Ryan Vera

CS 6375

Project 2 Report

1.) The Gradient Boosting decision tree yield the highest average score across all 15 data sets. This high performance can be attributed to the nature of the classifier. The scheme of adding weights to misclassified data points allows the decision tree to be flexible to its training data with introducing too much variance.

2.) Increasing the number of training examples improved the performance of all the classifiers by approximately 10 points between each number of examples: 100, 1000, 5000. Specifically, the basic decision tree saw an average of 11-point increase between the smallest and largest datasets. The Bagging Classifier saw an average increase of 13 points in accuracy between the smallest and largest datasets. The Random Forest Classifier saw an average increase of 8 points in accuracy between the smallest and largest dataset. The Gradient Boosting Classifier saw an average increase of 7 points in accuracy between the smallest and largest datasets.

3.) As the number of clauses increased, the accuracy of all the classifiers increased as well. This may not be expected since many clauses implies a higher complexity. However, this makes a decision tree is a good choice in classifiers of a problem space like this one. Decision trees tend to be very robust to complexity.

4.) The highest accuracy scoring classifier on the MNIST dataset was the Random Forest Classifier. This is because random forests have low bias, while still avoiding a high variance. This allows the random forest to be a strong learner, and still generalize well enough to perform highly on test data.

On the next pages, there are result tables for all the hyperparameters for each classifier and each dataset. The excel tables of the data below are in “classification_results.xlxs”.

# Decision Tree Classifier Results

| Data Set No | GridSearch DT | | | | | | |
|---|---|---|---|---|---|---|---|
| | Criterion | Max Depth | Random State | Splitter | Accuracy Score | F1 Score | Precision |
| 1 | gini | 6 | 7 | random | 0.66 | 0.676 | 0.645 |
| 2 | entropy | 6 | 5 | random | 0.686 | 0.706 | 0.664 |
| 3 | gini | 8 | 8 | best | 0.779 | 0.799 | 0.732 |
| 4 | entropy | 6 | 5 | random | 0.635 | 0.64 | 0.631 |
| 5 | gini | 6 | 5 | best | 0.706 | 0.719 | 0.689 |
| 6 | entropy | 9 | 5 | best | 0.793 | 0.804 | 0.763 |
| 7 | gini | 6 | 6 | random | 0.695 | 0.711 | 0.675 |
| 8 | entropy | 7 | 6 | best | 0.813 | 0.826 | 0.769 |
| 9 | entropy | 10 | 6 | random | 0.856 | 0.862 | 0.827 |
| 10 | gini | 6 | 5 | best | 0.855 | 0.861 | 0.825 |
| 11 | entropy | 6 | 5 | random | 0.92 | 0.922 | 0.894 |
| 12 | entropy | 13 | 8 | best | 0.953 | 0.954 | 0.942 |
| 13 | entropy | 6 | 9 | best | 0.895 | 0.901 | 0.9 |
| 14 | gini | 8 | 6 | best | 0.974 | 0.974 | 0.96 |
| 15 | entropy | 11 | 7 | best | 0.983 | 0.983 | 0.976 |
| | | | | Avg | 0.814 | 0.823 | 0.793 |

## Bagging Tree Classifier Results

| | GridSearch BDT | | | | | |
|---|---|---|---|---|---|---|
| Data Set No | Bootstrap | Boostrap Features | Max Samples | Accuracy Score | F1 Score | Precision |
| 1 | TRUE | FALSE | 80 | 0.655 | 0.555 | 0.782 |
| 2 | TRUE | FALSE | 750 | 0.703 | 0.674 | 0.75 |
| 3 | TRUE | FALSE | 4950 | 0.814 | 0.807 | 0.838 |
| 4 | TRUE | FALSE | 90 | 0.665 | 0.659 | 0.67 |
| 5 | TRUE | FALSE | 640 | 0.765 | 0.75 | 0.801 |
| 6 | TRUE | FALSE | 4850 | 0.857 | 0.851 | 0.888 |
| 7 | TRUE | FALSE | 70 | 0.805 | 0.794 | 0.843 |
| 8 | TRUE | FALSE | 650 | 0.9 | 0.898 | 0.913 |
| 9 | TRUE | FALSE | 4750 | 0.93 | 0.929 | 0.94 |
| 10 | TRUE | FALSE | 70 | 0.94 | 0.942 | 0.915 |
| 11 | TRUE | FALSE | 640 | 0.967 | 0.966 | 0.982 |
| 12 | TRUE | FALSE | 4950 | 0.983 | 0.983 | 0.989 |
| 13 | TRUE | FALSE | 50 | 0.955 | 0.956 | 0.933 |
| 14 | TRUE | FALSE | 220 | 0.987 | 0.987 | 0.987 |
| 15 | TRUE | FALSE | 4850 | 0.993 | 0.993 | 0.996 |
| | | | Avg | 0.861 | 0.850 | 0.882 |

