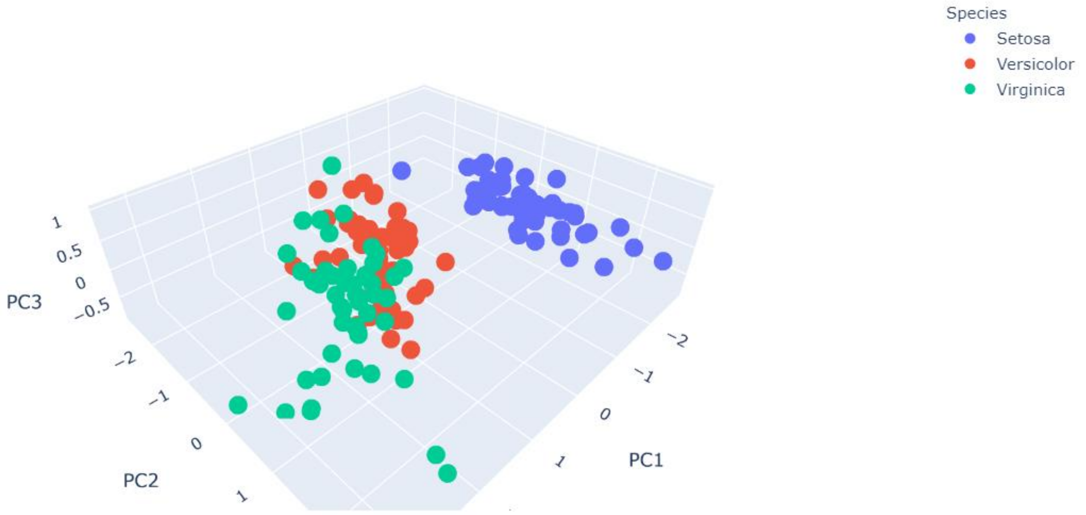
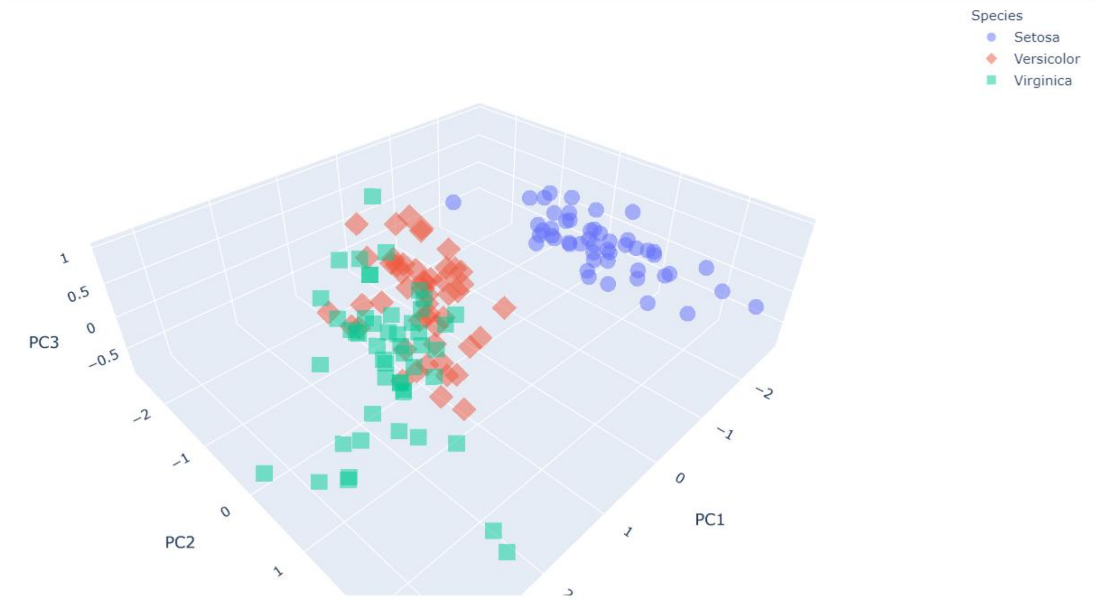
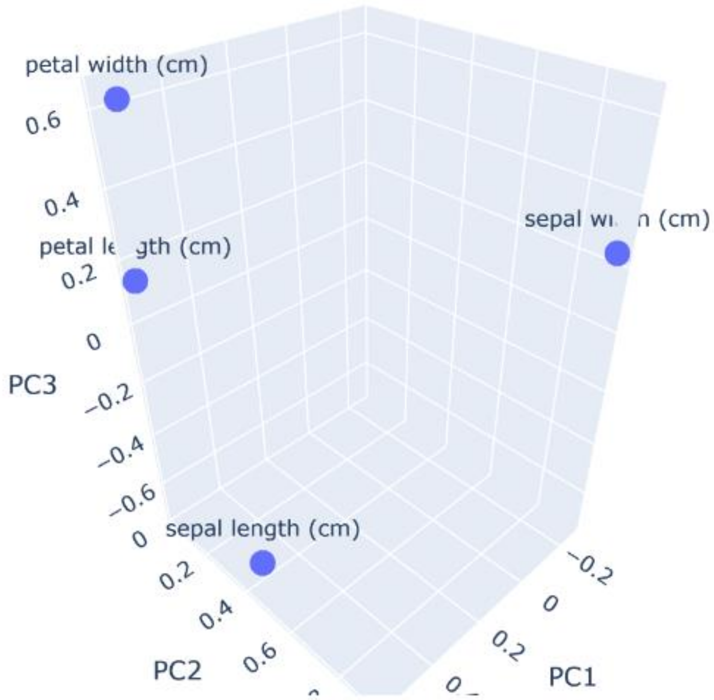
Project Name: Dimensionality Reduction with PCA and Plotly

