R Noah Padilla

10/26/2020

# Decision Tree Exercise Report

The purpose of this exercise was to modify parameters to optimize the accuracy or runtime of the decision tree that was made to classify MNIST samples. I found that when researching the most important changes to a decision tree was within the max depth and the minimum of samples per leaf. It states in [1] "The default values for the parameters controlling the size of the trees (e.g. `max_depth`, `min_samples_leaf`, etc.) lead to fully grown and unpruned trees which can potentially be very large on some data sets. To reduce memory consumption, the complexity and size of the trees should be controlled by setting those parameter values." Since this was advised I only modified these 2 parameters; I changed the parameters alone first then did both at the same time.

First and foremost, the most important change was to change the default function to measure a split from 'gini' to 'entropy' since we previously discussed entropy and information gain. The time taken to train was improved while testing time took longer but the accuracy increased.

```
----------Output without changes----------
Elapsed_time training  17.761801 secs
Accuracy on training set: 1.000000
Elapsed_time testing  0.027988 secs
Accuracy on test set: 0.861429
Depth:  47
Leaves:  5558
```

```
----------Output with 'entropy' instead of gini ----------
Elapsed_time training  11.643139 secs
Accuracy on training set: 1.000000
Elapsed_time testing  0.033979 secs
Accuracy on test set: 0.872429
Depth:  30
Leaves:  5082
```

CRITERION AND MAX DEPTH

| DEPTH | LEAVES | TRAIN TIME | TEST TIMES | ACC ON TRAIN | ACC ON TEST |
|---|---|---|---|---|---|
| 10 | 929 | 8.712839 secs | 0.020990 secs | 0.863032 | 0.831857 |
| 20 | 4900 | 11.484385 secs | 0.020983 secs | 0.996762 | 0.873000 |
| 30 | 5082 | 11.727436 secs | 0.021909 secs | 1 | 0.872429 |
| 40 | 5082 | 11.700648 secs | 0.026978 secs | 1 | 0.872429 |
| 50 | 5082 | 11.613911 secs | 0.020990 secs | 1 | 0.872429 |

CRITERION AND MIN SAMPLE LEAVES

| MIN SAMP LEAVES | LEAVES | TRAIN TIME | TEST TIMES | ACC ON TRAIN | ACC ON TEST |
|---|---|---|---|---|---|
| 1(DEFAULT) | 5082 | 12.268816 secs | 0.021992 secs | 1 | 0.872429 |
| 2 | 4249 | 11.552864 secs | 0.021982 secs | 0.974508 | 0.867571 |
| 4 | 3216 | 11.559186 secs | 0.021983 secs | 0.947952 | 0.867714 |
| 8 | 2236 | 11.021868 secs | 0.020991 secs | 0.914048 | 0.863571 |
| 16 | 1465 | 10.762563 secs | 0.020992 secs | 0.883365 | 0.847714 |

Although I modified the minimum number of samples required to be at a leaf node it didn't yield better results and the combination of the best of the two parameters is just the same as testing with only a max depth of 20, hence there is no need for a new table of results. A new test was made to find out if the max_leaf_nodes parameter influenced anything at all.

| DEPTH | MAX LEAF NODES | LEAVES | TRAIN TIME | TEST TIMES | ACC ON TRAIN | ACC ON TEST |
|---|---|---|---|---|---|---|
| 20 | 2000 | 2000 | 14.564323 secs | 14.722437 secs | 0.938714 | 0.873000 |
| 20 | 3000 | 3000 | 14.598027 secs | 0.020990 secs | 0.964159 | 0.870857 |
| 20 | 4000 | 4000 | 14.852098 secs | 0.028981 secs | 0.982746 | 0.870857 |
| 20 | 5000 | 4892 | 14.752832 secs | 0.022989 secs | 0.996762 | 0.871286 |
| 20 | 6000 | 4892 | 14.722437 secs | 0.021986 secs | 0.996762 | 0.871286 |

In conclusion the best change regarding runtime and accuracy was only modifying the max_depth parameter to be 20 with the criterion=’entropy’. Even though this is how I tested, I am sure there is better parameter modifications that can be made but I was unable to find an optimal way of testing.

## References

[1] “sklearn.tree.DecisionTreeClassifier¶.” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. [Accessed: 27-Oct-2020].