Note: Information Theory, Axiomatic Foundations, Connections to Statistics

36-350, Data Mining

12 September 2008

There was a question today in lecture about how to "derive" entropy and information theory. This (optional) note elaborates on my answer.

There are a number of approaches to justifying the use of the entropy formula

$$H[X] = -\sum_{x} \Pr(X = x) \log_2 \Pr(X = x)$$

and the information formula

$$I[X;Y] = H[Y] - H[Y|X]$$

for our purposes.

1. *Pragmatic.* Mathematically, this set of ideas lets us prove nice and interesting theorems, and empirically doing things like feature-selection by means of information content leads to good results.

2. *Analogical.* The entropy formula appears in statistical physics, where it measures (roughly speaking) the number of ways you can arrange molecules into configurations with given macroscopic properties (the probability distribution). One could think of this as something like the uncertainty in the molecular configuration.

   Of course, there are many analogies one *could* make, so the use of any *particular* analogy has to be justified by something else.

3. *Theoretical unification.* Though I did not go into this, it's possible to re-frame large parts of ordinary statistics, like maximum likelihood estimation and Bayesian updating, with an information-theoretic vocabulary. This is essentially because the entropy is the expected value of the log-likelihood under the true model. (More on this below.)

   Of course, there are many conceptual frameworks within which you could unify different bits of applied math, so the use of any *particular* set of unifying concepts and terms needs to be justified by something else.

4. *Axiomatics.* One could bgin by postulating a certain set of plausible-sounding axioms, in this case for a measure of information or uncertainty, and then show that the entropy is the unique object satisfying those axioms.

   I will say a bit more about this approach, and its limits, in the next section.

1 The Khinchin Axioms for Entropy

I got the set of axioms for entropy a bit wrong at the board. Here are the right ones (Khinchin, 1957).

Let X be a discrete random variable, taking k distinct values. Without loss of generality, let these be the integers $1, 2, \ldots k$, and abbreviate $\Pr(X = i)$ as p_i. We want to boil this distribution down to a single real number, $H[X]$, and assert that the latter must obey certain axioms. We define, by way of abbreviation, that $H[X, Y]$ is H of the two-component random variable (X, Y), that $H[Y|X = x]$ is the H of the conditional distribution $\Pr(Y|X = x)$.

1. $H[X]$ depends only on the probability distribution of X. (That is, we can change the labels of the events as much as we like without changing H.)

2. $H[X]$ is maximal, for a given k, when $p_i = 1/k$ for all i. (That is, the uniform distribution has maximal H.)

3. If Y is random variable on $1, 2, \ldots m$, where $m > k$, but $\Pr(Y = i) = p_i$ if $i \leq k$, and $\Pr(Y = i) = 0$ if $k < i \leq m$, then $H[Y] = H[X]$. (That is, notionally adding possibilities of probability zero does not change H.)

4. For any random variables X and Y,

$$H[X, Y] = H[X] + \sum_x \Pr(X = x) H[Y|X = x] \tag{1}$$

   (That is, our joint H is the sum of the H for one variable, plus the average value of the H of the other variable given the first.)

It can be shown (Khinchin, 1957, p. 9–13) that a function $H[X]$ satisfies these axioms if and only if it has the form

$$H[X] = -\sum_i p_i \log_b p_i \tag{2}$$

for some base $b > 1$. This is called the **Shannon** or **Gibbs-Shannon** entropy, where we generally chose either $b = 2$ (in computer science and information theory) or $b = e$ (in statistical mechanics and theoretical statistics).

If we think of these axioms as being about an indicator of *uncertainty* or *variability* in X, the first three seem reasonable, once we grant that it makes sense to measure the uncertainty of any random variable on a common numerical

scale with that of any other. They say that swapping labels or names of events doesn't really change how uncertain we are about the events[1], that we're most uncertain when there is no probabilistic basis for expecting any one outcome more than any other, and that considering possibilities we know won't happen doesn't actually make us more or less uncertain. So far so good, presuming, of course, that it makes sense to rank the "uncertainty" of *any* random variable against that of *any other*, on a common numerical scale.

