Raymond Liu and Kenneth Gwon

Lab 6 Analysis

Experiment 1 Analysis

1) News Dataset

Using the support vector machine model on the news dataset, the average test accuracy for the 5 folds was 49.6%.

Using the random forest model on the news dataset, the average test accuracy for the 5 folds was 60.3%

Support Vector Machine Accuracy on News Dataset

Fold Number	Test Accuracy
1	0.471429
2	0.492683
3	0.545000
4	0.481865
5	0.489583

Random Forest Accuracy on News Dataset

Fold Number	Test Accuracy
1	0.614286
2	0.595122
3	0.595000
4	0.601036
5	0.609375

We obtained a p-value of 0.0023 which is less than 0.05 thus we reject the null hypothesis that the methods have similar generalization error.

MNIST Dataset

Using the support vector machine model on the MNIST dataset, the average test accuracy for the 5 folds was 91.4%.

Using the random forest model on the MNIST dataset, the average test accuracy for the 5 folds was 90.4%

Support Vector Machine Accuracy on News Dataset

Fold Number	Test Accuracy
1	0.902439
2	0.896040
3	0.920000

4	0.923858
5	0.928571

Random Forest Accuracy on News Dataset

Fold Number	Test Accuracy
1	0.897059
2	0.896552
3	0.910000
4	0.908629
5	0.908163

We obtained a p-value of 0.0048 which is less than 0.05 thus we reject the null hypothesis that the methods have similar generalization error.

2) For the news dataset, we see that the results using the random forest model were slightly better because they produced an average test accuracy of 60.3% compared to 49.6% for the support vector machine model. From this we can conclude that random forest might be slightly better for this dataset. The test accuracy for both models was high considering the news dataset contains 20 possible classes.

   When we look at the MNIST dataset, we see the results using the SVM model were slightly better than the random forest model as the SVM model produced an average test accuracy of 91.4% whereas the random forest model produced an average test accuracy of 90.4%. However, the test accuracy for both models was very similar. From this we can conclude both models performed well considering they produced average test accuracy of over 90% when there were 10 possible classes.

3) For the random forest model run on the news dataset, the best hyperparameter for max_features was 0.1. All five folds had max_features equal to 0.1.

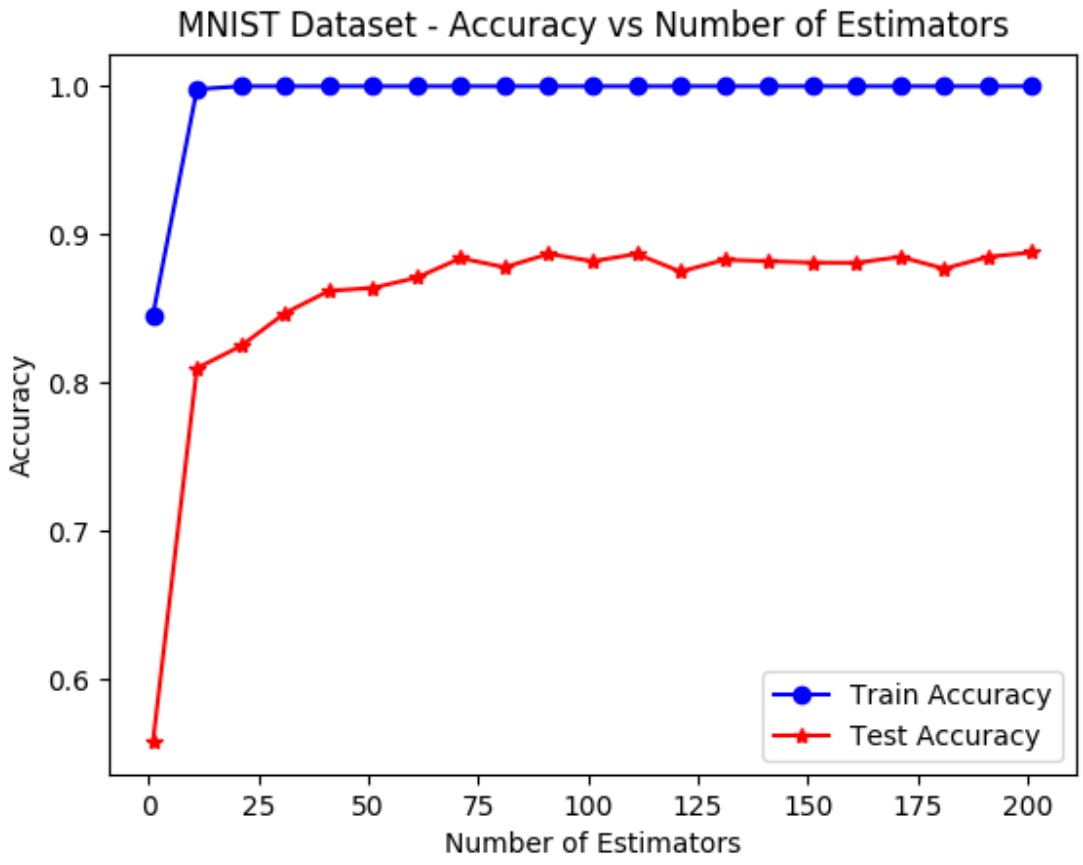
   For the support vector machine model run on the news dataset, the best hyperparameter for C was C = 100 for two of the folds and C = 1000 for three of the folds. The best value for gamma was .01 for three of the datasets and .1 for two of the datasets.

