Decision Trees

In this lecture. . .

- Decision trees for classification
- Minimizing entropy and maximizing information gain
- Decision trees for regression

Introduction

Decision trees are a supervised-learning technique.

Really they are just flowcharts.

We start with classified data points with multiple features. These features describe each data point according to their attributes (tall or short, man or woman, or numerical values).

The method uses a hierarchy of divisions of the data points according to these attributes.

It can also be used for regression if the data has associated numerical values.

What Is It Used For?

Decision trees are used for

- Classifying data according to attributes that can be categories or numerical
- Example: Predict the number of votes in the Eurovision Song Contest by looking at type of song, beats per minute, number of band members, etc.
- Example: Forecast IQ based on books read, type, subjects, authors, etc.

You will no doubt have seen decision trees before, although perhaps not in the context of machine learning.

You will definitely have played 20 Questions before. "Is the actor male?" Yes. "Is he under 50?" No. . . "Is he bald?" Yes. "Bruce Willis?" Yes.

If the answer to the first question had been No then you would have taken a different route through the tree.

In 20 Questions the trick is to figure out what are the best Yes/No questions to ask at each stage so as to get to the right answer as quickly as possible, or at least in 20 or fewer questions.

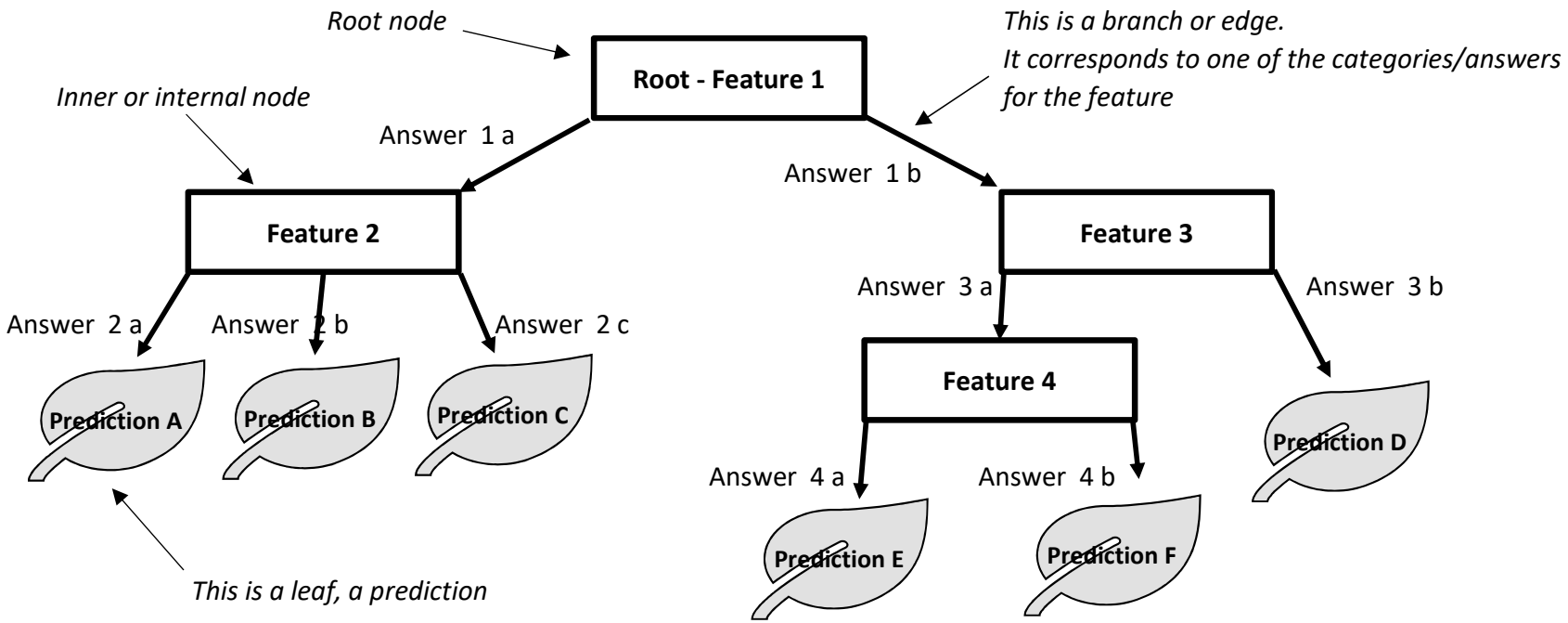
In machine learning the goal is similar.

