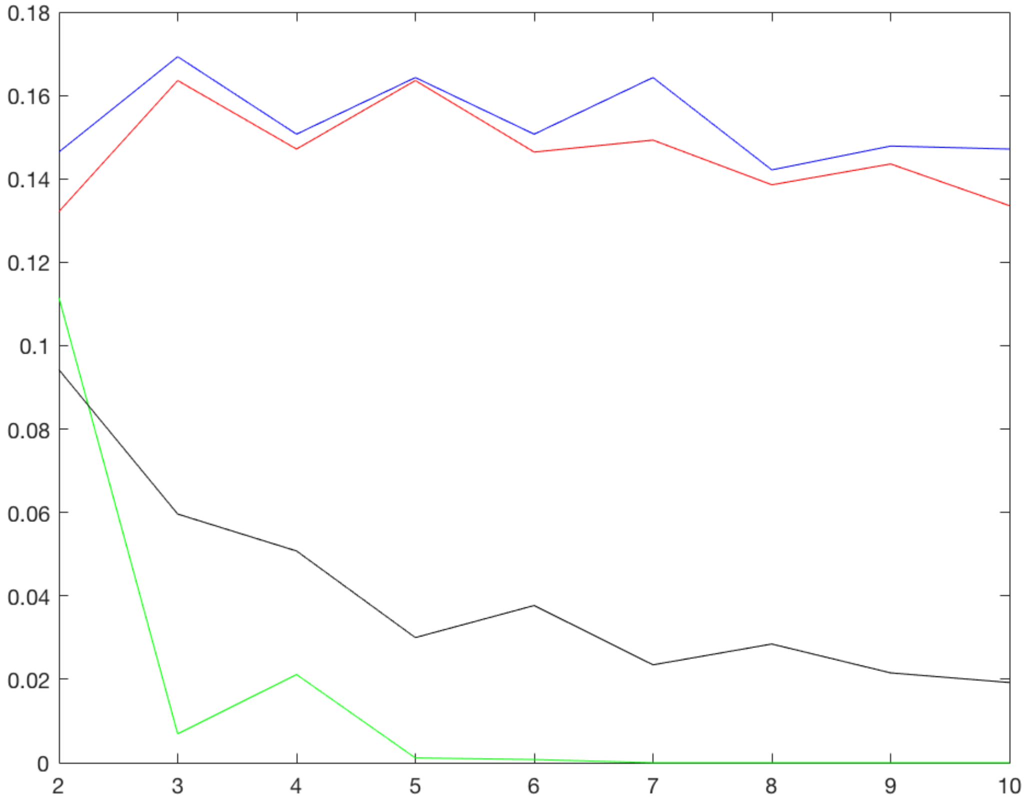
Dominic DiPasquale
CS1567 Introduction to Machine Learning
Professor Milos Hauskrecht

Problem Assignment 10

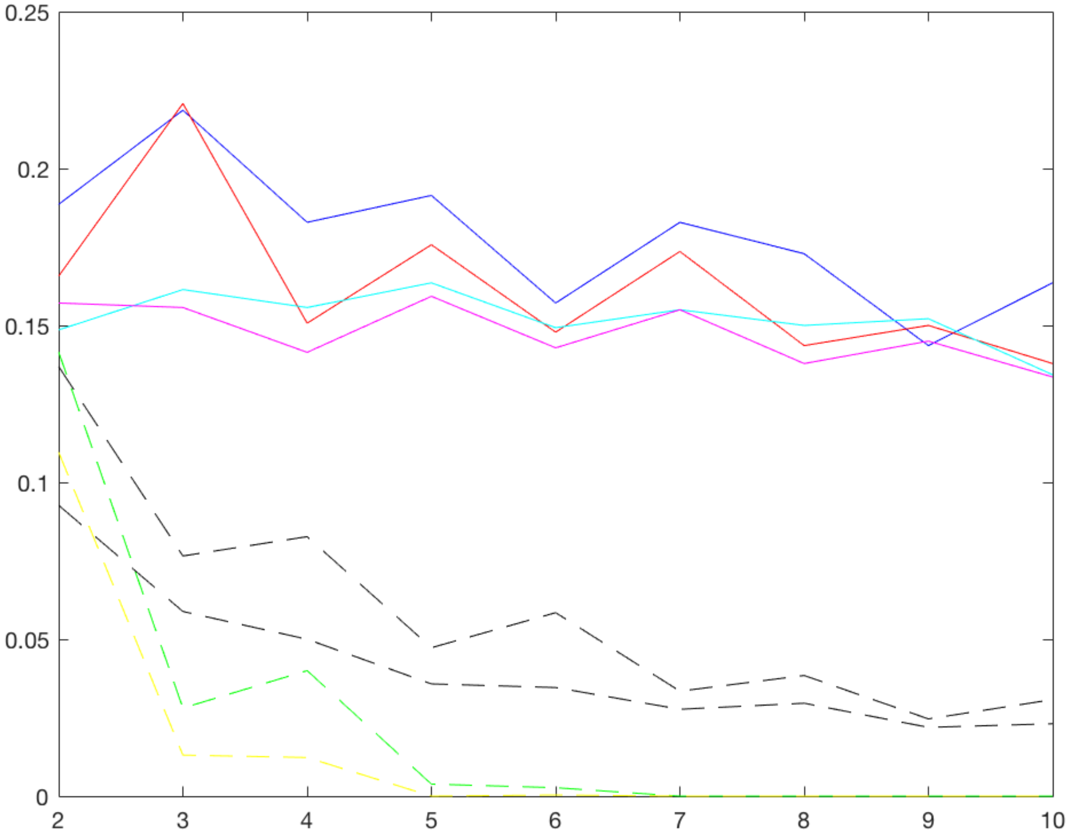
Problem 1



a)

