Decision Trees - Derivation Demystified

In the last post, we saw the application and use of decision tree. The link to the previous article is given below: https://codingmachinelearning.wordpress.com/2016/06/23/decision-tree-classifier-explanation-example-using-iris-dataset-6/

In this post, we will see how to build a decision tree and a little of its derivation. I will spare you the gory details, but the essence of how we choose a split, what kind of problems we face etc will be dealt in this post.

Say we have the data in the following format: $(x_1,x_2,...,x_n)$ and label y. We have the data this way for instance x_i of the data set.Essentially we have a data matrix from which we need to build tree.

How to decide on the node split?

This is one of the most asked questions in decision tree classifiers. How do we decide on which attribute are we going to split the data. In other words, which attribute gives the best split to the data. Splitting randomly may not always give good results. So we have various spitting criterions which we can choose from, say for example, gini,entropy etc
In this post we will deal with gini index as our splitting condition and proceed to describe the process in detail.

Gini Index formula : 1 - $\Sigma[p(j|t)]^2$
Where **p(j|t)** refers to the relative frequency of class j at node t

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1
Gini = 1 – P(C1)² – P(C2)² = 1 – 0 – 1 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6
Gini = 1 – (1/6)² – (5/6)² = 0.278

C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6
Gini = 1 – (2/6)² – (4/6)² = 0.444

This is a classic example of gini index computation. C1 and C2 are the 2 classes under consideration. We are looking at problems which have 2-way binary split.

Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset

Now we will do an example for the same, by hand and build a tree manually to get the hang of what happens. But before that we will see the steps in decision tree construction.

- Step 1: Find the best splitting criterion using suitable impurity measures
- Step 2: Partition the data into $Data_{Left}$ and $Data_{Right}$ based on the attribute splitted
- Repeat Step 1 and Step 2 till all the attributes are covered,

If we keep splitting till the very last attribute, we may be overfitting the data. So we have 2 choices:

1. Pre-Pruning
2. Post-Pruning

Pruning refers to cutting down (chopping down) unwanted branches in our decision tree due to overfitting concerns. Questions like How to select how many branches to prune, should the pruning be done after the construction of tree, or while constructing may arise, but they are out of scope for this blog. I will point you to the places where you can get answers for these questions.

Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
human	warm-blooded	yes	no	no	yes
pigeon	warm-blooded	no	no	no	no
elephant	warm-blooded	yes	yes	no	yes
leopard shark	cold-blooded	yes	no	no	no
turtle	cold-blooded	no	yes	no	no
penguin	cold-blooded	no	no	no	no
eel	cold-blooded	no	no	no	no
dolphin	warm-blooded	yes	no	no	yes
spiny anteater	warm-blooded	no	yes	yes	yes
gila monster	cold-blooded	no	yes	yes	no

Take the following table and let us find the best split for our decision tree. As said earlier, we will be using gini index as the measure of impurity.

We will draw a small table which does the computation for the same

Gini Index - Attribute chosen is **body temperature**

Temperature vs Class	**Yes**	**No**
Warm-blooded	4	1
Cold-blooded	0	5

Gini computation for binary attribute will be as follows.

Warm-blooded: $1 - (4/5)^2 - (1/5)^2 = \mathbf{0.32}$
Cold-blooded: $1 - (0/5)^2 - (5/5)^2 = \mathbf{0}$

Weighted gini index (using class weights - cold,warm blood = 5 each) will be: $(5/10) * 0.32 + (5/10) * 0 = \mathbf{.16}$

Once we are clear on this concept, we will try one more just to become familiar with the idea. This time we will split based on the attribute - **Gives birth**

We will have the following table if we fill in the values.

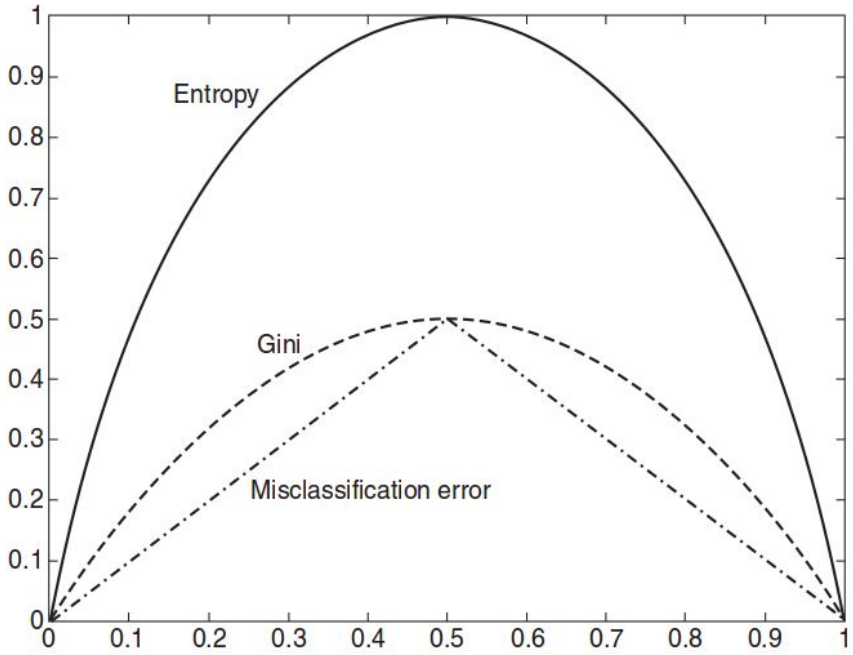
Gives Birth versus Class	**yes**	**no**
yes	3	1
no	1	5

Gini computation for binary attribute will be as follows.

Gives Birth Yes : $1 - (3/4)^2 - (1/4)^2 = \mathbf{0.375}$
Gives Birth No: $1 - (1/6)^2 - (5/6)^2 = \mathbf{0.28}$

Weighted gini index (using class weights) will be: $(4/10) * 0.375 + (6/10) * 0.280 \sim \mathbf{.318}$

Between .16 and .318, we *choose the attribute with minimum gini impurity* - Body temperature. Once we choose this attribute, then remove that columns from the data set. Partition the data into where the classes are yes and no - and continue this process on the left sub-data and right sub-data.



The range of values that gini, entropy and misclassification error metrics can take, is shown in the above diagram.

There is no significant reason as to why we choose one over another, just values, but in essence it measures impurity.

We have a fair idea of what is gini index, entropy is just same but for the formula of computation.

In the next post, we will see about SVM, what it does and how it works. What are kernels and their use etc. Till then, bye from Suriya.

References:
https://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf - Decision tree explained
http://scikit-learn.org/stable/modules/tree.html - Implementation aspect of Decision Trees
https://en.wikipedia.org/wiki/Decision_tree_learning - What and how of Decision Trees