Lecture 3: Joint entropy, conditional entropy, relative entropy, and mutual information

Biology 429
Carl Bergstrom

January 13, 2008

Sources: Today, we closely follow Chapter 2 from Cover and Thomas (1991). The definitions and theorems below are often quoted directly from that text.

Entropy in its basic form a measure of uncertainty rather than a measure of information. Specifically, the entropy of a random variable a measure of the uncertainty associated with that random variable. When the entropy of a random variable is large this means that the uncertainty as to the value of that random variable is large, and vice versa.

Here — in much of what I say about in this course — I'm talking about *random variables* rather than about the sorts of explicit *communication channels* that Shannon envisioned in the original 1948 paper. This provides a cleaner, more general formalism for talking about the subject.

The connection to information comes in when we start thinking about communication channels along which signals are sent. If you are the receiver at the end of a communication channel, and you don't know what symbol is coming next, that value of the next incoming symbol is a random variable to you. Recall that we defined information as that which eliminates (i.e., the negative of) uncertainty. Thus if a random variable has an entropy of α, and we then we get a signal that tells us the value of that random variable, we will have gotten rid of uncertainty α, or in other words, gained information α. When we turn to channel coding theory, we will look more at the explicit properties of transmission on communication channels. In the meantime, on to the definitions and properties of entropy and other related quantities.

Definition 1 *The* entropy *of a discrete random variable X is defined by*

$$H(X) = -\sum_{x\in\mathscr{X}} p(x)\log p(x).$$

Last lecture we motivated this definition in some detail, and looked at examples of entropy calculations. A few additional notes:

- Entropy is always positive. $H(X) \geq 0$ since $0 \leq p(x) \leq 1$ for all $p(x)$.
- We can change bases freely: $H_b(X) = (\log_b a)H_a(x)$ since $\log_b p = \log_b a \log_a p$.

Entropy measures the uncertainty inherent in the distribution of a random variable. Joint entropy and conditional entropy are simple extensions that measure the uncertainty in the joint distribution of a pair of random variables, and the uncertainty in the conditional distribution of a pair of random variables.

Definition 2 *The* joint entropy $H(X,Y)$ *of a pair of discrete random variables with a joint distribution $p(x,y)$ is defined as*

$$H(X,Y) = -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log p(x,y).$$

Definition 3 *The* conditional entropy $H(Y|X)$ *is defined as*

$$H(Y|X) = \sum_{x\in\mathscr{X}} p(x)H(Y|X=x).$$

Alternatively, we can write this as $H(X,Y) = -\sum_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} p(x,y)\log p(y|x)$. Note that the conditional entropy of Y conditional on X refers to the average entropy of Y conditional on the value of X, *averaged over all possible values of* X. This is different than conditioning on X taking one particular value x_i, though we can write the other as well: $H(Y|X = x_i)$.

The chain rule for joint entropy states that the total uncertainty about the value of X and Y is equal to the uncertainty about X plus the (average) uncertainty about Y once you know X.

Theorem 1 *The chain rule for joint entropy:*

$$H(X,Y) = H(X) + H(Y|X).$$

Now we turn to a new issue. How would one quantitatively compare the similarity or difference between two different distributions? The *relative entropy* provides a way with many pleasing statistical features. It can be thought as a measure of the "distance" (in the colloquial sense, not the mathematical one) between two distributions. So how do we do it? Draw two discrete distributions. How far apart are they?

Definition 4 *The* relative entropy $D(p \parallel q)$ *between two probability distributions is given by*

$$D(p \parallel q) = \sum_{x \in \mathscr{X}} p(x) \log \frac{p(x)}{q(x)}.$$

This gives us a measure of something like the distance between two different probability distributions, in the sense the relative entropy is always positive, it is zero if and only if the two distributions are the same, and increases as the distributions diverge. As we see when we turn to coding theory and gambling theory, the relative entropy also serves as a measure of how bad of a mistake it is to design a code or gambling strategy thinking that a distribution is q, when it is actually p. Note that relative entropy is not necessarily symmetric: $D(p \parallel q) \neq D(q \parallel p)$.

We can also ask how much one random variable tells us about another. How much does a cue or signal tell us about the state of the world? Here we turn to the *mutual information*, which of the most important information concepts for biology.

Definition 5 *The mutual information $I(X;Y)$ measures how much (on average) the realization of random variable Y tells us about the realization of X, i.e., how by how much the entropy of X is reduced if we know the realization of Y.*

$$I(X;Y) = H(X) - H(X|Y)$$

For example, the mutual information between a cue and the environment indicates us how much on average the cue tells us about the environment. The mutual information between a spike train and a sensory input tells us how much the spike train tells us about the sensory input.

If the cue is perfectly informative — if it tells us everything about the environment and nothing extra — then the mutual information between cue and environment is simply the entropy of the environment:

$$I(X;Y) = H(X) - H(X|Y) = H(X) - H(X|X) = H(X).$$

In other words, the mutual information between a random variable and itself is simply its entropy: $I(X;X) = H(X)$.

Surprisingly, mutual information is symmetric; X tells us exactly as much about Y as Y tells us about X.

Theorem 2 *Symmetry of mutual information*

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y,X).$$

We could write this out and show algebraic equality, but perhaps the most intuitive way to understand the symmetry of mutual information is to view it in an alternative formulation, as the relative entropy between the joint distribution of X and Y and the (independent) product distribution of X and Y:

$$I(X,Y) = \sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$

Seen in this form, clearly the relation between X and Y is symmetric. This form also recalls the earlier definition of relative entropy. Here we see that the mutual information between X and Y is the relative entropy between the joint distribution $p(x,y)$ and the product of the marginals $p(x)p(y)$. In other words, the mutual information measures the cost of assuming that two random variables are independent when fact they are not.

