Collin Gros
03-10-2021

HOMEWORK 4

AdaBoost ALGORITHM DETAILED CALCULATION/FINAL RESULTS

Index	**x**	**y**	**weights**	**yhat**	**Updated weights**
1	1.0	1	0.072	-1	**0.15**
2	2.0	1	0.072	1	**0.0468**
3	3.0	1	0.072	1	**0.0468**
4	4.0	-1	0.072	-1	**0.0468**
5	5.0	-1	0.072	-1	**0.0468**
6	6.0	-1	0.072	-1	**0.0468**
7	7.0	1	0.167	1	**0.11**
8	8.0	1	0.167	-1	**0.35**
9	9.0	1	0.167	1	**0.11**
10	10.0	-1	0.072	-1	**0.0468**

2.

c. Error rate = (0.072, 0.072, 0.072, 0.072, 0.072, 0.072, 0.167, 0.167, 0.167, 0.072) (1, 0, 0, 0, 0, 0, 0, 1, 0, 0)**^T** = 0.072 + 0.167 = 0.239

d. coefficient (alpha) = 0.5ln([1 - 0.239]/0.239) = 0.579

e. updated weights = [
0.072 * e^(-0.579 * 1 * -1) = 0.128,
0.072 * e^(-0.579 * 1 * 1) = 0.04,
0.072 * e^(-0.579 * 1 * 1) = 0.04,
0.072 * e^(-0.579* -1 * -1) = 0.04,
0.072 * e^(-0.579 * -1 * -1) = 0.04,
0.072 * e^(-0.579 * -1 * -1) = 0.04,
0.167 * e^(-0.579 * 1 * 1) = 0.094,
0.167 * e^(-0.579 * 1 * -1) = 0.298,
0.167 * e^(-0.579 * 1 * 1) = 0.094,
0.072 * e^(-0.579 * -1 * -1) = 0.04
]

f. normalized weights = updated weights / sum(updated weights) = updated weights / 0.854 = [

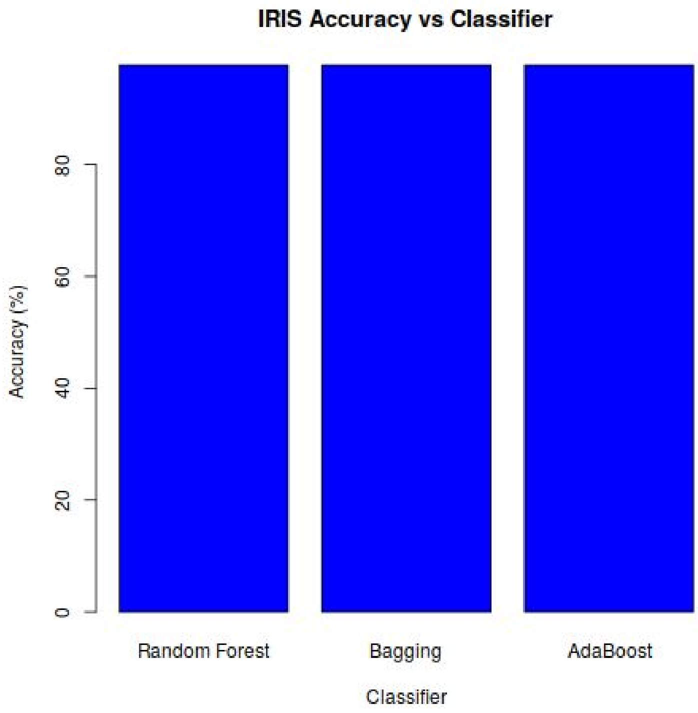
0.15
0.0468
0.0468
0.0468
0.0468
0.0468
0.11
0.35
0.11
0.0468

]