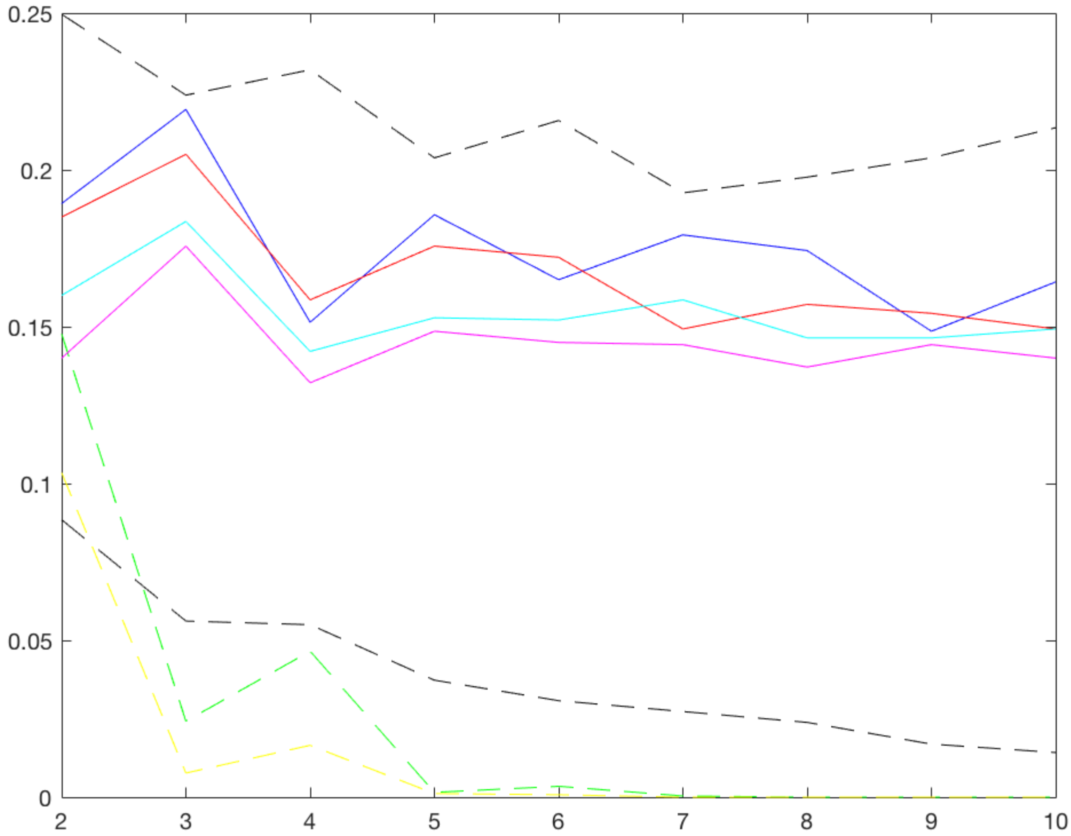
In the above graph the top two lines are the test boosting and bagging errors, the bottom to lines are the boost and bag training. The X axis represents the T value and the Y value represents the error. As expected we see that the training sets have much lower error in both cases. The test sets in general achieve a lower error as you increase the value of T.

b)



In the above graph all of the training set errors are dotted lines, as we expect they are considerably lower than any of the test sets hinting that perhaps we have done a little bit of overfitting. The blue line is the boost algorithm run with dt_full, while it follows the same pattern as all the others in this specific example it preformed the worst. It should be noted however that the performance difference seems to be negligible. The red line is the bagging algorithm run with the dt_full, it preforms very comparably to the SVM_base bag and boost runs that we saw in part a.

c)



As we can see again with this graph the SVML base is pretty consistent for both the train and testing sets in the same places as before. The difference is obviously the massive difference in the training set for the dt_simple. It has jumped from below all of the testing set runs to above them. This is a considerable change in average error. Surprisingly it does not seem to effect the testing set as much, perhaps implying that the simple dt implementation is as bad as the training set would suggest. It also might be that it is hard to overfit when you are only given 1 split (i.e. you have a simple model), but when using a boosting or bagging a simple model can be comparable to a more complex model without bagging.

Problem 2

a) The top 20 fisher scores are:

```
0.3192   48.0000
0.2140   25.0000
0.1910   21.0000
0.1892   70.0000
0.1693   65.0000
0.1673   40.0000
0.1650   29.0000
0.1402   19.0000
0.1255   57.0000
0.1212   20.0000
0.0995   24.0000
0.0950   30.0000
0.0858   12.0000
```

```
   0.0846   47.0000
   0.0607   61.0000
   0.0579   10.0000
   0.0527   34.0000
   0.0462   27.0000
   0.0461   39.0000
   0.0422   41.0000
b) The top 20 ROC scores are:
   0.7340   25.0000
   0.6837   29.0000
   0.6695   11.0000
   0.6661   47.0000
   0.6315   19.0000
   0.6174   34.0000
   0.6021   32.0000
   0.6021   30.0000
   0.6000    9.0000
   0.5971   56.0000
   0.5953   27.0000
   0.5929   60.0000
   0.5881   51.0000
   0.5874   26.0000
   0.5845   53.0000
   0.5797    7.0000
   0.5709   10.0000
   0.5686   61.0000
   0.5567   43.0000
   0.5422   44.0000
```