EE229A Lecture Notes

Arin Chang

Fall 2022 - Professor Venkat Anantharam

1 Introduction, 8/25

Information Theory starts with Claude Shannon. Shannon associated a number called entropy to any probability distribution on a discrete (finite or countably infinite) set. This is a non-negative number representing how surprising on average an observation from this distribution is. For a probability distribution on a finite set $(p_1, ..., p_n)$ the entropy is $H(p_1, ..., p_n) = -\sum_{i=1}^{n} p_i \log p_i$. We will typically take the log to base 2.

How did Shannon come up with this formula? one approach might have been axiomatic. Let's select a function over probability distributions $\Theta(p_1, ..., p_n)$ (think of the domain as some simplex).

Axiom 1. *(normalization):* $\Theta(1/2, 1/2) = 1$

Axiom 2. *for all* $n \geq 1$ *and for all* $(p_1, ..., p_n)$, $\Theta(p_1, ..., p_n) \leq \Theta(1/n, ..., 1/n)$

Axiom 3. *for all* $n \geq 1$ *and for all* $(p_1, ..., p_n)$, $\Theta(p_1, .., p_n, 0) = \Theta(p_1, ..., p_n)$.

Axiom 4. *given* $(p_1, ..., p_r)$ *and* $(q_r, ..., q_s)$. *for all* $1 \leq k \leq r$, $\Theta(p_1, p_2, ..., p_{k-1}, p_k q_1, p_k q_2, ..., p_k q_s, ..., p_r) = \Theta(p_1, .., p_r) + p_k \Theta(q_1, ..., q_s)$

Axiom 5. Θ *is continuous in* $(p_1, .., p_n)$ *for each* $n \geq 1$

These set of axioms naturally give rise to the formula for entropy. See Handout 1 for a theorem showing that Θ must be H. Another way Shannon might have come up with this definition of entropy is related to the idea that in order to represent uncertainty in efficient way, it's natural to pick more efficient representations for more commonly appearing symbols. Concretely, he may have studied the optimum (least expected length) prefix free (self parsing) binary data compression code associated to a probability distribution $(p_1, ..., p_n)$.

Definition 1. *binary data compression code* $c : \{1, ..., n\} \mapsto \{0, 1\}^* - \emptyset$. *for all* $x, x' \in \{1, .., n\}$, *c(x) should not be a prefix of c(x')*

The aim is to minimize the prefix code c, i.e. $\sum_{i=1}^{n} p_i length(c(i))$. First we notice that every prefix free code satisfies $\sum_{i=1}^{n} 2^{-length(c(i))} \leq 1$ (Kraft's inequality).

[Arin: TODO: draw a binary tree with colored nodes representing prefix free codes]

c from above is prefix free iff path from root to each orange node meets no other orange node.

associate mass $2^{-length(c(i))}$ to the orange node i. pull each mass to the root. This gives Kraft's inequality

Conversely given $l_1, ..., l_n$ such that $\sum_{i=1}^{n} 2^{-l_i} \leq 1$ there is a prefix free $c : [n] \mapsto \{0, 1\}^* - \emptyset$ with c(i) having length l_i. The problem now is to minimize $\sum_{i=1}^{n} p_i l_i$ s.t. $\sum_{i=1}^{n} 2^{-l_i} \leq 1$. Note that this problem is still hard: this is an integer programming problem. Let's relax the integer condition to allow the l's to be real valued. The relaxed problem is now $min \sum_{i=1}^{n} p_i y_i$ s.t. $\sum_{i=1}^{n} 2^{-y_i} \leq 1$ where y_i are real numbers. We make the following observations: 1) $y_i \geq 0$ is necessary. 2) We can assume equality at optimum. This implies we can use Lagrangian optimization techniques.

