* What is a decision tree?
* What is ID3 algorithm?
* What is information gain and entropy?
* What are the steps in ID3 algorithm?
* Using ID3 algorithm on a real data
* What are the characteristics of ID3 algorithm?

**What is a Decision Tree?**

A Supervised Machine Learning Algorithm, used to build classification and regression models in the form of a tree structure.

*A decision tree is a tree where each -*

* Node - a feature(attribute)
* Branch - a decision(rule)
* Leaf - an outcome(categorical or continuous)

There are many algorithms to build decision trees, here we are going to discuss ID3 algorithm with an example.

**What is an ID3 Algorithm?**

* ID3 stands for Iterative Dichotomiser 3
* It is a classification algorithm that follows a greedy approach by selecting a best attribute that yields maximum Information Gain (IG) or minimum Entropy (H), Gain.

**What is Entropy and Information gain?**

* **Entropy is a measure of the amount of uncertainty(randomness) in the dataset D. [Info(D) = H(D)]**



* **Information Gain I(A) tells us how much uncertainty in D was reduced after splitting set D on attribute A.**



* **Information gained by branching on attribute A**

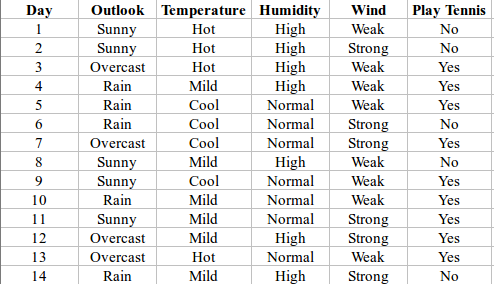
****

The attribute with the largest information gain is used to split the set S on that particular iteration.

**What are the steps in ID3 algorithm?**

1. Calculate entropy for dataset.
2. For each attribute/feature
   1. Calculate entropy for all its categorical values.
   2. Calculate information gain for the feature.
   3. Find the feature with maximum information gain.
   4. Repeat it until we get the desired tree.

**Use ID3 algorithm on the given data and generate decision tree.**



Complete entropy of dataset is -

Info (D) = H (D) = - p (yes) \* log2 (p (yes)) - p (no) \* log2 (p (no))

= - (9/14) \* log2 (9/14) - (5/14) \* log2 (5/14)

= - (-0.41) - (-0.53)

= 0.94

Let’s calculate Information of each attribute

**First Attribute – Outlook**



I (sunny) = - (2/5)\*log (2/5)-(3/5)\*log (3/5) =0.971

I (rain) = - (3/5)\*log (3/5)-(2/5)\*log (2/5) =0.971

I (overcast) = - (4/4)\*log (4/4)-0 = 0

= (5/14)\*0.971 + (5/14)\*0.971 + (4/14)\*0

= 0.693

Information Gain = I (D) – I (Outlook)

= 0.94 - 0.693

Gain (outlook) = 0.247

**Note: log21 = log220 = 0**

**Second Attribute – Temperature**

Categorical values - hot, mild, cool

I (Temperature) = p (hot)\*I (Temperature=hot) + p (mild)\*I(Temperature=mild) + p(cool)\*I(Temperature=cool)

I (Temperature=hot) = - (2/4)\*log (2/4)-(2/4)\*log (2/4) = 1

I (Temperature=cool) = - (3/4)\*log (3/4)-(1/4)\*log (1/4) = 0.811

I (Temperature=mild) = - (4/6)\*log (4/6)-(2/6)\*log (2/6) = 0.9179

I (Temperature) = p (hot)\*I (Temperature=hot) + p (mild)\*I(Temperature=mild) + p(cool)\*I(Temperature=cool)

= (4/14)\*1 + (6/14)\*0.9179 + (4/14)\*0.811

= 0.9108

Information Gain = I (D) - I(Temperature)

= 0.94 - 0.9108

Gain (Temperature) = 0.0292

**Third Attribute – Humidity**

**Categorical values - high, normal**

I (Humidity=high) = - (3/7)\*log (3/7)-(4/7)\*log (4/7) = 0.983

I (Humidity=normal) = - (6/7)\*log (6/7)-(1/7)\*log (1/7) = 0.591

I (Humidity) = p (high)\*I (Humidity=high) + p (normal)\*I (Humidity=normal)

