Project : Dimensionality Reduction for studying Diffuse Circumgalactic Medium

Organization: ML4SCI (Machine Learning for Science)

Google Summer of Code ‘21

About Me:

Name and Contact information

- **Name:** Aamir Miyajiwala
- **Country of Residence:** India
- **Time zone:** IST (UTC +5:30)
- **Github id:** github.com/aamirmiy/
- **Email:** aamir.miyajiwala@gmail.com
- **LinkedIn:**www.linkedin.com/in/aamir-miyajiwala-34516a1a7/
- **Mobile:** +91 9822019828
- **Primary Language:** English

Education:

- **University:** Pune University
- **College:** Pune Institute Of Computer Technology
- **Degree:** B.E(Bachelor of Engineering) with Honours in AI&ML.
- **Field of Study:** Computer Engineering
- **Current Year:** 3rd Year(6th semester ongoing)
- **CGPA:** 9.4/10
- **Expected Graduation:** 2022

Personal Background

Computers have fascinated me since early school days and I have been a programming geek since high school. In college I was introduced to Machine Learning and that has excited me enough to explore more and more in the past 2 years and even opt for an Honours in the field of AI&ML. NLP is especially an area of interest and I have worked on various projects on this subject.

To start with, I did a project on Sentiment analysis on tweets related to Covid-19 tweets. It included everything right from data collection to the creation of an app displaying different interactive visualizations along with inferences.It was deployed using Heroku.

I am at the end of my internship at **C-DAC** (Centre of Department for Advanced Computing) where I was conducting research on POS(Parts of Speech) tagging, surveying the different methods and techniques used and implementing the research for building POS tagger for Hindi language.

Apart from NLP I have also worked on algorithms for Time Series analysis and anomaly detection during my internship at **Sarvatra Technologies** (a fintech company), where I developed a forecasting model to forecast daily debit card ATM transactions over the 24 hours for real time anomaly detection.This was based on the approach of Predictive Analysis for Fraud Detection. For my work done, **I was awarded a monetary prize for best performing intern.**

All these projects and internships have ingrained the importance of clean code and documentation in me.

I am proficient in **C,C++ , Python** and in frameworks like **tensorflow** and **keras**.

Community Contributions.

I am part of the **Google Student Developer Club** and actively participate in the organization of events.

Being a part of **PICT ACM** Student Chapter which is a branch of **ACM(Association for Computing Machinery)**, I have conducted introductory sessions for the students of my college on **Machine Learning** and have also handled the responsibility of being a **team leader** for developing a centralized Event Management System for our organization's flagship event **PULZION2020**.

About the Project

Abstract: The history of a galaxy is encapsulated in the history of its gas cycling between the visible body of the galaxy and its harder-to-detect diffuse circumgalactic medium (CGM). The aim is to conduct research on different methods, to study the history of the CGM by developing computational and machine learning tools to use quasar absorption lines to determine the gas composition, temperature and density**.**

Focus : This project focuses on applying ML dimensionality reduction techniques for the dataset of simulated quasar absorption spectra.

Outcome: Reduced dimensional results that will still maintain a high level of accuracy compared to the original dataset.

Background:

To understand the problem and get familiar with the technical details related to astrophysics I read a few research papers including **The circum-galactic medium of quasars: transverse and line-of-sight absorptions** written by **A. Sandrinelli** *et al* and **The Structure of Circumgalactic Medium of Galaxies** written by **David V. Bowen** *et al* and have summarized below why quasars and their absorption spectra are studied to probe properties of CGM.

Problem related theory: The circumgalactic medium (CGM) is the diffuse gas spanning a few hundred kpc(kiloparsec) outside the galaxies and represents the most abundant reservoir of baryons(composite particles made of three quarks) of galaxies. It is the link between the interstellar and intergalactic media, so that gas exchanges between them have to go through it and necessarily leave here the proof of its passage.These exchanges play an important role in the evolution of a galaxy because they can favour or suppress different processes in the galaxy, notably star formation.

One way to probe the properties of the CGM of a galaxy is to study the absorption lines they cause in the spectra of background objects. These background

objects or source need to be angularly so close that the projected distance is of the order of hundreds kpc.Quasars are sources apt to this scope since they are bright (they are very luminous as they have active galactic nucleus(AGN)) and point like.Quasar emits energy in a wide spectrum of electromagnetic radiation.Between earth and quasar there are numerous gas clouds(CGM) containing different sorts of atoms and ions which absorb at certain characteristic wavelengths determined by type of ion and position on the sightline(imaginary line between observer and subject of interest).With detection of these absorption lines a unique opportunity has arisen to study the physical and chemical properties of regions of universe otherwise not accessible for astrophysical observations.QSO absorption lines thus provide an important gateway to infer observational constraints on galaxy information and evolution and to probe conditions in early universe as discussed in the description of the project.

