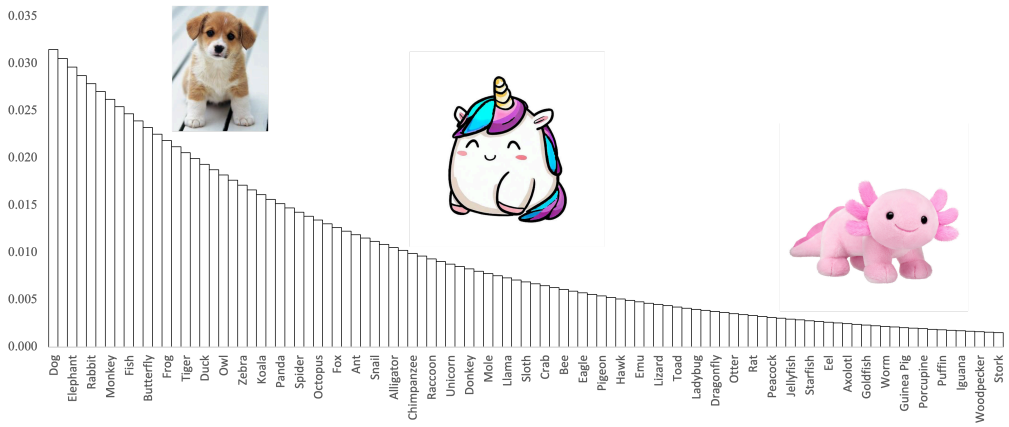
Information Theory

Information theory is an incredibly powerful perspective which plays a central role in a ton of algorithms, including Decision Trees, the WordleBot, Adaptive Tests, Compression of Data, Optimal Poker Play and even compression of data! It is a concept that is not too complicated, but can get confusing for some. The goal of this chapter is to balance showing off the awesome power of Information Theory while also keeping things as straight forward as possible. To that end, a great place to start is thinking about how you could write a bot that can play the question answering game of "Think of an Animal".

Think of an Animal?

The game of "Think of an Animal" goes like this. The human is going to be thinking of an animal. We assume that the distribution of how often they chose an animal is known (based off how popular the animal is to four year olds):

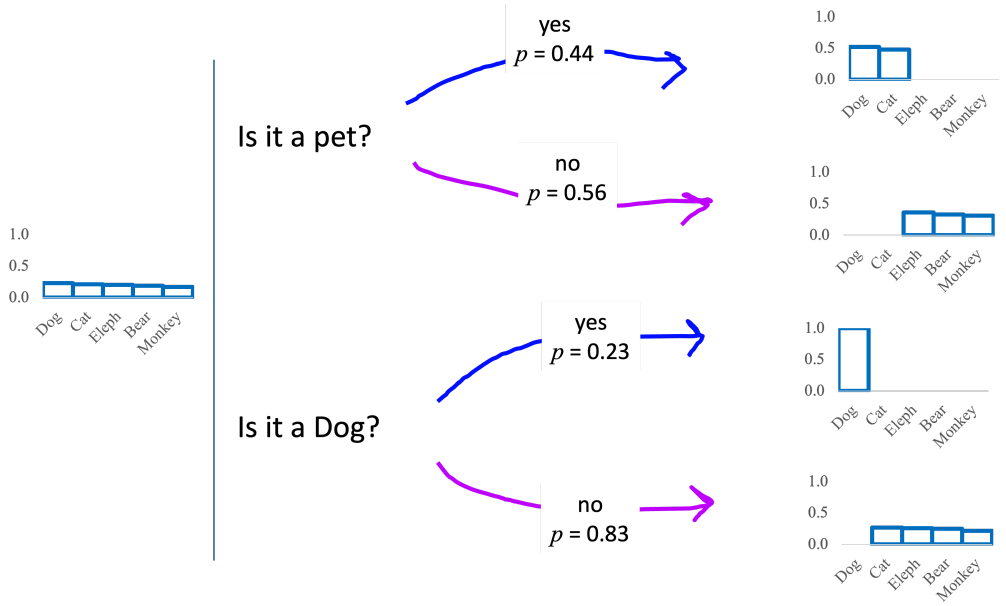


The task of your algorithm is to select which question to ask next. Assume you are given a bank of yes or no questions which include classics like:

- Is it a pet?
- Does it live in the water?
- Are you thinking of a dog?