You will have a training set of data that has been classified, and this is used to construct the best possible tree structure of questions and answers so that when you get a new item to classify it can be done quickly and accurately.

I would not recommend using Excel for decisions trees when you have a real problem, with lots of data. It's just too messy. That's because you don't know the tree structure before you start, making it difficult to lay things out on a spreadsheet.

First some jargon and conventions. The tree is drawn upside down, the 'root' (the first question) at the top. Each question is a 'condition' or an 'internal node' that splits features or 'attributes.'

From each node there will be 'branches' or 'edges,' representing the possible answers. When you get to the end of the path through the tree so that there are no more questions/decisions then you are at a 'leaf.' There are also the obvious concepts of parent and child branches and nodes.


Root node
Root - Feature 1
This is a branch or edge.
It corresponds to one of the categories/answers for the feature
Inner or internal node
Answer 1 a
Answer 1 b
Feature 2
Feature 3
Answer 2 a
Answer 2 b
Answer 2 c
Answer 3 a
Answer 3 b
Feature 4
Prediction A
Prediction B
Prediction C
Prediction D
Answer 4 a
Answer 4 b
Prediction E
Prediction F
This is a leaf, a prediction

You can use decision trees to classify data, such as whether or not a mushroom is edible given various physical features. Those features could have binary categories such with/without gills, or multiple categories, such as colour, or be numerical, height of mushroom, for example.

Decision trees also can be used for regression, when you have numerical data, how much is a car worth based on make, model, age, etc.

Building or growing a tree is all about choosing the order in which features of your data are looked at and what the conditions are.

Example: Magazine subscription

I am going to work with member data from my own website now. I shall be taking data for a small subset of members to figure out which people are likely to subscribe to our magazine.

Below I show the first few lines of the raw data with all distinguishing information redacted.

ID	Employment Status	Degree Level	CQF Alumnus	Wilmott Magazine Subscriber
1	Self Employed	Postgraduate	No	No
2	Self Employed	Postgraduate	Yes	Yes
3	Employed	Postgraduate	Yes	Yes
4	Student/Postdoc.	Postgraduate	No	Yes
5	Student/Postdoc.	Undergraduate	Yes	Yes
6	Student/Postdoc.	Undergraduate	Yes	No
7	Employed	Undergraduate	Yes	Yes
8	Self Employed	Postgraduate	No	No
9	Self Employed	Undergraduate	No	Yes
10	Student/Postdoc.	Undergraduate	Yes	No
11	Self Employed	Undergraduate	Yes	Yes
12	Employed	Postgraduate	Yes	Yes
13	Employed	Undergraduate	No	Yes
14	Student/Postdoc.	Postgraduate	Yes	No
15	Employed	Postgraduate	No	Yes
16	Student/Postdoc.	Postgraduate	No	No
17	Self Employed	Postgraduate	No	No

There are three features we will look at: Employment Status; Highest Degree Level; CQF Alumnus

And the classification is whether or not they are magazine subscribers.

The obvious use of the any results that we find is in helping decide who to target for magazine subscriptions. Out of these 17 members there are ten who are magazine subscribers.

I am only going to work with this small subset of the full magazine dataset, so just the 17 lines.

Just looking at this table we cannot immediately see any single question, whether they are employed etc., that will determine whether or not someone is a magazine subscriber. So we are going to have to ask several questions. But in what order should we ask the questions?

