Information Theory

Delon Shen

Notes on Information Theory. If you have any comments let me know at hi@delonshen.com.

Last updated on February 6, 2021

Note: Witten refers to Witten's *"A Mini-Introduction To Information Theory"*

Witten: Classical Information Theory

Started: February 01, 2021. Finished: February 06, 2021

Disaster! You've been struck by an acute and permanent case of locked in syndrome. In a tragic turn of events, your eyes begin to randomly move either up or down with probability p and $(1-p)$ every half a second. Even worse, you're on a conveyor belt that's slowly but steadily going towards a furnace (like that scene in Toy Story 3) and you know for certain that you'll be no more in 3 days. How inconvenient! For some reason you have some machines attached to you that records when you move your eyes up(denoted by H) and when you move your eyes down(denoted by A). Using this machine you can usually send messages to your loved ones that are of the form

$$HAHAHAHAHAHAHAAAHAAHAHAHAHAHHAA\ldots$$

But now it's just a random string generator. Given your deadline you guess that the resulting message will be N letters long. As you slowly inch towards your death you start to think about strange things. For example you think that when your loved ones look back on your final message that there will be around pN "H" characters and $(1-p)N$ "A" characters. How many messages with this combination of characters are there? Well from basic combinatorics we know that the number is

$$\frac{N!}{(pN)!((1-p)N)!} \approx \frac{N^N}{(pN)^{pN}((1-p)N)^{(1-p)N}} = \frac{1}{p^{p\times N}(1-p)^{(1-p)\times N}}$$

Now lets define a quantity called *Shannon entropy* S as

$$2^{NS} = \frac{1}{p^{p\times N}(1-p)^{(1-p)\times N}} \Rightarrow S = -p\log_2(p) - (1-p)\log_2(1-p)$$

Say we were sending the same kind of message that's N characters long but **we don't know the probabilities each letter would occur** then by combinatorics there would be 2^N possible messages. However we have more information than that fool, we know that H occurs with probability p and A occurs with probability $(1-p)$. That means we can send 2^{NS} messages. Or at least I think this is what we're saying. Just by knowing the probabilities each letter can occur has effectively (though not in reality) increased the length of our message by a factor of S! Each and every single letter in the string carries more information(or are maybe messages harder to decipher? I don't know yet). Lets extend this to the general case

Definition 1: (Shannon Entropy) Lets say we have a message that is composed of n different letters $\{a_1,\ldots,a_n\}$ where each letter occurs with probability $\{p_1,\ldots,p_n\}$. We define the Shannon Entropy S as

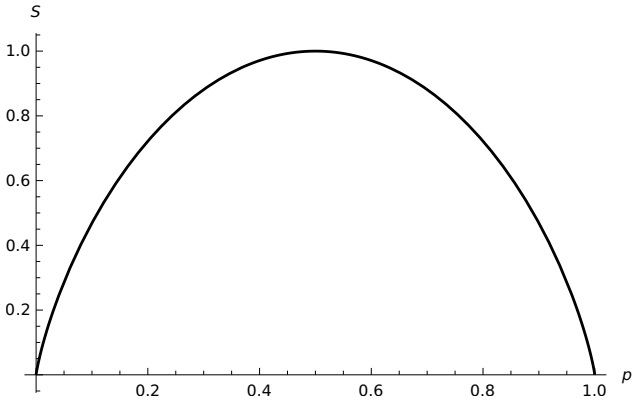
$$2^{NS} = \frac{N^N}{(p_1N)^{p_1N}\ldots(p_nN)^{p_nN}} \Rightarrow S = -\sum_{i=1}^{N} p_i\log_2(p_i)$$

Lets notice a few things about Shannon entropy. First of all since the numbere of messages we can send has to be at least 1 we know that $S \geq 0$. Now lets also ask, what's the maximum possible entropy for an alphabet of k letters.

Example 1: (Maximizing Shannon Entropy) Before we consider a general alphabet of k letters lets just consider our simple two alphabet case. We had that

$$S = -p\log_2(p) - (1-p)\log_2(1-p)$$

Now lets plot the Shannon entropy versus p and see if we can't qualitatively guess the maximum.



