School of Computer Science Engineering and Technology

Course- B. Tech	Type- Core
Course Code- CSET228	Course Name: DMPM
Year- 2024	Semester- odd
Date: 20-01-2024	Batch- ALL

Lab Assignment 3.1.1

CO-Mapping

Exp. No.	Name	CO1	CO2	CO3
1.		✓		
2.	PCA		✓	✓

Objective:

- Students will be able to learn data vectorization.
- To understand dimension reduction using the concept of principal component analysis

1. Read the ‘Employee_list’ file using pandas and implement following questions.
Write a program to print the percentages for Engineer vs. Doctor having total salary? Call this method Compare profession and return the result as a DataFrame with a row for Engineer and a row for doctor with the column "% of total Salary". The data is not arranged properly. Arrange the data in ascending order of employees age and save the details of first 5 younger employees in New_Data.csv. (30 minutes)

Introduction:

PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

We do PCA analysis using the following steps.

- Standardize the range of continuous initial variables
- Compute the covariance matrix to identify correlations
- Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
- Create a feature vector to decide which principal components to keep
- Recast the data along the principal components axes

Collect the data of breast cancer from the following link.

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

About Dataset

The data contains the following description,

Repository: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image,

resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

2. Implement PCA to reduce the feature dimension of the above-mentioned dataset. Follow the following steps.

Data Pre-processing (30)

- Import the necessary Libraries
- Read the dataset
- Check the shape of the dataset
- Print the first 5 rows of the dataset
- Check the presence of missing values. Handle it if present
- Selecting the feature i.e., Identify the Independent variables and perform the extraction. (Hint: Remove the Target Column as it is Unsupervised Learning Problem).

Finding the optimal number of features using the PCA method (30)

- Standardize the data using StandardScalar or MinMaxScaler
- Set the n-components
- Fit the scaled data to PCA algorithm
- Display the scatter plot of the reduced feature with respect to the target class.

Training The model using regression algorithm (30)

- Apply Any Classification problem and find the accuracy
- Calculate precision, recall for above dataset.

Suggested Platform: Python: Jupyter or Google Colab Notebook.