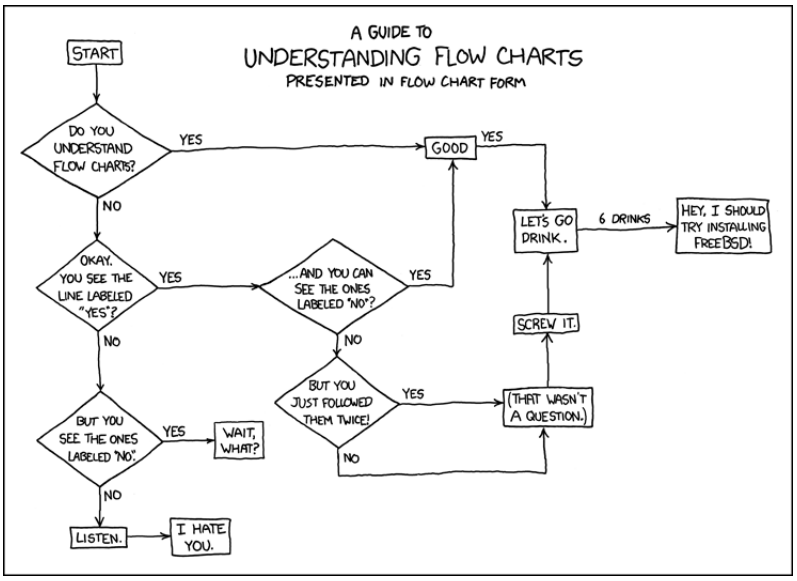
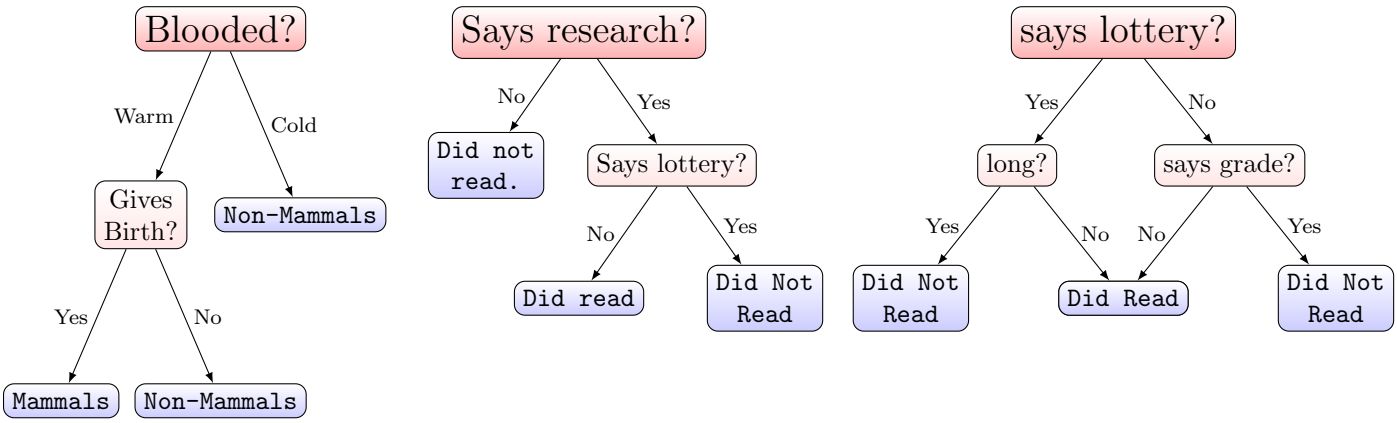


- XKCD # 518: Flow Charts. At 8 drinks, you switch the torrent from freeBSD to Microsoft Bob. C'mon, it'll be fun!
- Your professor is not endorsing underage drinking

What is a Decision Tree?

Consider the following **decision trees**:



Question 1. Identify the following key parts of the tree: root, leaf, internal nodes, path

Question 2. A flamingo is an animal that is warm-blooded and does not give live birth. Is it a mammal according to the above decision tree? How do you know?

