Classification of NASA Asteroid Dataset (August 2018)

Ashrya Agrawal, Machine Learning Intern, AITS

Link to AITS website : https://ai-techsystems.com/

***Abstract*— Asteroids can be classified depending on whether they are hazardous or not. Machine Learning can be used to do this. ML models like SVM can classify the asteroids to a significant level of accuracy. There is not much difference if we use raw data instead of principal components. Accuracy changes by less than 1 %. The training time on raw data is more than when we train on principal components .**

Keywords—SVM Classifier, Asteroids, Sklearn

I. Introduction

Asteroids can be a potential threat for life on Earth. Thus it is important to detect the asteroids that can potentially harm Earth, in order to take timely measures. As there are a large number of asteroids in space, it is essential that the process of finding (or classifying as) hazardous asteroids be done by computers based on the data from satellites. Machine Learning can be used for this. This project uses SVM to classify the asteroids .

II. Steps involved in making project

a. Gathering of Data

The dataset is taken from kaggle. Link : https://www.kaggle.com/shrutimehta/nasa-asteroids-classification. All the data of this kaggle dataset is taken from the (http://neo.jpl.nasa.gov/). This API is maintained by SpaceRocks Team: David Greenfield, Arezu Sarvestani, Jason English and Peter Baunach.

b. Exploring Dataset

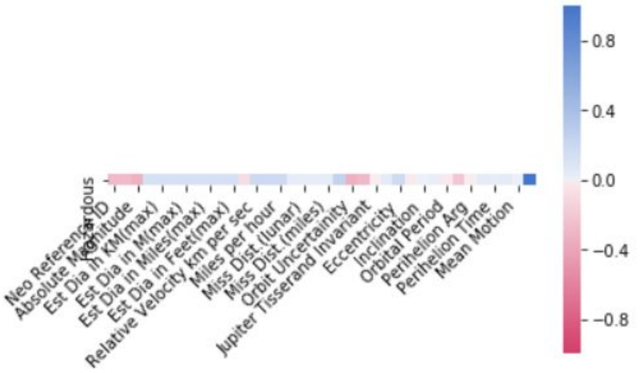
The data is about Asteroids - NeoWs. NeoWs (Near Earth Object Web Service) is a RESTful web service for near earth Asteroid information. With NeoWs a user can: search for Asteroids based on their closest approach date to Earth, lookup a specific Asteroid with its NASA JPL small body id, as well as browse the overall data-set.

The dataset is divided into 2 parts : Raw Data and Principal parts of data. Raw data is in the form of json files and principal data is in the form of csv.

Fields present in the dataset are :

1. Neo Reference ID
2. Name
3. Absolute Magnitude
4. Est Dia in KM(min)
5. Est Dia in KM(max)
6. Est Dia in M(min)
7. Est Dia in M(max)
8. Est Dia in Miles(min)
9. Est Dia in Miles(max)
10. Est Dia in Feet(min)
11. Est Dia in Feet(max)
12. Close Approach Date
13. Epoch Date Close Approach
14. Relative Velocity km per sec
15. Relative Velocity km per hr
16. Miles per hour
17. Miss Dist.(Astronomical)
18. Miss Dist.(lunar)
19. Miss Dist.(kilometers)
20. Miss Dist.(miles)
21. Orbiting Body
22. Orbit ID
23. Orbit Determination Date
24. Orbit Uncertainity
25. Minimum Orbit Intersection
26. Jupiter Tisserand Invariant
27. Epoch Osculation
28. Eccentricity
29. Semi Major Axis
30. Inclination
31. Asc Node Longitude
32. Orbital Period
33. Perihelion Distance
34. Perihelion Arg
35. Aphelion Dist
36. Perihelion Time
37. Mean Anomaly
38. Mean Motion
39. Equinox
40. Hazardous

Heat Map for input features :



c. Approach and algorithms

SVM classifier of sklearn is used to perform classifier on both raw data and principal components.Then, before using SVM dates are converted to a number of days (relative to a fixed day) which is compatible with sklearn SVC. The strings like “Equinox”, “Orbiting Body” have been converted to a dictionary and corresponding keys are used.

The features that don’t help in classification but instead reduce the performance are dropped.

After this MinMaxScaler is used to scale the data. This scaling of data is necessary to avoid exceptionally bad results. Heat map from seaborn is used to visualise the correlation of the features.
Note : Correlation is a measure of only linear relationship. It can not be used to conclude if output does not depend on a particular feature.

The target i.e. “is_hazardous” is one hot encoded to feed into the model.
Sklearn function train_test_split is used to create training and test dataset.
Linear classifier is used together with OnevsRestClassifier. After this we get accuracy and hyperplanes.
Matplotlib.pyplot is used to visualise hyperplane and the dataset.