## Random Forest Classifier Results

| | GridSearch RFT | | | | | |
|---|---|---|---|---|---|---|
| Data Set No | Criterion | Max Dpeth | Max Samples | Accuracy Score | F1 Score | Precision |
| 1 | gini | 12 | 71 | 0.7 | 0.7 | 0.696 |
| 2 | entropy | 5 | 951 | 0.84 | 0.843 | 0.829 |
| 3 | gini | 7 | 2751 | 0.89 | 0.893 | 0.867 |
| 4 | gini | 6 | 71 | 0.795 | 0.798 | 0.786 |
| 5 | entropy | 5 | 751 | 0.933 | 0.934 | 0.912 |
| 6 | entropy | 7 | 4001 | 0.933 | 0.934 | 0.921 |
| 7 | entropy | 12 | 81 | 0.955 | 0.954 | 0.96 |
| 8 | gini | 11 | 451 | 0.989 | 0.988 | 0.988 |
| 9 | gini | 13 | 2501 | 0.993 | 0.993 | 0.992 |
| 10 | gini | 5 | 51 | 1 | 1 | 1 |
| 11 | gini | 5 | 351 | 0.999 | 0.999 | 0.999 |
| 12 | gini | 13 | 3501 | 0.999 | 0.999 | 0.999 |
| 13 | gini | 5 | 11 | 0.985 | 0.985 | 0.97 |
| 14 | gini | 5 | 51 | 1 | 1 | 1 |
| 15 | gini | 5 | 501 | 1 | 1 | 1 |
| | | | Avg | 0.934 | 0.935 | 0.928 |

Gradient Boosting Classifier Results

| | GridSearch GBT | | | | | |
|---|---|---|---|---|---|---|
| Data Set No | Criterion | Learning Rate | Loss | Accuracy Score | F1 Score | Precision |
| 1 | Squared Err | 0.6 | Exponential | 0.81 | 0.817 | 0.787 |
| 2 | Friedman MSE | 0.3 | Log Loss | 0.992 | 0.992 | 0.983 |
| 3 | Friedman MSE | 0.4 | Log Loss | 0.997 | 0.997 | 0.995 |
| 4 | Friedman MSE | 0.7 | Log Loss | 0.89 | 0.891 | 0.882 |
| 5 | Friedman MSE | 0.4 | Exponential | 0.995 | 0.995 | 0.991 |
| 6 | Friedman MSE | 0.5 | Exponential | 0.998 | 0.998 | 0.996 |
| 7 | Friedman MSE | 0.8 | Log Loss | 0.975 | 0.975 | 0.979 |
| 8 | Friedman MSE | 0.7 | Exponential | 0.997 | 0.997 | 0.994 |
| 9 | Squared Err | 0.4 | Log Loss | 0.999 | 0.999 | 0.999 |
| 10 | Friedman MSE | 0.6 | Log Loss | 1 | 1 | 1 |
| 11 | Friedman MSE | 0.3 | Log Loss | 1 | 1 | 1 |
| 12 | Friedman MSE | 0.8 | Log Loss | 1 | 1 | 1 |
| 13 | Friedman MSE | 0.3 | Exponential | 0.995 | 0.995 | 0.99 |
| 14 | Friedman MSE | 0.1 | Log Loss | 1 | 1 | 1 |
| 15 | Friedman MSE | 0.2 | Log Loss | 1 | 1 | 1 |
| | | | Avg | 0.977 | 0.977 | 0.973 |