Machine Learning
Experiment 4

SAP ID : 60004200139 Name : Riya Bihani
Division : B Batch : B1

AIM

To implement PCA.

THEORY

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of a large dataset. It is a commonly used method in machine learning, data science, and other fields that deal with large datasets. This method was introduced by Karl Pearson. It works on a condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.

PCA works by identifying patterns in the data and then creating new variables that capture as much of the variation in the data as possible. These new variables, known as principal components, are linear combinations of the original variables in the dataset. It reduces the dimensionality of a data set by finding this new set of variables, smaller than the original set of variables, retains most of the sample's information and useful for the compression and classification of data.

The PCA algorithm is based on some mathematical concepts such as -

- Variance and Covariance
- Eigenvalues and Eigen factors

PCA can be used for a variety of purposes, including data visualization, feature selection, and data compression. In data visualization, PCA can be used to plot high-dimensional data in two or three dimensions, making it easier to interpret. In feature selection, PCA can be used to identify the most important variables in a dataset. In data compression, PCA can be used to reduce the size of a dataset without losing important information.

<u>Applications</u>

- PCA is mainly used as the dimensionality reduction technique in various AI applications such as computer vision, image compression, etc.
- It can also be used for finding hidden patterns if data has high dimensions. Some fields where PCA is used are Finance, data mining, Psychology, etc.

CODE

```
import pandas as pd
import numpy as np
from numpy.linalg import eig

df = pd.read_csv('Salary_Data.csv')

x = df.iloc[ : , 0]
y = df.iloc[ : , 1]

mew1 = round((sum(x)/len(x)), 2)
mew2 = round((sum(y)/len(y)), 2)

mean = [mew1, mew2]

print(mean)

x_new = list(map(lambda x1 : round((x1 - mew1), 2), x))
y_new = list(map(lambda y1 : round((y1 - mew2), 2), y))

cov = np.array([[0, 0], [0, 0]])

for i in range(0, len(x_new)):
  temp = np.array([[float(x_new[i])], [float(y_new[i])]])
  temp1 = temp.transpose()
  temp2 = temp@temp1
  cov = cov + temp2

cov = cov/len(x_new)

eigenvalues, eigenvectors = eig(cov)

feat = eigenvectors[ : , 1]

new_values = []

for i in range(0 , len(x_new)):
  temp = np.array([[float(x_new[i])], [float(y_new[i])]])
  temp1 = feat @ temp
  new_values.append(temp1.tolist())
```

print(new_values)

OUTPUT

```
[5.31, 76003.0]
[[36660.000238359156], [29798.000253290018], [38272.00018958758], [32478.000168662715], [36112.00012977673], [19361.00014477929], [15853.000152639528],
```

CONCLUSION

Thus, we have successfully implemented PCA Algorithm.