Question 3. A decision tree represents a *disjunction* of *conjunctions*. What does the above tree represent for mammals? For non-mammals? What about the other two, with respect to whether or not an email was read?

A decision tree is best suited to problems with the following characteristics:

- Attributes have discrete values
- Output is discrete
- Disjunctive descriptions
- Training data may contain errors
- Training data may contain missing attribute values

Introducing The ID3 Algorithm

The algorithm we will present today begins by selecting a *root node*[1] for the decision tree by *splitting* on a chosen attribute. The training data is then partitioned based on that attribute and the process is repeated in each sub-tree. One major decision made by any algorithm that follows this general procedure is *how* to select which attribute is used as a split. Another major decision to be made is when to stop.

Outlook	Temperature	Humidity	Wind	PlayTennis
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

One of these attributes (Outlook, Temperature, Humidity, Wind) will be the root of the decision tree.
Question 4. Suppose we wanted to have a rote learner for this problem. Each of tomorrow's attributes are chosen independently and uniformly at random. What is the probability tomorrow is contained in the training data?

Question 5. Looking at the data, which attribute *appears* to be a good choice for the root? Why do you think so?

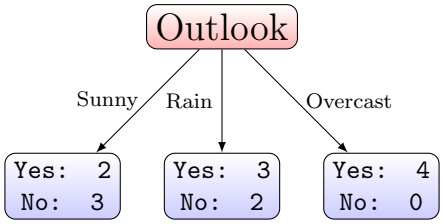
[1]To be more precise: the entire dataset is the root at the onset; we repeatedly select a leaf node that has both yes and no instances and replace that with a question node that will divide the set.

Question 6. To make an informed decision, and because a computer cannot just "eyeball" a selection, let's start by determining how many "yes" and "no" instances would be in each sub-tree for each possible choice of the root; using the above table, fill in these values.

Attribute	# Yes	# No
Outlook = Sunny		
Outlook = Overcast		
Outlook = Rain		
Temperature = Hot		
Temperature = Mild		
Temperature = Cool		

Attribute	# Yes	# No
Humidity = High		
Humidity = Normal		
Wind = Weak		
Wind = Strong		

We haven't done much math to make our decision yet. We are developing the intuition for the math.
Question 7. Suppose we decide the first question we ask will be about the outlook. What should the remaining question(s) be? They can depend on the answer to outlook. Draw the completed tree.



Using entropy and information gain to split attributes

Let's look at the definition of *entropy*; you can think of it as a measure of uncertainty. There is also the view that it measures impurity in a collection of training examples. The formula is as follows:

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

where p_i is the proportion of S belonging to class i.

Question 8. Suppose S has 14 examples, 9 positive and 5 negative. Calculate the entropy.

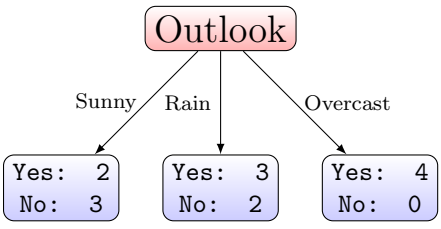
Question 9. What should entropy be if everything is one category?

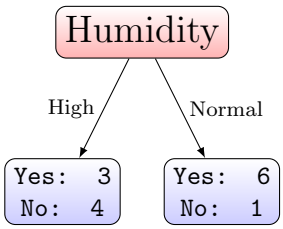
How can we use entropy to measure the effectiveness an attribute?

The **information gain** for selecting attribute A to split set S:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

Question 10. Suppose I choose "Outlook" as the root (left diagram, below). What is the information gained? What if I choose "Humidity" instead (right diagram, below)?





Question 11. Suppose I choose "Outlook" as the root. Within the sub-tree that corresponds to a Sunny Outlook, what is the information gain if I split *that subtree* at Humidity?

Practice / Reinforcement Exercise

Consider the following set of training data:

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

Use the ID3 algorithm to make a decision tree that will correctly classify each of these. Do this using information gain to decide splitting attributes as appropriate. You may (and are encouraged to) use a calculator (including one on your phone or laptop). Please choose your root, or at least attempt to do so, before discussing with your groupmates.