   For the random forest model run on the MNIST dataset, the best value for max_features was unclear as the best value was max_features = sqrt for three of the datasets and max_features = 0.1 for two of the datasets.

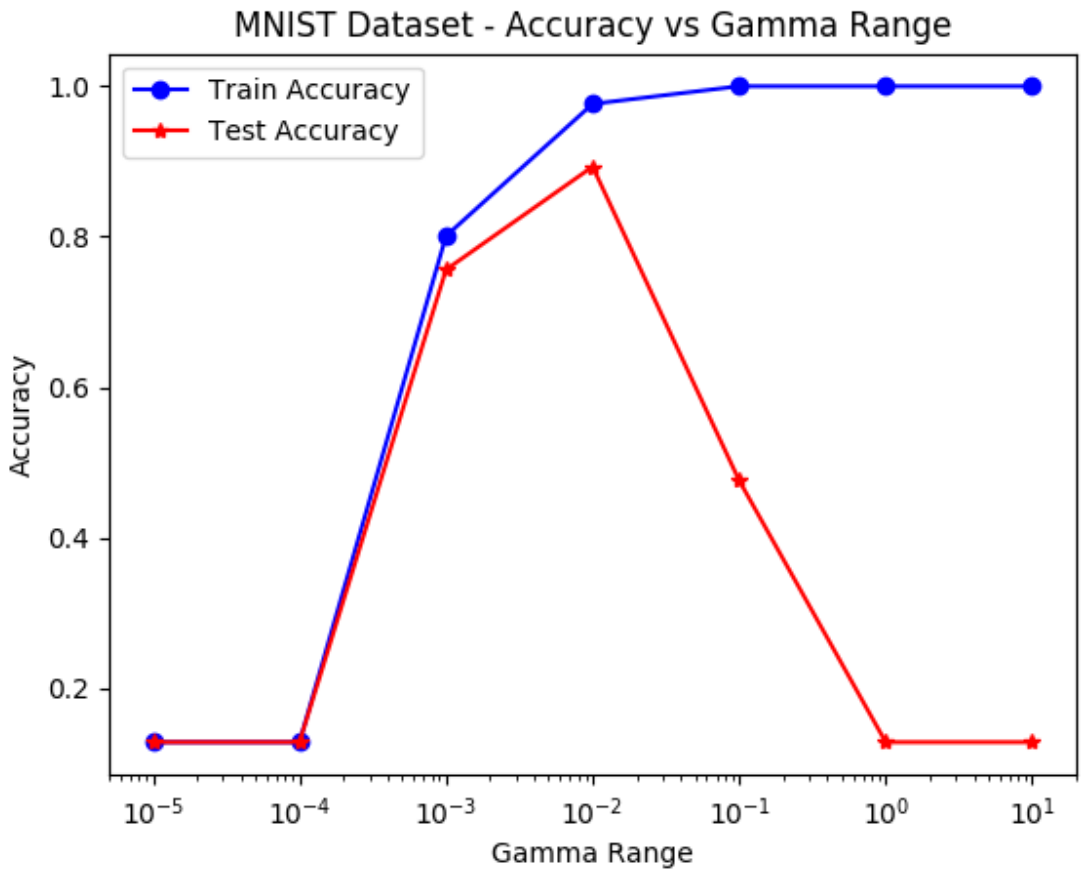
   For the support vector machine model run on the MNIST dataset, the best value for C is 10 as four of the five folds had C = 10 while the other fold had C = 1. The best value for gamma was .01 as all five folds had gamma = .01

