Information Theory and Predictability. Lecture 2: Important functionals and their properties

1. Elementary Properties of Entropy

We can write $H(S)$ as

$$H(S) = \sum_{s \in A} p(s) \log\left(\frac{1}{p(s)}\right)$$

and since by the definition of a probability function $0 \leq p(s) \leq 1$ this sum is term by term non-negative and hence $H(S) \geq 0$. Obviously the minimum possible value 0 is attained when $p(s) = 1$ for some $s \in A$ since all other $p(s) = 0$ and hence no other terms contribute. This case corresponds to certainty for the value of S.

What about the maximum value? We determine this (for the finite category case) using Lagrange multipliers because there is the constraint that $\sum p(s) = 1$. The appropriate Lagrange functional is

$$F(p(1), \ldots, p(N)) = H(S) - \lambda\left(\sum_{i=1}^{N} p(s) - 1\right)$$

and differentiating this shows critical points occur when the $p(s)$ are all identical. Elementary calculus tells us that if a (well behaved) function is defined on a closed domain then it attains a maximum either at the critical points or on the domain boundary. Here since we are dealing with probabilities, the boundary consists of points with a $p(s) = 0$ which is a problem with $N-1$ outcomes. It is easily checked that at the critical point identified we have $H = \log(N)$ which is increasing with outcome size. Thus this must be the absolute maximum. Thus entropy is maximized for the uniform probability function which is intuitive since such a scenario means maximum uncertainty for S. As stated the value for the uniform distribution of N categories is $\log(N)$ which is increasing with category number consistent with Shannons first axiom. This is also the definition of entropy for equilibrium statistical mechanics of a gas where a uniform probability distribution of particle states is a common assumption. (see future Lecture).

2. Extension to many random variables

If one has two or more random variables then the number of possible categories for all the random variables obviously goes up multiplicatively $(N_1 N_2 \ldots N_k)$ as new random variables are added. This suggests the natural definition

$$H(X_1, X_2, \ldots, X_k) \equiv -\sum\nolimits_{x_1, x_2, \ldots, x_k} p(x_1, x_2, \ldots, x_k) \log\left(p(x_1, x_2, \ldots, x_k)\right) \tag{2.1}$$

where the summation extends over all possible categories.

A subject of common interest is the interaction between different random variables and one often considers *conditional* probabilities. Here one fixes the value of several random variables and considers the effect of this on the probability of the unfixed variables. This conditional probability function is the quotient of the

full probability function divided by the marginal distribution of the variables held fixed. Symbolically we have

$$\begin{aligned} p(x_1,\ldots,x_j|x_{j+1},\ldots,x_n) &= \frac{p(x_1,\ldots,x_n)}{p(x_{j+1},\ldots,x_n)} \\ p(x_{j+1},\ldots,x_n) &\equiv \sum_{x_1,\ldots,x_j} p(x_1,\ldots,x_n) \end{aligned}$$

This concept can be extended naturally to the notion of entropy/uncertainty. If one fixes the final $n-j$ selected variables with certain values then the entropy of the remaining j variables can be calculated in the obvious way:

$$\begin{aligned} H(X_1,\ldots,X_j|X_{j+1} &= x_{j+1},\ldots,X_n=x_n) = \\ &- \sum_{x_1,\ldots,x_j} p(x_1,\ldots,x_j|x_{j+1},\ldots,x_n)\log\left(p(x_1,\ldots,x_j|x_{j+1},\ldots,x_n)\right) \end{aligned}$$

Now each choice of fixed values has a particular, marginal, probability associated with it so it is natural to weight the above entropies by these probabilities and form a sum to give the expected new entropy in the case of fixed variables. This is referred to as the *conditional* entropy:

$$\begin{aligned} H(X_1,\ldots,X_j|X_{j+1},\ldots,X_n) &\equiv \sum_{x_{j+1},\ldots,x_n} p(x_{j+1},\ldots,x_n)H(X_1,\ldots,X_j|X_{j+1}=x_{j+1},\ldots,X_n=x_n) \\ (2.2) \qquad &= -\sum_{x_1\ldots,x_n} p(x_1,\ldots,x_n)\log\left(p(x_1,\ldots,x_j|x_{j+1},\ldots,x_n)\right) \end{aligned}$$