The Lagrangian is $\sum_{i=1}^{n} p_i y_i + \lambda(\sum_{i=1}^{n} 2^{-y_i} - 1)$, where the lambda multiplication is a "prize" for violating the constraint. Differentiate with respect to y_i and set derivative to 0: $p_i - \lambda(\log_2) 2^{-y_i} = 0$. We have $y_i = -\log_2 p_i + C$ for some constant C. This constant is in fact zero, which can be seen by plugging into Kraft's inequality. Finally we have $\sum_{i=1}^{n} p_i y_i^* = H(p_1, ..., p_n)$, where y_i^* was the solution to the optimization problem above.

2 8/30

Recall the entropy formula. Entropy is non-negative because $-\log p_i \geq 0$ for $p_i \in [0,1]$. What about $p_i = 0$? but $x \log x \to 0$ as $x \to 0$. To see why, write $x = e^{-u}$ with $u \to \infty$. $x \log x$ becomes ue^{-u} by taking the log to base e. so $p_i \log p_i$ is interpreted as 0 for $p_i = 0$. Also, $H(p_1, ..., p_n)$ is a concave function of $(p_1, ..., p_n)$.

Definition 2. *A concave function ϕ is defined as one where $-\phi$ is convex*

Definition 3. *A convex function ϕ is a real-valued function whose domain is a convex set in some R^n satisfying $\phi(\lambda x + (1-\lambda)x') \leq \lambda\phi(x) + (1-\lambda)\phi(x')$ for all x, x' in domain D for $\lambda \in [0,1]$.*

Definition 4. *A convex set D is one such that for all x, x' in D, for all $\lambda \in [0,1]$ we have $\lambda x + (1-\lambda)x'$ in D.*

Basically, a convex set is a set where if you take any two points inside it and draw the line connecting them (set of convex combinations of those two points), the line is contained in the set. Similarly, a convex function lies beneath the secant line. A concave function lies above the secant line. A consequence of the definition of a convex function is that for any number of $x_1, ..., x_n$ we have $\phi(\sum_j r_j x_j) \leq \sum_j r_j \phi(x_j)$ for all distributions r_i.

To see that $H(p_1, ..., p_n)$ is concave, observe that each term $-p_{ii}$ is concave on $[0,1]$. because $x \mapsto x \log x$ is convex (a fact on the domain $[0, \infty)$). Check that $x \log x$ is convex: derivative of $x \log x$ is $\log x + 1$. Second derivative of $x \log x$ is $\frac{1}{x}$ which is indeed non-negative.

As an aside, the $x \log x$ function is extremely important in Information Theory. in this sense Information Theory is not a terribly natural subject since the log function is not a very natural function and it has all kinds of weird properties.

Let's look at the graph of $x \log x$. it decreases super fast in the beginning then starts increasing again and crosses the x axis at around 1. then it steadily increases to infinity. asymptotically it grows a bit faster than x.

[Arin: TODO: insert graph of xlogx]

Because $H(p_1, ..., p_n)$ is concave we get $H(p_1, .., p_n) \leq H(\frac{1}{n}, ..., \frac{1}{n}) = \log n$. This was also shown in the axioms handout. To see this, consider all permutations of $p_1, .., p_n$ and their average. i.e. $p_{\sigma_1}, .., p_{\sigma_n}$ where σ is a map from [n] to itself. The set of permutations is defined S_n which has size $n!$.

To properly write this down, we have $H(p_{\sigma_1}, .., p_{\sigma_n}) = H(p_1, ..., p_n)$. for all $\sigma \in S_n$. By concavity,

$$H(\frac{1}{n}, ..., \frac{1}{n}) = H(\frac{1}{n!} \sum_{\sigma \in S_n} (p_{\sigma_1}, .., p_{\sigma_n})) \geq \frac{1}{n!} \sum_{\sigma \in S_n} H(p_{\sigma_1}, .., p_{\sigma_n}) = H(p_1, .., p_n) \tag{1}$$

The first equality holds because each p_{σ_i} is in each slot in exactly $(n-1)!$ permutations, and the sum over all σ_i is 1. Thus in each slot, we have $\frac{(n-1)!}{n!} = \frac{1}{n}$. Relative entropy is a statistical way to measure how far some distribution $(p_1, ..., p_n)$ and $(q_1, ..., q_n)$.

