School of Computer Science Engineering and Technology

Course- B. Tech	Type- Core
Course Code- CSET301	Course Name: AIML
Year- 2022	Semester- odd
Date: 21-11-2022	Batch- ALL

Lab Assignment 12.1.2

CO-Mapping

Exp. No.	Name	CO1	CO2	CO3
12.1.2	PCA		✓	✓

Objective: To understand dimension reduction using the concept of principal component analysis

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

Voice Gender

Gender Recognition by Voice and Speech Analysis

This database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech. The dataset consists of 3,168 recorded voice samples, collected from male and female speakers. The voice samples are pre-processed by acoustic analysis in R using the seewave and tuneR packages, with an analyzed frequency range of 0hz-280hz (human vocal range).

Collect the data set from the following link,

https://www.kaggle.com/datasets/primaryobjects/voicegender

The following acoustic properties of each voice are measured and included within the CSV:

meanfreq: mean frequency (in kHz)

sd: standard deviation of frequency

median: median frequency (in kHz)

Q25: first quantile (in kHz)

Q75: third quantile (in kHz)

IQR: interquantile range (in kHz)

skew: skewness (see note in specprop description)

kurt: kurtosis (see note in specprop description)

sp.ent: spectral entropy

sfm: spectral flatness

mode: mode frequency

centroid: frequency centroid (see specprop)

peakf: peak frequency (frequency with highest energy)

meanfun: average of fundamental frequency measured across acoustic signal

minfun: minimum fundamental frequency measured across acoustic signal

maxfun: maximum fundamental frequency measured across acoustic signal

meandom: average of dominant frequency measured across acoustic signal

mindom: minimum of dominant frequency measured across acoustic signal

maxdom: maximum of dominant frequency measured across acoustic signal

dfrange: range of dominant frequency measured across acoustic signal

modindx: modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range

label: male or female

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 (30)

- Apply Decision tree and find the accuracy
- Calculate precision, recall for above dataset.

Suggested Platform: Python: Jupyter or Google Colab Notebook**.**