861

The first Principal Component accounts for nearly 92.5% of the variation in the data whereas the second Principal component accounts for only 5% of the variation in data.
The proportion of the variation helps in summarizing the data.
Explained Variance Ratio refers to the eigenvalues of the correlation/covariance matrix. It gives the "spread" of the data.

Factor Loadings:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Component 1	0.36158968	-0.08226889	0.85657211	0.35884393
Component 2	0.65653988	0.72971237	-0.1757674	-0.07470647



Comparison with a Published Result:

In the given article [1] , the dataset is reduced into 4 components and the resulting variance is computed as 72.77% for the first component compared to the 92.5% variance for my result.

The second principal component holds 23.03% of variance compared to the 5% of my analysis.

However the PCA plot and Variance ratio appears similar in both cases.

CODE

The code implemented in Python and provided below:

```
import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn.decomposition import PCA

iris = datasets.load_iris()

X = iris.data
y = iris.target
target_names = iris.target_names
pca = PCA(n_components=2)
X_r = pca.fit(X).transform(X)
plt.figure(figsize=(6, 4))
# Percentage of variance explained for each components
print('explained variance ratio (first two components): %s'
      % str(pca.explained_variance_ratio_))
print('explained variance (first two components): %s'
      % str(pca.explained_variance_))
plt.figure()
colors = ['blue', 'green', 'red']
lw = 1
for i,color, target_name in zip([0, 1, 2],colors, target_names):
    plt.scatter(X_r[y == i, 0], X_r[y == i, 1], color=color,lw=lw,label=target_name)
plt.legend(loc='best')
plt.gca().invert_yaxis()
plt.title('PCA on IRIS Dataset')
plt.xlabel('First Component')
plt.ylabel('Second Component')
plt.figure()
plt.show()
```

REFERENCES:

1.http://sebastianraschka.com/Articles/2015_pca_in_3_steps.html#exploratory-visualization

2.http://www4.ncsu.edu/~slrace/LinearAlgebra2016/RChapters/PCA.pdf