Another way to see this is by introducing the concept of relative entropy. This is the central concept in information theory, statistics, data analysis etc. Given $(p_1, ..., p_n)$ and $(q_1, ..., q_n)$ probability distributions on $\{1, ..., n\}$.

Definition 5. $D(p||q) = \sum_{i=1}^{n} p_i \log \frac{p_1}{q_i}$ *called the relative entropy of p with respect to q. Note $D(p||q)$ is extended real valued; it can take on the value ∞. This happens iff $q_i = 0$ and $p_i > 0$ for some i. Think about why it being infinity in this case makes sense in the perspective of something from p but think it s from q.*

Note that it's not symmetric in p and q. It says things of the nature, if i get something from distribution q what are the chances i think it's from distribution p.

It turns out that $D(p||q) \geq 0$ with equality iff $p = q$. This can be seen from the convexity of the function x with domain $[0, \infty)$ proved earlier. Can assume $D(p||q)$ is finite, so take the case when q_i is not 0. We can write $D(p||q) = \sum_i q_i (\frac{p_i}{q_i} \log \frac{p_i}{q_i}) \geq \sum_{i=1}^{n} q_i (\frac{p_i}{q_i} \log(\sum_i q_i \frac{p_i}{q_i})) = 0$. The inequality holds by convexity of $x \log x$ which we showed above. We can do this because $q_i = 0$ implies $p_i = 0$.

Now given $(p_1, ..., p_n)$ consider $D(p||u)$ where $u = (\frac{1}{n}, ..., \frac{1}{n})$. $D(p||q) = \sum_i p_i \log \frac{p_i}{\frac{1}{n}} = -H(p_1, ..., p_n) + \log n$. This shows that entropy is a manifestation of the relative entropy. So $D(p||u) \geq 0$ is the same as saying that $H(p_1, ..., p_n) <= \log n$. Another proof that shows that uniform distribution has largest entropy!

[Arin: TODO: think through these cases] There are lots of special cases and intricacies to consider. for example if pi and qi are 0.

Now we are going to build the concept of entropy and related quantities for random variables. In any applied probabilistic scenario we have many quantities (modeled as r.vs) each having some "information" content and related to each other.

If X is a random variable taking values in a set of size n [n] with distribution $(p_1, ..., p_n)$, we write $H(X)$ for $H(p_1, ..., p_n)$. Now given two random variables X $\in [n]$ and Y $\in [m]$ jointly distributed. $H(X, Y)$ is the notation used for the "joint entropy" of X and Y which is basically the entropy of the pair (X, Y) $\in [n]x[m]$, i.e. $H(X, Y) = -\sum_{x,y} P(X = x, Y = y) \log P(X = x, Y = y)$.

We could write this as $-\sum p_{xy}(x, y) \log p_{xy}(x, y)$ but we are going to write it as $-\sum_{x,y} p(x, y) \log p(x, y)$. This is a real problem because $H(X) = -\sum_x p(x) \log p(x)$ and $H(Y) = -\sum_y p(y) \log p(y)$. But we will just get used to this notation XD.

Intuitively $H(X, Y)$ should be at least as big as max(H(X), H(Y)). Is this true? Consider $H(X, Y) - H(X) = -\sum_{x,y} p(x, y) \log p(x, y) + \sum_x p(x) \log p(x)$. Note that second term is $\sum_{x,y} p(x, y) \log p(x)$. So we have $-\sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)}$. This is equal to $-\sum p(x, y) \log p(y|x) = \sum_x p(x)(-\sum_y p(y|x) \log p(y|x))$. This is magical! Inside the parentheses we are computing an entropy. We denote this as $H(Y|X = x)$. Called the conditional entropy of Y given that X=x. We denote this $H(Y|X)$ as the conditional entropy of Y given X (in Shannon's theory, this is just a number not a r.v.).