Related Work : Machine Learning has been used extensively in the field of astrophysics. Regarding this project, dimensionality reduction is not new to astrophysics. It has been previously implemented for Investigating Quasar Emission-Line Variance where a Variational Autoencoder was used. Another instance of its implementation can be found in the paper Reducing the Dimensionality of Data: Locally Linear Embedding of Sloan Galaxy Spectra by J. T. VanderPlas, A. J. Connolly where they used LLE(Locally Linear Embedding).

Technical details:

How Dimensionality reduction is used in quasar spectra?

Quasar spectra are quite diverse: while most quasar spectra show the same emission lines, those lines do not always appear in the same ratio or with the same shape. Characterizing the variations in the appearance of emission lines is important in order to understand the physical processes that drive those changes. One statistical technique commonly applied to large datasets is principal component analysis (PCA), which can identify variance in a dataset and reduce its dimensionality.An alternative to PCA is an autoencoder, which can function like a

nonlinear generalization of PCA.An autoencoder consists of two neural nets, an encoder which maps each input into a low-dimensional latent space, and a decoder, which maps points in the latent space back to the input space. By training the model to condense and then reconstruct given inputs, it can learn an efficient and informative lower-dimensional representation of the input space, in a nonlinear way.

I will be making use of these two techniques in the project along with some other algorithms ,since they are the most popular dimensionality reduction techniques.

Proposed Strategy:

After referring different articles ,examples and papers.I have come up with the following fundamental steps:

1)Understanding the data and learning the meaning of all the features.

2)Drop features which have no significance based on Step-1.

3)Plotting the correlation map and deciding candidates(features) for dimensionality reduction based on potential multicollinearity between the features.

4)Applying preprocessing techniques like one hot encoding(if required), logarithmic transforms, normalization depending on distribution of data.

5)Applying dimensionality reduction techniques,tuning them and comparing their performance.

Based on the above strategy I have prepared a demonstration below to give an idea of how I would be approaching the problem statement during the coding period.

I have used the dataset **DR14** from **Sloan Digital Sky Survey**
to perform classification of **Galaxies,Stars and Quasars**.I have implemented only **PCA** algorithm since this is purely for demonstrative purposes. (I am aware that this dataset is not similar to what I will be working on but since the steps involved will almost be the same I went ahead with this for a quick demonstration.)

Demonstration:

Data:

```
In [8]: df.columns
Out[8]: Index(['objid', 'ra', 'dec', 'u', 'g', 'r', 'i', 'z', 'run', 'rerun', 'camcol',
       'field', 'specobjid', 'class', 'redshift', 'plate', 'mjd', 'fiberid'],
```

Feature Description:

1**. objid ,ra, dec** are the camera features.
2. **ra (Right ascension)** and **dec (declination)** together are the astronomical coordinates which specify the direction of a point on the celestial sphere in the equatorial coordinate system.
3. **u,g,r,i,z** represent the response of the 5 bands of the telescope.They belong to the Thuan-Gunn astronomic magnitude system.
4. **Run,rerun,camcol** and **field** are features which describe a field within an image taken by the SDSS. A field is basically a part of the entire image corresponding to 2048 by 1489 pixels.

5. The **class** identifies an object to be either a galaxy,star or quasar.This will be the response variable which we will be trying to predict.

6. In physics **redshift** is an increase in the wavelength and corresponding decrease in frequency and photon energy of electromagnetic radiation.**Modified Julian Date (mjd)**, used to indicate the date that a given piece of SDSS data (image or spectrum) was taken.

