An Introduction to Information Theory: Notes

Jon Shlens | *jonshlens@ucsd.edu*

03 February 2003

1 Preliminaries

1.1 Goals

1. Define basic set-up of information theory
2. Derive why entropy is the measure of information capacity.
3. Discuss the basics of mutual information
4. Solve binary symmetric channel

1.2 Probability Theory

- **probability distribution relations**

$$\begin{aligned}
\textit{joint}: \quad & P(X,Y) && \neq P(X) \times P(Y) \\
\textit{marginal}: \quad & P(X) && = \sum_{y_i \epsilon Y} P(X, Y = y_i) \\
\textit{conditional}: \quad & P(X|Y) && = \sum_{y_i \epsilon Y} P(Y = y_i) P(X, Y = y_i) \\
\textit{Bayes' rule}: \quad & P(X,Y) && = P(X|Y) \times P(Y) = P(Y|X) \times P(X)
\end{aligned}$$

- **iid** = independently and identically distributed

$$P(X_1, X_2) = P(X) \times P(X)$$

2 A Simple Example

2.1 Situation A

Pretend we like to buy and sell a particular commodity - how about pork bellies at the Chicago Mercantile Exchange in 1860. We talk to our trader every day and tell him one action a day: **BUY** , **SELL** or **HOLD** .

One day we decide that we are going on a trip to Europe but we would like to keep trading. Because phones don't exist, we decide on a simple system. We use a telegraph line to send a Morse code signal of a *dot* (denoted 0) or a *dash* (denoted 1). Here is what we agree on:

- We will **BUY** and **SELL** exactly one half of the days; we will never **HOLD** .
- We will send a 0 repeatedly if it is a **BUY** and a 1 repeatedly if it is a **SELL** .

2.2 Qualtitative Analysis of A

- **question:** what does the trader learn by receiving a 0 or 1?
- **before signal:** equal chance of a **BUY** or a **SELL** but never **HOLD** .
- **after signal:** 0,1 denotes with 100% certainty to either **BUY** or **SELL**

2.3 Situation B

Same sitaution as above but now let's say that our telegraph machine is *noisy.* Most of the time that we press a 0 or 1, the trader receives a 0 or 1, respectively. But occasionally, say, 20% of the time, the trader receives the opposite.

2.4 Qualitative Analysis of B

- **question:** what does the trader learn by receiving a 0 or 1?
- **before signal:** same as situation A
- **after signal:** Not 100% certain what order was. However, the trader does have a good hunch.

2.5 Statements about Classical Information Theory

1. There exists a preset, agreed-upon model between the *sender* to *receiver.*
2. Information is usually measured in *bits.*
    - 1 question in the game of 20 questions
3. **Information is selection between possible alternatives.**
    - **deep point:** the quantity of information does **not** depend on the complexity of the preset alternatives.

3 Intuitive Example

Pretend we have a set of possible messages $X = \{x_1, x_2, \ldots, x_N\}$ all with equal probability $\{p : p_i = \mathtt{p} \text{ for } i = 1, 2, \ldots, N\}$. We plan to send only one message x_i through our channel.

3.1 A Simple Game

Each element of X is *labeled* with a number $j = 1, 2, \ldots, N$.

- Pretend that you are the *sender* and you are about to transmit one symbol x_i. Your friend will be the *receiver.*
- Let your friend try to guess which symbol you will send.
- This game is a formal version of *20 questions.*
- **conclusion:** how many questions does your friend need to select N equally probable numbers?

3.2 Define the uncertainty

If the answers to your questions are *yes* or *no*, then we attach an equation to this situations. Let H be the average minimum number of questions your friend needs to guess which symbol you will send.

$$\begin{aligned}2^H &= N \\ H &= \log_2 N \\ H &= -\log_2 \frac{1}{N} \\ H &= -\log_2 \mathtt{p}\end{aligned}$$

We define H as the Shannon *entropy*.

4 Entropy

1. The Shannon entropy is the one and the same from thermodynamics.
2. Entropy measures the number of possible states in a system.
   - equivalent to a measure of uncertainty, variability or even "concentration" in a pdf.
3. In base 2 the units of entropy are bits.
   - In many theoretical treatments, base e is measured in *gnats.*
4. The most general form of entropy for $X = \{x_1, x_2, \ldots, x_N\}$ and $P = \{p_1, p_2, \ldots, p_N\}$ (non-equal probabilities) is:

$$H(P(X)) \equiv \langle -\log_2 p_i \rangle = -\sum_i^N p_i \log_2 p_i$$

5. Entropies can generalize to continuous distributions.
   - **discrete distributions:** $H \geq 0$
   - **continuous distributions:** not well defined.