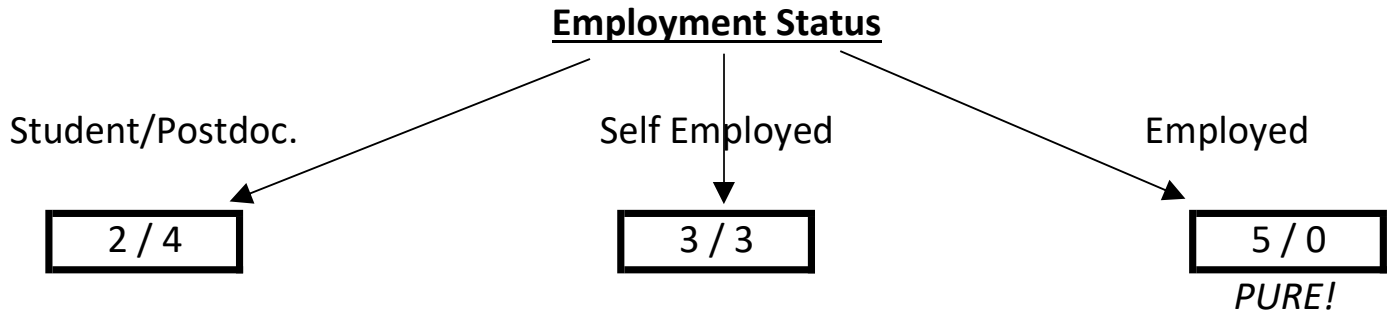
Now I must confess that I have tweaked the data a bit. That is so that I can get a single decision tree that shows as many possible outcomes as possible. You will shortly see what I mean.

I'm first going to explain how the final decision tree works, but how I constructed it in an optimal fashion will come a bit later.

Suppose I start by asking members of `wilmott.com` about their employment status and whether or not they are magazine subscribers. They can answer Employed, Self Employed or Student and Yes or No.

According to the data we would get results that could be represented as below. This is the root of our decision tree.

We read this as follows. In the boxes are two numbers. The number on the left is the number of people who are magazine subscribers and the number on the right those who aren't. Is this useful?



If they say they are Employed then this is very useful because five employed people say they are magazine subscribers and none say they aren't.

That is very useful information because as soon as we know someone is employed we know that they will be subscribers.

And no more questions need be asked.

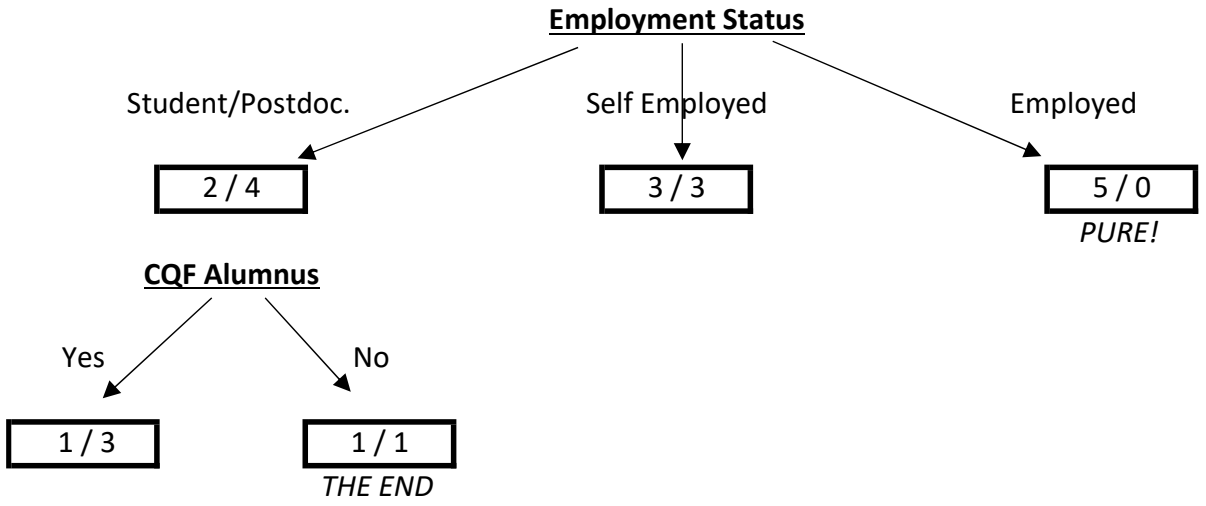
If only it were true that all employed people subscribed to the magazine. Sadly, this data has been massaged as I mentioned. In practice you would have a lot more data, you'd be unlikely to get such a clear-cut answer, but you'd have a lot more confidence in the results. However even with these numbers we can start to see some ideas developing.

