Information Theory

Curtis Liu & Sunny Liang

2023-12-03

1 Warm-up: Balance Puzzles

A mysterious criminal organization has been circulating counterfeit coins. They are identical to genuine coins except that they are very slightly lighter. Given a simple balance and a set of coins in which all are genuine except for one, you are tasked with finding the counterfeit using the balance as few times as possible.

1. Suppose you have 3 coins, one of which is counterfeit. What is the smallest number of times you would need to use the balance to guarantee that you find the counterfeit coin?

2. How about with 9 coins?

3. 243 coins?

4. Propose a way to determine the odd coin from 13 coins using only 3 weighings.

5. Suppose you have 12 coins with three possibilities: All are genuine, one is counterfeit and is lighter, or one is counterfeit and is heavier. Can you determine if there is a counterfeit in just three trials? If so, how can you find the counterfeit and tell whether it's lighter or heavier in just three trials?

2 Information and Entropy

Information theory was pioneered by researchers and mathematicians named Harry Nyquist, Ralph Hartley, and Claude Shannon. All three of them have a lot of cool things named after them so you might have heard their names before. Claude Shannon was especially cool and has a long Wikipedia page. His name comes up a lot.

So, what does it mean for something to have information? Let's consider the following statements:

- Harold has 10 toes.
- Harold has 11 toes.

6. Which above statement is more likely? Which statement if confirmed is more informative?

Consider an event x, we write the probability of x as $P(x)$ and the information gained by observing x as $I(x)$, which is a quantification of the information gained from x. For example, $I(x = \text{the sky is blue})$ is intuitively much less than $I(x = \text{the sky is green})$, and our goal is to find a way to model $I(x)$ numerically. $I(x)$ is called the *self-information*, *information content*, or *surprisal*, which are all highly-suggestive names. Shannon said that, intuitively, our measure of information should have the four following properties:

- $I(x) \geq 0$
- $P(x) = 1 \implies I(x) = 0$
- $P(x \cap y) = P(x) \cdot P(y) \implies I(x \cap y) = I(x) + I(y)$
- $I(x)$ and $P(x)$ are inversely related (i.e. if $P(x) > P(y)$ then $I(x) < I(y)$)

7. Discuss with the people around you what each of these axioms mean and why they make sense. Then, come up with a function $I(x)$ that satisfies these axioms.

   Hint 1: The third and fourth axioms suggest a very specific type of function.

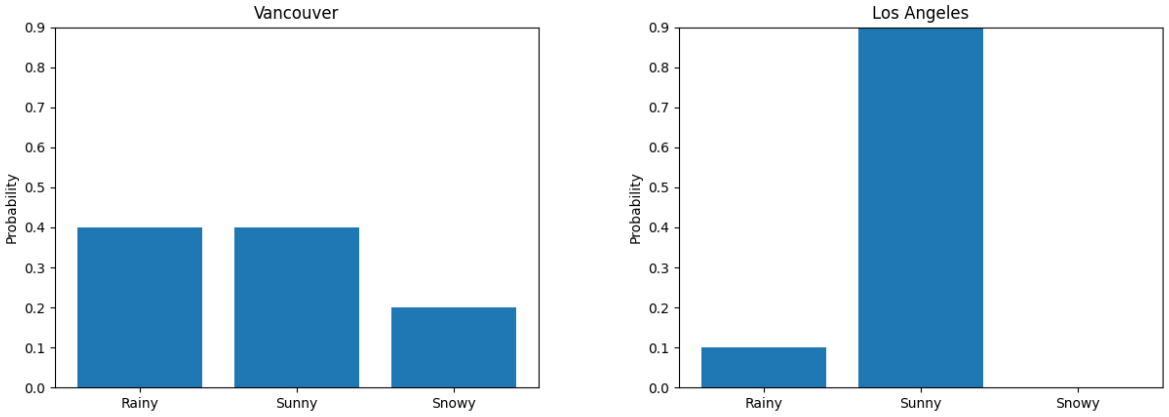
   Hint 2: There are actually infinitely many functions that work, though they are all of essentially the same form. Which one you choose merely changes the units of your function's output.

8. Using your answer from the previous question, evaluate the following in bits (ask for help if needed!):

   - $I(x = \text{a fair coin lands heads})$
   - $I(x = \text{you roll a 1 on a fair dice})$
   - $I(x = \text{Harold has 10 toes})$
   - $I(x = \text{Harold has more than 10 toes})$

   You can use Google to help estimate the last two.

Consider Figure 1 showing two probability distributions describing the weather in Vancouver and Los Angeles respectively. Clearly, the weather in Los Angeles is a lot more predictable than in Vancouver. So there is a sense in which knowing the weather in Vancouver is usually more informative. We can make this



(a) A graph.

(b) Another graph.

Figure 1: Two graphs.

concrete using a quantity that Shannon called *entropy*[1] at the suggestion of John Von Neumann (another pretty cool mathematician).

9. Entropy measures how informative events are on average for a given distribution. This takes into account both how likely each event is and how informative it is. Given this information, define entropy in terms of information $I(x)$ and probability $P(x)$.

   Hint: If you were here for the first two weeks this should be easy. Preferably, you would use the same sort of notation as we used with probability.

10. Calculate and compare the entropies for the weather in Vancouver and Los Angeles given the above distributions.

[1]Some of you may recognize this word from thermodynamics. It's not obvious what they have to do with each other, but the choice to name information entropy after thermodynamic entropy is not accidental!

