2019-03-11 Decision Trees

Monday, March 11, 2019 2:59 PM

- A decision tree is nothing more than an N-way tree
  - N-way tree is a tree that has at most N branches
- Each internal node represents a question
- Each leaf node represents an outcome
- A decision tree algorithm is a technique that constructs a decision tree from data
  - We are using the ID3 algorithm
- Given some data set, ID3 will construct an N-way tree that best fits the data
  - Best fit does not guarantee correctness in all situations

Basic Example

- Given a set of predictor variables (Outlook, Temperature, Humidity, Wind), can we mathematically construct a decision tree that accurately predicts whether or not we will play baseball?

Outlook	Temperature	Humidity	Wind	Play ball
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Outlook	Sunny	2 - YES	3 - NO
	Overcast	4	0
	Rain	3	2
Temperature	Hot	2	2
	Mild	4	2
	Cool	3	1
Humidity	High	3	4
	Normal	6	1
Wind	Weak	6	2
	Strong	3	3

- Terms:
  - Outcome variable: what we're trying to predict (dependent variable)
  - Predictor variable / predictor factor: what we're using in our prediction of outcome (independent variables)
- ID3 is a greedy algorithm, in which nodes of a lower depth (root is lowest) have a higher predicting capability than lower nodes.
- The ID3 algorithm uses math to determine predicting capability using entropy.
  - In this context entropy quantifies the amount of information present in a predictor variable
  - Entropy ranges from 0 (no information) to 1 (most information)
- Term: Information gain
  - How much entropy is gained from a predictor variable

Calculating Entropy

take this, expand out for each outcome

$$Entropy(S) = (-)S - p_i \log_2(p_i)$$

S = complete collection of outcomes for given predictor — goes away when all outcome variables are accounted for

p_i = proportion of S belonging to class i

Calculating Entropy of our entire observation set

- Play baseball 9 times, don't play baseball 5 times

$$\text{Entropy}(ALL) = -\underbrace{\left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right)}_{no} - \underbrace{\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right)}_{yes}$$

$$= -(-.53) - (-.41)$$

$$= .94$$

- Now that we know entropy is .94 for the system, what predictor variable contributes the most information gain?

Calculating Entropy for each predictor variable

Outlook	Sunny	2 - YES	3 - NO
	Overcast	4	0
	Rain	3	2
Temperature	Hot	2	2
	Mild	4	2
	Cool	3	1
Humidity	High	3	4
	Normal	6	1
Wind	Weak	6	2
	Strong	3	3

Gain for Outlook variable

$$\text{Entropy}(\text{Outlook}_{\text{Sunny}}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$$

$$= .44 + .53$$

$$= .97$$

$$\text{Entropy}(\text{Outlook}_{\text{Overcast}}) = -\frac{0}{4}\log_2\left(\frac{0}{4}\right) - \frac{4}{4}\log_2\left(\frac{4}{4}\right)$$

$$= 0$$

$$\text{Entropy}(\text{Outlook}_{\text{Rain}}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= .97$$

$$\text{Gain}_{\text{Outlook}} = \text{Entropy}(S) - \Sigma(\text{ratio}(\text{outlook}_c))$$

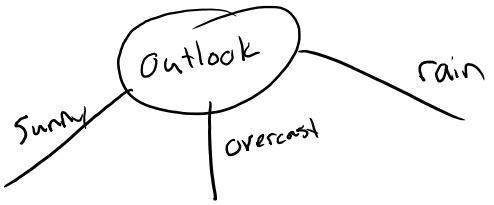
$$= .94 - \frac{5}{14}(.97) - \frac{4}{14}(.0) - \frac{5}{14}(.97)$$

$$= .25$$

Information gain for each predictor variable

- Outlook: .25
- Wind: 0.05
- Temperature: 0.03
- Humidity: 0.15

Based on the above results, we choose outlook as our root node



- Recursively go through and calculate information gain on the subset of data that matches the prior condition.

