Benjamin Limoges
Assignment 6

1.

(a) If a feature is strongly positively correlated with the class label, naturally it is a feature we want to include in our model. But if a feature is strongly negatively correlated with the class label, it is equally as good as the positively correlated one. Imagine that there was a strongly positively correlated feature and a strongly negatively correlated feature. Both would have good predictive power of the class label, since the strongly negatively correlated feature would tell you to go in “the opposite direction.”

(b) For the filter method, the value of M that yielded the same accuracy was M=1, with 755 correct classification. This equated to an error rate of 71.6%. This was for the non-normalized dataset (my classification was far more accurate on normalized data.)

(c) The main advantage in my implementation is that it is more computationally tractable than the wrapper method. While it did take longer in run time than the wrapper method (although barely), this was merely a coincidence of both the dataset and my own (poor) implementation. In general, the filter method will be quicker than the wrapper method, and will quickly find features that will strongly predict the class type.

2.

(a) My final accuracy over the final set of selected features is 1039/1055 or approximately a 98.5% accuracy. It does this over a small number of features (3).

(b) The advantage of the wrapper method is that it uses a much smaller number of features than the filter method. With the wrapper method, my accuracy is higher than the filter method and my number of features is much more reduced. The smaller feature set makes it faster to test my data. In addition, it is implicitly checking for joint predictive power; while the filter method uses a correlation matrix to determine the best predictors one on one, the wrapper method allows for there to be joint correlation. There may be cases where the data needs both variables (both of which are weakly correlated) to have true predictive power. A trivial example is a batter in baseball needs both bat speed and base running speed to be an effective batter. Both are weakly correlated with runs earned, but together they are one of the strongest predictors of success in MLB.

The main disadvantage is that it is extremely computationally intensive, since it is exhaustively searching for the “optimal” features to include in the model. The filter method's down and dirty use of correlation as a proxy makes it quicker to determine “good” variables.

3.

(a) My method was to somewhat combine the two methods. To make the calculation quicker (especially since my run time was so prohibitively long), I thought that reducing the feature space would help cut down on computation time. So I reduced the feature space to those features that had a correlation of 0.3 or higher in absolute value. I chose this cut off value from standard statistical theory, that uses 0.3 as a heuristic for moderate correlation.
After limiting the features, I used the wrapper method since it generally produces better results. I was hoping that my methodology would give me a "best of both worlds."

(b) Ultimately it only included two features, {2, 7}. (Note: I have 1 indexed the features as they were in the original ARFF file). This was surprising, not because it had a small number of features, but because there was no overlap with the pure wrapper method's features.

(c) There were 865 correct classifications out of 1055 possible. This translates to a 82.0% accuracy, which beats the filter method's run.