# Machine Learning Advanced Nanodegree

## Capstone Proposal

Sayan Biswas

September 04, 2018

### Proposal

Predicting pulsar stars in the universe

### Domain Background

Pulsars are member of a family of stars, called neutron stars. They are highly dense spherical objects about the size of a large city (e.g. 20 km in diameter) with mass of several times that of Sun, which is of about 1.4 million km in diameter. They are called pulsars because they radiate two focused beams of light in opposite directions, which usually don't align with their axis of rotation (similar to a lighthouse). The frequencies of these pulses are so accurate that they can rival atomic clocks which are the most precise clocks known to us. The reason behind choosing this topic is to gain a good understanding on which features are observed in a pulsar, so that we can get better at finding them, and also given a bunch of relevant features associated to celestial objects, we can classify them as pulsars.

#### Reference

1. https://www.space.com/32661-pulsars.html
2. https://www.universetoday.com/25376/pulsars

### Problem Statement

The objective here is to classify the celestial objects to two categories: pulsar and non-pulsar with very high accuracy. This is important because pulsars serve many scientific purposes like below:

1. They give information about the physics of the matter inside them. Under such incredible pressure (second to only black holes) matter behaves quite differently. Hence, it works as very good resource materials, because such extreme environments can't be practically created (yet) in a test lab on earth.
2. Pulsars are extremely accurate in emitting their pulses, as mentioned before. Hence observing the changes in a pulsar's blinking provides information about what's happening in its vicinity. Also, these can be used as a clock for scientific experiments which require precise timing.
3. Pulsars are useful in finding extrasolar planets. In fact, the first extrasolar planet was found orbiting a pulsar.
4. They help in testing aspects of Albert Einstein's theory of general relativity.

### Datasets and Inputs

Our dataset consists of a csv file with almost 18000 entries, and it's obtained from Kaggle dataset. The input feature set consists of 8 continuous variables mentioned below, whose names are also self-explanatory:

1. Mean of the integrated profile.

2. Standard deviation of the integrated profile.
3. Excess kurtosis of the integrated profile.
4. Skewness of the integrated profile.
5. Mean of the DM-SNR curve.
6. Standard deviation of the DM-SNR curve.
7. Excess kurtosis of the DM-SNR curve.
8. Skewness of the DM-SNR curve.

This is a binary classification problem. So based on these features we need to decide whether the object is a pulsar or not. This metric is given by the variable `class`, where 1 means pulsar, and 0 otherwise. Without these metrices, detecting real pulsars from false ones would become tricky, because many of the pulses detected through instruments are from noise and radio frequency interference.

## Solution Statement

The solution strategy is to come up with a classifier that gives the best outcome. It's a supervised learning scenario as we have labelled data. I'll use the 8 features as the input to the models and the `class` variable as target. In case I see, that not all features are relevant or there are dependent features, then I need to clean up the feature vector accordingly. I will test various classifiers such as Logistic Regression, SVM, AdaBoost etc. and then whichever model gives the best output, I'll choose that one, and fine tune the hyperparameters to get optimum result.

## Benchmark Model

As this is a classification problem, we can keep a naïve predictor (which considers all starts as pulsar or non-pulsar) and then compare our accuracy, precision/recall against that one. As we don't want to lose out on actual pulsars, so we'd be giving more emphasis on precision than recall.

## Evaluation Metrics

Our input data contains 17,898 total examples. 1,639 positive examples. 16,259 negative examples. So, only accuracy won't be a good indicator about the effectiveness of our model. For the reasons mentioned before, we'll calculate the goodness of our model using $F_\beta$ - score with $\beta$ as 0.5, because it gives more emphasis on precision.

## Project Design

1. From initial observation, the dataset seems to be clean, so I presume there won't be much need for data sanitization. We can directly use this data for exploration, visualization, discard outliers if any. Also, we'll split it up in training and testing data sets.
2. In training phase, we'll consider multiple supervised learning models, check their accuracy, $F_\beta$ - score, and training time, and then choose the best one. We'll use grid-search, cross validation in this case, and after choosing the final model, we'll further optimize it using hyperparameter tuning.