So right off the bat we guess that Shannon entropy is maximized when the probability is equally distributed to each letter in the alphabet. Lets see how we can prove this for a general alphabet. Witten prescribes using Lagrange multipliers with the constraint $\Sigma_i p_i = 1$. So we want to solve the system of equations

$$\frac{dS}{dp_i} = -\frac{\ln p_i + 1}{\ln 2} = \lambda = \frac{d(\text{Constraint})}{dp_i} \qquad \sum_{i=1}^{k} p_i = 1$$

The blue term gives us

$$\ln p_i = \ln p_j \Rightarrow p_i = p_j \ \forall i,j \Rightarrow \sum_{i=1}^{k} p_i = kp = 1 \Rightarrow \boxed{p_i = \frac{1}{k} \ \forall i}$$

The Shannon Entropy is maximized if each letter has the same probability of occuring!

Now lets talk about two good friend, Dan and Phil. They're sending messages to each other. Dan sends an instance of a discrete random variable x to Phil who recieves an instance of a discrete random variable y. What all this means physically is that Dan is speaking a random letter into a really bad microphone and Phil is listening on really shitty headphones and can't tell with 100% certainty what Dan is saying. Let $p(x_i, y_j)$ be the probability that Dan says x_i and Phil hears y_j. Now lets say for a second that Phil is certain that he just heard y_j on his

headphone. In this case the probability that Dan said x_i is given by Bayes rule

$$p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)} \Rightarrow S(x|y = y_j) = -\sum_i p(x_i|y_j)\log_2(p(x_i|y_j))$$

Knowing that Phil heard y_j for certain will make us more sure Phil can guess what Dan said and thus should lower the entropy. So we can guess that the Shannon entropy corresponding to the conditional probability distribution is the entropy of the message given that we heard y_j for certain. But in reality this will never happen since y is a random variable. But we can take a weighted sum of all these entropies to see on average what the entropy of the message Phil recieves is. To take this weighted sum we first need to reduce the dimension of the probability distribtuion to get $p(y)$

$$p(y) = \sum_i p(y, x_i)$$

And from here we can take a weighted sum

$$\begin{aligned}
\sum_j p(y_j)S(x|y = y_j) &= -\sum_j \sum_i p(y_j)p(x_i|y_j)\log_2(p(x_i|y_j)) \\
&= -\sum_i \sum_j p(y_j)\frac{p(x_i, y_j)}{p(y_j)}\log_2\left(\frac{p(x_i, y_j)}{p(y_j)}\right) \\
&= -\left(\sum_i \sum_j p(x_i, y_j)\log_2(p(x_i, y_j)) - \sum_i \sum_j p(x_i, y_j)\log_2(p(y_j))\right) \\
&= -\left(\underbrace{\sum_i \sum_j p(x_i, y_j)\log_2(p(x_i, y_j))}_{S_{XY}} - \underbrace{\sum_i p(y_j)\log_2(p(y_j))}_{S_Y}\right)
\end{aligned}$$

So we find that the average entropy of a message sent between Dan and Phil is the difference between the entropy of the joint probability distribution minus the entropy of the probability distribution of y. It turns out this little guy is so important that they have a name

Definition 2: (Conditional Entropy) The entropy of some probability distribution x given that we have observed y is

$$S_{XY} - S_Y$$

Conditional Entropy is related to another concept which is called **mutual information** denoted by $I(X;Y)$. This is the information that we can get about X given that we have observed Y. This would just be the difference of the actual entropy of x (S_X) and how much we don't know about x (S_{XY}).

Definition 3: (Mutual Information) Meaesurement of how much we know about a probability distribution x once we have observed y

$$I(X;Y) = S_X - S_{XY} + S_Y$$

Lets consider an example to motivate the definition of another cool guy.

Example 2: (Your Friend is a Sketchy Loser who likes Gambling and Dice) You and your friend are playing a game with two die. If the sum of the two die is > 7 you win a buck and if they're < 7 then your friend wins a buck. Rolling a 7 is a tie. We know that you both should have an equal chance of winning. However you're suspicious of your friend. He is a big chater after all. Even more suspicious, he insits on using his special die. He says he got them from a famous die maker on etsy for free after buying a space themed DnD dice set. First lets review what you'd expect to happen. There are 36 possible outcomes to this game.

$$\text{Your Friend Win} : (1,1)(1,2)(1,3)(1,4)(1,5)(2,1)(2,2)(2,3)(2,4)(3,1)(3,2)(3,3)(4,1)(4,2)(5,1)$$
$$\text{You Win} : (2,6)(3,5)(3,6)(4,4)(4,5)(4,6)(5,3)(5,4)(5,5)(5,6)(6,2)(6,3)(6,4)(6,5)(6,6)$$

