CS524 : Data Mining Assignment #1 Q1

Problem Statement : Apply PCA from scratch on Libras movement dataset and report eigenvalues and vectors.

Submitted By : Aman Bilaiya - 2018CSB1069

★ INTRODUCTION

Principal Component Analysis (PCA) is a dimensionality reduction technique that looks for a n-dimensional basis (aka # PCA components) that best captures the variance in the data. The direction with the largest projected variance is called the first principal component. The orthogonal direction that captures the second largest projected variance is called the second principal component, and so on. Also the direction that maximizes the variance is also the one that minimizes the mean squared error. PCA is a very useful data processing technique used in mostly all ML applications for efficient computation. Eigenvalues and vectors are the main fundamentals behind PCA.

★ DATASET DESCRIPTION

Libras Movement Dataset Link

- "movement_libras.data" is the parent data file and rest files are subsets of it. We will apply PCA to this dataset file. There are **90** features [say x1 to x90].

X_col9	X_col10	...	X_col82	X_col83	X_col84	X_col85	X_col86	X_col87	X_col88	X_col89	X_col90	Y_col
0.68472	0.57870	...	0.33796	0.15087	0.34954	0.13926	0.34491	0.13153	0.35185	0.12766	0.38194	12
0.41006	0.80324	...	0.27315	0.24565	0.27778	0.24565	0.28009	0.24758	0.27778	0.24758	0.27778	8
0.65764	0.78241	...	0.55093	0.57253	0.53935	0.56286	0.53241	0.55126	0.52546	0.54159	0.52083	15
0.40039	0.72917	...	0.18056	0.48162	0.17593	0.47776	0.17361	0.46615	0.16435	0.43907	0.15741	11
0.23017	0.52315	...	0.72685	0.46422	0.72454	0.47582	0.72222	0.48743	0.71991	0.50290	0.71759	14

If we train this dataset it will take a huge time as there are 90 dimensions. So we only want top k-important features which are contributing to the most of the variance in the data and discard other features, without much effect in accuracy. Also resulting in reduced training time.

- There are **15** unique target classes labelled as {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}. Below data info is generated using "dataInfo.ipynb" code.

movement_libras	360 rows * 91 cols
movement_libras_1	45 rows * 91 cols
movement_libras_5	90 rows * 91 cols
movement_libras_8	135 rows * 91 cols
movement_libras_9	45 rows * 91 cols
movement_libras_10	270 rows * 91 cols

★ CODING PART [Principal_Component_Analysis.ipynb]

- Implemented PCA from scratch using numpy & pandas (for maths operations & data processing) and matplotlib & seaborn (for plotting graphs).
- Steps involved as part of PCA implementation :-
  - Data processing
    - Preparing X and Y
    - Standardization of X data
  - Computing covariance matrix
  - Finding eigenvalues and eigenvectors from covariance matrix
  - Finding Principal Components with 2-components [Note #components can be changed in the code, I have used 2 for visualization ease]
    - Computing dimensionally reduced X
    - Plotting reduced dimension plot

Note : Refer "readme" file for instructions on how to run code.

★ RESULTS

Showing here a subset of eigenvalues & eigenvectors obtained. Also I have compared these results with PCA of sklearn lib and found them to be nearly similar.

NOTE : Kindly refer to the "Principal_Component_Analysis.ipynb" notebook for complete results as it is not possible to show here as the results size is large.

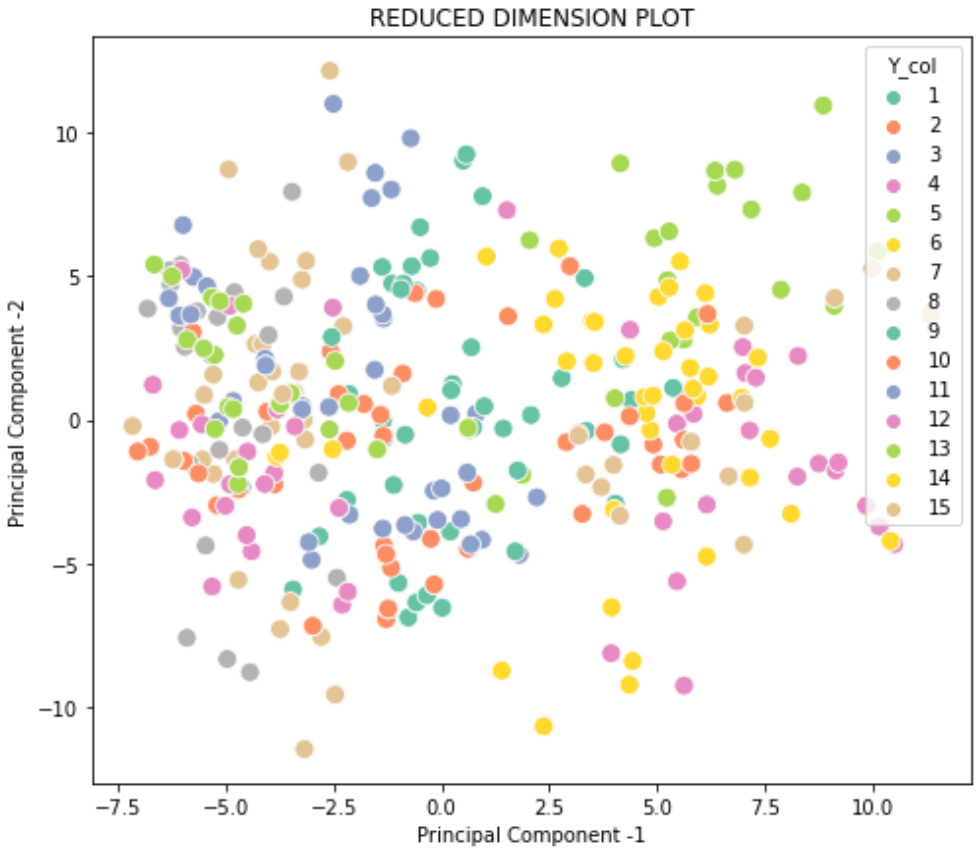
EIGENVALUES → [21.11374555 18.9491506 16.12252029 11.6711225 5.51957859]

EIGEN VECTORS → array ([[-0.13100049, 0.06285489], [-0.16883043, -0.05836495], [-0.13101386, 0.0633448], [-0.16894926, -0.05771133], [-0.13080585, 0.06406005]])

PC TABLE →

	Principal Component -1	Principal Component -2	Y_col
61	0.453196	-3.456530	3
189	-4.964112	-8.332198	8
249	-5.748476	4.966405	11
95	5.460189	-5.624988	4
97	2.044911	6.247680	5

REDUCED DIMENSION PLOT → with 2 PC's



REFERENCES :-

[1] Chapter 7 : Dimensionality Reduction, Mohammed J.Zaki, “Data Mining and Analysis: Fundamental Concepts and Algorithms”

[2] Numpy, seaborn, pandas and matplotlib modules documentation

[3] Lecture Slides