M.S.SANJAY

15BCE0517

L7+L8

MACHINE LEARNING LAB

EXPERIMENT 10

PRINCIPAL COMPONENT ANALYSIS

**DATASET**:

|  |  |
| --- | --- |
| |  | | --- | | Blood Transfusion Service Center Data Set | |

Given is the variable name, variable type, the measurement unit and a brief description. The "Blood Transfusion Service Center" is a classification problem. The order of this listing corresponds to the order of numerals along the rows of the database.   
  
R (Recency - months since last donation),   
F (Frequency - total number of donation),   
M (Monetary - total blood donated in c.c.),   
T (Time - months since first donation), and   
a binary variable representing whether he/she donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood).

**CODE:**

import numpy as np

from sklearn.decomposition import PCA

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.preprocessing import scale

#Load data set

data = pd.read\_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/blood-transfusion/transfusion.data',

sep= ',', header = 0 )

print ("Dataset Lenght: ", len(data))

print ("Dataset Shape: ", data.shape)

print ("Dataset: ",data.head())

#convert it to numpy arrays

X=data.values

#Scaling the values

X = scale(data)

pca = PCA(n\_components=5)

pca.fit(X)

#The amount of variance that each PC explains

var= pca.explained\_variance\_ratio\_

print("Variance of each component")

print(var)

#Cumulative Variance explains

var1=np.cumsum(np.round(pca.explained\_variance\_ratio\_, decimals=4)\*100)

print("Cumulative variance")

print(var1)

plt.plot(var1)

plt.show()

#first 3 pca cumulatively give 100 percent vriance.so 3 variables taken

pca = PCA(n\_components=3)

pca.fit(X)

X1=pca.fit\_transform(X)

print("Fitting and dimensionality reduction")

print(X1)

**OUTPUT**:

Dataset Lenght: 748

Dataset Shape: (748, 5)

Dataset: Recency (months) Frequency (times) Monetary (c.c. blood) Time (months) \

0 2 50 12500 98

1 0 13 3250 28

2 1 16 4000 35

3 2 20 5000 45

4 1 24 6000 77

whether he/she donated blood in March 2007

0 1

1 1

2 1

3 1

4 0

Variance of each component

[ 5.18177571e-01 2.67784120e-01 1.44473099e-01 6.95652095e-02

1.53187729e-31]

Cumulative variance

[ 51.82 78.6 93.05 100.01 100.01]

Fitting and dimensionality reduction

[[ 10.89631782 -0.6148538 -0.16472073]

[ 1.89749972 -1.98503362 0.41542014]

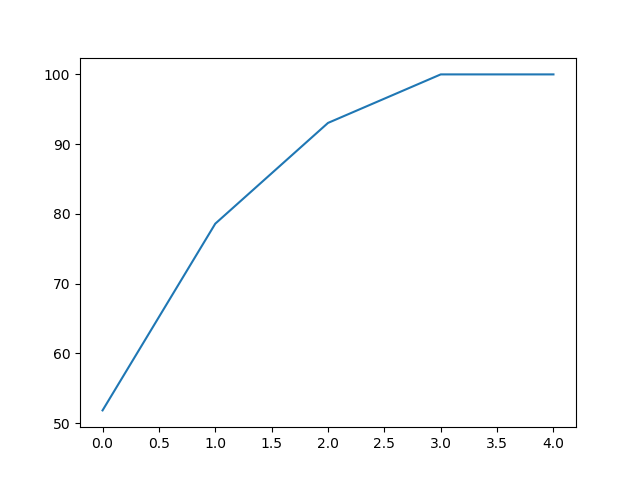
[ 2.63976772 -1.78093093 0.4377017 ]

...,

[ -0.2956787 1.93867757 0.77066109]

[ -1.39634638 2.91460984 2.01147407]

[ -1.26633931 6.26131324 4.68891592]]



Note: in the graph, numbering starts from 0 instead of 1 and so 2 means cumulative variance till 3rd PCA.

**EXPLANATION:**

As there are 5 attributes, initially all 5 variables are taken to find variance of all attributes.From the graph and cumulative variance it is evident that first 3 PCA components are enough to represent whole of the dataset upto 93% accuracy.So in the second step first 3 PCA components are taken.

SCREENSHOT:

