Nathanael Roy

Statistics 208 - Final Project

Classification Methods: Pulsar Stars

Introduction

In this project, I will be using a variety of Statistical Data Mining methods on a data set of pulsar stars. The main question of interest is using different attribute characteristics within the data set to attempt to classify the target attribute which is binary. The main programming language used on this data set is R, although I have also attempted a couple of the methods (tree based) in python as well to learn some of the functionality though this is not discussed here since the results are identical. The main concern of interest is which algorithm does the best job at classifying the data with secondary interests including which algorithms run faster.

Description of the Data

This data is “HTRU2” which I pulled from Kaggle, and it comes from the High Time Resolution Universe Survey (R. J. Lyon). This is labeled data, containing mostly negative examples of 16,259 rows of features associated with candidates deemed not to be pulsar stars. About 9% of the examples, 1,639 are examples of pulsar stars. In this data set there are eight continuous variable features along with the class variable labeling it as a pulsar star or not. The first four variables are statistics about the integrated pulse profile (IP) data associated with a given possible star- referring to the mean, standard deviation, kurtosis, and skewness- while the last four characteristics are from the Dispersion Measure- Signal to Noise Ratio Curve (DM-SNR) which is considered the second best feature for classifying pulsar stars (Lyon).

This data set has been extensively covered by physicists in the interests of attempting to classify candidate pulsar stars ahead of time from the millions of candidate possibilities generated by an ever expanding explosion of data. The methods discussed in this current review of methods would likely fail on modern data streams by being too simplistic, ignoring the details of astrophysics, and returning far too many false positives. However, due to the domain, special attention will be paid to the false positive rate.

The table below shows some descriptive statistics of the eight variables of interest (each row corresponds to one of the variables) along with their mean and standard deviation for the whole data set and the data set split up by mean and standard deviation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Mean | SD | Mean + | SD + | Mean - | SD - |
| Mean IP | 111.1 | 25.65 | 56.69 | 30.01 | 116.6 | 17.48 |
| Standard Deviation IP | 46.55 | 6.843 | 38.71 | 8.034 | 47.34 | 6.183 |
| Excess Kurtosis IP | .4779 | 1.064 | 3.121 | 1.873 | .2104 | .3346 |
| Skewness IP | 1.770 | 6.168 | 15.55 | 14.00 | .3808 | 1.028 |
| Mean DM-SNR | 12.61 | 29.47 | 49.83 | 45.29 | 8.863 | 24.41 |
| Standard Deviation DM-SNR | 26.33 | 19.47 | 56.47 | 19.73 | 23.29 | 16.65 |
| Excess Kurtosis DM-SNR | 8.304 | 4.506 | 2.757 | 3.106 | 8.863 | 4.239 |
| Skewness DM-SNR | 104.9 | 106.5 | 17.93 | 50.90 | 113.6 | 106.7 |

We can see right away that the positive and negative examples may be separable based on these descriptive statistics of the incoming data since the columns for the means of the positive and negative examples look quite a bit different from each other.

Comparing Methods

I performed five methods on this data set. For most of the methods I attempted some amount of tuning as long as time permitted and measured the error rate as well as the amount of time needed to perform the method

The first method attempted was a support vector classifier, using the svm function in R and utilizing a linear classifier. In terms of length of run time this was the most involved method and parameter tuning wasn’t that large a priority because of the amount of time it took to do even one run would have meant that tuning over a grid of parameters to optimize would have taken a great deal of time, although scaling the data seems to have sped up the svm a great deal for some reason. I attempted to fit an svm with a quadratic boundary to classify the data as well but after leaving the code running for a very long time it still would not run.

The second method I attempted was K-Nearest-Neighbors. This method is one of the most intuitive machine learning methods where the use of majority voting is used by the nearest labeled methods. I applied tuning across a number of different values of k and found that between 1 and 10 a value of k=7 performed best.

The third method involved a simple logistic regression. For this we are simply fitting a general linear model which does not have hyperparameters in general. Future considerations may include applying LASSO or Ridge like shrinkage methods to the parameters in logistic regression. Looking at the coefficients all but 2 of the estimates were significant in the logistic regression model.

