Vibhav Jha

BE 700 HW1

**Part I. Various K-NN with 10 Fold Cross Validation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **K-NN** | **1** | **3** | **5** | **9** |
| **Correctly Classified %** | 87.3563 | 88.5057 | 83.908 | 86.2069 |
| **Incorrectly Classified %** | 12.6437 | 11.4943 | 16.092 | 13.7931 |
| **Mean Absolute Error** | 0.1357 | 0.179 | 0.2084 | 0.2179 |
| **Root Mean Squared Error** | 0.3513 | 0.3153 | 0.3348 | 0.3139 |
| **Relative Absolute Error %** | 32.9246 | 43.4125 | 50.5455 | 52.8564 |
| **Root Relative Squared Error %** | 77.5302 | 69.5854 | 73.8837 | 69.2712 |
| **ROC Area** | 0.892 | 0.949 | 0.942 | 0.971 |
| **Confusion Matrix** |  |  |  |  |

On this data set of 87 subjects, I was surprised to see 1-NN perform as well as it did. Likewise, I initially had thought 5-NN would perform the best in terms of correct classifications. While 5-NN did perform the worst in correct classifications, it had a lower RMSE. Theoretically, if sorted by RMSE, 9-NN performed the best, closely followed by 3-NN, then 5-NN and 1-NN. It makes sense that 9-NN would have the best RMSE as searching for 9 neighbors will make for the best fit for this specific data. However, with so many neighbors or essentially looking at 10 subjects each iteration, there is a risk of over-fitting or high specificity. Consistently each of the models were able to classify poor prognosis with a high true positive rate, with 3 and 9-NN having TPRs of 1 for poor prognosis. (Shown by the bottom left entry in the confusion matrix).

In terms of the experience, WEKA is very simple, and the fact that robust classifiers can be obtained with just a few clicks is remarkable.

**Part 2. T-Test**

Results from t-test available in attached CSV. Matlab code is also attached.

For the t-test I selected Matlab’s ttest2 function, which is the two-sample t-test. The alternative was to use the paired or one sample t-test function in matlab, however that required sets of equal size. In class we had hypothesized that there should be a correlation between neg log p values and the mean ratio. The points in the figure below don’t indicate a high correlation, but you can see the trend that in general the higher the log p value the lower the mean ratio, with a few outliers. For future models it would make sense to exclude values where the means were very similar.