5 Properties of Entropy

5.1 Sending two symbols

The same equiprobable situation as the previous example. However, this time we will send two symbols x_i and x_j. What is the entropy of sending two symbols x_i and x_j?

$$\begin{aligned}H(x_i, x_j) &= -\sum_{i,j=1}^N \frac{1}{N^2} \log_2 \frac{1}{N^2} \\ H(x_i, x_j) &= \log_2 N^2 \\ H(x_i, x_j) &= 2\log_2 N \\ H(x_i, x_j) &= \log_2 N + \log_2 N\end{aligned}$$

5.2 Conclusion

1. Information is additive.
2. Entropy grows as more symbols sent.
    - Entropy is an *extensive* quantity.

6 Mutual Information

- The goal is to formally quantify the **reduction in uncertainty** by examining the appropriate subtraction of entropies.
- Let us first look at the probability distributions of the *receiver* before and after one symbol is sent.

  beforehand: $P(X)$
  afterwards: $P(X|Y) = \sum_i P(Y = y_i)P(X|Y = y_i)$

- By our definition of uncertainty, the reduction in entropy between the two probability states is defined as the *mutual information.*

$$\begin{aligned} I(X;Y) &= H(P(X)) - H(P(X|Y)) \\ &\textit{or} \\ I(X;Y) &= H(X) - H(X|Y) \end{aligned}$$

- Mutual information is also measure in *bits.*

6.1 Relations between entropies

I will just not justify these statements but it is easy to work out. Regardless of whether one remembers the details of these equations, it is much easier to remember the Venn diagram in Figure 1.

- mutual information is symmetric $I(X;Y) = I(Y;X)$
- mutual information can be defined many ways

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ I(X;Y) &= H(Y) - H(Y|X) \\ I(X;Y) &= H(X) + H(X) - H(X,Y) \end{aligned}$$

7 Simple Examples, Returned

We will now return full circle and calculate the mutual information I in the two beginning examples. In other words, we will formally quantify our previous qualitative notions.

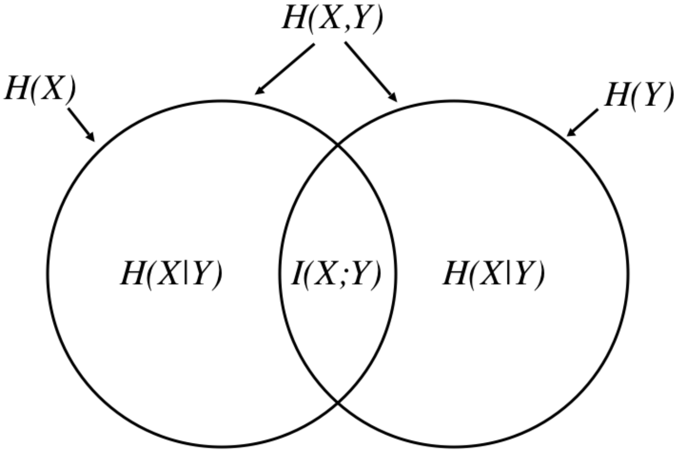

Figure 1: Venn diagram of relations between 2 variable entropies

7.1 Example A: Returned

$$X = \{\textbf{BUY}, \textbf{SELL}, \textbf{HOLD}\}, P(X) = \left\{\frac{1}{2}, \frac{1}{2}, 0)\right\}$$

The entropy beforehand $H(X)$.

$$H(X) = -\left[\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2} + 0\log_2 0\right]$$

But notice that $0\log_2 0$ is not finite. This brings up the *complicated* issue of *support* which authors go a great length to address. The simple, ad-hoc way avoiding these proofs is just to state in the context of information theory $0\log_b 0 \equiv 0$.

$$\begin{aligned} H(X) &= -\left[\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2} + 0\right] \\ &= \frac{1}{2}[1+1] \\ &= 1 \end{aligned}$$

Before calculating $H(X|Y)$, we need to compute $P(X|Y)$.

$$\begin{aligned} P(X|Y=0) &= \{1,0,0\} \\ P(X|Y=1) &= \{0,1,0\} \end{aligned}$$

Now we can compute the associated entropy.

$$H(X|Y) = P(Y=0)H(X|Y=0) + P(Y=1)H(X|Y=1) = 0 + 0 = 0$$

Therefore, the mutual information is $I(X;Y) = H(X) - H(X|Y) = 1 - 0 = 1$ *bit*.