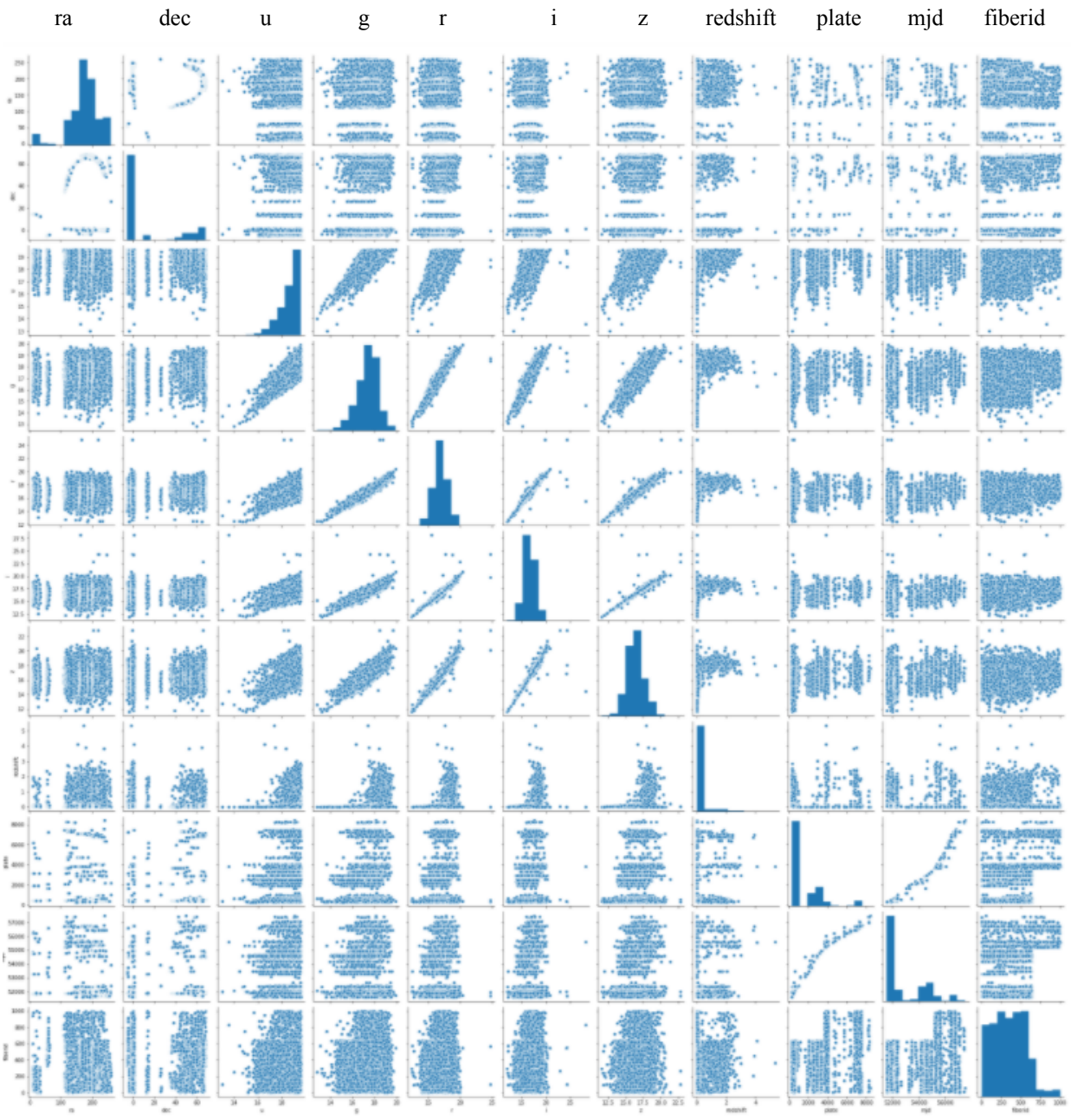
Analysis:

```
In [15]: df.isnull().sum()

Out[15]: objid        0
         ra           0
         dec          0
         u            0
         g            0
         r            0
         i            0
         z            0
         run          0
         rerun        0
         camcol       0
         field        0
         specobjid    0
         class        0
         redshift     0
         plate        0
         mjd          0
         fiberid      0
```

