Lecture 22

Name: Harsh Sanjay Roniyar

Roll Number: 22B3942

`Assignment Deadline extended to 3rd Nov`

`Assignment Doubts Discussed`

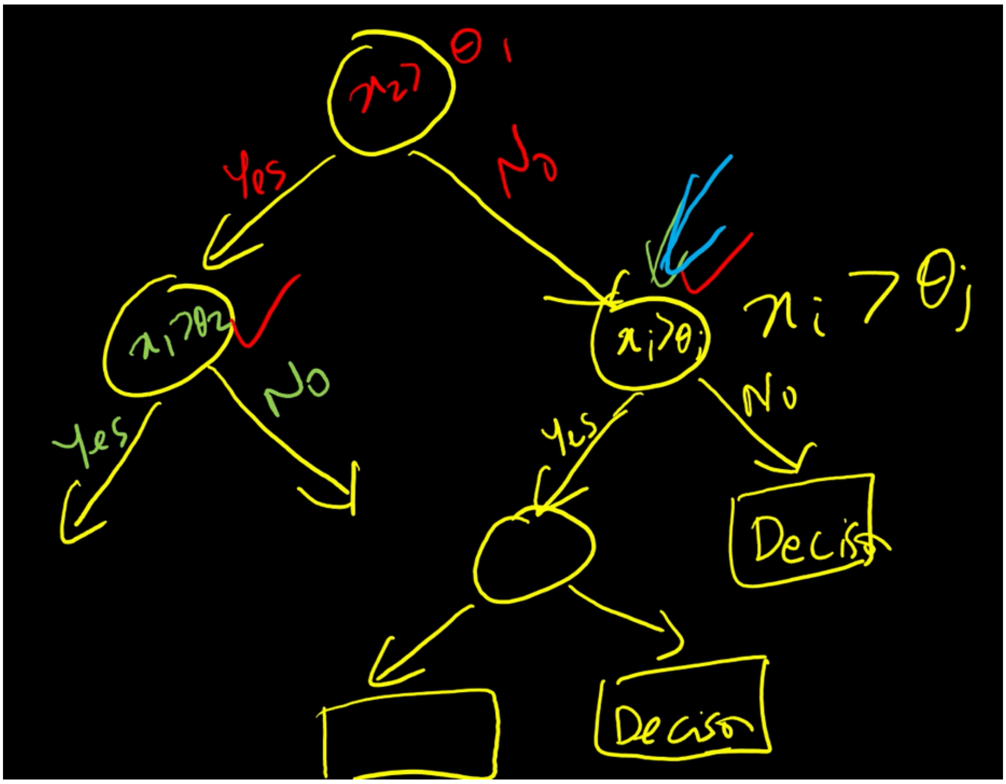
Cascade of Models (Decision Trees)

Binary Tree

Root
Internal nodes
Leaf nodes

Parent Node
Child Node

Threshold-based Decision Tree



Q1. How to choose a variable for a decision node?
Q2. How to select a threshold?

Q3. When to insert leaf nodes?

O(N) thresholds for each variables and hence O(ND) total number of thresholds, where N is the number of training points and D is the dimension.

Criteria to compare threshold classifiers

Maximal reduction in uncertainty

Greedy Tree Learning is the approach used.

- Among all thresholds evaluated, pick the one that maximally reduces the uncertainty in the data going to each of its children
- Splitting of subsets

Example: Weighted average entropy of D_1 and D_2 is minimized

Classification: $D \equiv D_0$ having label 0 or 1 with counts n_{00} and n_{01} respectively.
$|D_0| = n_0 = n_{00} + n_{01}$

The entropy is thus

$$\text{Entropy}[D_0] = -\sum_{c\in\{0,1\}} p_c \log(p_c) = -\left[\frac{n_{00}}{n_0}\log\left(\frac{n_{00}}{n_0}\right) + \frac{n_{01}}{n_0}\log\left(\frac{n_{01}}{n_0}\right)\right]$$

<u>Criteria</u>
Weighted average Entropy:

$$\frac{|D_1|\text{Ent}[D_1] + |D_2|\text{Ent}[D_2]}{|D_1| + |D_2|}$$

```
NodeSplit[D_i]:

If all D_i have same label then return None
Else
        For all dims n
        {
                For all pts in D_i
                {
                        Split D_i according to threshold into D_j and D_k
                        Compute wt. avg. entropy of the split D's
                }
        }
        Best combination <- arg min_{n,m} (x_n, theta_nm)
Return Best combo, corresponding D_j and D_k
```

```
TreeLearn[D_i]:
        Recursively split D_i
```

```
        Until [None] is returned
        Make LeafNode
```

Maximum number of leaf nodes = N
Maximum depth of tree = N

Balanced - Unbalanced Binary Trees

Deep trees can overfit
Regularization:

1. Stop splitting below a certain entropy
2. Prune trees beyond a certain depth

Purity/Impurity Criteria

For Classification:

1. Entropy: $-\sum_c p_c \log p_c$
2. Gini Index: $1 - \sum_c p_c^2$
3. $1 - \max p_c$
4. $p_1(1 - p_1)$

For Regression:

1. Variance of labels (target values)

Random Forest

It is an ensemble of decision trees

Randomize:

1. **Bagging**: Training on random subsets of samples. Out-of-bag samples (OOB) are not used
2. At each node consider only a random subset of features/dimensions/variables

Our hyperparameters are:

- number of trees
- number of features to consider at each node
- max depth
- percent of samples in a bag

Properties of random forests

1. **Can get good idea of generalization**
   - For each tree, identify OOB (out of bag) samples and their predicted label
   - Average OOB labels across trees
   - Converges to theoretical generalization error due to law of large numbers
2. **Guarantee against overfitting**
   - Since each sample is OOB for some trees, OOB generalization estimate gives a good idea of test accuracy
3. **Give feature importance**
   - Randomly permute a feature across samples and measure drop in accuracy
4. **Can handle both discrete and continuous variable**

`Random Forest - Brieman`

`Now, while I was planning on making text boxes in my notes like the one below, I stumbled upon this gem`

On the importance of sentence length

This sentence has five words. Here are five more words. Five-word sentences are fine. But several together bocome monotonous. Listen to what is happening. The writing is getting boring. The sound of it drones. It's like a stuck record. The ear demands some variety.

Now listen. I vary the sentence length, and I create music. Music. The writing sings. It has a pleasent rhythm, a lilt, a harmony. I use short sentences. And I use sentences of medium length. And sometimes when I am certain the reader is rested, I will engage him with a sentence of considerable length, a sentence that burns with energy and builds with all the impetus of a crescendo, the roll of the drums, the crash of the cymbals -- sounds that say listen to this, it is important.

- Gary Provost *(100 Ways to Improve Your Writing, 1985)*