When you get an answer that gives perfect predictability like this it is called 'pure.' That is the best possible outcome for a response.

The worst possible outcome is if you get the answer Self Employed because there is no information contained in that answer, it is 50:50 whether or not a self-employed person is a subscriber. In fact we seem to have gone backwards, at least before we had asked any questions we knew that ten out of 17 people were subscribers, now we are at a coin toss.

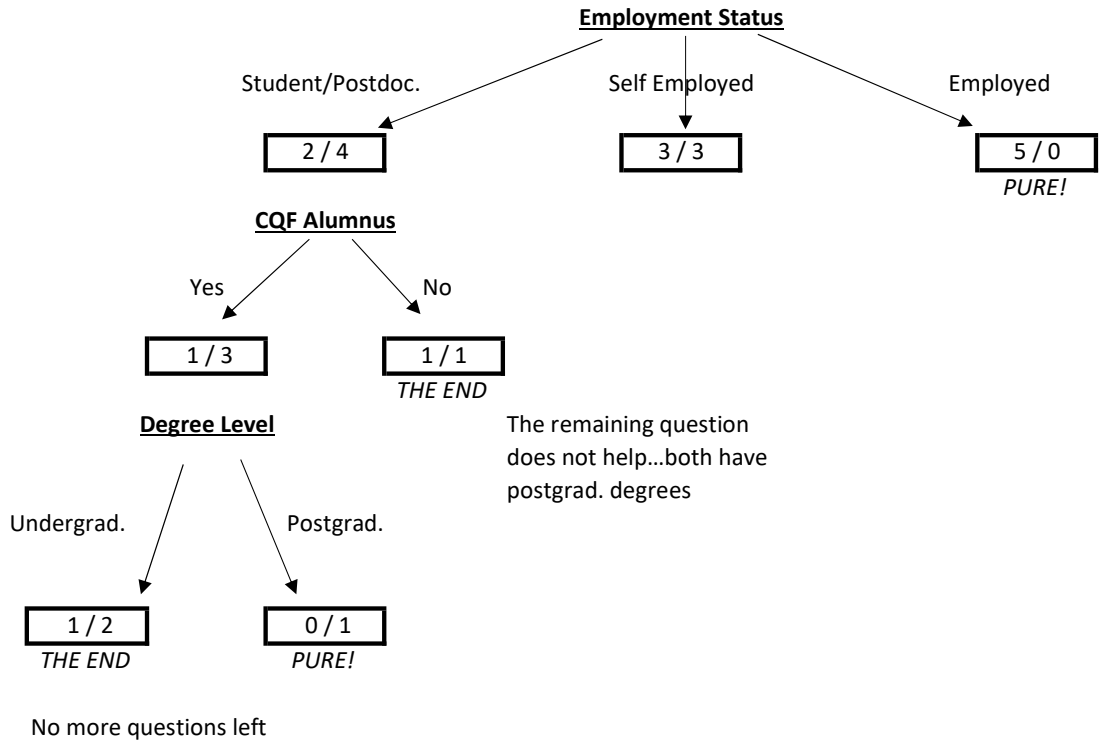
If someone is a Student/Postdoc. then two out of six are subscribers.