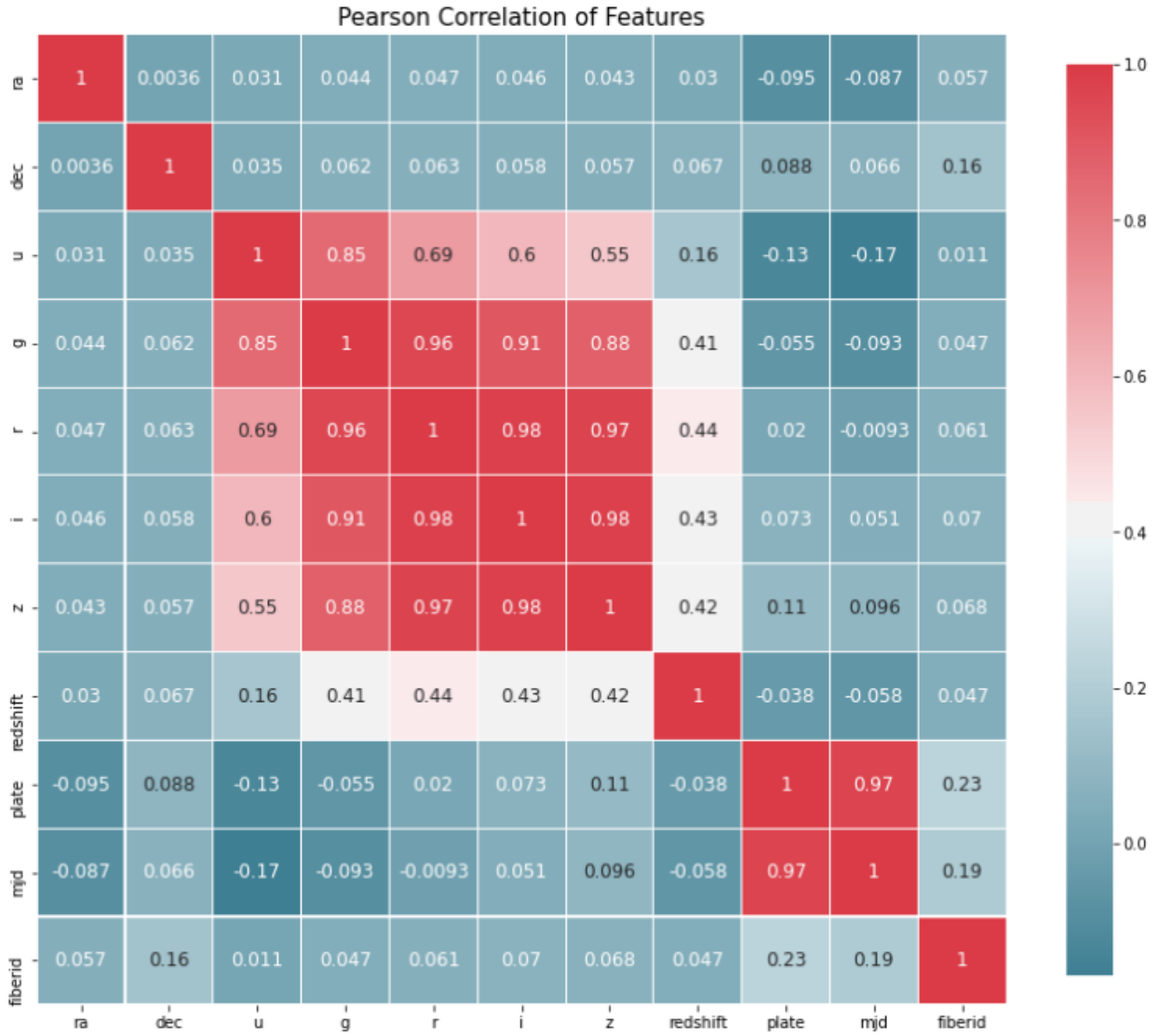
No null values which is great. We will also be dropping columns like objid, specobjid since they are random object identifiers.Also, field, run, rerun and camcol are features representing camera positions and don't have any significance with astronomical objects we are observing.

Below I have plotted a pair plot to look at the relationships between the features for a clear understanding.