The above process is done both, while working on principal components and raw data.
But while working with raw data one additional step of loading data from various json files is also done.

d. Support Vector Machines (SVM)

Support Vector Machine (SVM) is a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data. SVM is particularly suited to analyzing data with very large numbers (for example, thousands) of predictor fields.
SVM has applications in many disciplines, including customer relationship management (CRM), facial and other image recognition, bioinformatics, text mining concept extraction, intrusion detection, protein structure prediction, and voice and speech recognition.

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new data can be used to predict the group to which a new record should belong.

The advantages of support vector machines are:

1. Effective in high dimensional spaces.
2. Still effective in cases where number of dimensions is greater than the number of samples.
3. Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
4. Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

1. If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
2. SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).

e. MinMax Scaler

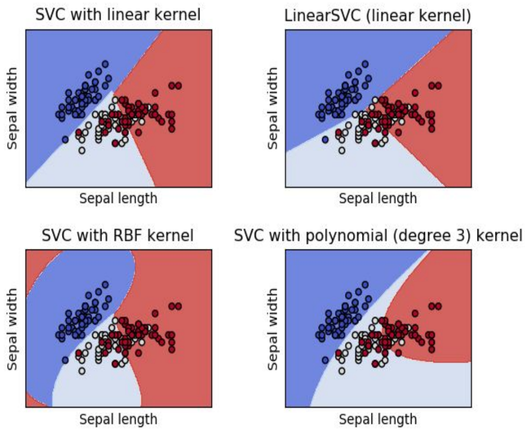
Transforms features by scaling each feature to a given range.This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

The transformation is given by:

X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))

X_scaled = X_std * (max - min) + min

f. Figures and Tables



Various kernels used in SVM

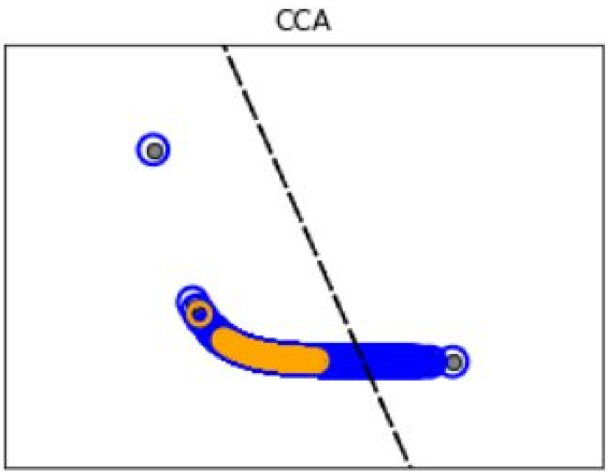
g. Observations

The accuracy and training time :

Using principal components of data :

Accuracy : 94.88%

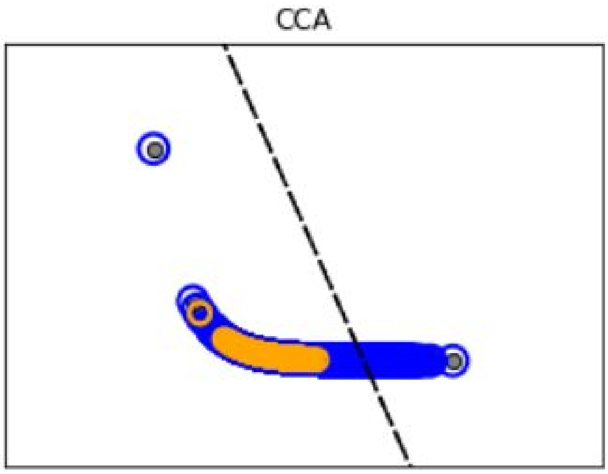
Training time : 582ms



Using raw data :

Accuracy : 94.67%

Training time : 717ms



H. Conclusion

There is not much difference in accuracy in the two cases. The training time is slightly higher when we use raw data.

III. Scope in Future

Machine Learning can definitely help in speeding up the process of identification of hazardous asteroids and space material in future. These objects on being detected and classified as hazardous, can be destroyed or their path can also be changed to avoid destruction on earth

IV. Acknowledgment

Data-set: 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. The problem statement is taken from kaggle. I would also thank AI-Techsystems for continuously monitoring the work and providing the support.

V. References

[1] http://neo.jpl.nasa.gov/

[2]https://www.kaggle.com/shrutimehta/nasa-asteroids-classification

[3] https://scikit-learn.org/stable/modules/svm.html