Let the outcome of the i^{th} game be denoted by x_i. If your friend isn't being a dick and cheating then we'd have the probability of $G(x_i)$ occuring as

$$G(x_i) = \begin{cases} \text{You Win} & \frac{15}{36} \\ \text{You Lose} & \frac{15}{36} \\ \text{Tie} & \frac{5}{36} \end{cases}$$

You and your friend play N games. For large N we'd expect you to have won $\frac{15}{36} \times N$, your friend to have won the same number of games, and to have tied $\frac{5}{36} \times N$ times. The number of sequences of N games with this many of each ending state is

$$\frac{N!}{\left(\frac{15}{36}N\right)!\left(\frac{15}{36}N\right)!\left(\frac{5}{36}N\right)!}$$

And thus we can judge the probability of what we've seen by considering [a]

$$\mathcal{P} = \underbrace{\left(\underbrace{\left(\frac{15}{36}\right)^{N\times\frac{15}{36}}}_{\text{You Win}}\underbrace{\left(\frac{15}{36}\right)^{N\times\frac{15}{36}}}_{\text{Your Friend Wins}}\underbrace{\left(\frac{5}{36}\right)^{N\times\frac{5}{36}}}_{\text{Tie}}\right)}_{\text{Probability you expect}}\underbrace{\frac{N!}{\left(\frac{15}{36}N\right)!\left(\frac{15}{36}N\right)!\left(\frac{5}{36}N\right)!}}_{\text{What you see}}$$

The first big parenthesis corresponds to the probability of the given sequence occuring. Each time x_i occurs we multiply by a factor of $G(x_i)$. Since for large N we expect x_i to occur $N \times G(x_i)$ times the resulting probability will be the first term. The second term is the number of sequences where each x_i occurs $N \times G(x_i)$ times. But what if the die are

rigged (presumably in your friends favor unles he's just a real interesting guy)? What if instead your friend can manipulate the die so that the probability of each outcome $P(x_i)$ is

$$P(x_i) = \begin{cases} \text{You Win} & \frac{10}{36} \\ \text{You Lose} & \frac{22}{36} \\ \text{Tie} & \frac{3}{36} \end{cases}$$

You'd still expect the same probability based on $G(x_i)$ if you go into this thinking your friend is honest but after large N what you see will be different and will instead be caused by $P(x_i)$.

$$\mathcal{P} = \underbrace{\left(\underbrace{\left(\frac{15}{36}\right)^{N\times\frac{10}{36}}}_{\text{You Win}} \underbrace{\left(\frac{15}{36}\right)^{N\times\frac{22}{36}}}_{\text{Your Friend Wins}} \underbrace{\left(\frac{5}{36}\right)^{N\times\frac{3}{36}}}_{\text{Tie}} \right)}_{\text{Probability you expect}} \underbrace{\frac{N!}{\left(\frac{10}{36}N\right)!\left(\frac{22}{36}N\right)!\left(\frac{3}{36}N\right)!}}_{\text{What you see}}$$

How can we quantify this discrepency? That's where **relative entropy** comes in

[a] I'm still not too certain on what Witten means by "judge." I think it has something to do with trying to see the "distance" between two probability distributions but from what I can tell it seems like he just multiplied two probability distribution dependent quantities together and called it a day. TODO intuition.

Definition 4: (Relative Entropy) Let X be a random variable which denotes the outcome of an experiment. We have some theory that predicts the probability distribution for our outcome to be Q_X. However if our theory isn't quite on the mark and the actual probability distribution if P_X how could we guess that Q_X is wrong. Let x_i denote the final state of the i^{th} expierment. Also let there be s possible final states. We do N experiments where N is large. If we go in thinking that Q_X is correct then we will judge the probability of what we have seen with

$$\mathcal{P} = \underbrace{\left(\prod_{i=1}^{s} Q_X(x_i)^{P_X(x_i)\times N} \right)}_{\text{Probability you expect}} \times \underbrace{\frac{N!}{\prod_{j=1}^{s}(P_X(x_i)\times N)!}}_{\text{What you see}}$$

When defining shannon entropy we saw that the second term could be rewritten as

$$\underbrace{\frac{N!}{\prod_{j=1}^{s}(P_X(x_i)\times N)!}}_{\text{What you see}} \approx 2^{-N\sum_i P_X(x_i)\log_2(P_X(x_i))}$$