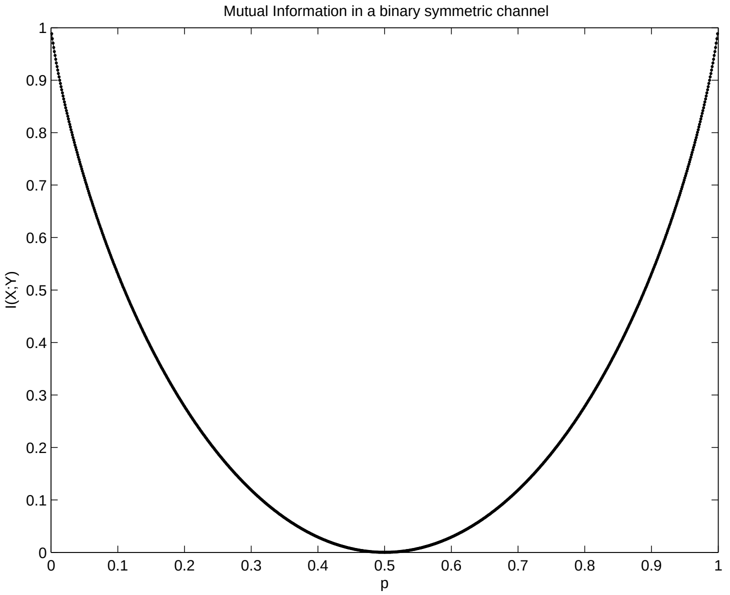

Figure 2: Mutual information in a binary symmetric channel

7.2 Example B: Returned

First of all, I need to state beforehand that this problem is a famous first chapter problem in any information theory textbook. It is often called the **binary symmetric channel** or more colloquially the **noisy typewriter**.

Let's just state all of the probability distributions before calculating the entorpies. Let the variable $p = 0.2$ be the probability of incorrect transmission.

$$\begin{aligned} P(X) &= \left\{\frac{1}{2}, \frac{1}{2}, 0\right\} \\ P(X|Y=0) &= \{1-p, p, 0\} \\ P(X|Y=1) &= \{p, 1-p, 0\} \end{aligned}$$

We can now calculate all of the entropies.

$$\begin{aligned} H(X) &= 1 \\ H(X|Y=0) &= -\left[(1-p)\log_2(1-p) + p\log_2(p)\right] = (1-p)\log_2\frac{1}{1-p} + p\log_2\frac{1}{p} \\ H(X|Y=1) &= -\left[(1-p)\log_2(1-p) + p\log_2(p)\right] = (1-p)\log_2\frac{1}{1-p} + p\log_2\frac{1}{p} \end{aligned}$$

Finally we can calculate the mutual information.

$$\begin{aligned} I(X;Y) &= H(X) - [P(Y=0)H(X|Y=0) + P(Y=1)H(Y=1)] \\ I(X;Y) &= 1 - \left[\left(\frac{1}{2}\right)\left((1-p)\log_2\frac{1}{1-p} + p\log_2\frac{1}{p}\right) + \left(\frac{1}{2}\right)\left((1-p)\log_2\frac{1}{1-p} + p\log_2\frac{1}{p}\right)\right] \\ I(X;Y) &= 1 - \left[(1-p)\log_2\frac{1}{1-p} + p\log_2\frac{1}{p}\right] \end{aligned}$$

As a consistency check, notice that if $p = 0$, we recover the solution for Example A of 1 bit. This function is plotted in figure 2.

8 Conclusions

1. Classical information theory requires a set probability model.

2. Information is selection between possibilities.

3. Entropy is an extensive measure of uncertainty.

4. **food for thought**: if entropy is an extensive quantity, what is an *invariant* of a system?

9 References

1. Cover T and Thomas J (1991) *Elements of Information Theory.* New York: John Wiley and Sons.

2. Rieke et al (1997) *Spikes: exploring the neural code.* Cambridge, MA: MIT Press.

3. Mackay, David (2003) *Information Theory, Inference and Learning Algorithms*, online `http://www.inference.phy.cam.ac.uk/mackay/itprnn/book.html`