Machine Learning Engineer Nanodegree

Capstone Proposal

Classification of Galaxies, Stars and Quasars Based on Photometric Data

Sara A. Alhamed
January 24th, 2019

Proposal

Domain Background

Scientific and technological developments in the fields of Astronomy in the last decades have brought with them huge amount of data about our universe. Powerful and sophisticated telescopes were built to either be sent to space like the famous Hubble Space Telescope or designed to be installed here on Earth like those used in The Southern European Observatory, Gemini Observatory, Mauna Kea Observatory, Apache Point Observatory and many others. Telescopes used in the Apache Point Observatory in New Mexico, US, in 2000 began collecting data about the night sky in project Sloan Digital Sky Survey (www.SDSS.org) named after Alfred P. Sloan Foundation who generously have contributed to the project's funding. Since its beginning, SDSS using an optical telescope gathered photometric data of 500 million objects and spectrometric data for more than 3 million objects. The data SDSS collects about Galaxies, Stars, Exoplanets (planets outside our solar system), Black Holes and many other mysterious objects is extremely valuable providing scientists and astronomers a rich content to study and research.

Astronomy and outer space have always been subjects of interest to me since a young age and as a computer scientist I find problems in Astronomy that I can work on by using knowledge and skills from my field, such as machine learning and data analytics techniques, to be very fascinating and even more interesting than the usual Astronomy I used to read about in the past.

Problem Statement

Data collected by SDSS is huge and is increasing annually as more and more parts of the sky are surveyed. Analyzing data from space for a certain purpose is something Astronomers have always been doing to classify celestial bodies, understand the nature of stars by analyzing their light waves, for navigation and mapping purposes, ...etc. Classification using Machine Learning however, is bringing this process of Astronomical data analysis to whole new levels, it is faster dealing with thousands if not millions of records of numerical data or images or maps, picking up patterns and relations better than any time before, applying complex statistical and mathematical operations. The problem I want to work on in this project is the classification of three celestial bodies: Galaxies, Stars and Quasars (a massive and extremely remote celestial object, emitting exceptionally large amounts of energy, and typically having a star-like image in a telescope).

Datasets and Inputs

The Sloan Digital Sky Survey releases their collected data every two to three years and their latest Data Release is DR15 containing Astronomical observations from the beginning of this survey in 2000 through July of 2017 as the survey now is in its fourth phase. SDSS data can be accessed through SQL queries on any of their databases, downloaded as bulk data and other ways. The dataset I will be working on will has 10,000 entries taken from the table *dbo.SpecPhotoAll* in the SDSS DR 14 database shared on Kaggle (www.kaggle.com/lucidlenn/sloan-digital-sky-survey/home).

The dataset is imbalanced most of the entries are galaxies then stars and only 8.5% are quasars:

GALAXY	4998
STAR	4152
QSO	850

The original dataset has 18 columns, 14 of which are features. Most of the features are numeric continuous data since they are measurements except Camcol feature which is of a categorical numerical type. The 10,000 objects to be classified are described by basic photometric and spectrometric data about galaxies, quasars and stars which will be used as features in this classification:

- Redshift: the displacement of the spectrum of the object toward longer (red) wavelengths meaning that the greater the redshift manifested by light emanating from such an object, the greater the distance of the object and the larger its recessional velocity making it such an important feature when it comes to identify the celestial objects we have in this project.
- Thuan-Gunn astronomic magnitude system: u, g, r, i, z : representing the response of the 5 bands of the telescope as a way of characterizing the brightness of astronomical sources.
- Run, rerun, camcol and field: features which describe a field within the object image taken by the SDSS.
- Right Ascension and Declination: astronomical coordinates specify the direction of a point on the celestial sphere in the equatorial coordinate system.
- Plate: each spectroscopic exposure employs a large, thin, circular metal plate that positions optical fibers via holes drilled at the locations of the images in the telescope focal plane.
- MJD: Modified Julian Date, used to indicate the date that a given piece of SDSS data (image or spectrum) was taken.
SDSS Glossary where there is a description of their categories and columns:
(www.sdss.org/dr12/help/glossary).

Solution Statement

Based on photometric data from SDSS, the machine learning techniques I will be applying to classify the objects will predict the object's type far more accurately than by merely looking at their images or through a telescope in a clear night. After the pre-processing step (detailed below) which will deal with the imbalanced dataset by implementing K-fold Cross Validation I plan to select classification algorithms that would yield high accuracy such as Support Vector Machines and Random Forest classifiers.

Benchmark Model

I will be classifying the data with different supervised machine learning algorithms and compare their results against results from (www.kaggle.com/farazrahman/predicting-star-galaxy-quasar-with-svm). I will also be implementing one unsupervised learning algorithm to see what results that yields.

Evaluation Metrics

Four common evaluation metrics for supervised learning I will be using:

• Accuracy = $\frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$	• Precision = $\frac{tp}{tp + fp}$
• Recall = $\frac{tp}{tp + fn}$	• F1-Score = $(1 + \beta^2)\frac{\text{precision} \times \text{recall}}{\beta^2 \text{precision} + \text{recall}}$

Unsupervised learning evaluation metrics:

- Silhouette:

  A technique provides a succinct graphical representation of how well each object l cluster.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Project Design

A summarized theoretical initial workflow as some steps and points might change as results come:

1- **Data Visualization:** visual the data in two to three ways to make it easy to notice patterns and get a good view of the data.
2- **Data Pre-processing:** Pre-processing will include checking if the data is balanced or imbalanced will solve that by implementing K-fold Cross Validation, excluding entries with missing values if there is any (we cant do anything about missing observational data but to exclude them), dropping duplicate records if there is, drop unique columns from the dataset that the model may cheat from, drop irrelevant features, if any exists, by implementing a PCA.
3- **Model Selection:** I'm considering two to three supervised learning algorithms and one unsupervised learning algorithm for the models. I would like to try some Neural Network architecture for this problem, however, it will be depending on the size of the dataset I will getting from SciServer because for a high NN accuracy a large dataset needed and, most importantly time, if I have enough time before the deadline it be fun to working on a NN as I wish to be able to apply most of what I have learned in this Nanodegree.
4- **Model Tuning:** improve initial results by fine tuning the model and make sure there is no overfitting and optimize parameters of some algorithms.
5- **Testing and Evaluation:** test the models and evaluate them using the evaluation metrics specified above.