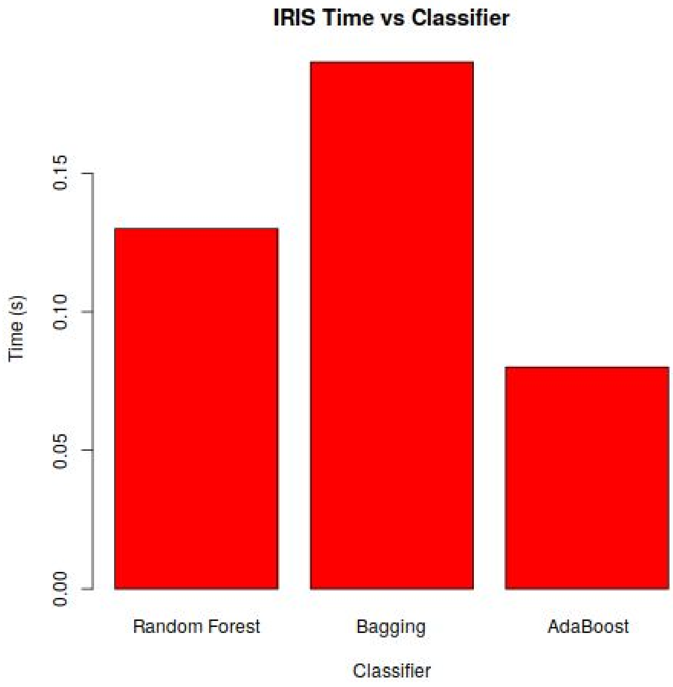
SCIKIT-LEARN ENSEMBLE PERFORMANCE AND ANALYSIS

All datasets were trained and tested on split data from the *IRIS (SCIKIT-LEARN), DIGITS (SCIKIT-LEARN),* and *Chronic Kidney Disease (UCI repository)* data sets. Some data in the *Chronic Kidney Disease* dataset was pre-processed for use with the algorithms: any nominal value (good/poor, yes/no) was replaced with (1, -1). Any '?' value was dropped from the data. Following are graphs showcasing the accuracy and time values each classifier obtained from training/testing on the above three datasets:

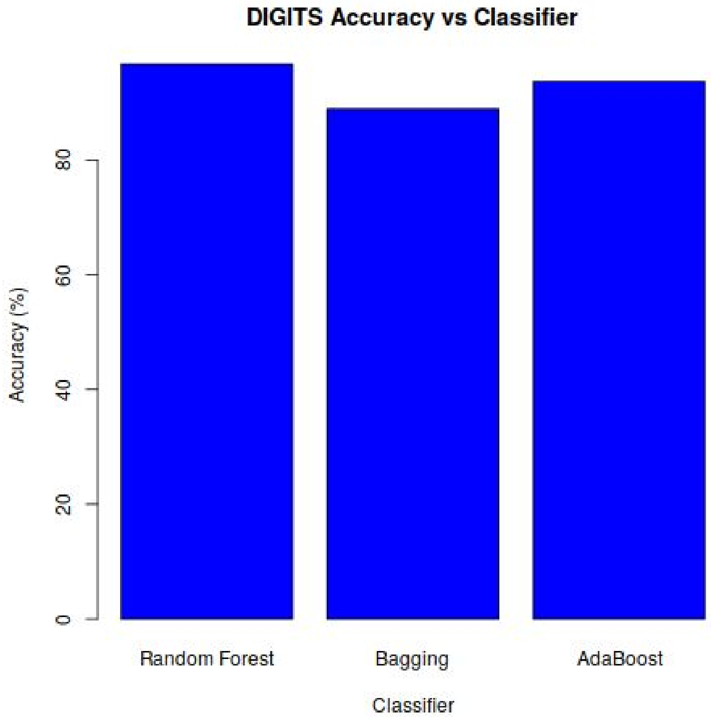
IRIS ACCURACY:



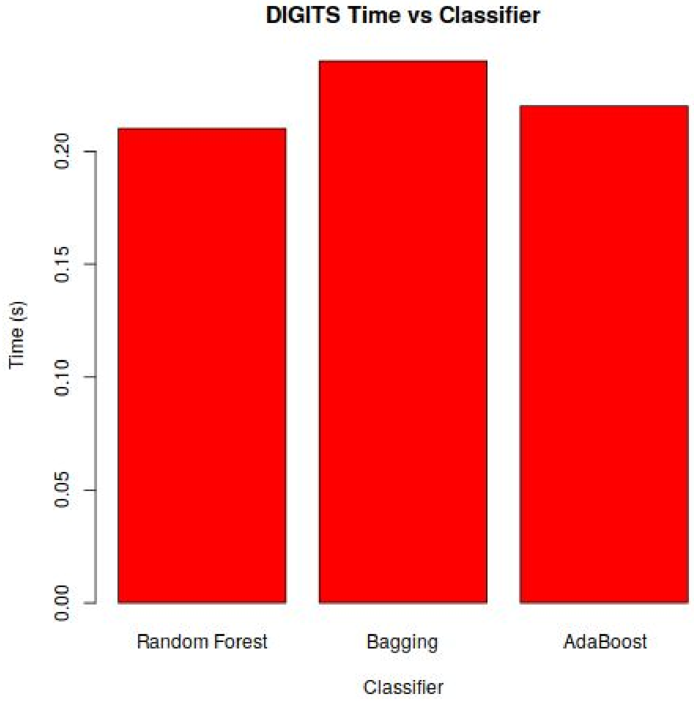
IRIS TIME:



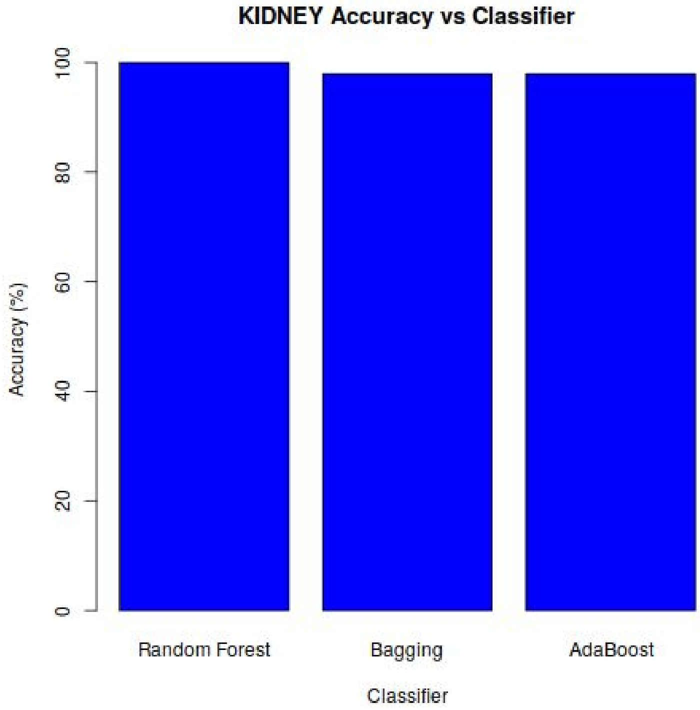
DIGITS ACCURACY:



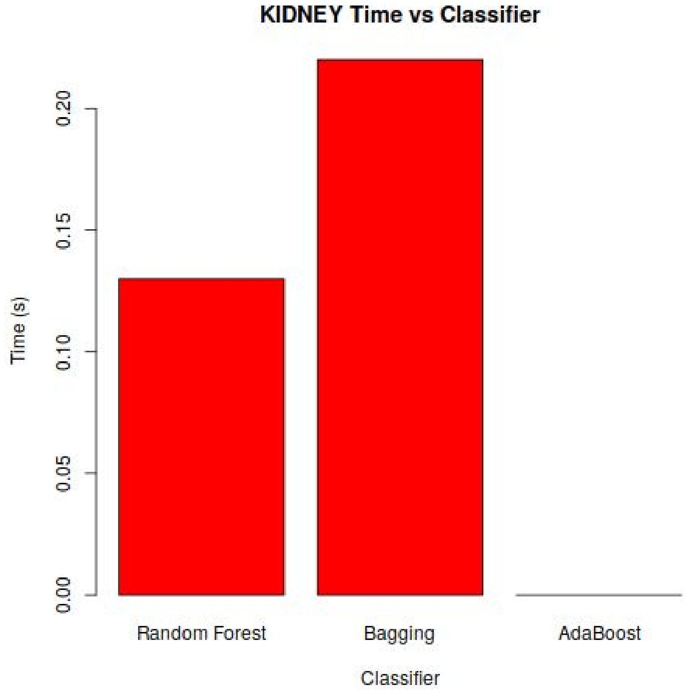
DIGITS TIME:



KIDNEY ACCURACY:



KIDNEY TIME:



ANALYSIS

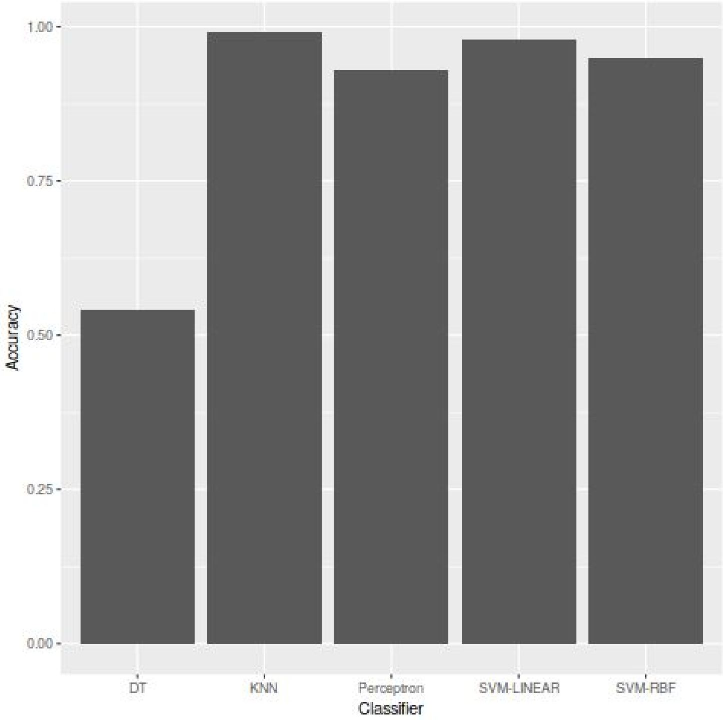
For comparison between the ensemble approaches and the base classifiers, I have included the accuracy and time graphs for the base classifiers below.

From the DIGITS Time data, it is clear that the ensemble approaches used more than twice as much time as all the base classifiers (besides SVM-RBF). From an accuracy perspective, using the DIGITS Accuracy data, it seems that the ensemble approaches performed nearly the same as the base classifiers. There are some differences, as different default values were used (e.g., 'criterion' was switched from 'gini' to 'entropy', though the differences are small.

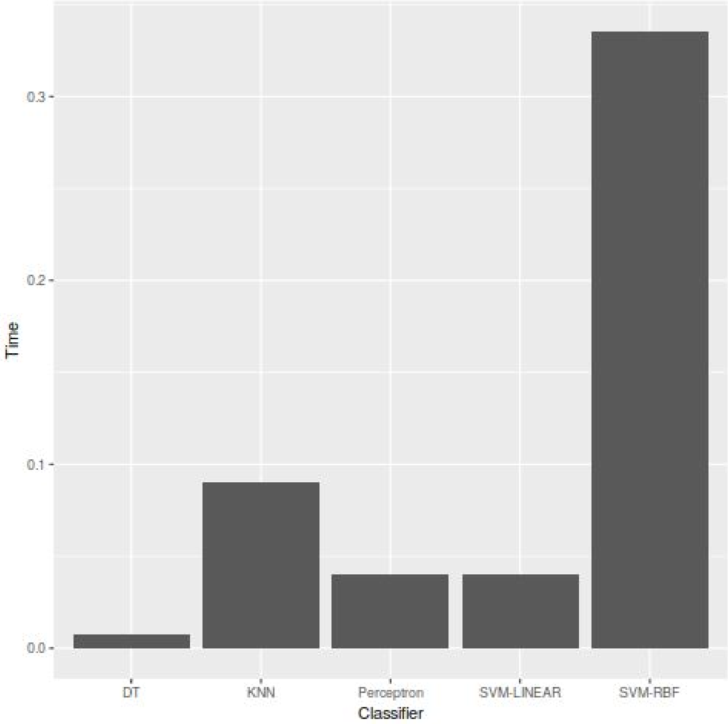
After testing different default values for AdaBoost, I noticed that none of the changes really had any affect on the time or the accuracy for analysis with the Kidney database. It seems like it quits prematurely as it has found a good set of weights almost immediately.

