Building Decision Trees and its Math

Rakend Dubba
Follow
Jun 10, 2018 · 7 min read

Decision Tree from Scratch (Photo by Evie Shaffer on Unsplash)

A very basic introduction to Decision Trees and its Math.

What is covered :

1. Decision Trees
2. ID3 algorithm
3. Mathematical calculations for calculating Entropy and Information Gain.

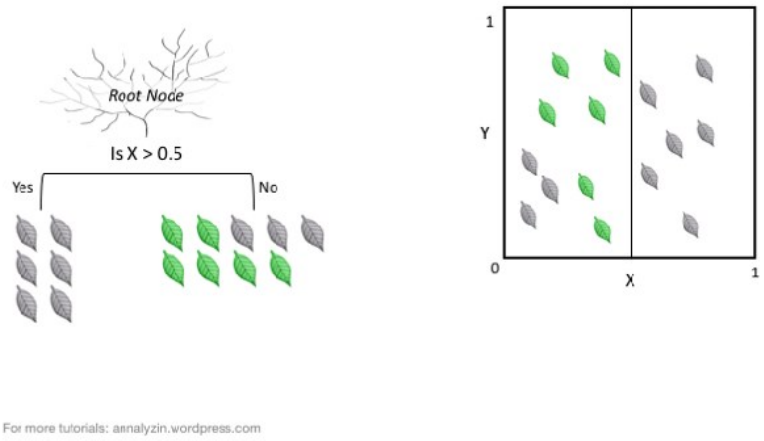
What is a Decision tree:

Decision trees are extensively used classifiers in data mining to predict the target variable for a given input variable.

It provides wonderful transparency for human interpretation by graphically representing the process of decision-making. Using this, we can predict both categorical variable (classification tree) and a continuous variable

(regression tree).

It does this by creating a flow chart like structure, and an if-else condition is applied at the nodes on the attributes (if male or female?). From here the condition outcome is represented by branches and class label is represented by the leaf node. So, a rule is created at the decision nodes and a result is delivered at the leaf node. The entire flow is from root to leaf.



Decision Trees animation. (source)

As shown in the top right side, we divide the predictor space into N distinct and non-overlapping regions. For any test observation, if we want to predict the target variable, we will simply consider the mean of the training values that fall in this region. This is for the regression tree.

For classification tree, we take the mode.

Their types:

There is a wide variety of decision trees, such as ID3 (Iterative Dichotomiser 3), CART (Classification and Regression Tree), CHAID (Chi-squared Automatic Interaction Detector), etc. Not to get intimidated by their names, they've small variations, like one is the updated version of the other (ID3 and C4.5) and can handle numeric variables.

How does it works?

1. Starts at the root node
2. Splits data into groups (based on some criteria, we will see this later)
3. Set a decision at node
4. Move the data along the respective branches
5. Repeat the process until a stopping criterion is met (max levels/depth reached, min samples left to split, nothing left to split, etc)

How to choose root node:

This is slightly different in regression and classification trees.

In regression trees, we chose a splitting point such that there is the greatest reduction in RSS (Residual Sum of Squares).

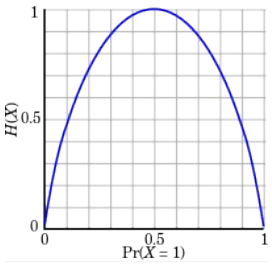
$$\sum (y_i - \hat{y}_{R_1})^2 + \sum (y_i - \hat{y}_{R_2})^2$$

Or we can calculate standard deviation reduction of the feature with respect to the training data. Here YR1 and YR2 are mean responses of region 1 and 2. Once we train the tree, we predict the response for a test data using the mean of the training observations in that group.

In Classification trees:

We use Entropy and Information Gain (in ID3). Gini Index for classification in the CART.

Entropy (Shannon's Entropy) quantifies the uncertainty of chaos in the group. Higher entropy means higher the disorder. It is denoted by H(x), where x is a vector with probabilities p1, p2, p3…..



Entropy vs Probability

From the adjacent figure, we can see that the entropy (uncertainty) is highest (1) when the probability is 0.5, i.e. 50–50 chances. And entropy is lowest when the probability is 0 or 1, i.e. there is no uncertainty or high chance of occurrence.

So, entropy is maximum if in a class there are an equal number of objects from different attributes (like the group has 50 cats and 50 dogs), and this is minimum if the node is pure (like the group has only 100 cats or only 100 dogs). We ultimately want to have minimum entropy for the tree, i.e. pure or uniform classes at the leaf nodes.

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i,$$

Entropy calculation

S — Current group for which we are interested in calculating entropy.