It turns out as one would expect that $H(Y|X) \leq H(Y)$:

$$H(Y) - H(Y|X) = -\sum_y p(y) \log p(y) + \sum_{x,y} p(x, y) \log p(y|x) \tag{2}$$

$$= \sum p(x, y) \log \frac{p(y|x)}{p(y)} = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = D(p(x, y)||p(x)p(y)) \geq 0 \tag{3}$$

This is denoted $I(X; Y) = I(Y; X)$ called the mutual information between X and Y. The mutual information is symmetric about X and Y, which is quite surprising since one wouldn't expect that the information that X tells about Y is the same as the information Y tells about X.

3 9/1

We review the notations/definitions of entropy, joint entropy of random variables, conditional entropy of X given Y=y, and $H(X|Y)$. $H(X|Y = y) = -\sum_x p(x|y) \log p(x|y)$, $H(X|Y) = \sum_y p(y) H(X|Y = y) = -\sum_{x,y} p(x, y) \log p(x|y)$. The mutual information is $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.

Let's call $\log p(x)$ as a function of x to be the "entropy density of x". Similarly $\log p(X|Y)$ as a function of (x,y) is the conditional entropy density of Y given X. $\log \frac{p(x,y)}{p(x)p(y)}$ as a function of (x,y) the "information density of the pair (x,y)". We can now write $H(X) = E[\log \frac{1}{p(X)}]$ i.e. expectation of the entropy density. Note that this entropy density is a very strange function of a random variable (it's defined in terms of the r.v.s distribution!). Similarly we can write $H(X|Y) = E[\log \frac{1}{p(X|Y)}]$ and $I(X; Y) = E[\log \frac{p(X,Y)}{P(X)P(Y)}]$. We have $H(X, Y) = H(X) + H(Y|X)$. This is just $E[\log \frac{1}{p(x,y)}] = E[\log \frac{1}{p(x)}] + E[\log \frac{1}{p(y|x)}]$. We can also write the mutual information in terms of the entropy density: $I(X; Y) = H(Y) - H(Y|X) = E[\log \frac{p(Y|X)}{p(Y)}] = E[\log \frac{1}{p(Y)}] - E[\log \frac{1}{p(Y|X)}]$.

The chain rule for entropy is: $H(X_1, ..., X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + H(X_4|X_1, X_2, X_3) + ... + H(X_N|X_1, ..., X_{n-1})$. Note this is just $E[\log \frac{1}{p(X_1,...,X_n)}] = E[\log \frac{1}{p(X_1)}] + E[\log \frac{1}{p(X_2|X_1)}] + ... + E[\log \frac{1}{p(X_n|X_1,...,X_n)}]$. Similarly, we have the chain rule for mutual information $I(X; Y_1, ..., Y_n)$. This is how much you learn about X from seeing all of $Y_1, ..., Y_n$. The formula itself follows intuitively: $I(X; Y_1, ..., Y_n) = I(X; Y_1) + I(X; Y2|Y1) + I(X; Y_3|Y_1, Y_2) + ... + I(X; Y_n|Y_1, ..., Y_{n-1})$. It's saying that the amount we learn about X is decomposed as the amount we learn about X from seeing Y_1 plus the amount we learn about X from seeing Y_2 given that we've seen Y_1 and so on. Here $I(X; Y|Z) = \sum_z p(z) I(X; Y|Z = z)$ where $I(X; Y|Z = z) = \sum_{x,y} p(x, y|Z = z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$