The last axiom, however, is not so reasonable looking. It says that uncertainty is additive, and it says that it is additive in a particular strict way, that the means of conditional uncertainties add up to total uncertainty. Notice that it is strictly stronger than asserting that

$$H[X,Y] = H[X] + H[Y] \tag{3}$$

when X and Y are independent. (Exercise: Show that Eq. 1 implies Eq. 3 when X and Y are independent.) That weaker axiom would say that if we had two independent coin-tosses, we are twice as uncertain about the pair as we are about any one of them, which doesn't sound too outlandish.[2] But the actual fourth axiom, once again, is stronger than this, and asserts a lot about the treatment of dependent events. We might instead consider a rule where

$$H[X,Y] = H[X] + \max_x H[Y|X=x] \tag{4}$$

where our joint uncertainty depends on the *maximum* conditional uncertainty. (Exercise: Show that Eq. 4 implies Eq. 3 when X and Y are independent.)

(Problem: Can you find a functional which satisfies the other four axioms and Eq. 4? If not, can you prove that it's impossible?)

The stronger fourth axiom is needed to derive the Shannon form of the entropy, Eq. 2. If we just use Eq. 3 instead, any function of the form

$$H_\alpha[X] = \frac{1}{1-\alpha}\log\sum_{i=1}^{k} p_i^\alpha \tag{5}$$

where $\alpha \geq 0$, will satisfy the axioms. (Once again, I got the form wrong at the board.) These are known as the Rényi entropies, and there are, clearly, infinitely many different ones. Since Khinchin's fourth axiom is a special case of the additive-if-independent rule, the Shannon entropy should be a special case of the Rényi entropy, and it is, where $\alpha = 1$. (As $\alpha \to 1$, Eq. 5 goes to $0/0$; you

[1] Suppose that my favorite pen is either in my house or my office, which are a mile apart, and these are equally likely; and that my house-keys are either with me in the immigration line in Newark, or on the night-stand in the Hotel Pulaski in Brussels, and again these are equally likely. Do I *really* have the same degree of uncertainty whether the margin of error is a mile or 3700 miles?

[2] But is it actually so luminously certain and correct that anyone must be thirty-two times as uncertain about the outcome of thirty-two independent and identically-distributed coin-tosses as they are about one? If I claimed that I was only *five* times as uncertain, would that *really* force me to run around going "Error! Error! Does not compute!" until my head exploded?

recover the Shannon formula (2) by using L'Hopital's rule.) Starting from the Rényi entropy, one can define Rényi information,

$$I_\alpha = H_\alpha[X] + H_\alpha[Y] - H_\alpha[X, Y]$$

and so forth, through the rest of the formal structure of information theory. Crucially, however, a lot of the connections to coding theory, to statistics, and to the limit theorems ("large deviations principles") of probability theory grow weaker, or even break down. There are situations in dynamical systems theory where the Rényi entropies and informations are very useful (Badii and Politi, 1997; Beck and Schlögl, 1993), but the stronger and less "natural" axiom leads to a much more fruitful theory.

1.1 Remarks on Attempts at Axiomatic Foundations

Many people find themselves more comfortable with pieces of mathematics if they can be derived from a small set of axioms, in the way Khinchin derived entropy from his axioms. As I have tried to suggest, I think this is a mistake. This only provides more security if the axioms themselves are secure. Said another way, it only pushes the problem back a stage, from "why should I use this math?" to "why should I accept *these* axioms, and not others?" The point was well-made by Alfred North Whitehead and Bertrand Russell, two people who certainly understood axiomatic systems, almost a century ago (Whitehead and Russell, 1925–27):

> ... the chief reason in favour of any theory on the principles of mathematics must always be inductive, i.e., it must lie in the fact that the theory in question enables us to deduce ordinary mathematics. In mathematics, the greatest degree of self-evidence is usually not to be found quite at the beginning, but at some later point; hence the early deductions, until they reach this point, give reasons rather for believing the premises because true consequences follow from them, than for believing the consequences because they follow from the premises.

The same point is made, at some length, by Herbert Simon in *The Sciences of the Artificial* (Simon, 1996), which as I said in lecture is the best book to ever come out of this university and something you should all read as soon as possible.