Isolating the ensemble approach results, Bagging took the longest to complete in all cases, and was also less accurate than the other two classifiers. AdaBoost quickly exited during the Kidney dataset test, but achieved a high accuracy score. In all cases except the Digits case, AdaBoost was the fastest, followed by Random Forest.

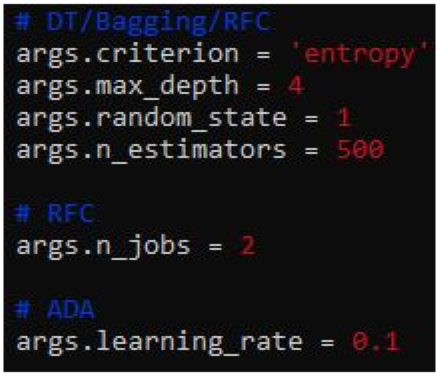
ACCURACY-DIGITS (BASE CLASSIFIERS):



TIME-DIGITS (BASE CLASSIFIERS):



DEFAULT VALUES USED WITH ENSEMBLE CLASSIFIERS:

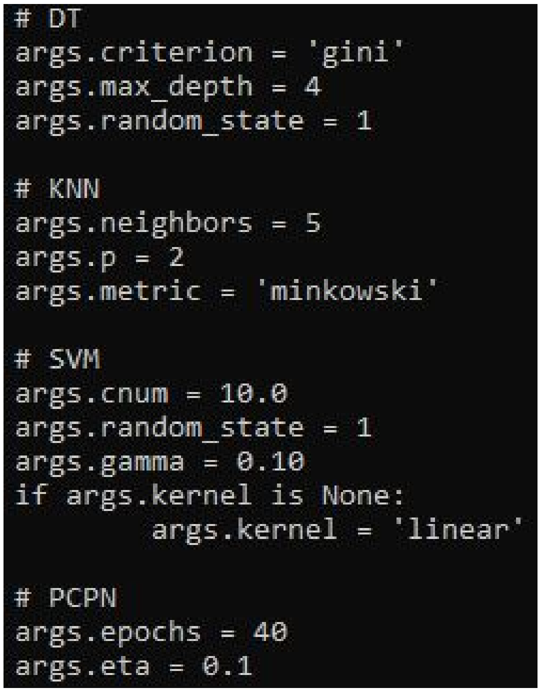


```
# DT/Bagging/RFC
args.criterion = 'entropy'
args.max_depth = 4
args.random_state = 1
args.n_estimators = 500

# RFC
args.n_jobs = 2

# ADA
args.learning_rate = 0.1
```

DEFAULT VALUES USED WITH BASE CLASSIFIERS:



```
# DT
args.criterion = 'gini'
args.max_depth = 4
args.random_state = 1

# KNN
args.neighbors = 5
args.p = 2
args.metric = 'minkowski'

# SVM
args.cnum = 10.0
args.random_state = 1
args.gamma = 0.10
if args.kernel is None:
        args.kernel = 'linear'

# PCPN
args.epochs = 40
args.eta = 0.1
```