Deciding on the Decision Tree

We've run into an important question: how can we be sure that we're constructing the best possible decision tree?

We've seen that we want the "most important question" to be at the top of our tree, but how can we determine what this is? The "most important question" will be the one that reduces uncertainty the most. Let's return to our soccer game example.

Temperature	Rain	Humidity	Play
90	no	high	no
73	no	low	yes
81	yes	high	no
67	no	high	yes
72	yes	high	yes
77	no	low	yes
96	no	low	no
81	yes	high	no
58	yes	high	yes
72	no	medium	yes

1. Looking at the days with high humidity, how many games were played? How many games weren't played?

2. Based on the data, if the humidity is high, what is the probability that the game is played?

3. Looking at the days with low humidity, how many games were played? How many games weren't played?

4. Based on the data, if the humidity is low, what is the probability that the game is played?

5. Based on your answers above, which situation reduces uncertainty more, if the humidity is low or if it's high? Why?

In mathematics, uncertainty is measured using something called *entropy*[1]. If the probability of an event happening is p, then the probability of the same event *not* happening is $1-p$, and the *entropy* of this split is

$$-p\log_2(p) - (1-p)\log_2(1-p).$$

For example, let's look at when the temperature is above 75°. In this case, one game was played and four were not, so the probability of a game being played is $\frac{1}{5}$. So, the entropy is

$$-\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \left(1-\frac{1}{5}\right)\log_2\left(1-\frac{1}{5}\right) \approx 0.7219.$$

Entropy is always between 0 and 1, and larger values correspond to more uncertainty.

1. If the humidity is high, what is the probability that the game is played? What is the entropy?

2. If the humidity is low, what is the probability that the game is played? What is the entropy?

3. In which situation is there more uncertainty: when the humidity is high, or when it is low?

4. If the humidity is *not* high (so it's either low or medium), what is the probability that the game is played? What is the entropy?

5. If the humidity is *not* low (so it's either medium or high), what is the probability that the game is played? What is the entropy?

6. In which situation is there more uncertainty: when the humidity is not high, or when it is not low?

[1] You might have heard of entropy in physics, or some other context. Entry is used to describe uncertainty in many different contexts, including the disorder in the universe!

We would like to construct our decision tree in the way that reduces uncertainty the most[2]. First, we need to consider entropy of the entire system. Ignoring the weather, 6 games were played and 4 games were not. So the probability of a game being played is $\frac{6}{10}$. The entropy is then

$$-\frac{6}{10}\log_2\left(\frac{6}{10}\right)-\left(1-\frac{6}{10}\right)\log_2\left(1-\frac{6}{10}\right)\approx 0.9710.$$

That's a lot of uncertainty!

We want to figure out what question will split the data in the way that reduces uncertainty the most. We'll measure the uncertainty by computing the entropy of each branch, and averaging the entropies.

Let's start by splitting on whether or not the humidity is high. If the humidity is high, 3 games were played and 3 games were not. This means that if the humidity is high, then the probability that a game is played is $\frac{1}{2}$, and the entropy is

$$-\frac{1}{2}\log_2\left(\frac{1}{2}\right)-\left(1-\frac{1}{2}\right)\log_2\left(1-\frac{1}{2}\right)=1.$$

Next we need to find the entropy of the other branch, which is when the humidity is not high. If the humidity was low or medium, 3 games were played, and 1 was not. So, if the humidity is not high, then the probability that a game is played is $\frac{3}{4}$, and the entropy is

$$-\frac{3}{4}\log_2\left(\frac{3}{4}\right)-\left(1-\frac{3}{4}\right)\log_2\left(1-\frac{3}{4}\right)\approx 0.8113.$$

Now, we need to average our two entropies. However, we need to take into account that it's more likely that humidity is high than that it's not. Because of this, we compute the weighted average, using the probability that the humidity is high, $\frac{6}{10}$, and the probability that it's not, $\frac{4}{10}$. This weighted average is

$$\frac{6}{10}\cdot 1+\frac{4}{10}\cdot 0.8113=0.925.$$

From this weighted average, we can see that splitting on high humidity decreases entropy by

$$0.9710-0.925=0.046.$$

So we've decreased uncertainty by a bit, but hopefully we can do better! Let's try some different splits to try to find the best one.

1. What is the decrease in entropy if you split on whether or not the humidity is low?

[2] An alternative way to construct a decision tree is using something called the Gini index or Gini impurity. Gini impurity measures how likely a mistake is if you stopped at a branch and just randomly picked some outcome from that division. We won't cover this in our camp, but feel free to read about it if you're interested!

2. What is the decrease in entropy if you split on whether or not it rains?

3. Can you find the split that decreases entropy the most?

Once we find the split that reduces uncertainty the most, we can repeat this process for each of our branches, further reducing the uncertainty. Fortunately, we don't need to do this by hand! We'll be able to use Python to quickly construct decision trees.