There *are* advantages to axiomatic characterizations of bits of math. The first is that of clarity and summarization: the axioms are *just enough* to give the results; one does not *need* to assume more, and one *has* to assume that much. The second is abstraction (again): any system where the axioms hold (under some interpretation of their terms) is one where the theory is true (under that interpretation). This lets us study for example, linear systems theory without worrying too much whether the variables in the linear system are voltages and currents, or chemical fluxes and potentials, or positions and velocities of mechanical elements, etc., etc.

Nonetheless, ultimately axiom systems derive their value from what they let us prove, and so from the value of the math they formalize, rather than anything else. The universe, even the universe of mathematics, is under no obligation to make plausible- or beautiful- sounding axioms *relevant*. If anything, mathematicians re-shape their ideas of what is beautiful

2 Entropy, Relative Entropy, and Likelihood

In addition to the entropy, the other basic quantity of information theory is the **relative entropy**, also known as the **Kullback-Leibler (KL) divergence**: for two distributions P and Q over the same set of values,

$$D(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \tag{6}$$

It is not hard to show (there are a couple of ways of doing so!) that $D(P\|Q) \geq 0$, and that $D(P\|Q) = 0$ if and only if $P = Q$. One interpretation of $D(P\|Q)$ is therefore how *different* the distributions P and Q are.

Ordinary entropy can be defined in terms of the relative entropy (though not vice versa). If P is the distribution of X, and U is the uniform distribution on the same k values, then

$$H[X] = \log k - D(U\|P)$$

(Exercise: Show this.) The mutual information can also be directly defined in terms of the relative entropy. Let P be the marginal distribution of X, Q the marginal distribution of Y, J the actual joint distribution, and $P \otimes Q$ the product of the marginal distributions. Then

$$I[X;Y] = D(J\|P \otimes Q)$$

(Exercise: Show this.)

The divergence $D(P\|Q)$ is sometimes interpreted as the *gain in information* if we think the distribution is Q and we learn it is actually P; this would make the mutual information how much we learn if we realize that X and Y are not actually independent. Once again, however, it is not exactly obvious that this is the right way to formalize "gain in information", as opposed to, e.g.,

$$\sum_x |P(x) - Q(x)|$$

A different interpretation of relative entropy comes from thinking about description lengths and coding. Remember that $H[X]$ is the average number of bits needed to describe the value of X, when we chose our coding scheme to minimize the description length. We code the value x using $-\log_2 \Pr(X = x)$ bits. Suppose we think the distribution is $Q(x)$ and chose our coding scheme accordingly. Then the *actual* expected description length will be

$$-\sum_x P(x) \log_2 Q(x) = H[X] + D(P\|Q) \tag{7}$$

That is, we will waste $D(P\|Q)$ bits, on average. The quantity on the left-hand side of Eq. 7 is also known as the **cross-entropy.** (Exercise: Prove Eq. 7.) Minimizing our mean description length is the same task as minimizing $D(P\|Q)$.

Finally, we can connect the relative entropy and the cross-entropy to the likelihood. If we use a statistical model to get our probability distribution, this model will generally have some parameter θ, making the distribution $Q_\theta(x)$. If the true distribution is $P(x)$, then the *expected* log-likelihood will be

$$\mathbf{E}[L(\theta)] = \sum_x P(x) \log Q_\theta(x)$$

which is just minus the cross-entropy. (Or minus the cross entropy times a constant if we take the log of the likelihood to a base other than 2. Assume we use base 2.) Since log is an increasing function, maximizing the likelihood is the same as maximizing the log-likelihood. So maximizing expected likelihood is the same as *minimizing* $D(P\|Q_\theta)$, or minimizing expected description length. Since $D(P\|P) = 0$, the entropy is the expected log-likelihood of the *true* model.