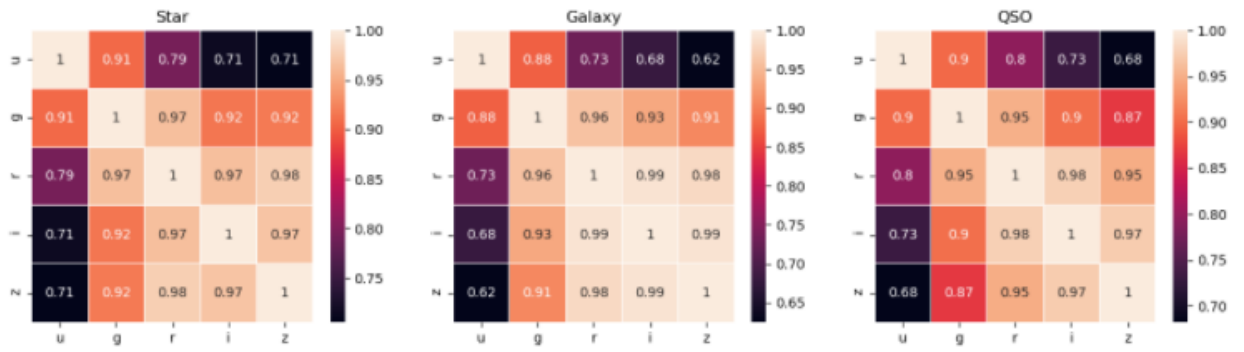
- There looks to be a strong linear relationship between the 5 bands of the telescope. Because of potential multicollinearity these 5 are good candidates for dimensionality reduction.
- mjd and plate also look to be highly correlated and are also good candidates for dimensionality reduction.

Let's have a look at the correlation matrix for the dataset.

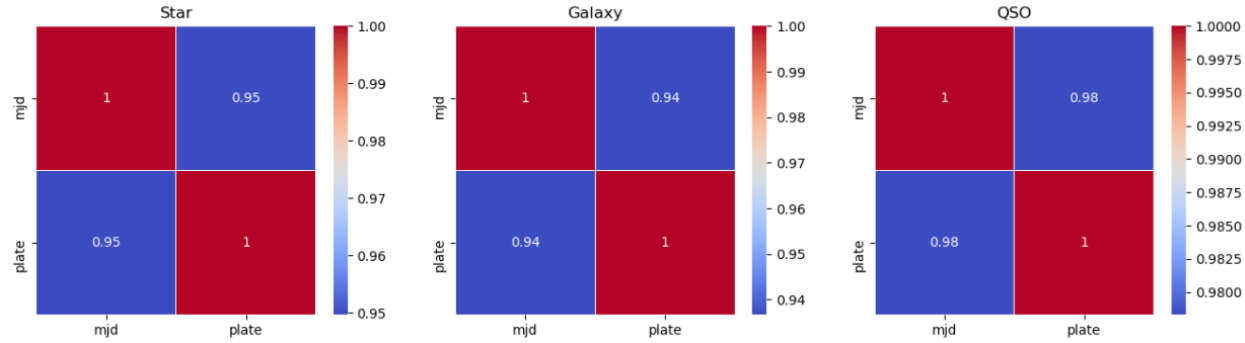


- We see that the 5 bands of the telescope are highly correlated.
- As are mjd and plate.

Let’s explore the distributions of the three classes separately



The correlation between the 5 bands of the telescope is evenly distributed amongst the three classes of objects.



The correlation between mjd and plate is also evenly distributed.

Now we will perform dimensionality reduction on the telescope bands ,mjd and plate features using PCA which is one of the most popular dimensionality reduction techniques.

```
In [29]: bands = df[['u','g','r','i','z']].values

In [37]: from sklearn.decomposition import PCA
         pca=PCA().fit(bands)
         cumsum = np.cumsum(pca.explained_variance_ratio_)
         d=np.argmax(cumsum>=0.95)+1

In [38]: d

Out[38]: 2
```

Here I combined the 5 bands into one dataframe and then calculated the number of dimensions required to maintain 95% variance which is stored in variable d. The value calculated was 2. Hence the information provided by those 5 features can be provided by only 2 components also thereby reducing the dimensions.

PCA has also been applied to mjd and plate features,reducing the dimension from 2 to 1.

```
In [40]: pca1=PCA(n_components=1)
         mjdplate = pca.fit_transform(df[['mjd','plate']])
         mjdplate

Out[40]: array([[ 2685.26540812,  -328.26110819],
                [-1725.85982404,   285.34478015],
                [-1491.00753634,   -50.22101513],
                ...,
                [ 7090.7160119 ,   641.42697446],
                [-1462.39779001,   164.48222754],
                [-1462.39779001,   164.48222754]])
```

