Principal Component Analysis on Asteroids Dataset

Saransh Kumar Karan
Machine Learning Intern
AI Technology and Systems
saranshkaran0@gmail.com
www.ai-techsystems.com

Abstract - The dataset contains features with the purpose of classifying an asteroid as hazardous or not. The aim of this report is to find the minimum number of principal components that contain 50% and 90% of the information. Principal Component Analysis is a dimensionality reduction technique used for identification of a smaller number of uncorrelated variables known as principal components from a larger set of data. It was found that 3 principal components hold 50% of the information, whereas, 9 principal components hold 90 % of the information.

Keywords - exploratory data analysis, dimensionality reduction, principal component analysis, preprocessing, Pandas. sklearn

I. INTRODUCTION

Principal components analysis is essentially just a coordinate transformation. It is calculated by solving an algebraic eigenvalue problem: finding the eigenvectors (PC's) of the covariance matrix of your original dataset. The resulting PC corresponding to the largest eigenvalue is the line of highest variance. The PC’s can then be used for exploratory data analysis and prediction such as classification problems. Thus, reducing the amount of data as well as the noise in the data. The categorical values where not considered. A total of 33 features where considered for the study. The built in PCA function from the ‘sklearn’ library was used to find the information stored by each components/feature. Out of 33 components 3 were found to hold 50% of the information and 9 were found to hold 90 % of the information.

II. RELATED WORK

PCA is mostly used as a tool in exploratory data analysis and for making predictive models. It is often used to visualize genetic distance and relatedness between populations. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after a normalization step of the initial data.

III. METHODOLOGY

A. Dataset

All the data is from the (http://neo.jpl.nasa.gov/). This API is maintained by SpaceRocks Team: David Greenfield, Arezu Sarvestani, Jason English and Peter Baunach.Exploratory Data Analysis.

B. Code:

The PCA function from the ‘preprocessing’ module of the ‘sklearn’ library was used to find the information held by each component.
The sklearn.preprocessing package
provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators. It was stored in a Pandas dataframe. The following is the first 15 rows of the dataframe:

	Percentage_info	Components
1	24.53	1
2	19.68	2
3	13.87	3
4	10.46	4
5	6.61	5
6	4.39	6
7	3.72	7
8	3.55	8
9	3.39	9
10	2.91	10
11	2.30	11
12	1.79	12
13	1.01	13
14	0.91	14
15	0.57	15

From the above table, it is observed that four components hold about 75 % of the information.

A function was written that takes two inputs:

The list containing the percent of information stored by each component. The first column the above table, under the name ‘Percentage_info’ is the required values of the list. This is the first argument.

The second argument is the percentage information corresponding to which you want the minimum number of principal components required to represent them. In our study, it is 50% and 90%.

The function returns the minimum number of components that contain the provided percent of information (second argument).

The following is the code for the function in python:

```
def no_of_components(arr, info_percent):

    comps = 0

    sum_info = 0

    info_percent = float(info_percent)

    for i in range(len(arr)):

        if sum_info >= info_percent:

            break

        sum_info = sum_info + arr[i]

        comps += 1



    return comps
```

IV. CONCLUSION

The purpose of the project to perform principal component analysis on the asteroid dataset. Furthermore, find out the minimum components that contain specific percentage of the information.

It was found that that 3 principal components contains about 50 % of the information whereas 9 principal components contain 90% of the information.

V. REFERENCES

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

https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

https://www.coursera.org/learn/machine-learning/lecture/GBFTt/principal-component-analysis-problem-formulation

https://scikit-learn.org/stable/modules/preprocessing.html