CS 4980/6980: Introduction to Data Science © Spring 2018

Lecture 13: Decision Trees

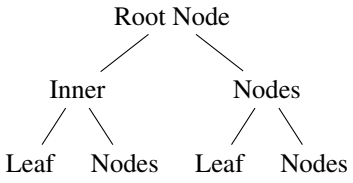
INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN Scribed by: Safin Salih and Russell Hornbuckle

13.1 Decision Trees

This is preliminary work and has not been reviewed by instructor. If you have comments about typos, errors, notation inconsistencies, etc., please email the scribers.

13.1.1 Tree Overview

Let's look at the structure of a tree. On the top is the Root node. Children of the root node that have children of their own are called internal nodes (or inner nodes.) Nodes with no children of their own are called leaf nodes (or terminal nodes.)

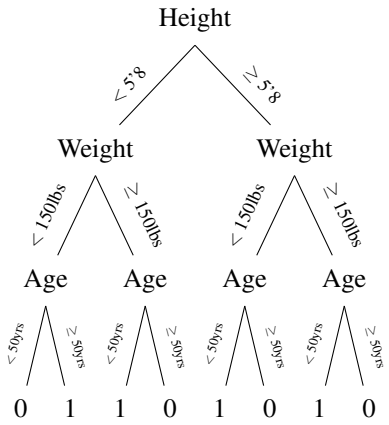


A decision tree is a tree structure where each internal node represents a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a classifying label (the decision of the tree.) The paths from root to leaf represent classification rules.

Here we have a test set of samples **X**, each with 3 attributes and a variable of interest, **Y**:

X = Attributes	Person 1	Person 2	Person 3	···	Person N
Height	5'10"	5'7"	6'1"	···	5'5"
Weight	146	160	155	···	130
Age	46	51	33	···	57
Y = Cancer?	0	1	0	···	0

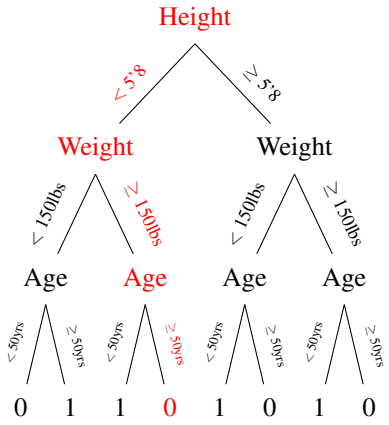
Let's build a rudimentary decision tree for classifying our sample data:



Let's say that we have a new sample, Person **x**:

$$\mathbf{x} = \begin{bmatrix} 5'7'' \\ 163 \\ 70 \\ ? \end{bmatrix}$$

We don't know whether or not person *x* has cancer so we'll use our decision tree to find out.



Our decision tree has classified person **x** as not having cancer.

The question to ask now is: Which attribute should be at the root, and in what order should the other attributes follow it? This is where the concept of entropy from information theory grants powerful insight.

13.1.2 Entropy Review

Entropy is a measure of the average uncertainty in the random variable x. You can also think of entropy as the maximum potential sum of information that the random variable x can provide.
The entropy of a data set is defined as:

$$H(x) = N\sum_{\mathbf{X}} P(x) \times log_2(\frac{1}{P(x)})$$

13.1.3 Information Gain

By design, a decision tree recursively splits your training set. At each descending level of a decision tree, the optimal split is one where each slice of training data going into the children of the current node is as clustered from the other slices as possible. You can achieve this by measuring the entropy of each feature vector. The feature vector with the highest entropy will provide higher quality clustered slices than the other feature vectors. At every new node within the tree, you must remeasure the entropy of the remaining feature vectors present for your slice of the training set to achieve optimal splits.

13.1.4 Downsides to Decision Trees

- Decision Trees have high variance. A single mistake within the feature data can produce a wildly different outcome for the relevant sample.

- Decision Trees are notoriously bad for over-fitting to their training data. In the next section, we'll discuss a variant model that alleviates some of these shortcomings

13.2 Random Forests

A Random Forest is a variant on the decision tree that can circumvent some of the issues faced by decision trees and somewhat mitigate the model's tendency to over-fit.

(1) Randomly select n Samples from $\mathbf{X}$ out of a maximum size of N (n= 0.8N is a good size)

(2) Obtain a Decision Tree with those n samples

(3) Repeat steps 1 & 2 to obtain trees: $T_1, T_2, T_3, T_4, ..., T_t$

(4) Given a new sample, x, "classify" it according to $T_1(x), T_2(x), T_3(x), ..., T_t(x)$. We'll call it T($x$).

(5) Final Decision:

$$T(x) = \frac{1}{t}\sum_{i=1}^{t} T_i(x)$$

This can also be stated as: T(x)=consensus of $(T_1(x), ..., T_t(x))$

Example 13.1 (Practical Example: Infidelity)**.** Is my girlfriend/boyfriend/whatever cheating on me?

X=Variables	1	2	3	5	6	7	8	9	10	11	12	13	14
Age	25	20	33	23	25	32	27	21	25	27	26	29	47
Length	$\frac{1}{2}$	2	7	5	6	10	3	4	$\frac{3}{2}$	3	4	$\frac{1}{3}$	1
# of children	0	0	2	1	0	3	0	0	0	0	1	0	6
# of hours away	20	7	10	7	13	2	8	7	13	11	8	4	12
# of trips per year	2	3	5	2	2	5	10	4	3	2	0	2	6
Previous offenses	2	1	0	2	0	0	2	3	0	4	0	0	13
Y=Cheating (?)	1	0	1	1	0	1	0	0	0	0	1	1	1

We will compute this Random Forest next lecture.

13.3 Wrap-up

Decision Trees

- Pick the most informative features first.

- Caveat: Has a tendency to overfit and cannot handle errors within the data set.

Random Forests

- Pick subsets of data randomly.

- Make a decision tree for each subset of data.

- The resulting consensus of your forest will somewhat mitigate the challenges faced by decision trees.