After reducing the dimensions a new dataframe was made

```
In [46]: df_pca
```

Out[46]:

	ra	dec	class	redshift	plate	mjd	fiberid	bands_1	bands_2	mjdplate_1
0	183.531326	0.089693	2	-0.000009	3306	54922	491	-1.507202	-1.377293	2685.265408
1	183.598371	0.135285	2	-0.000055	323	51615	541	-0.195758	-0.028410	-1725.859824
2	183.680207	0.126185	0	0.123111	287	52023	513	1.297604	-0.590023	-1491.007536
3	183.870529	0.049911	2	-0.000111	3306	54922	510	-1.446117	0.566685	2685.265408
4	183.883288	0.102557	2	0.000590	3306	54922	512	-0.849271	1.287505	2685.265408
...	...	...	...	...	...	...	...	...	...	...
9995	131.316413	51.539547	0	0.027583	447	51877	246	0.222959	-0.134301	-1462.397790
9996	131.306083	51.671341	0	0.117772	447	51877	228	0.259171	0.415333	-1462.397790
9997	131.552562	51.666986	2	-0.000402	7303	57013	622	1.480725	0.388717	7090.716012
9998	131.477151	51.753068	0	0.014019	447	51877	229	1.392088	0.117004	-1462.397790
9999	131.665012	51.805307	0	0.118417	447	51877	233	-0.936205	-1.113215	-1462.397790

10000 rows × 10 columns

Now the dataset is split into X and y, scaled using Standard Scaler and split into train and test set. I assessed the performance by applying Logistic Regression on it and have provided the confusion matrix along with the different metrics.

In [52]:
```
from sklearn.linear_model import LogisticRegression

LG = LogisticRegression(penalty="l2")
LG.fit(X_train, y_train)
```

```
C:\Users\aamir\anaconda3\lib\site-packages\sklearn\utils\validation.py:760: Dat
when a 1d array was expected. Please change the shape of y to (n_samples, ), fo
  y = column_or_1d(y, warn=True)
```

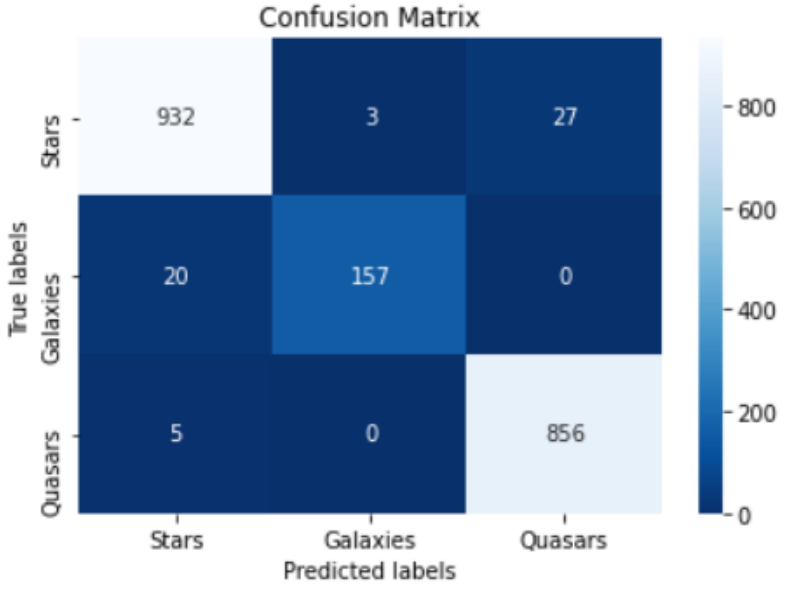
Out[52]:
```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

In [53]:
```
y_preds = LG.predict(X_test)
accuracy = LG.score(X_test,y_test)

print(f"Accuracy of Logistic Model: {accuracy*100}%")
```

```
Accuracy of Logistic Model: 97.25%
```

In [54]:
```
from sklearn.metrics import confusion_matrix
import seaborn as sns
cm = confusion_matrix(y_test, y_preds)
ax = plt.subplot()
sns.heatmap(cm, annot=True, ax = ax, fmt="g", cmap="Blues_r");
ax.set_xlabel('Predicted labels');
ax.set_ylabel('True labels');
ax.set_title('Confusion Matrix');
ax.xaxis.set_ticklabels(['Stars', 'Galaxies','Quasars']);
ax.yaxis.set_ticklabels(['Stars', 'Galaxies','Quasars']);
```



```
In [55]: from sklearn.metrics import precision_score, recall_score, f1_score
precision = precision_score(y_test,y_preds, average="micro")
recall = recall_score(y_test, y_preds, average="micro")
f1 = f1_score(y_test, y_preds, average="micro")

print(f"Precision Score: {precision}\n\
Recall Score: {recall}\n\
F1 Score: {f1}")