Pi — Probability of finding that system in ith state, or this turns to the proportion of a number of elements in that split group to the number of elements in the group before splitting(parent group).

In this classification tree, while splitting the tree we select those attributes that achieves the greatest reduction in entropy. Now, this reduction (or change) in entropy is measured by Information Gain

$$IG(Q) = S_O - \sum_{i=1}^{q} \frac{N_i}{N} S_i,$$

Example:

We consider some toy problem. The problem is about predicting

Information Gain calculation

whether some kid is going to eat a particular type of food given that kids prior eating habits.

	Taste	Temperature	Texture	Eat
0	Salty	Hot	Soft	No
1	Spicy	Hot	Soft	No
2	Spicy	Hot	Hard	Yes
3	Spicy	Cold	Hard	No
4	Spicy	Hot	Hard	Yes
5	Sweet	Cold	Soft	Yes
6	Salty	Cold	Soft	No
7	Sweet	Hot	Soft	Yes
8	Spicy	Cold	Soft	Yes
9	Salty	Hot	Hard	Yes

Eating habits of a kid

(As a side note, there is no taste like spicy, but just let's consider that in this stupid kid's bizarre eating habits preferences. _)

Solution:

From the above chart, we can see that the food preferences Taste, Temperature and Texture are exploratory variables and Eat (Yes/No) is target variable.

Now, we need to construct a top-down decision tree that splits the dataset and finally form a pure group, so we can predict for a new test variable if the kid eats or not.

We are going to use the ID3 algorithm for this. First, we look at the math involved in this, later we develop an algorithm from scratch.

The Math behind scenes:

Entropy of the total attribute at the root node

$$E_o = \sum_{i=1}^{2} \left[-P_i log_2(P_i) \right]$$
$$= \frac{-4}{10} log_2(\frac{4}{10}) - \frac{-6}{10} log_2(\frac{6}{10})$$
$$= 0.971$$

No. of 'NO' → 4

No. of 'YES' → 6

No. of objects → 10

For Taste:

$$IG = E_o - \sum \frac{N_i}{N} S_i$$

Ni→No. of ith elements in group

N→Total no. of elements in group

$$E_{Salty} = -\frac{N_1}{N} S_1 \quad \overbrace{\text{Yes}} \quad \overbrace{\text{NO}}$$
$$= -\frac{3}{10} \left[\overbrace{\frac{1}{3} log_2(\frac{1}{3})}^{} + \overbrace{\frac{2}{3} log_2(\frac{2}{3})}^{} \right]$$

$$= 0.2754$$

$$E_{Spicy} = -\frac{N_2}{N} S_2$$
$$= -\frac{5}{10}\left[\frac{3}{5} log_2\left(\frac{3}{5}\right) + \frac{2}{5} log_2\left(\frac{2}{5}\right)\right]$$
$$= 0.4854$$

$$E_{Sweet} = -\frac{N_3}{N} S_3$$
$$= -\frac{2}{10}\left[\frac{2}{2} log_2\left(\frac{2}{2}\right)\right]$$
$$= 0$$

$$E_{Taste} = E_{Salty} + E_{Spicy} + E_{Sweet}$$
$$= 0.7608$$

$$IG_{Taste} = E_o - E_{Taste}$$
$$= 0.971 - 0.7608$$
$$= 0.21$$

For Temperature:

For E_Cold, same number of 'YES' and 'NO' are there. So, entropy will be very high → 1.

For Texture:

$$E_{Hot} = -\frac{N_1}{N} S_1$$
$$= -\frac{6}{10}\left[\frac{4}{6} log_2\left(\frac{4}{6}\right) + \frac{2}{6} log_2\left(\frac{2}{6}\right)\right]$$
$$= 0.5509$$

$$E_{Cold} = -\frac{N_2}{N} S_2$$
$$= -\frac{4}{10}\left[\frac{2}{4} log_2\left(\frac{2}{4}\right) + \frac{2}{4} log_2\left(\frac{2}{4}\right)\right]$$
$$= 0.4$$

$$E_{Temp.} = E_{Hot} + E_{Cold}$$
$$= 0.9509$$

$$IG_{Temp.} = E_o - E_{Temp.}$$
$$= 0.971 - 0.9509$$
$$= 0.02$$

$$E_{Soft} = -\frac{N_1}{N} S_1$$
$$= -\frac{6}{10}\left[\frac{3}{6} log_2\left(\frac{3}{6}\right) + \frac{3}{6} log_2\left(\frac{3}{6}\right)\right]$$
$$= 0.6$$

