

School of Computer Science Engineering and Technology
Assignment-06

Course- B.Tech
Course Code-
Year- 2023-2024
Date- 14-08-2023

Type- Core
Course Name- Statistical Machine Learning
Semester- odd
Batch- AIML-B, D

1 Implement Dimensionality Reduction using Principal Component Analysis (PCA)

Part 1 – Import the required Python, Pandas, Matplotlib, Seaborn packages

Problem: you have a multidimensional set of data (such as a set of hidden unit activations) and you want to see which points are closest to others. PCA allows you to identify the dimensions of greatest variance, to the dimensions of least variance. PCA1 has greatest variance.

Step 1 –Download Iris Flower Dataset and verify its features

The dataset would be obtained from UCI Machine Learning Repository. To do that, you can right-click on the link below and save a copy of the dataset to your local drive. Iris Flower Data set link: "<https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data>"

Step 2 –Preview the Iris Flower Dataset

Verify the Iris dataset and its features to analyze the principal components.

Part 2 : Perform Scaling on the Data

This means that we need to center and scale the data. In this way the average value of each record would be 0 and the variance for each record would be 1.

To scale our data, we would use StandardScaler which is available in sklearn.

Note that we are only going to scale the features and not the target. So to do this, we

- first import StandardScaler
- separate the features from the target
- scale the features

Step 1 –Perform PCA model

Load the PCA from `sklearn` module to implement the principal components into dataframe.

Step 2 –Combine target and principal components

Remember that the original data has five columns: four features and one target column. Now after performing PCA, we have just two columns for the features. The target dataset y was not touched. Therefore, we attached back the target column to the new set of principal components.

Step 3 : Visualize Iris Flower PCA using Scatter Plots

Generate a scatter plot of the PCA new dataset to see how it differs from original data.

Part 3: Inference and Variance to PCA Analysis

Apply *explained_variance_ratio_* to predict the PCA for PC1 and PC2 variance