Sharik Purkar

Machine Learning

Assignment 4

Question 2

I decided to run Naïve Bayes, Linear Regression and Logistic regression on the SUSY dataset.

- The dataset was split into 7 sections and
- Each section then used 10 fold cross validation.
- Applied all three algorithms for different values of lamda
- Used the best accuracy from each section to perform the t test

The following are my results

Accuracy set : {'Naive Bayes': [79.16666666666666, 85.0, 83.33333333333334, 78.33333333333333, 83.33333333333334, 79.16666666666666, 80.0],

'Logistic Regression': [83.33333333333334, 85.0, 83.33333333333334, 81.66666666666667, 84.16666666666667, 85.0, 83.33333333333334],

'Linear Regression': [83.33333333333334, 85.0, 83.33333333333334, 82.5, 84.16666666666667, 85.0, 83.33333333333334]}

Metaparameter set : {'Naive Bayes': [1e-05, 1e-05, 1e-05, 1e-05, 1e-05, 1e-05, 1e-05],

'Logistic Regression': [1e-05, 1e-05, 1e-05, 1, 1e-05, 1e-05, 1],

'Linear Regression': [1e-05, 1e-05, 1e-05, 1e-05, 1e-05, 1e-05, 1]}

Mean Accuracy Linear : 83.8095238095

Mean Accuracy Logistic : 83.6904761905

Mean Accuracy Bayes : 81.1904761905

0.765306122449

1.16213151927

5.92403628118

statistic = 0.21004 and pvalue = 0.8373062147

statistic = 0.21004 and pvalue = 0.8371588563

The assignment asks for this experiment to conclude which algorithm is ultimately better but that can be a tricky answer considering how each one has its own particular merits and one dataset in one experiment can certainly not be the end all be all answer to which algorithm. That being said, in the narrow confines of this experiment, linear regression was clearly the winner. Linear Regression not only had the highest mean accuracy with 83.8095238095, but it also has the smallest variance of all three algorithms, consistently making good predictions. As mentioned previously, there are too may variables in play here to claim that this means linear regression is better in general, but it certainly was in this particular experiment.