Table of Content

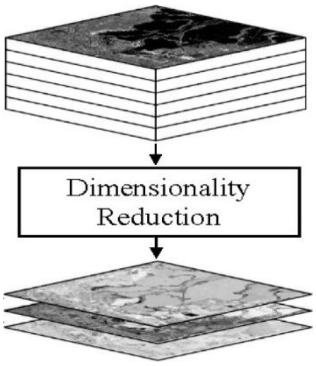
Demo







Overview



This is a Dimensionality Reduction means transformation of data from high-dimensional space into a low-dimensional space using PCA algorithm and Plotly.

This repository contains the code for Dimensionality Reduction using python’s various libraries.
It used Numpy, Pandas, Sklearn libraries.
These libraries help to perform individually one particular transformation.
Numpy is used for working with arrays. It stands for Numerical Python.
Pandas objects rely heavily on Numpy objects.
Sklearn has 100 to 200 models.
The purpose of creating this repository is space is less required, less computation needed.
These python libraries raised knowledge to hands on with this algorithm.
It leads to growth in my ML repository.
The screenshots will help you to understand flow of output.

Motivation

The reason behind building this project is, these algorithm captures essence of data. Simple logic is that even we as humans understands quickly if taught in simple words rather than complex words and ambiguous sentence structure. Similarly, reducing its dimension presents data in compact form. Another reason is that while model building, some algorithms do not perform well when we have large dimensions. Major necessity to reduce dimension is, space required to store data is reduced as the number of dimensions comes down and another is less dimensions lead to less computation and training time. At the same time, it maintains most of the key information. It reduces overfitting. As this is import for companies because tremendous increase in the way sensors are being used in the industry and these sensors continuously record data and because of respective errors in recording, problem arises of high unwanted dimensions and so the need of dimension reduction. For example, bike riders' movements get measured by GPS, set top box collects data about program preferences, casinos capturing data using cameras.

Technical Aspect

Sklearn is known as scikit learn. It provides many ML libraries and algorithms for it. It provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.

Numpy contains a multi-dimensional array and matrix data structures. It works with the numerical data. Numpy is faster because is densely packed in memory due to its homogeneous type. It also frees the memory faster.

Pandas module mainly works with the tabular data. It contains Data Frame and Series. Pandas is 18 to 20 times slower than Numpy. Pandas is seriously a game changer when it comes to cleaning, transforming, manipulating and analyzing data.

PCA is a technique used as a data preparation technique to create a projection of a dataset prior to fitting a model. First, standardize, second, covariance matrix computation performs, third compute eigenvectors and eigenvalues of covariance matrix to identify the principal components.