= (7/14)\*0.983 + (7/14)\*0.591

= 0.787

Information Gain = I(D) - I(Humidity)

= 0.94 - 0.787

Gain (Humidity) = 0.153

**Fourth Attribute – Wind**

**Categorical values - weak, strong**

I (Wind=weak) = - 6/8)\*log (6/8)-(2/8)\*log (2/8) = 0.811

I (Wind=strong) = - (3/6)\*log (3/6)-(3/6)\*log (3/6) = 1

I (Wind) = p (weak)\*I (Wind=weak) + p (strong)\*I (Wind=strong)

= (8/14)\*0.811 + (6/14)\*1

= 0.892

Information Gain = I (D) – I (Wind)

= 0.94 - 0.892

Gain (wind) = 0.048

Gain (outlook) = 0.247

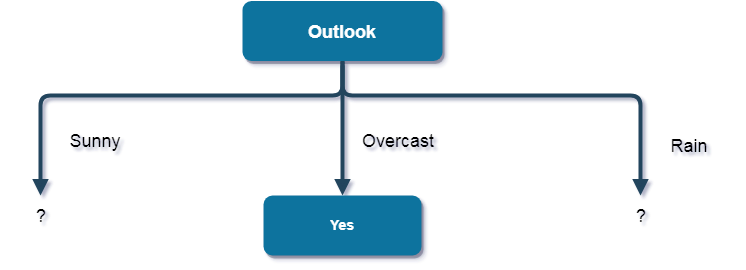
Gain (Humidity) = 0.153

Gain (wind) = 0.048

Gain (Temperature) = 0.0292

Here, the attribute with maximum information gain is Outlook.

So, the decision tree built so far is



Here, when Outlook == overcast, it is of pure class (Yes).  
Now, we have to repeat same procedure for the data with rows consist of Outlook value as Sunny and then for Outlook value as Rain.

Now, finding the best attribute for splitting the data with **Outlook=Sunny values**{ Dataset rows = [1, 2, 8, 9, 11]}.

Complete entropy of Sunny is -

I (D) = - p (yes) \* log2 (p(yes)) - p(no) \* log2(p(no))

= - (2/5) \* log2 (2/5) - (3/5) \* log2 (3/5)

= 0.971

**First Attribute - Temperature**

Categorical values - hot, mild, cool

I(Sunny, Temperature=hot) = -0-(2/2)\*log(2/2) = 0

I(Sunny, Temperature=cool) = -(1)\*log(1)- 0 = 0

I(Sunny, Temperature=mild) = -(1/2)\*log(1/2)-(1/2)\*log(1/2) = 1

Average Entropy Information for Temperature -

I(Sunny, Temperature) = p(Sunny, hot)\*I(Sunny, Temperature=hot) + p(Sunny, mild)\*I(Sunny, Temperature=mild) + p(Sunny, cool)\*I(Sunny, Temperature=cool)

= (2/5)\*0 + (1/5)\*0 + (2/5)\*1

= 0.4

Information Gain = I(Sunny) - I(Sunny, Temperature)

= 0.971 - 0.4

Gain(sunny,temperature)= 0.571

**Second Attribute - Humidity**

Categorical values - high, normal

I(Sunny, Humidity=high) = - 0 - (3/3)\*log(3/3) = 0

I(Sunny, Humidity=normal) = -(2/2)\*log(2/2)-0 = 0

Average Entropy Information for Humidity -

I(Sunny, Humidity) = p(Sunny, high)\*I(Sunny, Humidity=high) + p(Sunny, normal)\*I(Sunny, Humidity=normal)

= (3/5)\*0 + (2/5)\*0

= 0

Information Gain = I(Sunny) - I(Sunny, Humidity)

= 0.971 - 0

Gain(Sunny, Humidity) = 0.971

**Third Attribute - Wind**

**Categorical values - weak, strong**

I(Sunny, Wind=weak) = -(1/3)\*log(1/3)-(2/3)\*log(2/3) = 0.918

I(Sunny, Wind=strong) = -(1/2)\*log(1/2)-(1/2)\*log(1/2) = 1

Average Entropy Information for Wind -

I(Sunny, Wind) = p(Sunny, weak)\*I(Sunny, Wind=weak) + p(Sunny, strong)\*I(Sunny, Wind=strong)

