Entropy and Information

Robert deCarvalho

1 Introduction

Entropy is one of those words that many people use, but few really understand. This might be because the concept is invoked in seemingly very different contexts. For example, your physics teacher might have told you, “The entropy of the universe is increasing!” Or perhaps you encountered the word while reading about compression. “This algorithm compressed our file to within 10% of the entropy limit.” Or maybe you vaguely remember entropy as something your chemistry teacher told you about and maybe it had something to do with reactions, but you can’t quite remember. Well, hopefully by the end of this chapter you will have a decent grasp of what entropy is and how you can make practical use of it.

2 Why Should You Care About Entropy

Before we get into what entropy is, let’s talk a bit about why we even need such a concept, and how it is useful.

1. Working with probability, want to be honest about minimizing what you know.
2. Can measure population diversity.
3. Help you reason effectively in the face of uncertainty. Take into account what you know, but make sure you are honest about what you don’t know.

3 A Working Definition of Entropy

Our quest to understand entropy will start with a simple definition.

Definition 1. *Entropy is the amount of information you do not know.*

That’s a pretty short definition, but why would you possibly care about how much information you don’t know. Shouldn’t you care much more about you do know? Also, what does it even mean to talk about an “amount of information?” You probably have an intuitive sense of what “information” means, but how is it quantified? Would you be able to identify when you double the information you have on some topic As a guide to understanding how information can be measured, we turn to a contrived example that will illustrate how information can be quantified.

3.1 The Library of Information

Imagine you work at a library containing exactly $1,000$ books. Instead of the Dewey Decimal System, the books have been given sequential labels from 000 to 999. The books are arranged onto 100 shelves each holding 10 books and labeled with two digits corresponding the first two digits of all the books it contains. When patrons enter the library, they present you with slips of paper containing the three-digit number of the book they’d like to check out. It is your job to retrieve the correct book from the shelves.

A patron arrives and hands you a slip of paper with the number 248. Let’s pause and consider the information you have at your disposal to retrieve the desired book. You know that every book in the library has been labeled with a number, and that the books are ordered sequentially onto appropriately labeled shelves. By looking at the first two digits on the patron’s slip of paper, you can navigate directly to shelf 24. By looking at the last digit, 8, you know immediately that you want the second to last book on that shelf. Because of the way the library has been organized, you can walk directly to the desired book and retrieve it on the first try. You have perfect information on where the book will be, and there is no missing information to impede your search.

Now consider a modified scenario in which you arrive at the library early one morning only to discover that a prankster has come in overnight and covered all the books with identical dust jackets. There is no way to distinguish between the books because they all look the same. To further complicate your life, the miscreant has gone through each shelf and scrambled the order of the books. Fortunately, he was kind enough to ensure that each book remained on the appropriate shelf, but its location on that shelf is now a complete mystery. Now your first patron of the day arrives bearing a ticket with the number 369. What information do you have now to aid your search for the book. Because of the prank, all you know is that the book is on shelf 36, but you don’t know where on that shelf it is. In order to actually find the book, you will need to remove the dust cover each of each book until you locate book 369. You no longer have perfect information for locating the book. You’re going to have to make at least one guess, and you may need to make up to ten guesses in order to retrieve it. There is now missing information – you don’t know how the books on shelf 36 are arranged.

Now imagine a third scenario in which the prankster has abandoned all civility and has completely shuffled all the books in the entire library, moving books from shelf to shelf in a completely random way. In this scenario, when a patron hands you a ticket, the only thing you know is that the book is somewhere in the library, but you have no idea where. You have no information to help you in your search, so your only option is to one by one remove the dust cover on each book until you find the what you're looking for.

In each of these three scenarios, you have a different amount of information available to aid in locating a specific book. Let's lay the groundwork for determining how much information you have in each case. To do so we consider the labeling system. Each book has a three digit number associated with it. If you were to write down the call number of each book in the library on a piece of paper, you would end up writing $3 \times 1,000 = 3,000$ digits. It turns out that the number of digits you need to write down to catalog each book is an excellent way to measure information. So, for example, if had $10,000$ books in your library instead of $1,000$, you would end up needing $4 \times 10,000 = 40,000$ digits to account for all the books. This means that a sequentially organized library of $10,000$ books contains $40,000$ digits of information. Note that the information I'm talking about here has to do with the location of the books and that's all. If we were talking about any other property of the books other than their labels and locations in the library, then we would have to increase the scope of what we mean by the word "information".

Let's now think about each of the scenarios and determine how many digits of information you know in each of them.

In the first scenario, the library was perfectly ordered with $1,000$ books, so with the three-digit labels on each book, you had $3 \times 1,000 = 3,000$ digits of information. You knew everything you needed to locate any book, so we'll say you had 0 digits of missing information.

In the second scenario, the ordering of books on each shelf was scrambled, but you still knew which of the 100 shelves contained any book you were looking for. So let's do the math. You know the first two digits of all the books on a given shelf, so thats $2 \times 10 = 20$ digits of information you know. What you don't know is the last digit of the books, so that's $1 \times 10 = 10$ digits of missing information. There are 100 shelves, so in total, you know $20 \times 100 = 2,000$ digits of information and you are missing $10 \times 100 = 1,000$ digits. Notice that the total amount of information remains the same. The effect of the prankster was to remove $1,000$ digits of known information.

Finally, in the third scenario, you had no idea where any book might be located. You have 0 digits of information. You are missing all the information, which is three digits for each of the $1,000$ books, or $3 \times 1,000 = 3,000$ digits.

3.2 Entropy In The Library of Information

Let's take the information accounting we did in the last section and apply our definition of entropy to identify the entropy in each of the three library scenarios we presented.

In the first scenario, we knew all the information, and no information was missing, so there was 0 digits of entropy. In the second scenario, the amount of information we didn't know (i.e. the entropy) was $1,000$ digits. And finally in our last scenario, all the information was unknown, so there was $3,000$ digits of entropy.