Chosing a Best Question

How can we chose the best question to ask? Consider a simplified game with five animals (Dog, Cat, Elephant, Bear and Monkey) and only two questions to chose from "Is it a pet?" and "Is it a Dog?" For each question we can use the Law of Total Probability to think through the probability the response will be "yes" or "no". Even better! We can think through the resulting probability mass function of the animal random variable based on each possible answer to each possible question:



We are so close! If we could only quantify how much uncertainty the four resulting probability mass functions have, we could simply chose the question that minimizes our expected uncertainty as to what animal the person is thinking of. This quantification of uncertainty is formally called "Entropy" and it is the key concept in Information Theory.

Measure of Uncertainty from a High Level

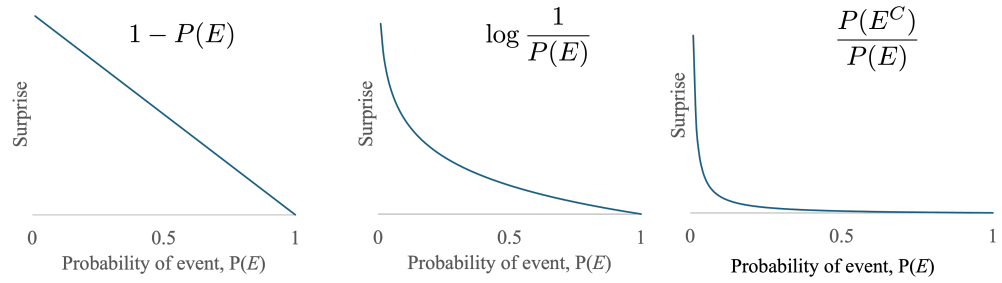
Let X be any random variable. A very elegant way to quantify how much uncertainty the Probability Mass Function of X represents is to think through all the values that X could take on and for each value calculate how surprised you would be if it turned out that X in fact took on that value. If you calculate a weighted sum over these surprise values, you would get the expected Surprise of the random variable

$$\begin{aligned}\text{Uncertainty}(X) &= E[\text{Surprise}(X)] \\ &= \sum_x \text{Surprise}(x) \cdot P(X = x)\end{aligned}$$

That is the big idea! The main remaining todo is to define what we mean by "Surprise" of an event.

Measure of Surprise of an Event

How should we quantify the degree to which we would be surprised if we were told that X took on the value x? There are many ways one could quantify Surprise, all of which are based on the probability of the event $\text{P}(X = x)$. Here are three reasonble Surprise functions:



All of these functions hit the following desiderata:

- Low probability events are surprising
- High probability events are not surprising
- Surprise is a monotonically decreasing function of probability

For many reasons, Information Theory defines Surprise using a variation of the middle equation

$$\text{Surprise}(E) = \log_2 \frac{1}{P(E)}$$

Looking at the relationship between $P(E)$ and the value of $Surprise(E)$ we can observe some of those intuitive relationships:

Probability of Event $P(E)$	**Surprise of Event** $\text{Surprise}(E) = \log_2 \frac{1}{P(E)}$
1/2	1
1/4	2
1/8	3
1/16	4
1/32	5
1/64	6
1/128	7

In other words, if an event with probability $P(E) = 1/16$ were to occur we would be four times as "Surprised" as if an event with $P(E) = 1/2$ were to occur. That feels nice!

Definition: Surprise of an Event (Information Content)

The information content, also called the surprisal or self-information, of an event E is a function which increases as the probability of an event decreases. When the probability is close to 1, the surprisal of the event is low, but if the probability is close to 0, the surprisal of the event occuring is high. This relationship is described by the function:

$$\text{Surprise}(E) = \log_2 \left(\frac{1}{\text{P}(E)} \right)$$

This can be written (in a more confusing way) as:

$$\begin{aligned} \text{Surprise}(E) &= \log_2 \left(\frac{1}{\text{P}(E)} \right) \\ &= \log_2 \text{P}(E)^{-1} \\ &= -\log_2 \text{P}(E) \end{aligned}$$

In the initial definition of Surprise of an Event, the function name I was used as shorthand for Surprise. I stands for "Information Content" or "Self Information", two other names for Surprise of an event.

