Principal Component Analysis application on Iris dataset

AKHILESH KUMAR

Department of Engineering Physics

***Abstract*—** In this paper, we introduced a concept widely used by statisticians, the principal component analysis technique. Principal Component Analysis (PCA) is a simple yet popular and useful linear transformation technique that is used in numerous applications such as stock market predictions, the analysis of gene expression data, and many more. This PCA methods were tested using measurements made on various kinds of flowers (IRIS data).

***Key words*—** *Data Visualization*, PCA algorithm , Big Data Analysis, Data Science, .

I. INTRODUCTION

While the internet applications are widely used, the amount of data information has a rapid growth. Data Science analysis has become a popular issue nowadays. Datasets with a large number of features are called high-dimensional datasets. By using principal component analysis we convert this higher dimensional data to lower dimension by reducing the number of features etc. The sheer size of data in the modern age is not only a challenge for computer hardware but also a main bottleneck for the performance of many machine learning algorithms. The main goal of a PCA analysis is to identify patterns in data; PCA aims to detect the correlation between variables. If a strong correlation between variables exists, the attempt to reduce the dimensionality only makes sense. In a nutshell, this is what PCA is all about: Finding the directions of maximum variance in high-dimensional data and project it onto a smaller dimensional subspace while retaining most of the information.

Reducing the dimensionality of a dataset can be useful in different ways. For example, our ability to visualize data is generally limited to large dimensions. Lower dimension can sometimes significantly reduce the computational time of some numerical algorithms. Besides, many statistical models suffer from high correlation between covariates, and PCA

can be used to produce linear combinations of the covariates that are uncorrelated between each other.PCA was introduced by K. Pearson in early 20th century . It linearly

Principal Component Analysis application on Iris dataset

transforms data into a lower-dimensional subspace by obtaining the maximized variance of the data in a low-dimensional representation.

2.PCA and Dimensionality Reduction

Often, the desired goal is to reduce the dimensions of a *d*-dimensional dataset by projecting it onto a (*k*)-dimensional subspace (where $k<d$) in order to increase the computational efficiency while retaining most of the information.

computed eigenvectors (the principal components) of a dataset and collect them in a projection matrix. Each of those eigenvectors is associated with an eigenvalue which can be interpreted as the “length” or “magnitude” of the corresponding eigenvector. If some eigenvalues have a significantly larger magnitude than others that the reduction of the dataset via PCA onto a smaller dimensional subspace by dropping the “less informative” eigenpairs is reasonable.

Summary of the PCA Approach :

. Standardize the data.

- Obtain the Eigenvectors and Eigenvalues from the covariance matrix or correlation matrix, or perform Singular Vector Decomposition.
- Sort eigenvalues in descending order and choose the *k* eigenvectors that correspond to the *k*largest eigenvalues where *k* is the number of dimensions of the new feature subspace ($k\leq d$).
- Construct the projection matrix **W** from the selected *k* eigenvectors.
- Transform the original dataset **X** via **W** to obtain a *k*-dimensional feature subspace **Y**.