And the first term can be trivially rewritten as

$$\underbrace{\left(\prod_{i=1}^{s} Q_X(x_i)^{P_X(x_i)\times N} \right)}_{\text{Probability you expect}} = 2^{N\sum_i P_X(x_i)\log_2(Q_X(x_i))}$$

All together this gives us

$$\mathcal{P} \approx 2^{-N\sum_i P_X(x_i)\left(\log_2\left(\frac{P_X(x_i)}{Q_X(x_i)}\right)\right)} = 2^{-N\,S(P_X\|Q_X)}$$

Where the red term is what we define as **relative entropy** (or Kullback-Liebler divergence if you're in the mood for a mouthfull.)

$$S(P_X \parallel Q_X) = \sum_i P_X(x_i) \times \log_2\left(\frac{P_X(x_i)}{Q_X(x_i)}\right)$$

The relative entropy has a few properties. First we note that if $P_X = Q_X$ then relative entropy is zero, else it's positive. We'll also notice that if $\mathcal{P}$ decreases and N is fixed then our $S(P_X \parallel Q_X)$ is increasing. From this we can guess that larger relative entropy means the more sure our initial hypothesis Q_X is wrong. So what can relative entropy do for us?

Example 3: (What Relative Entropy can tell us about Mutual Information) Something that relative entropy can tell us is that mutual information is positive. Consider a joint probability distribution $P_{X,Y}(x,y)$. We can reduce the distribution to a single variable in the normal way

$$P_X(x) = \int P_{XY}(x,y)dy = \sum_j P_{XY}(x,y_j) \qquad P_Y(y) = \int P_{XY}(x,y)dx = \sum_j P_{XY}(x_j,y)$$

And we'll define a new probability distribution $Q_{XY}(x,y) = P_X(x)P_Y(y)$. So what have we done? What Q_{XY} is saying is that $P_X(x)$ and $P_Y(y)$ are statistically independent. But we don't know if this actually true or not. But recall that intuitivley, relative entropy is sort of like the distance between two probability distributions. So if we find that the relative entropy between Q_{XY} and P_{XY} is zero we can say that $P_X(x)$ and $P_Y(y)$ are statistically independent. So lets calculate the relative entropy

$$\begin{aligned}
&S(P_{XY} \parallel Q_{XY}) = \\
&= \sum_{i,j} P_{XY}(x_i,y_j) \times \log_2\left(\frac{P_{XY}(x_i,y_j)}{Q_{XY}(x_i,y_j)}\right) \\
&= \sum_{i,j} P_{XY}(x_i,y_j) \times \left(\log_2 P_{XY}(x_i,y_j) - \log_2 P_X(x_i) - \log_2 P_Y(y_j)\right) \\
&= \underbrace{\left(\sum_{i,j} P_{XY}(x_i,y_j)\log_2 P_{XY}(x_i,y_j)\right)}_{-S_{XY}} - \underbrace{\left(\sum_{i,j} P_{XY}(x_i,y_j)\log_2 P_X(x_i)\right)}_{-S_X} - \underbrace{\left(\sum_{i,j} P_{XY}(x_i,y_j)\log_2 P_Y(y_j)\right)}_{S_Y} \\
&= S_X - S_{XY} + S_Y = I(x;y)
\end{aligned}$$

So we see that the mutual information is the relative entropy between a joint distribution P_{XY} and the hypothesis that P_X and P_Y are statistically independent and thus **must be positive.**

$$I(X;Y) = S_X + S_Y - S_{XY} \geq 0$$

Relative entropy has another interesting property called the **monoticity of relative entropy.**

Definition 5: (Monoticity of Relative Entropy) Lets say we have two joint probability distribution P_{XY} and Q_{XY} where P_{XY} is the "true" probability distribution and Q_{XY} is the hypothesized probability distribution. After N measurements we can quantify our certainty or uncertainty that our hypotheszied distribtuion Q_{XY} is correct with the relative entropy $S(P_{XY} \parallel Q_{XY})$. But lets say we're down bad. We can only measure X. So reducing to one variable gives us P_X and Q_X and again we can asses our hypothesis with the relative entropy $S(P_X \parallel Q_X)$. But lets think very hard for a second. Would $S(P_X \parallel Q_X)$ be greater than or less than $S(P_{XY} \parallel Q_{XY})$. Or physically we want to know, is our assesment of our hypothesis going to be more or less optimistic given that we've reduced the number of random variables[a].