= (3/5)\*0.918 + (2/5)\*1

= 0.9508

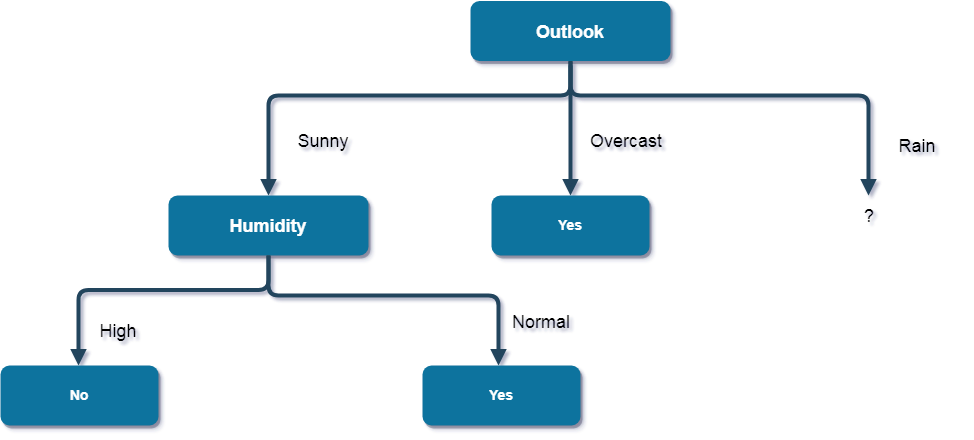
Information Gain = I(Sunny) - I(Sunny, Wind)

= 0.971 - 0.9508

Gain (Sunny, Wind) = 0.0202

Gain (sunny, temperature) = 0.571

Gain (Sunny, Humidity) = 0.971

Here, the attribute with maximum information gain is Humidity. So, the decision tree built so far -  


Here, when Outlook = Sunny and Humidity = High, it is a pure class of category "no". And When Outlook = Sunny and Humidity = Normal, it is again a pure class of category "yes". Therefore, we don't need to do further calculations.

Now, finding the best attribute for splitting the data with Outlook=Rain values{ Dataset rows = [4, 5, 6, 10, 14]}.

Complete entropy of Rain is -

I(D) = - p(yes) \* log2(p(yes)) - p(no) \* log2(p(no))

= - (3/5) \* log(3/5) - (2/5) \* log(2/5)

= 0.971

**First Attribute - Temperature**

**Categorical values - mild, cool**

I(Rain, Temperature=cool) = -(1/2)\*log(1/2)- (1/2)\*log(1/2) = 1

I(Rain, Temperature=mild) = -(2/3)\*log(2/3)-(1/3)\*log(1/3) = 0.918

Average Entropy Information for Temperature -

I(Rain, Temperature) = p(Rain, mild)\*I(Rain, Temperature=mild) + p(Rain, cool)\*I(Rain, Temperature=cool)

= (2/5)\*1 + (3/5)\*0.918

= 0.9508

Information Gain = I(Rain) - I(Rain, Temperature)

= 0.971 - 0.9508

Gain(Rain, Temperature) = 0.0202

**Second Attribute - Wind**

**Categorical values - weak, strong**

I(Wind=weak) = -(3/3)\*log(3/3)-0 = 0

I(Wind=strong) = 0-(2/2)\*log(2/2) = 0

Average Entropy Information for Wind -

I(Wind) = p(Rain, weak)\*I(Rain, Wind=weak) + p(Rain, strong)\*I(Rain, Wind=strong)