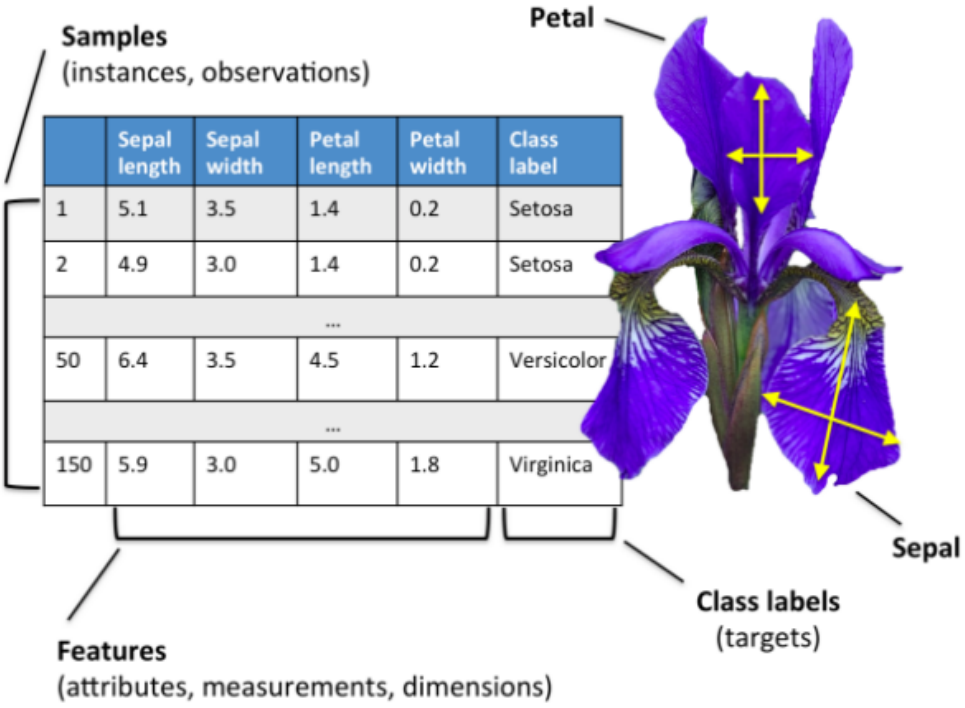
3. About Iris Dataset

we worked on with the famous “Iris” dataset that took from machine learning repository.

Principal Component Analysis application on Iris dataset

Dataset link:- https://archive.ics.uci.edu/ml/datasets/Iris

IRIS dataset is commonly used for a study of pattern classification, which was originally used in Fisher's experiment . It consists of 3 IRIS flowers: Setosa, Versicolor, and Virginica, each class contains 50 samples of four features which are the measurements of the sepal length (in cm), sepal width (in cm), petal length (in cm), and petal width(in cm).



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

4.Dimensionality Reduction on IRIS Data

Figure shows : PCA performs on iris data

Principal Component Analysis application on Iris dataset

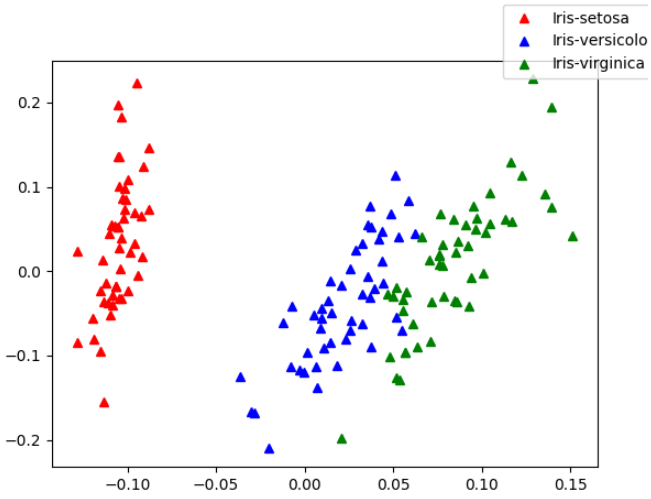


fig. (a)

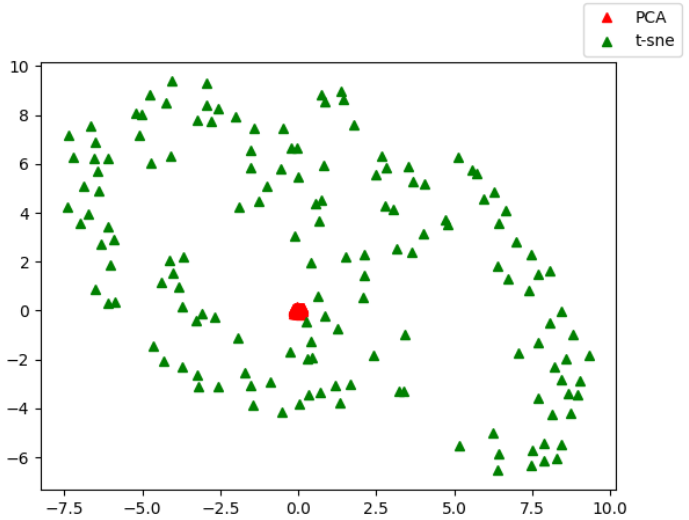


fig. (c)