Principal Component Analysis (PCA) is an unsupervised linear transformation technique that is widely used across different fields, most prominently for feature extraction and dimensionality reduction. Other popular applications of PCA include exploratory data analyses and de-noising of signals in stock market trading, and the analysis of genome data and gene expression levels in the field of bioinformatics.

Reason for selecting PCA is because PCA helps us to identify patterns in data based on the correlation between features. That is, in a nutshell, PCA aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one.

Note that the PCA directions are highly sensitive to data scaling, and we need to standardize the features *prior* to PCA if the features were measured on different scales and we want to assign equal importance to all features.

Before looking at the PCA algorithm for dimensionality reduction in more detail, let's summarize the approach in a few simple steps:

1. Standardize the *d*-dimensional dataset.
2. Construct the covariance matrix.
3. Decompose the covariance matrix into its eigenvectors and eigenvalues.
4. Sort the eigenvalues by decreasing order to rank the corresponding eigenvectors.
5. Select *k* eigenvectors which correspond to the *k* largest eigenvalues, where *k* is the dimensionality of the new feature subspace ($k \le d$).
6. Construct a projection matrix *W* from the “top” *k* eigenvectors.
7. Transform the *d*-dimensional input dataset *X* using the projection matrix *W* to obtain the new *k*-dimensional feature subspace.

Other techniques for dimensionality reduction are Linear Discriminant Analysis (LDA)
and Kernel PCA (used for non-linearly separable data).

Plotly is graphing libray makes interactive, publication-quality graphs. Plotly’s graphs are hosted using an online web service, so you will first have to setup a free account online to store your plots. Plotly Express is good for EDA to create best visualizations possible.

Reason for selecting Plotly for this is, Plotly is a more sophisticated data visualization tool that is better suited for creating elaborate plots more efficiently, which helped me to understand given set’s visualization in detail. And also, it allows me to create superior interactive graphs than Matplotlib and Seaborn.

Installation

Using intel core i5 9th generation with NVIDIA GFORECE GTX1650.
Windows 10 Environment Used.
Already Installed Anaconda Navigator for Python 3.x
The Code is written in Python 3.8.
If you don't have Python installed then please install Anaconda Navigator from its official site.
If you are using a lower version of Python you can upgrade using the pip package, ensuring you have the latest version of pip, *python* -m pip install --*upgrade pip and press Enter.*

Run/How to Use/Steps

Keep your internet connection on while running or accessing files and throughout too.
Follow this when you want to perform from scratch.

Open Anaconda Prompt, Perform the following steps:

```
cd <PATH>
pip install numpy
pip install pandas
pip install sklearn
pip install plotly
```

You can also create requirement.txt file as, pip freeze > requirements.txt

run files.

Follow this when you want to just perform on local machine.
Download ZIP File.
Right-Click on ZIP file in download section and select Extract file option, which will unzip file.
Move unzip folder to desired folder/location be it D drive or desktop etc.
Open Anaconda Prompt, write cd <PATH> and press Enter.
eg: cd C:\Users\Monica\Desktop\Projects\Python Projects 1\7)Dimensionality_Reduction
In Anconda Prompt, pip install -r requirements.txt to install all packages.
Open in Jupyter Notebook, Dimensionality_Reduction_With_PCA_2.ipynb
Please be careful with spellings or numbers while typing filename and easier is just copy filename and then run it to avoid any silly errors.
Note: cd <PATH>
[Go to Folder where file is. Select the path from top and right-click and select copy option and paste it next to cd one space <path> and press enter, then you can access all files of that folder] [cd means change directory]

Directory Tree/Structure of Project

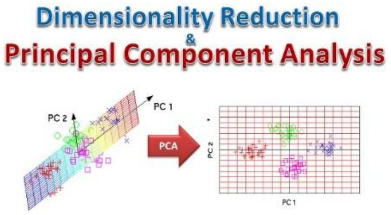
Folder: 7)Dimensionality_Reduction
Dimensionality_Reduction_With_PCA_2.ipynb

To Do/Future Scope

Can try with other dimensionality reduction techniques.

Technologies Used/System Requirement/Tech Stack

NumPy pandas $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$ scikit learn



plotly | Dash

Credits

Data Professors channel