Precision Score: 0.9725
Recall Score: 0.9725
F1 Score: 0.9725
```

We can see that the results are extremely promising.We reduced the dimensions from 17 (excluding class) to 10 and achieved an accuracy of 97%.

Note:PCA might not prove to be the best for the real data and hence other techniques like autoencoders,variational autoencoders,LLE,spectral embedding etc. will be applied depending on the type and count of features.

Actual data can contain millions of samples, hence neural networks will be the preferable solution instead of pre-built sklearn algorithms.

Timeline:

Community Bonding period(May 17- June 7):

The most important task I will be committing to during this time is to get to know the community and mentors.Discuss the details about the project and what outcome is to be expected. I will constantly stay in touch with the mentors and the community and will make sure I clear my doubts regarding any issues I face that time as well as discuss the factors that can improve my project.Also, I will be studying the theory and concepts related to this project for a comprehensive understanding of the data.

By the end of this period, my aims are to:

- Learn about the theory related to quasars, their absorption spectra, spectrum properties etc. and refer to the paper Quasar Spectrum Classification

with Principal Component Analysis:Emission Lines in Lyα Forest written by **Nao Suzuki** for developing a strong approach.

- Discuss with mentors about milestones to be set for the coding period.
- Align on the ways of working and project tracking with mentors.
- Learn thoroughly about the best practices used for dimensionality reduction.

Coding Timeline (June7-August16):

Tools used: Python,Tensorflow,Keras (or Pytorch depending on what is preferred),sklearn

Week1 -Week2 (June7 - June 20):

Understand the dataset clearly and define all the features and their significance and refer some articles or papers if needed. Based on the information gained from reading the papers mentioned, I assume the dataset will contain features in the form of wavelengths, relative flux ,absorption by different metals causing emissions ,redshifts,and degrees of freedom to fit the data. These features will be studied in detail and visualizations will be made to look at the different spectra.

Week3-Week4 (June21-July4):

Based on previous week's analysis some features which are useless will be noted and not removed immediately.Further analysis will be done to look for linearity between features by plotting a pair plot and displaying the correlation map which will show the correlation coefficients between the variables.All the features showing strong collinearity will be noted as they will be excellent candidates for dimensionality reduction.Other preprocessing steps like one hot encoding, log transforms etc. may be performed if required along with normalization.

Week5-Week6 (July5-July18):

Based on problems faced and communications with the mentors steps will be taken to incorporate changes and implement additional things. A few baseline models will be built on all the features for performance comparison in the future and to

check the effectiveness of the dimensionality reduction and the technique used for that. Preferably a Neural network will be built and cross validation will be used to evaluate the performance. The next phase would include implementation of dimensionality reduction techniques on the features noted during previous week's analysis. I will be implementing PCA, LLE,VAE(variational autoencoder) at first and check how well they do in preserving the variance of data while reducing the dimensions.This is done by checking how well they do in reconstructing the inputs.

Week7-Week8 (July19-August1):

Inputs will be reconstructed and fed into another model and performance will be checked against the baseline model. Results will be conveyed to the mentors and guidance will be taken to make changes and get few insights. Other techniques will be implemented and data will be retrained again and tested rigorously with the help of cross validation.

Week9(August2-August8):

Finalizing the models and hyperparameter tuning of the best performing techniques and preparation of a report of all the algorithms used, their limitations, number of dimensions reduced, features that were dropped and classification report of the metrics used to evaluate the performance.

Week10(August9-August16):

Spare week in case of some work getting delayed, in case of any emergency or otherwise.

Coding Period Ends.

Apart from the above schedule I will be maintaining a log about the progress of each week so that the mentors can get an overall summary of each week's work and maintain constant communication with my mentors and updating them about the work done regularly.Also, after Gsoc I would love to continue to work on other interesting research projects and contribute as much as possible.

GSoC:

Have you participated previously in GSoC? When? Under which project?

No, I have not participated in GSoC before. This is the first time I am participating in GSoC.

Are you also applying to other projects?

No, I am not applying for any other project in any other organization.

Commitments

I may have my end-semester exams in June (tentative due to CoronaVirus Pandemic) for a week so I will not be available around that period. I will compensate for that period before-hand. Even if there is a change in the exam timetables I will notify my mentors and shift my schedule accordingly without hampering my progress in the project.

Eligibility

I am eligible to participate in the Google Summer of Code. For any queries, clarifications or further explanations, feel free to contact aamir.miyajiwala@gmail.com

****