The fourth method involved using a neural network with one hidden layer. As this is a more complicated model it naturally takes a bit longer than logistic regression but because of the collapse to logistic regression in the case where there are 0 hidden layers I expected to do a little bit better. Initially the differences were not substantial, and just attempting a random size didn’t yield substantially different results. Tuning the size of the hidden layer led to a hidden layer of size 7 which did manage to make the neural network the best model in terms of false positive rates.

The fifth and final method involved fitting a tree-based method. An initial fit of a tree to the data yielded the same very simple two node split with whether Excess Kurtosis was above .66 (or .64 in one of the cross-validation methods, both of these are scaled values). This was a very simple split but yielded a surprisingly competitive model to the other methods explored. An attempt was made to tune the minimum split parameters yielding no new model. However, then the cost function for the amount needed in order to add a split was modified and substantially larger trees appeared.

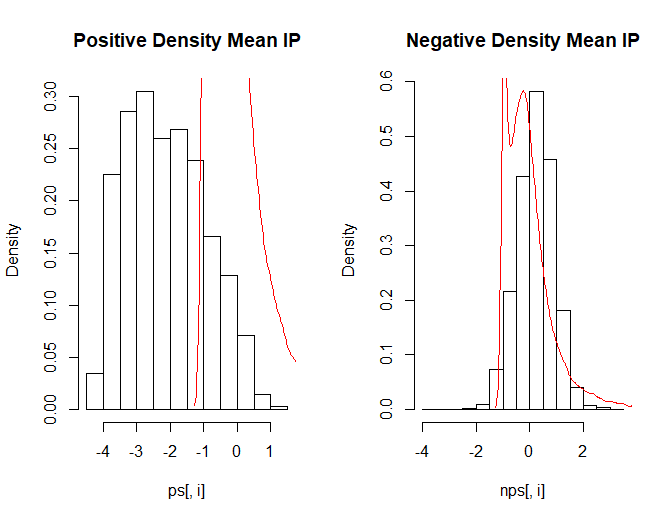
Here is a table summarizing the results:

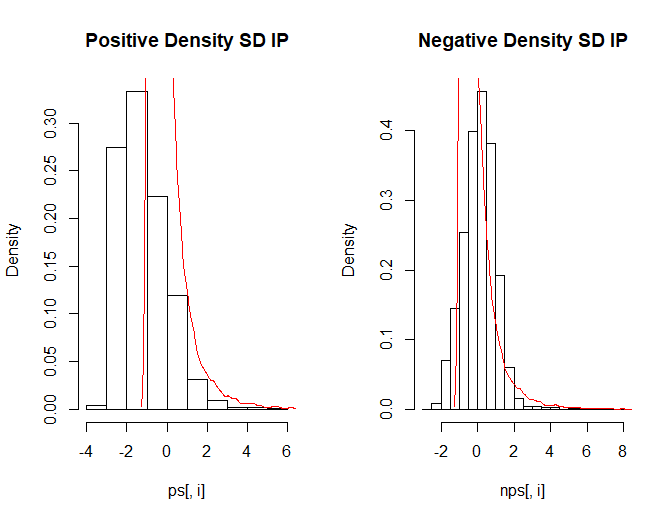
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method: | False Positive | False Negative | Overall Misclassification Rate | Time (seconds) |
| SVM | .0176 | .0050 | .0208 | 3.28 |
| KNN | .0176 | .0061 | .0217 | 1.61 |
| Logistic Regression | .0174 | .0054 | .0174 | .4 |
| Neural Network | .0140 | .0069 | .0191 | 13 |
| Tree Classification (Simple) | .0176 | .0067 | .0222 | .97 |
| Tree Classification (Complex) | .0142 | .0062 | .0187 | 1.81 |

