CEEL 82B, Data Science, 2022

Lab 2: Implementation of Principal Component Analysis

**Objective:** To implement and use Principal Component Analysis using Python Platform

We can use PCA to calculate a projection of a dataset and select a number of dimensions or principal components of the projection to use as input to a model.

Case studies to consider:

You can consider any dataset from following link

<https://archive.ics.uci.edu/ml/datasets.php>

Refer the paper placed on edmodo named ‘Principal\_Component\_Analysis.pdf’ showing famous example of PCA application on *Painted turtle dataset* . This reference will help you apply PCA to various dataset presentated on above given link

**Introduction:**

Principal Component Analysis (PCA) is an exploratory approach to reduce the data set's dimensionality to 2D or 3D, used in exploratory data analysis for making predictive models. Principal Component Analysis is a linear transformation of data set that defines a new coordinate rule such that:

1. The highest variance by any projection of the data set appears to laze on the first axis.
2. The second biggest variance on the second axis, and so on.

We can use principal component analysis (PCA) for the following purposes:

1. To reduce the number of dimensions in the dataset.
2. To find patterns in the high-dimensional dataset
3. To visualize the data of high dimensionality
4. To ignore noise
5. To improve classification
6. To gets a compact description
7. To captures as much of the original variance in the data as possible

In summary, we can define principal component analysis (PCA) as the transformation of any high number of variables into a smaller number of uncorrelated variables called principal components (PCs), developed to capture as much of the data’s variance as possible.

Mathematically the main objective of PCA is to:

1. Find an orthonormal basis for the data.
2. Sort dimensions in the order of importance.
3. Discard the low significance dimensions.
4. Focus on uncorrelated and Gaussian components.

Steps involved in PCA

1. Standardize the PCA.
2. Calculate the covariance matrix.
3. Find the eigenvalues and eigenvectors for the covariance matrix.
4. Plot the vectors on the scaled data.

PCA are used to identify an underlying structure or pattern beneath a set of multivariate data.

**Outcomes:**

1. To learn how PCA is to identify an underlying structure or pattern beneath a set of multivariate data.
2. To interpret the results of evaluated components

**System Requirements:** Linux/MaC/Windows OS with Anconda platform with *Pandas, numpy, scipy, matplotlib, seaborn* and *scikit-learn* ML library.

**General Steps:**

1. Load the dataset (Use pandas )
2. Data Preprocessing (Handling NA values)
3. Exploratory Data Analysis (understanding the relationships between the variables with help of plot, scatter-plot, enery-plot etc) Use matplotlib (What variables to include, sample size)
4. Apply Nomalization and Generate Covariance Matrix
5. Apply PCA
6. Build the model, summarize it and evaluate the model and Summarize
7. Tuning the model

**Conclusion:** (Write in own words)

**Note:** Complete your write-up with conclusion and upload your outputs on your github account