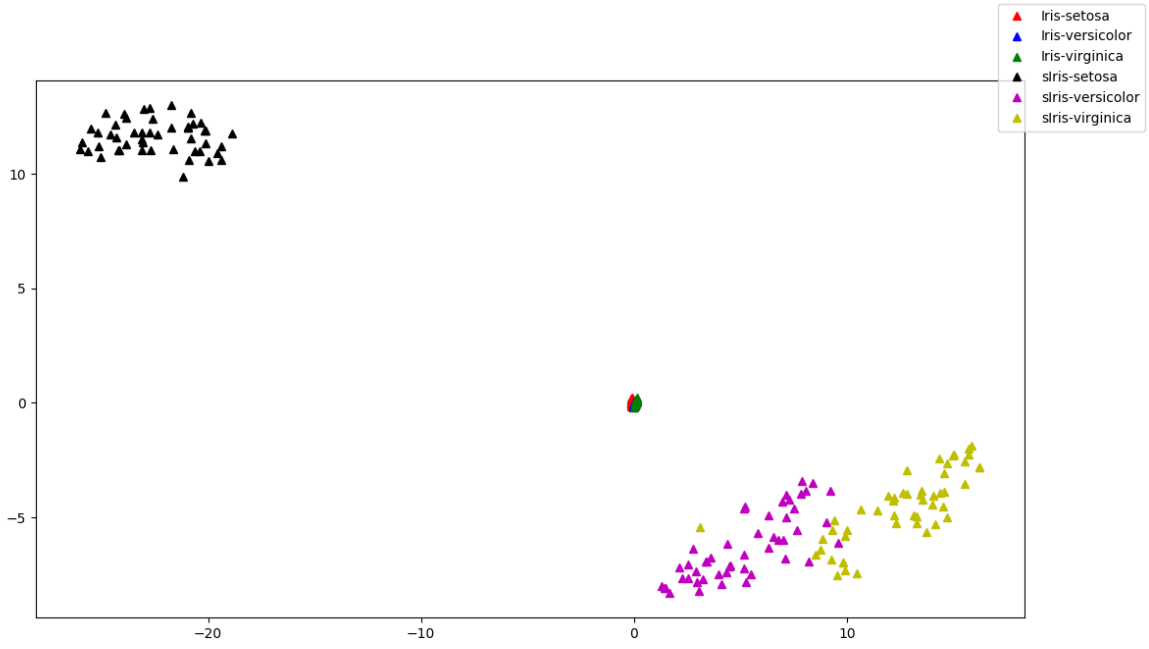


fig. (c)

figure (a) represents the points of the iris data set after the dimensional reduction using the own implementation of PCA

figure (b) represents the comparision between the points of the PCA implementation using matheatical constraints and the dimesionality reduction module t-sne in sklearn

figure (c) represents all the points of own implementation and t-sne module implementation, we can see that the points are scattered more in case of the inbuild implementation

5. Conclusion and Discussion

PCA is one of the good technique for dimensionality redudction. We performed the PCA on iris data we found that the dimensionality reduction made it easy to represent the points on a 2-D graph and also we can see from the graphs that it didn't loose the clustering property. Since all the points from the data set are clustered statisfying all the properties of the clusters and according to the kind of flower. We can similarly go for any dataset dimensionality reduction on the astropaticle physics datasets to reduce the cost of processing.

REFERENCES :

1. https://en.wikipedia.org/wiki/Principal_component_analysis

2. https://archive.ics.uci.edu/ml/datasets/iris

3. http://sebastianraschka.com/Articles/2014_pca_step_by_step.html

4. http://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

Principal Component Analysis application on Iris dataset