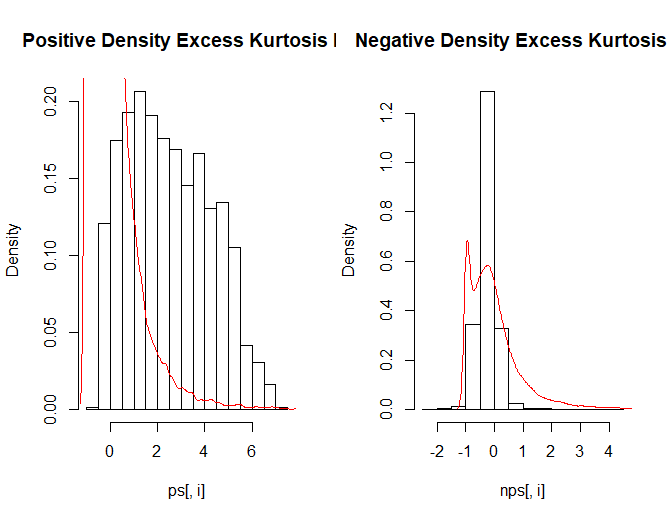
From the table above, the neural network takes by far the most amount of time and the complex tree classification seems to do the best job at predicting. Most of the methods have similar false positive rates of around 1.8% with the neural network and complex tree classifier doing a better job of around 1.4%. The overall error rate is also clustered around 2% for most of the methods with the logistic regression being the best performing method despite not doing as well on the false positive rate.

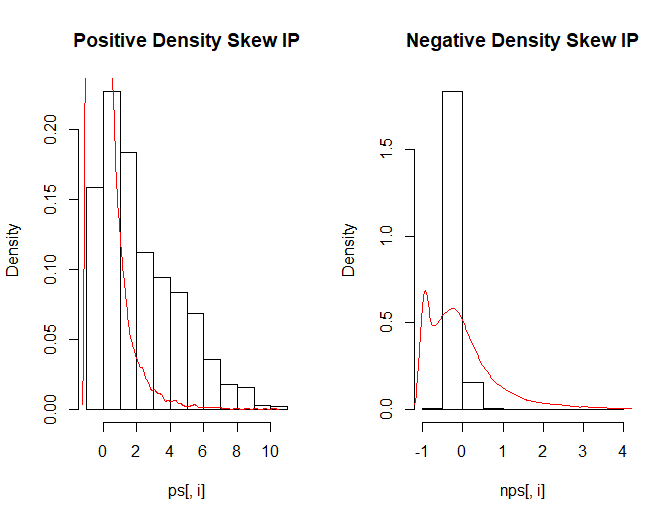
I would suggest that a tree based method may be the best especially given that in the reference “Why Are Pulsars Hard to Find?” the main method used to classify stars as pulsar stars is a tree based method. Since false positives are the main concern as well trees can easily be modified in order to create a larger cost for false positives than for false negatives to the potentially desired point on an ROC curve.

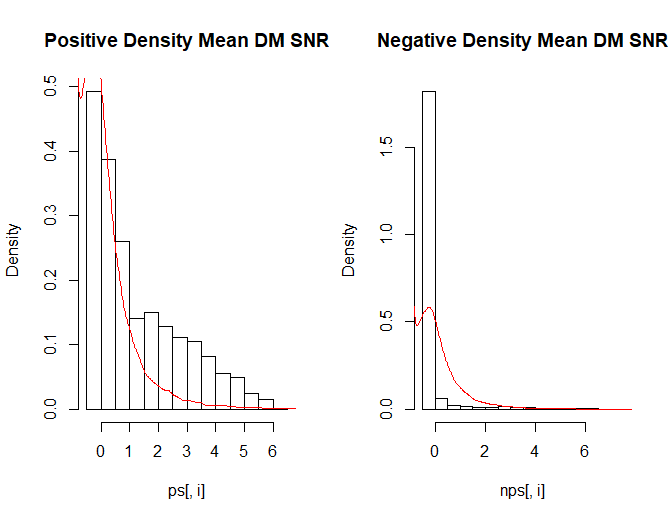
# Appendix: Density Graphs of Covariates

