CS487 - HW4

Cyrus Baker
March 15, 2020

Datasets

For this assignment I experimented with different classifiers using two datasets. The required dataset is the Digits dataset from the Sklearn library and the other dataset I chose is the Mammographic Masses dataset from the UCI Repository. The Mammographic dataset contains missing and improper data so I cleaned this dataset by dropping rows with missing or incorrect values.

Analysis of Classifiers

For each classifier I tested them with a range of values for different parameters. It would be impractical to report on every single run because I compared a very large number or runs but in this report I will include information for the best run of each classifier and the parameters which returned the best results.

The classifier I used to compare against the ensemble methods was the decision tree. I tested decision tree on both datasets with all combinations of these parameters:

criterion - [gini, entropy]
max_depth [5, 10, 20]

Decision Tree	Digits	Mammographic
Accuracy	0.8667	0.7831
Runtime	88.9 ms	8.7 ms
Criterion	entropy	gini
Max Depth	10	5
Train Runtime	13.9 ms	1.1 ms
Train Accuracy	1	0.8483

For the random forest classifier I used the following parameters:

n_estimators - [10, 50, 100]
max_depth - [5, 32, None]
criterion - [gini, entropy]
min_samples_split - [2, 5, 10]
min_samples_leaf - [1, 5, 10]

Radom Forest	Digits	Mammographic
Accuracy	0.9741	0.8153
Runtime	795.8 ms	385.5 ms
N Estimators	50	50
Criterion	gini	entropy
Max Depth	32	5
Min Samples Split	2	5
Min Samples Leaf	1	5
Train Runtime	120.8 ms	53.3 ms
Train Accuracy	1	0.8310

For bagging I used the following parameters:

n_estimators - [10, 50, 100]
max_samples - [1, 5, 10]
max_features - [1, 2, 3]
bootstrap - [True, False]

Bagging	Digits	Mammographic
Accuracy	0.7241	0.7952
Runtime	773.4 ms	362.6 ms
N Estimators	100	50
Max Samples	10	5
Max Features	3	3
Bootstrap	FALSE	FALSE
Train Runtime	112.0 ms	53.9 ms
Train Accuracy	0.7255	0.8224

For adaboost I used the following parameters:

n_estimators - [10, 50, 100]
learning_rate - [0.01, 0.1, 0.2]

Adaboost	Digits	Mammographic
Accuracy	0.85	0.7711
Runtime	93.6 ms	547.6 ms
N Estimators	50	50
Learning Rate	0.01	0.01
Train Runtime	13.9 ms	77.8 ms
Train Accuracy	1	0.9207

I discovered that random forest gave me the best accuracy for both datasets but the runtime is also fairly large so there is a tradeoff there. If I were to run these experiments again I might spend time fine tuning the parameters as well as tweaking other parameters to optimize the accuracy.

Problem 1:

2(c) e = w * (y != $\hat{y}$)

weights	y != $\hat{y}$	product
0.072	0	0
0.072	0	0
0.072	0	0
0.072	0	0
0.072	0	0
0.072	0	0
0.167	0	0
0.167	1	0.167
0.167	1	0.167
0.072	0	0
Sum	e =	0.334

2(d) a = 0.5ln((1-e) / e)

a = 0.5ln(.666/.334) = .345

2(e) w = w * exp(-a * y * $\hat{y}$)

weights	exp(-a * y * $\hat{y}$)	product
0.072	0.708	0.051
0.072	0.708	0.051
0.072	0.708	0.051
0.072	0.708	0.051
0.072	0.708	0.051
0.072	0.708	0.051
0.167	0.708	0.118
0.167	1.412	0.236
0.167	1.412	0.236
0.072	0.708	0.051

2(f) w = w/sum(wi)

sum(wi) = 7 * 0.051 + 2 * .236 + 0.118 = .947

0.051/0.947 = 0.054
0.118/0.947 = 0.125
0.236/0.947 = 0.249

Index	x	y	weights	$\hat{y}$	Updated weights
1	1.0	1	0.072	1	0.054
2	2.0	1	0.072	1	0.054
3	3.0	1	0.072	1	0.054
4	4.0	-1	0.072	-1	0.054
5	5.0	-1	0.072	-1	0.054
6	6.0	-1	0.072	-1	0.054
7	7.0	1	0.167	1	0.125
8	8.0	1	0.167	-1	0.249
9	9.0	1	0.167	-1	0.249
10	10.0	-1	0.072	-1	0.054