Does the tree match the one you believe you would have constructed from this data by intuition rather than by information gain? If so, what do we learn about decision trees from that situation?

Question 12. What hypothesis space does ID3 search?

Question 13. How many hypotheses does ID3 maintain? Can we use it to get a set of consistent hypotheses?

Question 14. What types of trees does ID3 produce?

Question 15. How do the preferences of ID3 and Candidate-Elimination differ?

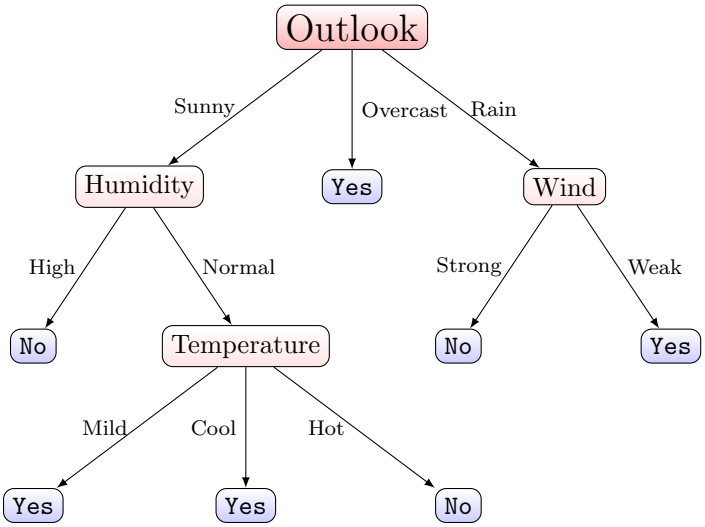
Question 16. ID3 can [almost] always achieve 100% training accuracy. Is this always desirable? Why or why not?

Question 17. What should cause us to suspect overfitting in a decision tree?

Suppose we add the following row to the training data for the PlayTennis? example:

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	Normal	Strong	No

If we had that entry and build the tree from the beginning, we end up with this decision tree:



Question 18. We saw that larger trees are suspected of overfitting. What can we change about ID3 to combat this?

Question 19. How can we decide branches as candidates for pruning?

Question 20. What are three reasons for the popularity of decision trees?

Exercise/Reinforcement

The following data set comes from *Artificial Intelligence: A Modern Approach* by Russell and Norvig. This is a commonly-assigned textbook for university courses in Artificial Intelligence. The general goal is to determine if we will wait at a restaurant, given the various conditions when we enter (are there alternates, is there a bar, what's the price range, food type, is it raining, did we make a reservation, etc).

Alt	Bar	Fri	Hun	Patrons	Price	Rain	Res	Type	Est-Wait	Will Wait
Yes	No	No	Yes	Some	$$$	No	Yes	French	0-10	Yes
Yes	No	No	Yes	Full	$	No	No	Thai	30-60	No
No	Yes	No	No	Some	$	No	No	Burger	0-10	Yes
Yes	No	Yes	Yes	Full	$	Yes	No	Thai	10-30	Yes
Yes	No	Yes	No	Full	$$$	No	Yes	French	> 60	No
No	Yes	No	Yes	Some	$$	Yes	Yes	Italian	0-10	Yes
No	Yes	No	No	None	$	Yes	No	Burger	0-10	No
No	No	No	Yes	Some	$$	Yes	Yes	Thai	0-10	Yes
No	Yes	Yes	No	Full	$	Yes	No	Burger	> 60	No
Yes	Yes	Yes	Yes	Full	$$$	No	Yes	Italian	10-30	No
No	No	No	No	None	$	No	No	Thai	0-10	No
Yes	Yes	Yes	Yes	Full	$	No	No	Burger	30-60	Yes

Suppose you wanted to build a decision tree for this data.

Question 21. Would attribute "type" be a good root attribute? Try to determine this initially without any numerical computations.

Question 22. Would attribute "Patrons" be better, worse, or the same quality root as type? Try to determine this initially without any numerical computations.

Question 23. Build a decision tree for this data.