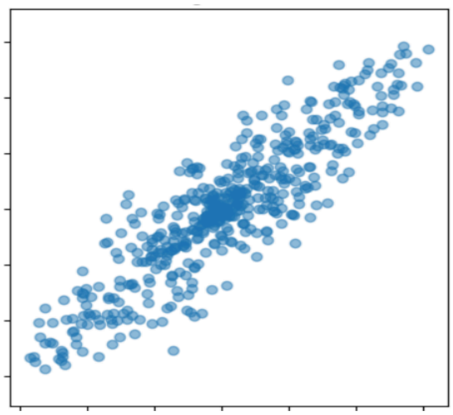
Information Retrieval CSF469

Lab Session - 7
Date - 10/04/2024
Marks: 10

Objective:
To perform different tasks related to PCA (Principal Component Analysis) to understand the working process of PCA and its use cases.

Dataset: Three different types of data have been provided.

TASK 1:



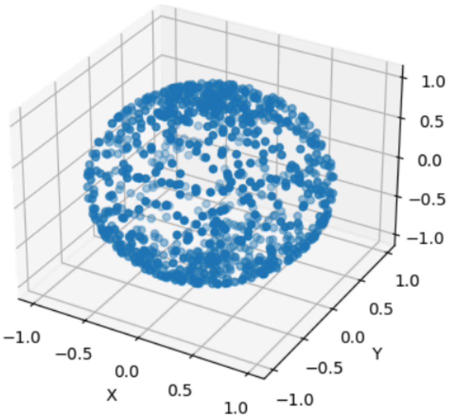
1. Generate data in the form as shown in the figure (code provided).
2. Implement in PCA from the sklearn library. Compute the explained_variance_ratio.
3. Now, implement the code for PCA manually and replicate the results of step 2 above by following the below steps:
    a. Normalise the dataset by subtracting the mean from each data point.
    b. Calculate the covariance matrix.
    c. Compute the eigenvalues and eigenvectors of the covariance matrix.
    d. The eigenvalues will correspond to the variance along each principal component, and the eigenvectors represent the direction of principal components.
    e. Compute the explained variance ratio using the formula.

    $$Explained\ variance\ by\ eigenvalue\ i = \frac{Eigenvalue\ i}{Sum\ of\ all\ eigenvalues}$$

    This gives the ratio of each eigenvalue to the cumulative eigenvalues. I.e. percentage contribution of variance in data of each principal component.

TASK 2:

1. Prepare a spherical dataset (code provided)
2. Repeat the PCA implementation on this data.
3. Compare the results with the previous dataset. (Hint: All the principal components will contribute equally.)



TASK 3:

PCA for visualisation.

The provided dataset is a toy dataset with 198 data points belonging to two classes. Each data point has a feature of size 768. Do the following operations:

a. Find the number of most important dimensions out of 768. (Hint: Plot number of dimensions vs cumulative explained variance)
b. Reduce the number of dimensions from 768 to 2 and print explained_variance_ratio.
c. Plot the newly transformed data and see the different clusters in the dataset.