Decision Tree

By Tate Sakai

Datasets Used

- Credit Approval Data Set
  - 15 Attributes
  - 2 Classes: approved or rejected
  - 690 instances
  - http://archive.ics.uci.edu/ml/datasets/Credit+Approval
- Mushroom Data Set
  - 22 Attributes
  - 2 Classes: Edible or Poisonous
  - 8124 Instances
  - http://archive.ics.uci.edu/ml/datasets/Mushroom

Overall Implementation

- Read in CSV
- Basic Train/Test split
- Recursive Implementation
  - Base: If its pure || less data left than minimum specified for new subtree || == maxDepth
  - Return class that appears most in subtree
  - Recursive: Go through all potential splits and decide based on entropy/info gain, then split the data
- Classify an example of test data against the tree
  - Repeated and calculated mean for accuracy numbers
- Couldn’t figure out post-pruning, so did pre-pruning

Credit Card Approval Data Set Accuracies (Avg of 3 Trials)

These results seem to say that pre-pruning/ early-stopping based on Min-Samples per leaf, lower train accuracy (obviously) and don’t seem to have much of an effect on test-accuracy

Overfitting.
Train Acc >> Test Acc

Max Depth	Max	Max	Max	Max	Max
Min Samples	2	4	8	12	16
Train Accuracy	1.0	.9893	.9684	.9522	.9425
Test Accuracy	0.8430	0.8124	0.8278	0.8425	0.8324

Mushroom Edibility Data Set Accuracies (Avg of 3 Trials)

These results seem to say that my algorithm is wrong. I believe it's having trouble due to there being a large amount of ? values for one of the columns. I dropped the column and also tried assigning the ? to the mode of the column but the data was very similar

Too unreliable to tell if overfitting or not, but strictly from the results, it's not overfitting

Max Depth	Max	Max	Max	7	3
Min Samples	2	4	15	2	2
Train Accuracy	1.0	1.0	1.0	1.0	.9602
Test Accuracy	1.0	1.0	1.0	1.0	0.9654

Sources

- https://automaticaddison.com/iterative-dichotomiser-3-id3-algorithm-from-scratch/#implementation
- https://github.com/nanditkhosa/ID3-Decision-Tree-Using-Python/blob/master/Decision_tree_id3_implementation_without_graphviz_textual_tree_representation.py
- https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/
- https://www.youtube.com/watch?v=y6DmpG_PtN0&list=PLPOTBrypY74xS3WD0G_uzqPjCQfU6IRK-