The problem with statistical inference is that we don't know the actual distribution P, but just have a sample from it, $x_1, x_2, \ldots x_n$. We can turn this into an **empirical distribution** $\widehat{P}_n$, where probability is proportional to the number of samples:

$$\widehat{P}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_x(x_i) \tag{8}$$

Assuming the samples are IID, we can write the actual log-likelihood in terms of the empirical distribution:

$$L_n(\theta) = \sum_{i=1}^{n} \log Q_\theta(x_i) = n \sum_x \widehat{P}_n(x) \log Q_\theta(x) \tag{9}$$

(Exercise: prove this.) Using our previous result,

$$-L_n(\theta) = nH[\widehat{P}_n] + nD(\widehat{P}_n\|Q_\theta) \tag{10}$$

So maximizing the likelihood will work well when $D(\widehat{P}_n\|Q_\theta) \approx D(P\|Q_\theta)$. More exactly, if

$$\operatorname*{argmin}_\theta D(\widehat{P}_n\|Q_\theta) \to \operatorname*{argmin}_\theta D(P\|Q_\theta) \tag{11}$$

then the maximum likelihood *estimate* will converge on the optimal value of θ. If we make a few assumptions about P (basically, that it is not too nasty) and about Q_θ (basically, that it is not too crazy, and changes smoothly with θ), then the law of large numbers gives us Eq. 11. (For purists: with convergence in probability under the weak law of large numbers, and almost sure convergence under the strong.)

Another way to frame all this is to measure distance between probability distributions with the relative entropy, and think about the geometry this induces. Maximizing the likelihood then becomes finding the value of θ such that

Q_θ is as close as possible to $\widehat{P}_n$. That is, the maximum likelihood estimate is the **projection** of the data on to our family of distributions. As n grows, the law of large numbers forces $\widehat{P}_n$ to be closer and closer to P, so the projections of these points should become closer and closer as well. The more a small change in θ leads to a big change in Q_θ — the more *curved* the family of probability distributions is — the more precisely we can estimate θ. Stated precisely, this turns out to be the Cramér-Rao inequality of theoretical statistics. This is the idea of **information geometry**.

However, there is nothing especially magical (as opposed to convenient) about IID samples. Basically, whenever we can show that $D(\widehat{P}_n \| Q_\theta) \to D(P\|Q_\theta)$ for a big enough range of θ, likelihood procedures will end up working well. For example, you can do the same things for Markov processes, using the ergodic theorem instead of the law of large numbers (Kulhavý, 1996). People in applied fields sometimes have the impression that independent samples are required for statistical procedures to work. They are wrong.

Further reading. There are a number of good books on the connections between information theory and statistics. Maybe the easiest one to read is one on information geometry and its extensions to things like Markov processes: Kulhavý (1996). Kullback's own book, Kullback (1968), like most books on information geometry (Kass and Vos, 1997; Amari and Nagaoka, 1993/2000), assumes an advanced knowledge of theoretical statistics.

References

Amari, Shun-ichi and Hiroshi Nagaoka (1993/2000). *Methods of Information Geometry*. Providence, Rhode Island: American Mathematical Society. Translated by Daishi Harada. As *Joho Kika no Hoho*, Tokyo: Iwanami Shoten Publishers.

Badii, Remo and Antonio Politi (1997). *Complexity: Hierarchical Structures and Scaling in Physics*. Cambridge, England: Cambridge University Press.

Beck, Christian and Friedrich Schlögl (1993). *Thermodynamics of Chaotic Systems: An Introduction*. Cambridge, England: Cambr idge University Press.

Kass, Robert E. and Paul W. Vos (1997). *Geometrical Foundations of Asymptotic Inference*. Wiley Series in Probability and Statistics. New York: Wiley.

Khinchin, Aleksandr Iakovlevich (1957). *Mathematical Foundations of Information Theory*. New York: Dover. Translated by R. A. Silverman and M. D. Friedman from two Russian articles in Uspekhi Matematicheskikh Nauk, **7** (1953): 3–20 and **9** (1956): 17–75.

Kulhavý, Rudolf (1996). *Recursive Nonlinear Estimation: A Geometric Approach*, vol. 216 of *Lecture Notes in Control and Information Sciences*. Berlin: Springer-Verlag.

Kullback, Solomon (1968). *Information Theory and Statistics*. New York: Dover Books, 2nd edn.

Simon, Herbert A. (1996). *The Sciences of the Artificial*. Cambridge, Massachusetts: MIT Press, 3rd edn. First edition 1969.

Whitehead, Alfred North and Bertrand Russell (1925–27). *Principia Mathematica*. Cambridge, England: Cambridge University Press, 2nd edn.