We can then ask whether they are a CQF Alumnus. Out of the six Students/Postdocs there are four who are alumni and two who aren't. Out of the four alumni there is only one magazine subscriber. And the two non alumni are equally split between subscribing and not:

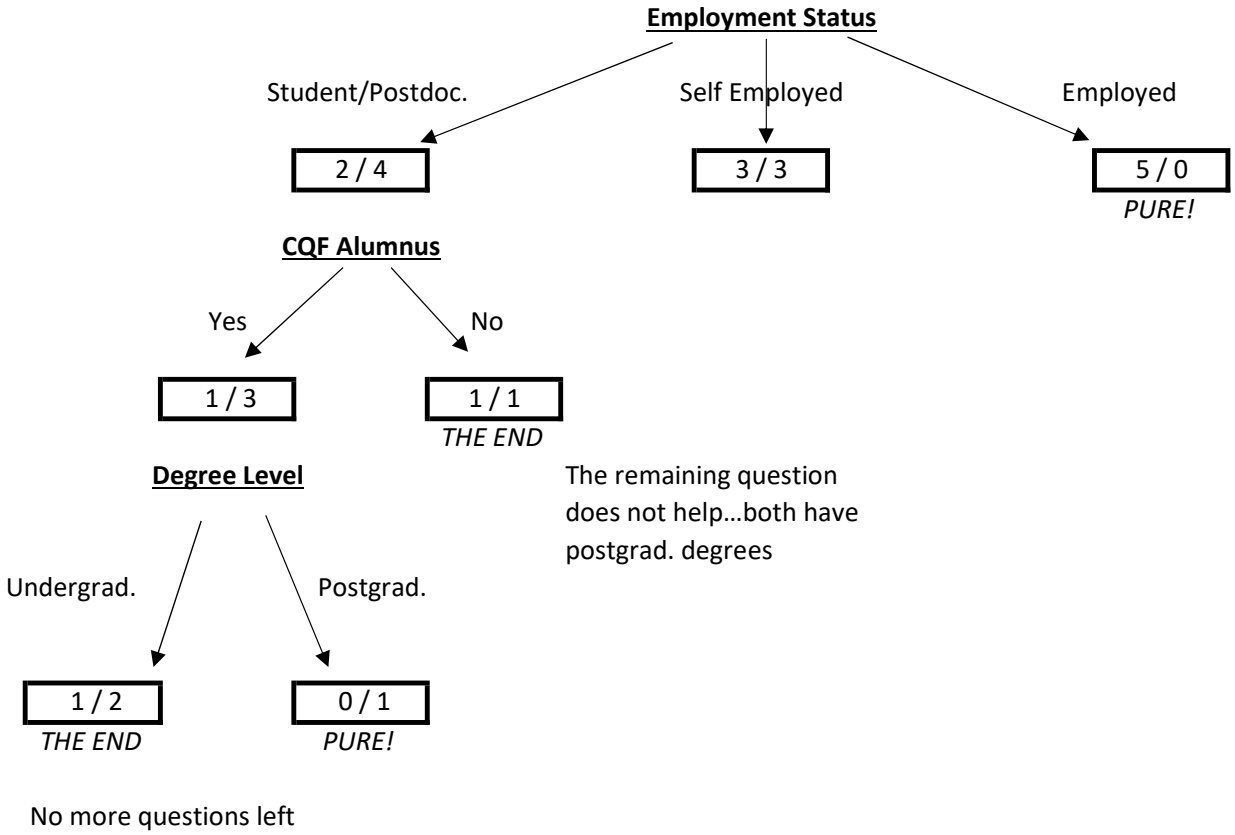


The remaining question does not help...both have postgrad. degrees

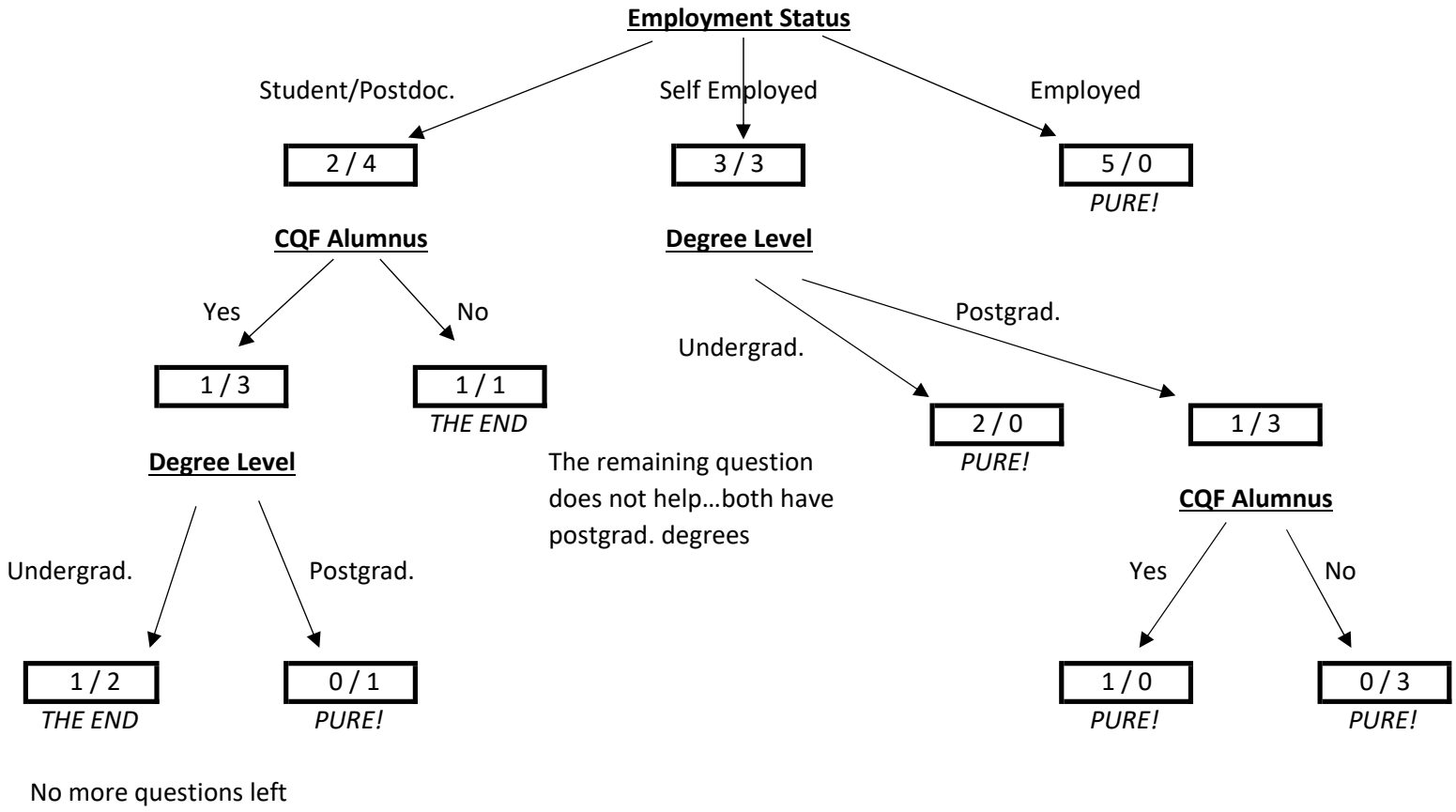
We can move further down the tree. We can ask those two non-alumni Students/Postdocs what their highest degree is. Unfortunately that does not help here because both have the same level, they have postgraduate degrees. There's nothing we can do with their answers to the three questions that will separate these two individuals.



Moving to the CQF Alumnus Students/Postdocs we look at the answers to the question about their highest degree.