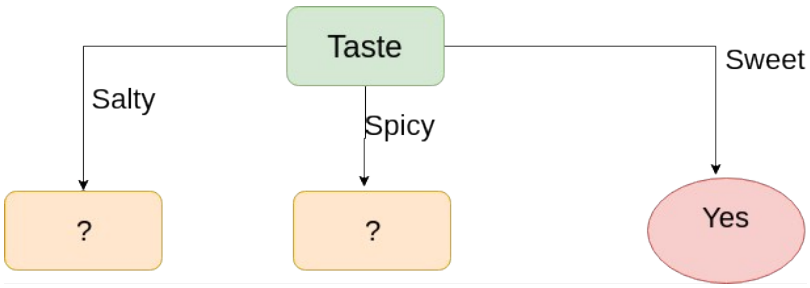
$$E_{Hard} = -\frac{N_2}{N} S_2$$
$$= -\frac{4}{10}\left[\frac{1}{4} log_2\left(\frac{1}{4}\right) + \frac{3}{4} log_2\left(\frac{3}{4}\right)\right]$$
$$= 0.3245$$

$$E_{Temp.} = E_{Soft} + E_{Hard}$$
$$= 0.9245$$

So, after our calculations, it

$$IG_{Temp.} = E_o - E_{Text.}$$
$$= 0.971 - 0.9245$$
$$= 0.05$$

seems that we should split the data based on Taste, as it has highest information gain.



Now we need to find the attributes to split at second level nodes for the Salty and the Spicy branches.

	Temperature	Texture	Eat
0	Hot	Soft	No
1	Cold	Soft	No
2	Hot	Hard	Yes

Table after the Salty branch

For Temperature(2nd level) in the Salty branch:

$$E_{1a} = \frac{2}{3}log_2(\frac{2}{3}) + \frac{1}{3}log_2(\frac{1}{3})$$
$$= 0.9182$$

$$E_{Hot} = -\frac{2}{3}\left[\frac{1}{2}log_2(\frac{1}{2}) + \frac{1}{2}log_2(\frac{1}{2})\right]$$
$$= 0.67$$

$$E_{Cold} = -\frac{1}{3}\left[\frac{1}{1}log_2(\frac{1}{1}))\right]$$
$$= 0$$

$$E_{Temp.} = 0.67$$

$$IG_{Temp.} = E_{1a} - E_{Temp.}$$
$$= 0.9182 - 0.67$$
$$= 0.2482$$

For Texture (2nd level) in the Salty branch:

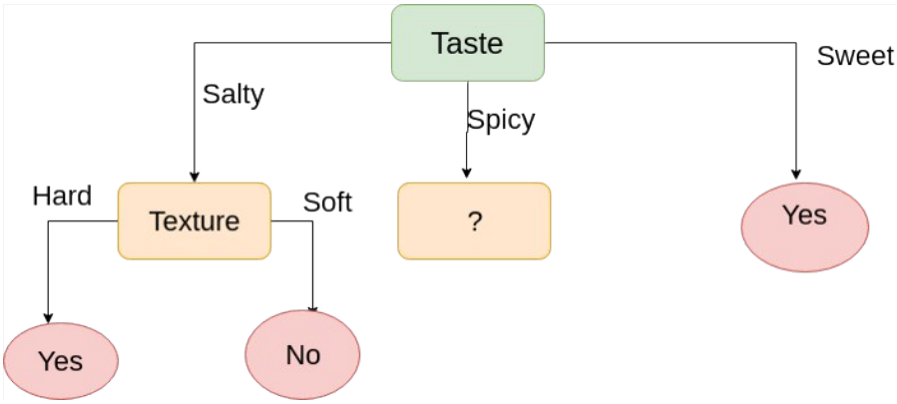
$$E_{Soft} = -\frac{2}{3}\left[\frac{2}{2}log_2(\frac{2}{2}))\right]$$
$$= 0$$

$$E_{Hard} = -\frac{1}{3}\left[\frac{1}{1}log_2(\frac{1}{1}))\right]$$
$$= 0$$

$$IG_{Text.} = E_{1a} - E_{Text.}$$
$$= 0.9182 - 0$$
$$= 0.9182$$

So, splitting based on texture sounds a good option, as it has higher

information gain.



