EEE-485 STATİSTİCAL LEARNİNG AND DATA ANALYTİCS PROJECT PROPOSAL FALL 2022-2023

Galactic Object Detection

Melih Berk Yılmaz 21803702 **Fuat Arslan 21803108**

In this project, the aim is to classify different types of galactic objects such as "GALAXY, STAR, QSO". To accomplish this task, we will use two supervised learning and one unsupervised learning type machine learning algorithms. Algorithms will be trained on a dataset which found on Kaggle named as "Sloan Sky Survey SDSS-DR16" (www.kaggle.com/datasets/rockdeldiablo/sloan-digital-sky-survey-dr16-70k?resource=download).

Dataset Description

Data consists of three classes, 70,000 samples and 17 features. There is a big imbalance between each class's sample amount. For example, GALAXY has 49,690 samples (71%) , STAR 13,494 samples (19%), QSO 6,816 samples (10%). Features:

Objid : Unique SDSS identifier
RA : Right Ascension
Dec : Declination
psfMag_u : PSF (Point Spread Function) flux in the u band
psfMag_g : PSF flux in the g band
psfMag_r: PSF flux in the r band
psfMag_i : PSF flux in the i band
psfMag_z : PSF flux in the z band
run : identifies the specific scan
rerun : A reprocessing of an imaging run
camcol : identify the scanline within the run
field:Field number
class : object class (galaxy, star or quasar object)
redshift : the Redshift of the object
plate : ID of the plate used for the telescope at the time the image was taken
mjd : modified Julian Date, i.e. the date at which the data were taken
fiberid : fiber ID

We are planning to make feature selection due to some features seem not giving proper information.

Algorithms

Following ML algorithms will be used for classification.

- K-means Clustering
- XGBoost
- Multi Layer Perceptron (MLP) (Neural Network)

We decided on these algorithms since we want to get deeper experience on theoretical and practical aspects of the algorithms that are commonly used in the industry. Also, we pay special attention to decide different type of algorithms. K-means is an unsupervised, XGBoost and MLP are supervised. XGBoost is chosen to make one of them more challenging than others. Furthermore, these algorithms expected to work on these data well.

Challenges

Before proposing this project, we statistically analyzed data. Data contains lots of extreme outliers, is imbalanced as explained before. Some features have very low variance near to zero which makes them unbeneficial for classification. Coding XGBoost expected to be very challenging.