We can continue like this to fill in the rest of the tree, as shown below. Luckily the rest of the tree results in pure splits.



If we are given a new data point, such as Self Employed with a Postgraduate Degree and who is a CQF Alumnus then we just run through our tree and we will see that such a person will be, like all the best people, a magazine subscriber. Although with so little data we would not be confident of that conclusion.

Some observations. . .

- Ideally you will end with a leaf that is a pure set, which classifies perfectly
- However if you have an impure set you might find that the remaining questions do not help to classify
- Or you might run out of questions
- You should always keep track of the numbers (in the boxes) because that will give you a probabilistic classification and a degree of confidence

- Notice how the same questions can appear at different places in the tree, although there will be no path along which you get asked the same question twice

- Questions don't have to be binary, yes/no

- And there don't have to be just two classifications (here whether or not they are magazine subscribers)

That's what decision trees look like and how they work.

But you should be asking how did I construct the tree?

Or more specifically how did I know in which order to ask the questions?

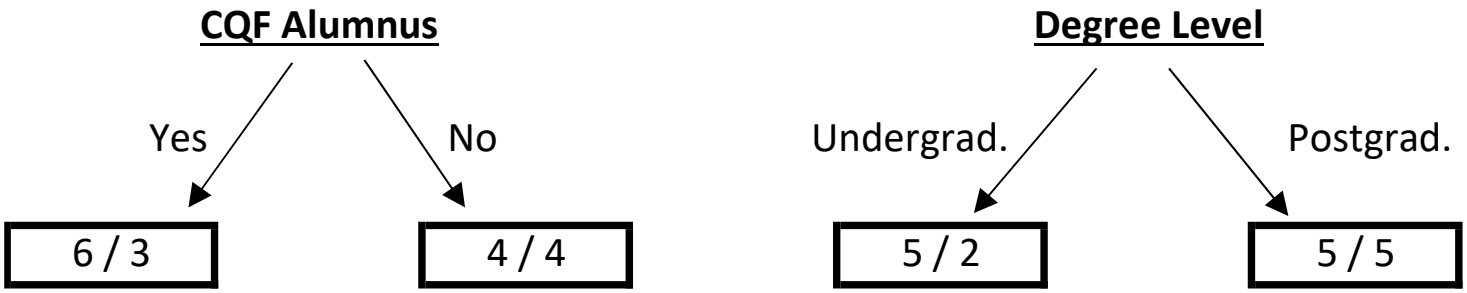
Why did I start with Employment Status as the root, and not another attribute? Which attributes best split the data? There was a clue in my occasional use of the word 'information' above. We are going to take inspiration from Information Theory.

Entropy

In layman's terms you want to find the attribute that gives you the highest information gain. I reckon that out of Employment Status, Highest Degree Level and whether or not a person is a CQF Alumnus the attribute that does the splitting the best, that gives the highest information gain, is Employment Status. But what makes me think that?

We need some way of looking at which attribute is best for splitting the data based on the numbers in each class. For example, we need some numerical measure that gives us how efficient the split is based on the numbers '2 / 4,' '3 / 3' and '5 / 0.'

But first we need to see what the Employment Status attribute is competing with. Below we see the splits you get from having a root that is either CQF Alumnus or Highest Degree Level.



It looks like both of these splits do a very poor job. Both of them have one branch that is a coin toss. And in both cases the other branch is not much better. But we need to quantify this.

The uncertainty in classification is going to be measured by the entropy.

All we need to do to determine the best attribute to start with, the root of our tree, is to measure the entropy for each attribute and choose the one with the lowest value. Usually this is done by measuring the 'information gain.'

To measure the gain we need a starting point to measure gain relative to. And that would be the raw data before it is split. We calculate the entropy for the original data, that is ten magazine subscribers and seven non subscribers:

$$-\sum p \log_2(p) =$$

$$-\frac{10}{10+7}\log_2\left(\frac{10}{10+7}\right) - \frac{7}{10+7}\log_2\left(\frac{7}{10+7}\right)$$

$$= 0.977.$$

It could be worse, i.e. closer to 1, but not much.