Sunny Condition

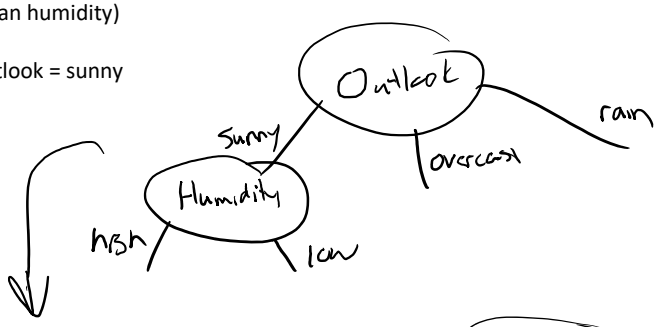
Outlook	Temperature	Humidity	Wind	Play ball
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

Overall entropy: $-\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5}) = .97$

(Class calculated)

- Temperature: (guaranteed to be less than humidity)
- Humidity: .97
- Wind: (guaranteed to be less than humidity)

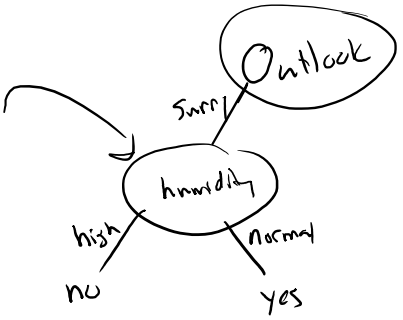
Therefore, we pick humidity when outlook = sunny



Data table when outlook = sunny and humidity = high

Outlook	Temperature	Humidity	Wind	Play ball
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No

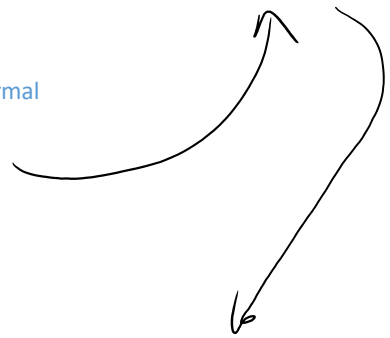
Gain for temperature: 0
Gain for wind: 0
No new information. Done - give prediction (leaf node)



Data table when outlook = sunny and humidity = normal

Outlook	Temperature	Humidity	Wind	Play ball
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

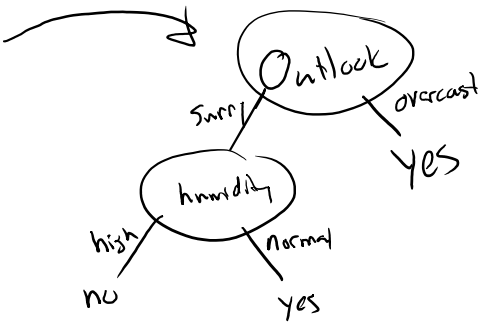
- Again, no information gain, done.



Recursively bubble up, visit next condition when outlook = overcast

Outlook	Temperature	Humidity	Wind	Play ball
Overcast	Hot	High	Weak	Yes
Overcast	Cool	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes

Gain: 0. So we're done with this branch

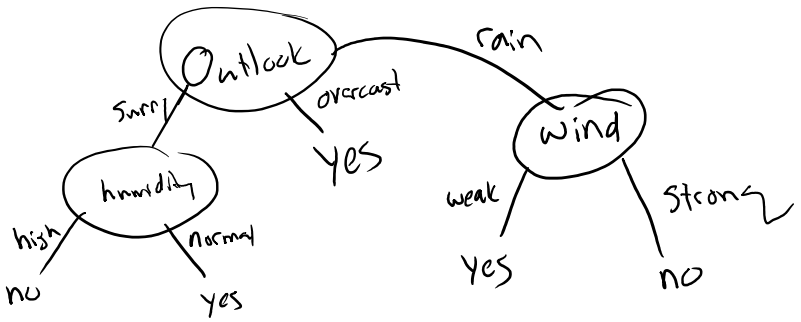


Finally, calculate when outlook = rain

Outlook	Temperature	Humidity	Wind	Play ball
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Rain	Mild	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Entropy when Outlook = rain: .97

Wind gain: .97
Temperature and Humidity gain are less than .97, thus choose wind



Outlook	Temperature	Humidity	Wind	Play ball
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes

Outlook	Temperature	Humidity	Wind	Play ball
Rain	Cool	Normal	Strong	No
Rain	Mild	High	Strong	No