Note that $I(X; Y|Z) = H(X|Z) - H(X|Y, Z)$ because starting from the RHS we have

$$\sum_{x,y} p(x, z) \log \frac{1}{p(x|z)} - \sum_{x,y,z} p(x, y, z) \log \frac{1}{p(x|y, z)} = \sum_{x,y,z} p(x, y, z) \log \frac{p(x|y, z)}{p(x|z)} \tag{4}$$

$$= \sum_z p(z) \sum_{x,y} p(x, y|z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} \tag{5}$$

Now, $I(X;Y|Z) = E[\log \frac{p(X,Y|Z)}{P(X|Z)P(Y|Z)}]$ just like $I(X;Y) = E[\log \frac{p(X,Y)}{p(X)p(Y)}]$. So the chain rule of mutual information can be proved as follows:

$$I(X;Y_1,...,Y_n) = E[\log \frac{p(X|Y_1,..,Y_n)}{p(X)p(Y_1,..,Y_n)}] \tag{6}$$

$$= E[\log \frac{p(X,Y_1)}{p(X)p(Y_1)}] + E[\log \frac{p(X,Y_2|Y_1)}{P(X|Y_1)p(Y_2|Y_1)}] + E[\log \frac{p(X,Y_3|Y_1,Y_2)}{p(X|Y_1,Y_2)p(Y_3|Y_1,Y_2)} \tag{7}$$

$$+ ... + E[\log \frac{p(X_3,Y_n|Y_1,...,Y_{n-1})}{p(X|Y_1,...,Y_{n-1})P(Y_n|Y_1,...,Y_{n-1})} \tag{8}$$

$$= E[\log \frac{p(X|Y_1)p(X|Y_1,Y_2)p(X|Y_1,Y_2,Y_3)...p(X|Y_1,Y_2,...,Y_n)}{p(X)p(X|Y_1)p(X|Y_1,Y_2)...p(X|Y_1,...,Y_{n-1})}] \tag{9}$$

We can see that multiplying all the conditional probabilities together results in the first expression.

We also saw the relative entropy $D(p||q)$ for $p = (p_1,...,p_n), q = (q_1,...,q_n)$ defined as $\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$. Venkat says this is **the** information theory quantity. Here $D(p||q) = \infty \iff$ there exists i such that $p_i > 0$ but $q_i = 0$. We saw $D(p||q) \geq 0$ based on the convexity of the $x \mapsto x \log x$ function. Since this function is actually strictly convex, it turns out $D(p||q) = 0 \iff p = q$.

Definition 6. *A convex function $\phi : D \mapsto \mathbb{R}$ (where D is a convex subset of $\mathbb{R}^n$) is **strictly convex** if for all x,x' $\in$ D, $\lambda \in [0,1]$, $\phi(\lambda x + (1-\lambda)x') = \lambda\phi(x) + (1-\lambda)\phi(x')$ implies either x=x', $\lambda = 0$ or $\lambda = 1$.*

For ϕ convex, if ϕ is twice differentiable and second derivative is strictly positive on the interior of D. We saw the second derivative of $x \log x = \frac{1}{x} > 0$ for $x > 0$. We also saw that $I(X;Y) = D(p(x,y)||p(x)p(y))$. Hence $I(X;Y) = 0$ iff X and Y are independent. This shows how profound of a concept mutual information is! Since many probability theorists would say that independence is one of the most greatest concepts of probability.

We can also define the conditional relative entropy (shows up a lot in data science). Given two probability distributions of the form $p(x)a(y|x)$ and $p(x)b(y|x)$. Define $D(b(y|x)||a(y|x)|p(x))$ as $\sum_x p(x) \sum_y b(y|x) \log \frac{b(y|x)}{a(y|x)}$. So given $p(x,y)$ and $q(x,y)$, we have

$$D(p(x,y)||q(x,y)) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} = \sum_{x,y} p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} = \sum_x p(x) \log \frac{p(y)}{p(x)} \tag{10}$$

$$= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \tag{11}$$

$$= D(p(x)||q(x) + D(p(y|x)||q(y|x)|p(x)) \tag{12}$$

which is the chain rule for conditional relative entropy.