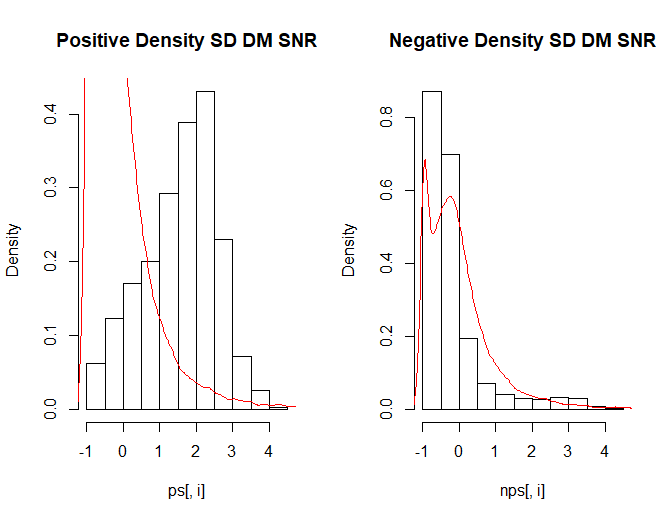


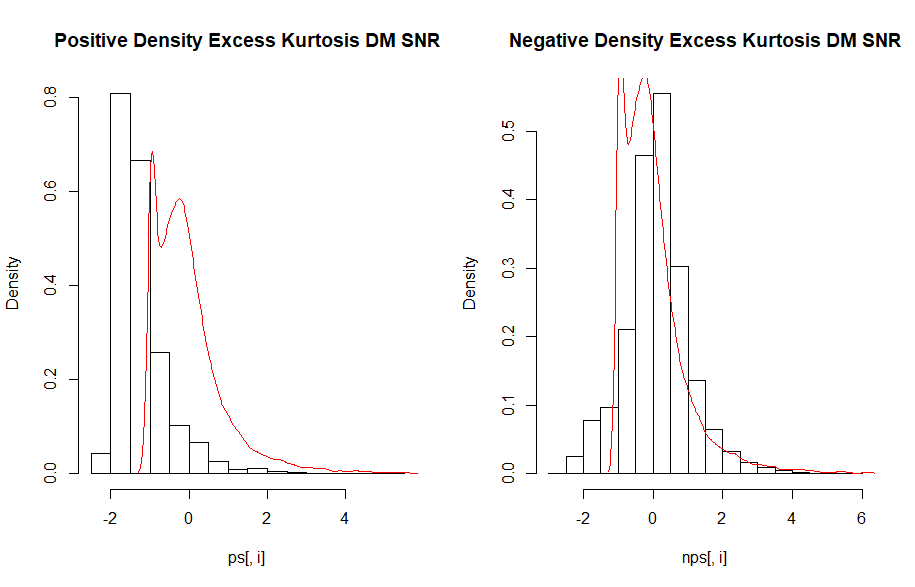


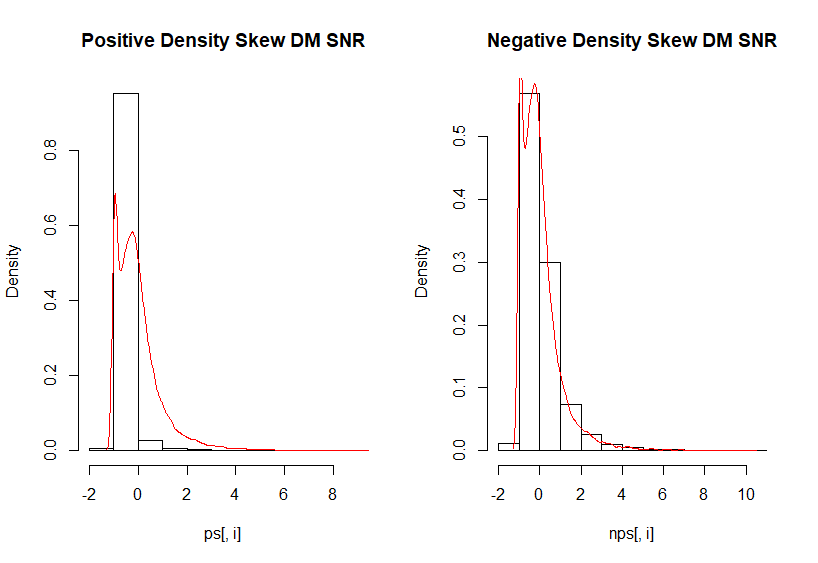












# References

Lyon, R. J. "'Why Are Pulsars Hard To Find?'." *PhD Thesis*. University of Manchester, 2016.

R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles. "Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach." *Monthly Notices of the Royal Astronomical Society* 459 (1) (2010): 1104-1123.