Now we go to the data and measure the entropy for each branch of each attribute.

For the Student/Postdoc. branch of Employment Status we have

$$-\sum p \log_2(p) = -\frac{2}{2+4}\log_2\left(\frac{2}{2+4}\right) - \frac{4}{2+4}\log_2\left(\frac{4}{2+4}\right) = 0.918.$$

For the Self Employed branch of Employment Status we have a rather obvious 1, since there is an equal number of subscribers and non.

And for the Employed branch the entropy is zero since the split is pure.

Now we measure the average entropy for Employment Status as

$$\frac{6}{17} \times 0.918 + \frac{6}{17} \times 1 + \frac{5}{17} \times 0 = 0.677.$$

That's because six of the 17 are Student/Postdoc., another six of the 17 are Self Employed and five are Employed.

Thus the information gain thanks to this split is

$$\text{Information gain for Employment Status} = 0.977 - 0.677 = 0.300.$$

We do exactly the same for the Highest Degree Level and CQF Alumnus attributes:

$$\text{Information gain for Highest Degree Level} = 0.034,$$

$$\text{Information gain for CQF Alumnus} = 0.021.$$

And thus Employment Status having an information gain of 0.300 is the easy victor, beating Highest Degree Level and CQF Alumnus. And so it becomes the root of the decision tree.

Having established the root attribute we now repeat this process at each branch.

The Decision Tree Algorithm

Step 1: Pick an attribute

Take one of the (remaining) attributes and branches and split the data.

Step 2: Calculate the entropy

For each branch calculate the entropy. Then calculate the average entropy over both/all branches for that attribute.

Return to Step 1 until all attributes have been examined.

Step 3: Set attribute to minimize entropy

Choose as the node the attribute that minimizes the entropy (or equivalently maximizes the information gain relative to the unsplit data).

Move down/across the tree and return to Step 1.

Numerical Features

What can we do if the answers to our questions are not categories but numerical? Suppose that we have data for people's heights, and whether or not they are magazine subscribers. The height data here is entirely made up!

ID	Employment Status	Degree Level	CQF Alumnus	Heights	Wilmott Magazine Subscriber
1	Self Employed	Postgraduate	No	174.6	No
2	Self Employed	Postgraduate	Yes	173.0	Yes
3	Employed	Postgraduate	Yes	185.2	Yes
4	Student/Postdoc.	Postgraduate	No	169.5	Yes
5	Student/Postdoc.	Undergraduate	Yes	179.9	Yes
6	Student/Postdoc.	Undergraduate	Yes	167.9	No
7	Employed	Undergraduate	Yes	177.7	Yes
8	Self Employed	Postgraduate	No	175.7	No
9	Self Employed	Undergraduate	No	176.5	Yes
10	Student/Postdoc.	Undergraduate	Yes	162.5	No
11	Self Employed	Undergraduate	Yes	178.4	Yes
12	Employed	Postgraduate	Yes	178.2	Yes
13	Employed	Undergraduate	No	173.0	Yes
14	Student/Postdoc.	Postgraduate	Yes	174.7	No
15	Employed	Postgraduate	No	188.0	Yes
16	Student/Postdoc.	Postgraduate	No	163.2	No
17	Self Employed	Postgraduate	No	159.5	No

One simple way of tackling this classification problem within a decision tree is to **choose a threshold s for the height so that being above or below that threshold determines the branch**. **The level of s can be chosen to maximize the information gain.** With this data, and if we used the height as the root attribute, then the information gain is maximized by a threshold of 176cm.

In principle you could have something more complicated than a simple threshold for determining the branch.

Now you can mix categories and numerical data and all of the above explanation for building your decision tree carries over.

Summary

Please take away the following important ideas

- Decision trees are sophisticated versions of 20 questions
- Use entropy minimization to build the most efficient tree