= (3/5)\*0 + (2/5)\*0

= 0

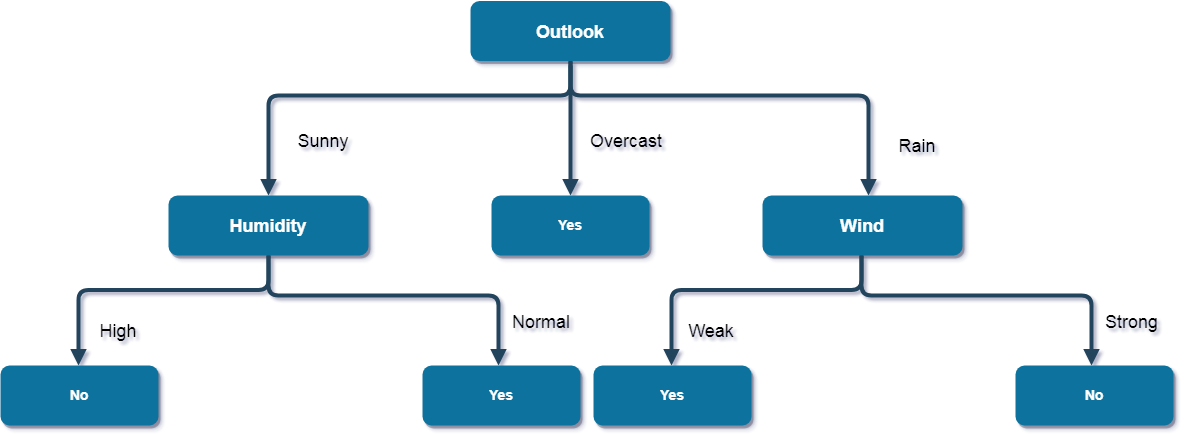
Information Gain = I(Rain) - I(Rain, Wind)

= 0.971 - 0

Gain(Rain, Wind)= 0.971

Gain(Rain, Temperature) = 0.0202

Here, the attribute with maximum information gain is Wind.

So, the decision tree built so far -  


Here, when Outlook = Rain and Wind = Strong, it is a pure class of category "no".

And When Outlook = Rain and Wind = Weak, it is again a pure class of category "yes".  
And this is our final desired tree for the given dataset.

What are the characteristics of ID3 algorithm?

Characteristics of ID3 Algorithm are -

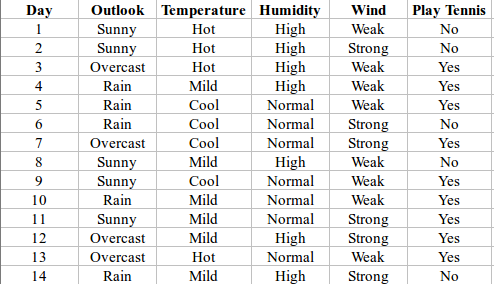
ID3 uses a greedy approach that's why it does not guarantee an optimal solution; it can get stuck in local optimums.

ID3 can overfit to the training data (to avoid overfitting, smaller decision trees should be preferred over larger ones).

This algorithm usually produces small trees, but it does not always produce the smallest possible tree.

ID3 is harder to use on continuous data (if the values of any given attribute is continuous, then there are many more places to split the data on this attribute, and searching for the best value to split by can be time consuming).

**Gain Ratio**



Gain (outlook) = 0.247

Gain (Humidity) = 0.153

Gain (wind) = 0.048

Gain (Temperature) = 0.0292





splitInfo(Humidity) = -7/14\*log2(7/14)-7/14log2(7/14)

Gain Ratio (Temperature) = Gain (Temperature)/ SplitInfo (Temperature)

= 0.0292/ 0.926

= 0.031

Choose that attribute as root node whose gain ratio is the highest and follow the process iteratively for other nodes at different levels.

**Gini Index**

****

****

****

**gini(D) = 1- (9/14)2-(5/14)2**

**= 1- 0.413 – 0.127**

**= 0.46**

**gini temperature(D) = 4/14[1-(2/4)2-(2/4)2]+6/14[1-(4/6)2-(2/6)2]+4/14[1-(3/4)2-(1/4)2]**

**=0.28[1-0.25-0.25]+0.42[1-0.44-0.11]+0.28[1-0.56-0.0625]**

**= 0.28 (0.5) + 0.42( 0.45)+0.28(0.3775)**

**=0.14+ 0.189 + 0.1057**

**=0.4347**

**Gini(Temperature) = gini(D)- gini temperature(D)**

**=0.46-0.4347= 0.0253**

Choose that attribute as root node whose reduction in impurity is the lowest and follow the process iteratively for other nodes at different levels.