$$S(P_{XY} \parallel Q_{XY}) \geq S(P_X \parallel Q_X)$$

We call this the monotonicity of relative entropy. To get a feel for why this is true. Consider observing a sequence of outcomes $\{x_{i_1}, \ldots, x_{i_n}\}$. For this sequence of outcomes there should also exists some $\{y_{i_1}, \ldots, y_{i_n}\}$ that minimizes the relative entropy. But any sequence of y we view will at best be equal to the relative entropy of only viewing one dimension of the joint probability distribution and will probably be increase the relative entropy. So the relative entropy of being able to measure one degree of freedom of the joint probability distribution must be lower than being able to measure the entire joint probability distribution.

[a] Witten just says "it is harder to disprove the inital hyptohesis if we observe only X" like it's clear but unfortunately I don't have any inspired thoughts that make this statement obvious intuitivley. My best guess is that reducing the number of random variables helps us cheat a bit sort of like looking at the first few lines of the solution to a problem. We've sort of assumed that we're good for one variable and are just trying to find the distance between two simpler distributions and thus $S(P_{XY} \parallel Q_{XY})$ is more chaotic then its single variable counterpart. Or maybe I'm overthinking it and there isn't an intuition for it. In other word's Witten is just quoting the result of the proof we're about to do without saying we should know this intuitively. Edit: nevermind he literally gives the intuition in the next paragraph.

If my ham-handed explanations and Witten's very nice intuition for the monotinicity of relative entropy don't satisfy you, don't fret! We'll also show it the brain dead way with some algebra. We can restate the monoticity of relative entropy as

$$S(P_{XY} \parallel Q_{XY}) - S(P_X \parallel Q_X) \geq 0$$

Which we can rewrite as

$$S(P_{XY} \,\|\, Q_{XY}) - S(P_X \,\|\, Q_X) = \sum_{i,j} P_{XY}(x_i, y_j) \times \left(\log_2 \left[\frac{P_{XY}(x_i, y_j)}{Q_{XY}(x_i, y_j)} \right] - \log_2 \left[\frac{P_X(x_i)}{Q_X(x_i)} \right] \right)$$

$$= \sum_{i,j} \underbrace{\frac{P_{XY}(x_i, y_j)}{P_X(x_i)} \times P_X(x_i)}_{\substack{P_X(x_i)P(y_j|x_i) \\ \text{by Bayes rule}}} \times \log_2 \left[\underbrace{\frac{P_{XY}(x_i, y_j)}{P_X(x_i)}}_{\substack{P(y_j|x_i) \\ \text{by Bayes rule}}} \times \underbrace{\frac{Q_X(x_i)}{Q_{XY}(x_i, y_j)}}_{\substack{Q(y_j|x_i)^{-1} \\ \text{by Bayes rule}}} \right]$$

$$= \sum_i P_X(x_i) \underbrace{\sum_j P_X(y_j|x_i) \log_2 \left[\frac{P(y_j|x_i)}{Q(y_j|x_i)} \right]}_{\substack{S(P_{Y|X=x_i} \| Q_{Y|X=x_i}) \\ \geq 0 \text{ since it's a} \\ \text{relative entropy}}}$$

$$= \sum_i P_X(x_i) S(P_{Y|X=x_i} \,\|\, Q_{Y|X=x_i}) \geq 0$$

There we go! Monoticity of relative entropy can show us something very interesting. Consider some probability distributions $P_{XYZ}(x_i, y_j, z_k)$ and $Q_{XYZ} = P_X(x_i)P_{YZ}(y_j, z_k)$ where we integrate out variables in the normal way. From the monoticity of relative entropy we know that

$$S(P_{XYZ} \,\|\, Q_{XYZ}) \geq S(P_{XY}|Q_{XY})$$

When we were proving that mutal information is positive we showed that

$$S(P_{XY} \,\|\, Q_{XY}) = S_X - S_{XY} + S_Y$$

We can apply this result here

$$\cancel{S_X} - S_{XYZ} + S_{YZ} \geq \cancel{S_X} - S_{XY} + S_Y$$

Giving us the result

$$S_{XY} + S_{YZ} \geq S_{XYZ} + S_Y$$

We call this strong subadditivity. From our definition of mutual information this is equivalent to saying that

$$I(X; YZ) \geq I(X; Y)$$

Which makes sense because being able to view the whole distribution will give you more information about X than only viewing a slice of the probability distribution.