Finally, we can express mutual information in one additional way by invoking the chain rule.

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

Now we move on to a series of important inequalities that help us understand the structure of the entropy and related quantities.

Definition 6 *A function is* convex *(concave) if the midpoint of every chord of that function lies at or above (below) the function itself.*

Convexity or concavity is determined by the second derivative of a function. Strictly convex (concave) functions have strictly positive (negative) second derivatives. We can now state Jensen's inequality.

Theorem 3 *Jensen's inequality. For any convex function f and random variable X,*

$$E[f(x)] \geq f(E[X]).$$

Using Jensen's inequality, along with the fact that log is a strictly concave function, we can prove a number of facts:

- The relative entropy is non-negative: $D(p||q) \geq 0$ for any two probability distribution functions p, q.
- The mutual information is positive: $I(X,Y) \geq 0$ for any two random variables X, Y. This follows immediately from the non-negativity of relative entropy and the fact that the mutual information is the relative entropy between a joint distribution and the product of its marginals.
- Uniform maximizes entropy (in the absence of constraints). If the random variable X takes on possible values $\mathcal{X}$, the entropy of X is not greater than the log of the size of the set $\mathcal{X}$. That is, $H(X) \leq \log |\mathcal{X}|$, where equality is achieved only if X is a uniform distribution. We'll

use this entropy-maximizing property of the uniform distribution when it comes time to move to continuous (differential) entropy in Chapter 8. In statistical mechanics, people often look at entropy-maximizing distributions subject to various constraints; this is treated in Chapter 12.

- Additional information never increases entropy.

  $$H(X|Y) \leq H(X).$$

  Notice that this is true averaging across all possible values of Y as the definitions of conditional entropy requires. It does not necessarily hold for every realization of Y, i.e., it is not true that $H(X|Y = y) \leq H(X)$.

- Joint entropy is maximized by independent random variables. If $X_1, X_2, \ldots, X_k$ are drawn from some distribution p,

  $$H(X_1, X_2, \ldots, X_k) \leq \sum_{i=1}^{k} H(X_i).$$

  This makes intuitive sense. We have the maximal uncertainty about a sequence of random variables if they are uninformative about one another. Proof sketch: Consider $k = 3$ By the chain rule for conditional entropy, $H(X_1, X_2, X_3) = H(X_1|X_2, X_3) + H(X_2|X_3) + H(X_3)$. But by the result above that information never increases entropy, the first two terms are each less or equal to the corresponding unconditional entropies $H(X_1)$ and $H(X_2)$. The proof extends in obvious ways to larger k.

- The entropy $H(p)$ is a concave function of the distribution function p. What does this mean? Suppose we have two discete distribution functions p_1 and p_2. (These are just vectors with vector sums of one). Then the entropy of any average of these two distributions is greater than the average of the entropies of those distributions. Where $0 \leq \alpha \leq 1$ is the relative weight of each distribution in the average, $H(\alpha p_1 + (1 - \alpha)p_2) \geq \alpha H(p_1) + (1 - \alpha)H(p_2)$.

All of this brings us to a very important idea in information theory, the data processing inequality. The data processing inequality states that if a random variable Y tells us something about another random variable X, we cannot extract further information about X by performing additional processing operations — random or deterministic — on Y. That is, the mutual information between X and Y is greater or equal to the mutual information between X and any function of Y alone.

Theorem 4 *The data processing inequality. If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.*

This inequality follows from the symmetry in expressing mutual informations:

$$I(X;Z) + I(X;Y|Z) = I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$$

We note that when $X \to Y \to Z$, X and Z are conditionally independent given Y, i.e. $I(X;Z|Y) = 0$. So $[I(X;Z) + I(X;Y|Z) = I(X;Y)$. By the non-negativity of mutual information, $I(X;Y|Z) \geq 0$ and it follows immediately that $I(X;Y) \geq I(X;Z)$

The data processing inequality is the foundation for the idea of sufficient statistics. Suppose you have observations $x_1, x_2, x_3, \ldots, x_n$ of a random variable X distributed according to $f_\theta(x)$. A statistic $T(X)$ extracts some of the information in your observed sample: $X \to T(X)$. By the data processing inequality, $I(\theta, X) \geq I(\theta, T(X))$. If equality obtains, T is a *sufficient statistic* for θ. In other words, a sufficient statistics for some distribution $f_\theta(x)$ extracts *all* of the information within your data $x_1, x_2, x_3, \ldots, x_n$ about the value of θ.

Notice that a significant statistic near not be a scalar; it can be the full original observed data, for example. This doesn't help us much if we want to extract the information into as tight of a form as possible. Here we seek a *minimally sufficient statistic*: a minimally sufficient statistic T is one that is a function of every other sufficient statistic U. As Cover and Thomas note, "a minimal sufficient statistic maximally compresses [without loss] the information about θ in the sample."

This is our first hint of optimal compression in the course; we'll be looking at compression a great deal in future lectures.

Homework for Ch 2–3 of Cover and Thomas.

Due Wednesday January 23rd at the start of class.

C&T Problems 2.2, 2.3, 2.10a, 2.44, 2.47
C&T Problems 3.3, 3.7a,b.
Extra credit: C&T Problem 2.7 both parts. (No partial credit).