Experiment 2 Analysis

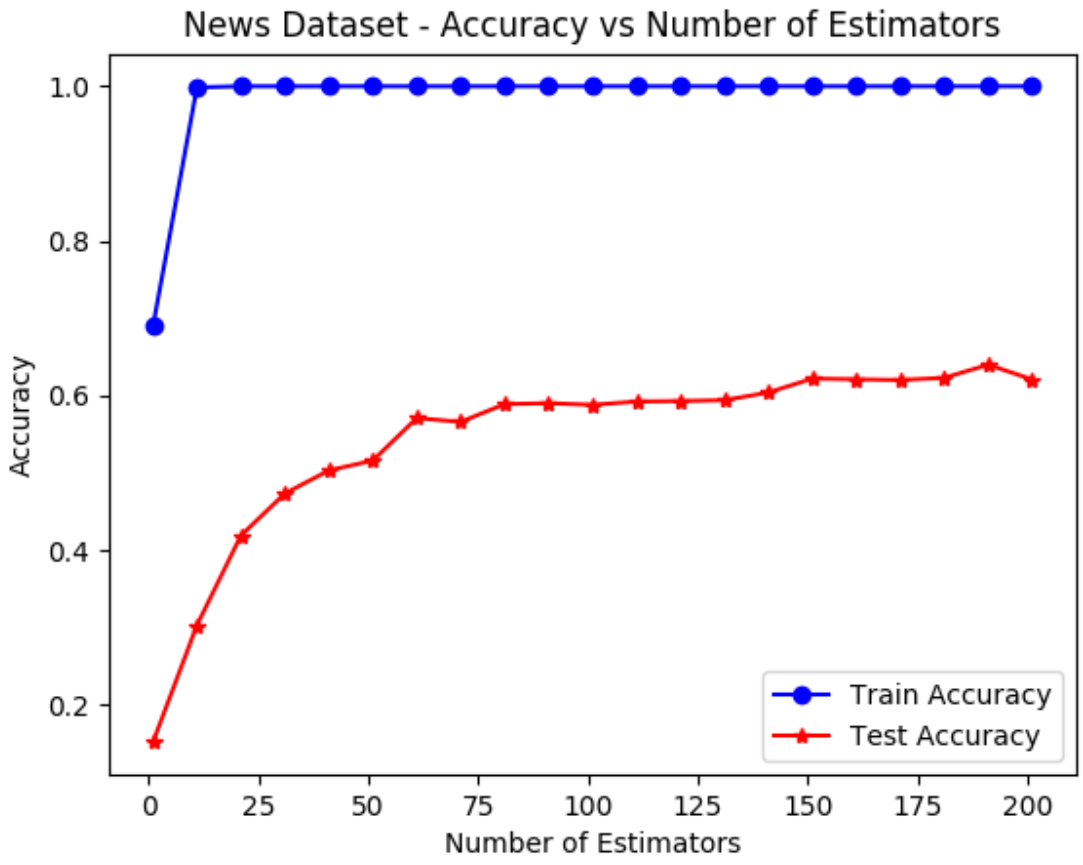
For the following graph the accuracy using the random forest model on the MNIST dataset, we see that as the number of estimators increases, our accuracy increases then levels off. The model has low bias as it as able to achieve around 90% accuracy but it has high variance as the difference between the train accuracy and test accuracy is nearly 10%. The model never overfits as the test accuracy never peaks it stays about the same or even slightly increases as the number of estimators increases.



For the following graph depicting the accuracy of the SVM model on the MNIST dataset, we see that bias is high for gamma = 10^{-5} and 10^{-4} as test accuracy is lower than 20% but for gamma = 10^{-3} and 10^{-2}, we have low bias because test accuracy is around 80%. Then as gamma increases, we have high bias again because test accuracy decreases. So as gamma increases, bias is first high, becomes low, the becomes high again. We can also see variance increases as gamma increases. We can measure variance as the distance between the train accuracy and the test accuracy. As we increase gamma, the distance between train accuracy and test accuracy increases meaning the model does not adapt do different datasets well as gamma increases. We also see that we have overfitting for gamma values greater than 10^{-2} as our test accuracy starts to decrease. We also have underfitting for gamma values of 10^{-4} or less as we see a huge spike in test accuracy after gamma = 10^{-4}.



For the following graph depicting the accuracy of the random forest model on the news dataset, we see bias decreases as the number of estimators increases and then levels off because our test accuracy increases then levels off. The variance decreases slightly as the number of estimators increases as the distance between train accuracy and test accuracy slightly decreases. We do not over fit as test accuracy is always increasing.



For the following graph depicting the accuracy of the SVM model on the news dataset, the model has high bias for all gamma values other than 10^0 as accuracy is under 20%. However for gamma = 10^0 accuracy is about 40% so the model has low bias. As gamma increases, variance increases as distance between train accuracy and test accuracy increases meaning the model does not generalize as well. We have underfitting for values less than or equal to 10^{-1} and overfitting for values greater than or equal to 10^1.