Lab 1: Decision Trees

Naman Goyal 2015csb1021@iitrpr.ac.in
Department of Computer Science and Engineering
Indian Institute of Technology Ropar

Abstract

Experimenting with decision tree based inductive classifier using early stopping, noise, post pruning, random forests and their effect on training and test accuracy.
Keywords: Decision Tree, Early Stopping, Post Pruning, Random Forest

1. Effects of Early Stopping

A threshold on information gain is used as an early stopping criterion.

Threshold	0	0.1	0.07	0.05	0.03	0.01
Training Accuracy	0.899	0.5	0.5	0.646	0.747	0.899
Test Accuracy	0.639	0.5	0.5	0.593	0.649	0.639
Total Nodes	201	1	1	11	67	199
Terminal Nodes	101	1	1	6	34	100

Number of times an attribute is used as the splitting function

Original ID3 (Threshold = 0)	
Words used 1 time	!, ?, an, anything, as, awful, bad, book, boring, but, could, doesn't, end, even, feel, films, good, has, i, if, in, know, long, love, loved, made, major, me, movie, music, no, one, only, or, pretty, saw, say, scene, script, seen, she, should, that, they, time, true, very, was, waste, what, whether, with, worst, your
Words used 2 times	a, not, so, this
Words used 3 times	and, he, to
Words used 4 times	is
Words used 5 times	it
Words used 13 times	the

Early Stopping (Threshold = 0.05)	
Words used 1 time	bad, it, the, time, very

Early Stopping (Threshold = 0.03)	
Words used 1 time	!, a, an, awful, bad, book, but, good, is, know, love, made, movie, one, saw, scene, script, time, true, very, was, waste, worst, your
Words used 2 times	it
Words used 3 times	of
Words used 4 times	the

Early Stopping (Threshold = 0.01)	
Words used 1 time	!, ?, an, anything, as, awful, bad, book, boring, but, could, doesn't, end, even, feel, films, good, has, i, if, in, know, long, love, loved, made, major, me, movie, music, no, one, only, or, pretty, saw, say, scene, script, seen, she, should, that, they, time, true, very, was, waste, what, whether, with, worst, your
Words used 2 times	a, not, so, this
Words used 3 times	and, he, to
Words used 4 times	is
Words used 5 times	it
Words used 7 times	of
Words used 12 times	the

1.1 Observations

1. The terminal nodes reduce when threshold for early stopping is increases.

2. The training accuracy decreases as threshold for early stopping is increased.

3. The test accuracy first increases (as over-fitted nodes are relaxed) and then decreases as threshold for early stopping is increased.

4. The word **bad** serves as a good classifier. It is used for classification in all tress even when height of tree is very small.

2. Effect of Noise in training data

Noise is added by randomly switching the label of a proportion of data points in training data.

Noise	0	0.5	1	5	10
Training Accuracy	0.899	0.898	0.901	0.899	0.897
Test Accuracy	0.639	0.638	0.636	0.603	0.622
Total Nodes	201	203	197	201	205
Terminal Nodes	101	102	99	101	103

2.1 Observations

1. The total nodes in the learned tree increase as more noise is added. This is direct effect of over-fitting.

2. The training accuracy remains almost constant with a slight decrease with addition of more noise.

3. The test accuracy significantly decreases as more noise is added.

3. Effect of Post Pruning

The Post Pruning strategy tries pruning i.e. removes sub-tree rooted a node (except the node itself) for each non-terminal node of tree; if accuracy over validation set remains same or increase. The node corroding to maximum accuracy over the validation set is pruned. Also if two nodes can be both pruned with same change in accuracy over validation set, the node as smaller depth form root is selected for pruning. This process is repeated until no node can be found which could be pruned.

Stage	Before Pruning	Pruning Stage 1	Pruning Stage 2	Pruning Stage 3	Pruning Stage 4	Pruning Stage 5
Training Accuracy	0.899	0.857	0.705	0.704	0.703	0.703
Test Accuracy	0.639	0.659	0.669	0.675	0.676	0.675
Total Nodes	201	151	19	13	11	9
Terminal Nodes	101	76	10	7	6	5

3.1 Observations

1. The total nodes in the learned tree decreases as the tree is pruned.
2. The training accuracy decreases as we prune further; this is expected as decision tree now reflects average case and is not biased towards training data.
3. The test accuracy increases as we prune further; this is expected as decision tree now reflects average case.

4. Effect of number of trees in the forest on train and test accuracies

For constructing Random Forest; *sqrt*(size attribute list) i.e. around 300 are randomly sampled from whole attribute list.

Number of Trees	Training Accuracy	Test Accuracy
1	0.551	0.508
2	0.521	0.507
4	0.523	0.523
8	0.638	0.559
16	0.638	0.606
32	0.572	0.523
64	0.768	0.641
128	0.581	0.54

4.1 Observations

The training and test accuracy increases as number of trees in forest increase.