11. Here's an application of Shannon's entropy in chemistry. Define W as the number of microstates of a molecule, or the number of different possible arrangements. Suppose further that all the microstates have an equal probability, say p_i. To model the entropy expression, the coefficient of your proposed answer is known as the *Boltzmann Constant* k_B, which appears quite frequently in chemistry. Write an expression for entropy S in terms of k_B and W.

3 Data Compression

The goal of data compression is to take information and encode it in a way that uses as few bits as possible. We will be dealing specifically with *lossless compression* which retains all information, as opposed to *lossy compression* which approximates the original data with far fewer bits at the expense of some information[2]. Suppose you are on a popular television game show. You are presented with four doors and you will be rewarded with whatever is behind the door of your choice. Three of the doors have a goat behind them and one of them has a car.

12. In lieu of any further information, the car is equally likely to be behind any of the four doors. So, what is the entropy of the situation? To clarify, you have four possible outcomes in your even distribution of probabilities, one for each door the car could be behind.

13. You are allowed to ask the host, Monty Hall, a series of yes-or-no questions. For some reason, Monty will smash a window for every question you ask, so you'd better not ask too many. What is the optimal strategy? That is, find the strategy that minimizes the number of questions asked on average.

[2]Have you ever noticed how blurry YouTube videos are with slower internet?

14. Now suppose that Monty tells you that the there is a 50% chance the car is behind the first door, 25% behind the second door, and 12.5% for the third and fourth doors. What is the entropy now?

15. What is the new optimal strategy? What is the average number of questions asked for this strategy?

In the previous thought experiment, there were four relevant outcomes corresponding to each of the doors the car could have been behind. Each question we asked gave us a bit of information, so we might assign 1 for yes and 0 for no. Then each outcome can be encoded in binary according to the sequence of answers we got from our questions.

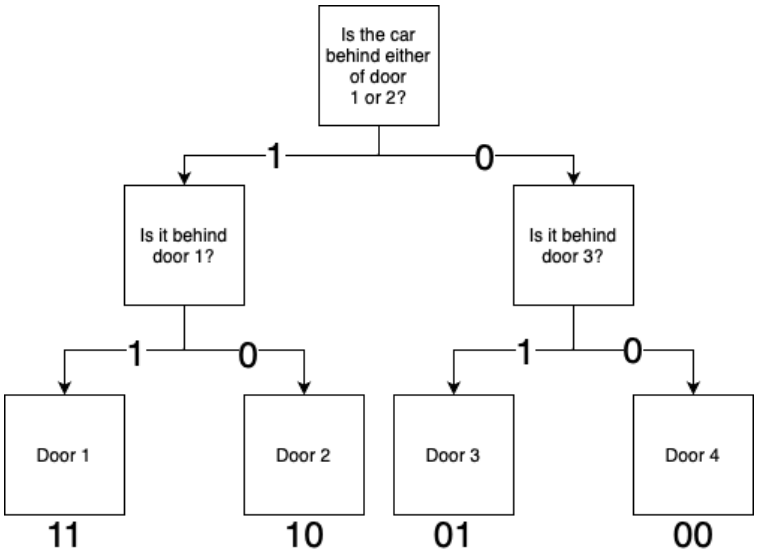


Figure 2: Flowchart showing how we might encode the case where each door is equally likely to have the car. This type of structure is called a *binary tree* and is very useful in computer science.

16. Analogous to Figure 2, draw a similar flowchart and encoding for the optimal strategy found in problem 9. Define the *symbol length* to be the number of digits encoded to a specific outcome. What is the average length of the symbols you used to encode the four outcomes weighted by their probability?

Clearly, the number of questions asked and symbol length are synonymous under this encoding scheme. Notice how an optimal encoding strategy assigns shorter symbols to the most likely outcomes, therefore minimizing the average symbol length. This is one of the most important ways that data compression actually works on things like documents, images, and videos.

Also, notice how the entropy seems to give us a lower bound on the expected symbol length? It turns out that this is always true, it's called *Shannon's source coding theorem*. In the examples I picked, the optimal encoding happens to exactly equal the entropy[3], but this is not always the case.

17. Suppose you are to encode a message consisting of only the first five letters of the alphabet. You are given the following probability distribution:

$$P(A) = 0.17,\ P(B) = 0.35,\ P(C) = 0.17,\ P(D) = 0.15,\ P(E) = 0.16.$$

The most straightforward way to encode this is just to assign each letter a sequential 3-bit binary number (so $A = 000$, $B = 001$, and so on). Your task is to improve on this, thereby compressing the data.

First, calculate the entropy of this distribution, then devise an encoding with an expected symbol length shorter than three bits. Compare your symbol length with some other people in the class. I'd be very impressed if any of you manage to get 2.30.

[3]These are called *dyadic* distributions. The word *dyadic* comes up elsewhere in math and other fields.

4 Bonus: Bulls and Cows

Bulls and Cows is basically an older and simpler version of Wordle. You can play it with a partner:

- Player A writes down a 4-digit code with no repeated digits. Do not show it to Player B, that would defeat the purpose.
- Player B guesses a 4-digit code.
- Player A responds with the number of bulls and cows. A bull is a correct digit in the correct place. A cow is a correct digit in an incorrect place.
- Player B keeps guessing until they get the answer.

Find somebody to play this with. Can you apply what you just learned about information theory to optimize your guesses?

What if, instead of using 4-digit codes, we used letters to form real words? Try playing this variation with a friend. Can you describe your most optimal strategy?