	Temperature	Texture	Eat
0	Hot	Soft	No
1	Hot	Hard	Yes
2	Cold	Hard	No
3	Hot	Hard	Yes
4	Cold	Soft	Yes

Table after the Spicy branch

For Temperature(2nd level) in the Spicy branch:

$$E_{1b} = -\frac{2}{5}log_2(\frac{2}{5}) - \frac{3}{5}log_2(\frac{3}{5})$$
$$= 0.9709$$

$$E_{Hot} = -\frac{3}{5}\left[\frac{1}{3}log_2(\frac{1}{3}) + \frac{2}{3}log_2(\frac{2}{3})\right]$$
$$= 0.5509$$

$$E_{Cold} = -\frac{2}{5}\left[\frac{1}{2}log_2(\frac{1}{2}) + \frac{1}{2}log_2(\frac{1}{2})\right]$$
$$= 0.4$$

$$E_{Temp.} = 0.9509$$

$$IG_{Temp.} = E_{1b} - E_{Temp.}$$
$$= 0.9709 - 0.9509$$
$$= 0.02$$

For Texture(2nd level) in the Spicy branch:

$$E_{Soft} = -\frac{2}{5}\left[\frac{1}{2}log_2(\frac{1}{2}) + \frac{1}{2}log_2(\frac{1}{2})\right]$$
$$= 0.4$$

$$E_{Hard} = -\frac{3}{5}\left[\frac{1}{3}log_2(\frac{1}{3}) + \frac{2}{3}log_2(\frac{2}{3})\right]$$
$$= 0.5509$$

$$E_{Text.} = 0.9509$$

$$IG_{Text.} = E_{1b} - E_{Text.}$$
$$= 0.9709 - 0.9509$$
$$= 0.02$$

Both the attributes generated same Information Gain. So, we can split with any attribute.

We'll choose Temperature to split.

Then the only option left is Texture to split.

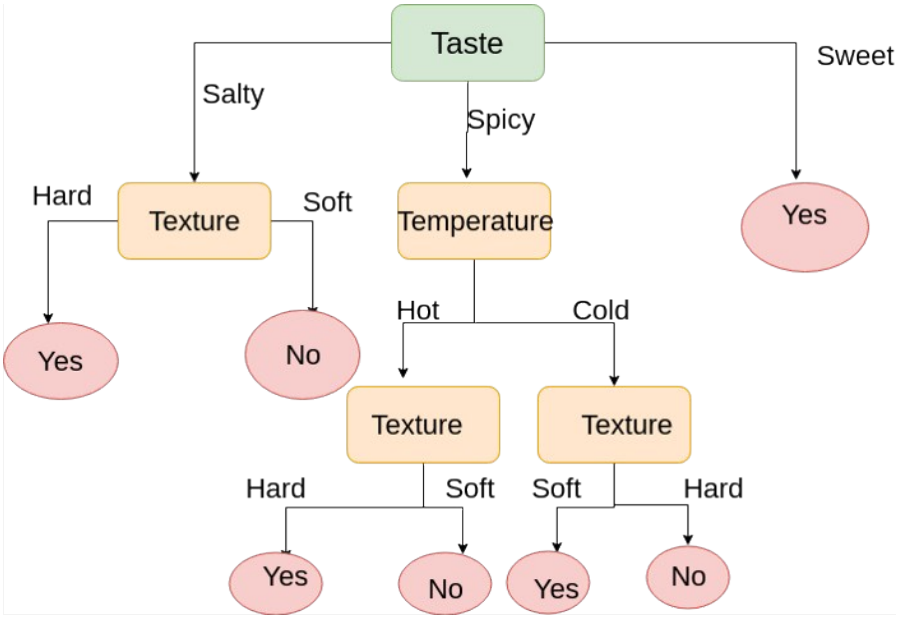
Tables left after Temperature split, for both the branches after Texture split it will be pure homogeneous groups.

	Texture	Eat
0	Soft	No
1	Hard	Yes
2	Hard	Yes

Table : Spicy-Temperature-Hot path

	Texture	Eat
0	Hard	No
1	Soft	Yes

Table : Spicy-Temperature-Cold path



Finally, what we can conclude is, if the food is sweet the kid is not caring about its Temperature or Texture, he is eating.

If the food is Salty he is eating only if the texture is hard. And if the food is Spicy he eats if it is Hot and Hard or Cold and Soft. Bizarre Kid. _

Next post will write building this tree using Python. Stay tuned.
Thank you for your patience _.

Decision Tree | Id3 | Machine Learning | Algorithms | Entropy

260 claps

WRITTEN BY

Rakend Dubba

Follow

Machine Learning Engineer | Data Scientist

See responses (3)

More From Medium

Related reads

LOGISTIC REGRESSION CLASSIFIER

Caglar Subasi in Towards Data Science

Mar 4 · 8 min read ★

181

Related reads

Linear Discriminant Analysis

Srishti Sawla

Jun 5, 2018 · 3 min read

467

Related reads

Gradient Descent: Simply Explained?

Koo Ping Shung in Towards Data Science

Apr 2, 2018 · 5 min read

390