There are many other stories that we could tell for why $\log_2 \frac{1}{\text{P}(E)}$ is a great choice for our measure of Surprise. Claude Shannon, the father of information theory, chose the base 2 logarithm because it would allow you to express your amount of surprise in bits (as in 0, 1 values used in a computer). Information theory has many applications, but it was first invented when he was trying to come up with a way to optimally compress text based data!

Uncertainty of a Random Variable (Entropy)

Now that we have a formal definition of Surprise we can revisit the computation of Uncertainty of a random variable.

Definition: Uncertainty of an Random Variable (Entropy)

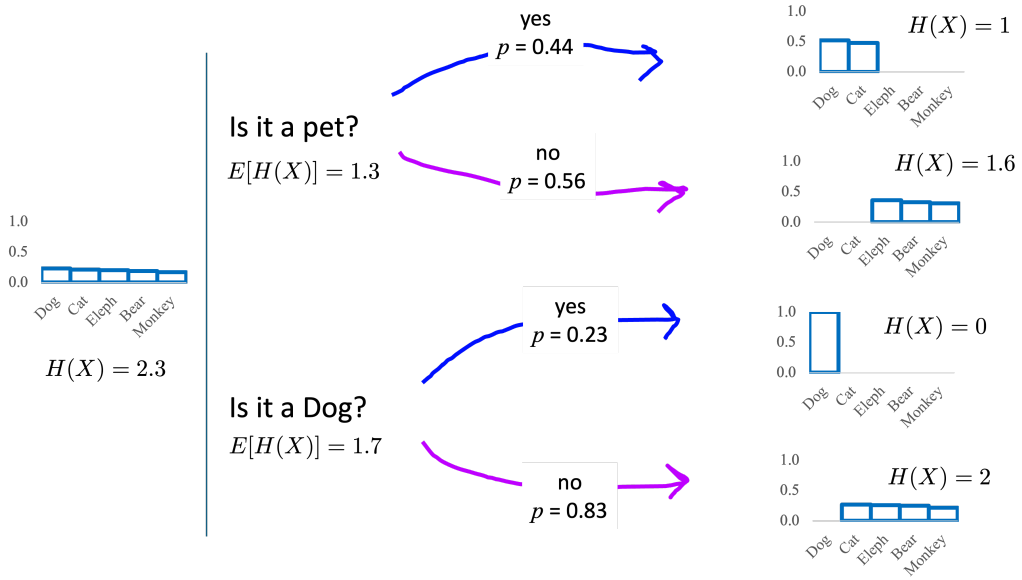
Let H be our measure of how much uncertainty we have about a random variable X. Define H to be the expected surprise of observing the assignment to X. $H(X) = E[\text{Surprise}(X)]$. Using the Law of Unconcious Statistician, and the definition of Surprise of an event, we can expand out the formula for H as:

$$H(X) = \sum_{x \in X} \log_2 \left(\frac{1}{\mathrm{P}(X = x)} \right) \cdot \mathrm{P}(X = x)$$

Uncertainty (H) can also be rewritten (in a more confusing way) as:

$$\begin{aligned}
\mathrm{H}(X) &= \sum_{x \in X} \text{Surprise}(X = x) \cdot P(X = x) \\
&= \sum_{x \in X} \log_2 \frac{1}{P(X = x)} \cdot P(X = x) \\
&= \sum_{x \in X} \log_2 P(X = x)^{-1} \cdot P(X = x) \\
&= \sum_{x \in X} -\log_2 P(X = x) \cdot P(X = x) \\
&= -\sum_{x \in X} \log_2 P(X = x) \cdot P(X = x) \\
&= -\sum_{x \in X} \log_2 P(X = x) \cdot P(X = x)
\end{aligned}$$

We now have all the theoretical tools we need to select our best question in the game of "Think of an Animal". For each possible resulting Probability Mass Function, we can calculate the Uncertainty (H) of that PMF:



The question "Is it a pet?" has an expected uncertainty of 1.3. The question "Is it a Dog?" has an expected uncertainty of 1.7. As such we would be less uncertain about what animal our friend is thinking about, in expectation, if we were to ask the question "Is it a pet?"

This is one of many applications of Information Theory!