Simple manipulation of this definition, the properties of conditional and marginal probabilities given above as well as equation (2.1) allows the derivation of the following "chain rule" for entropy:

(2.3)
$$H(X_1,\ldots,X_j,X_{j+1},\ldots,X_n) = H(X_{j+1},\ldots,X_n) + H(X_1,\ldots,X_j|X_{j+1},\ldots,X_n)$$

This has the straightforward interpretation that *on average* fixing several random variables reduces the uncertainty of the complete system by the uncertainty associated with these fixed variables. Note that the entropy of the many variable system is not order dependent so switching order of variables in the chain rule gives rise to a whole series of further identities which we shall use below.

Another interesting entropy chain rule is obtained iteratively from equation (2.3). Firstly we have

$$H(X_1,\ldots,X_n) = H(X_n|X_{n-1},\ldots,X_1) + H(X_1,\ldots,X_{n-1})$$

and then repeatedly breaking down the second remainder term on the right side in the same way we obtain

$$H(X_1,\ldots,X_n) = \sum_{k=1}^{n} H(X_k|X_{k-1},\ldots,X_1) \tag{2.4}$$

3. RELATIVE ENTROPY AND MUTUAL INFORMATION

3.1. **Definitions.** It is often useful to compare probability functions and then a concept of "distance" between distributions becomes useful. One place where this plays an important role is in the theory of learning: Suppose that before a learning experience occurs, our best estimate about a particular random variable X, based

on all previous learning, is that it has a probability function of q. Following another learning experience we revise our estimate to p. The change in the probability function as a result of this experience is clearly a measure of the amount of learning that has occurred.

Suppose that we attempt to encode, in the way discussed in the previous Lecture, assuming that X has a probability function q when in fact it turns out to have a function p. Obviously this will mean that we will use a code that it is longer than it needs to be. The extra length in the code is the *relative entropy* of p and q and will be shown in Lecture 4 to be given by:

$$D(p||q) \equiv \sum_{s\in A} p(s) \log \left(p(s)/q(s)\right) \tag{3.1}$$

In order that this be defined we need to assume that terms in the sum for which $p(s) = 0$ do not contribute and if there exist an s such that $q(s) = 0$ when $p(s) \neq 0$ then the relative entropy is undefined i.e. we require that $p(s) \neq 0 \ \Rightarrow q(s) \neq 0$ which means that the set of outcomes with non-zero p is contained in that with non-zero q. Relative entropy is always non-negative as we shall see below and is used to measure learning quantitatively. The learning paradigm described here derives from Bayesian statistics where q is called the prior while p is referred to as the posterior.

Note that the relative entropy is not symmetric ($D(p||q) \neq D(q||p)$) nor does it satisfy the triangle inequality (i.e. it is not the case necessarily that $D(p||q) \leq D(p||r) + D(r||q)$) so it is not a distance function in the usual sense (it defines a pre-metric not a metric). On the other hand as $p \to q$ it does satisfy these relations to a very good approximation . This reflects the fact that there is a very natural connection of information theory with differential geometry and in that one can consider relative entropy as a generalized distance (a divergence) in the (curved) space of probability distributions (see [1]).

Another place where the "distance" between probability functions is important is in statistics. Here one is often interested in the relationship between two (random) variables. Introductory courses in this field spend considerable time on the concept of correlation as a measure of relationship but this is really only a complete measure when the functions are Gaussian. For general functions the relevant concept is *mutual information* which is defined as the relative entropy between the joint function of the two variables and that which would apply if they were completely independent:

$$I(X;Y) \equiv \sum_{s\in A}\sum_{t\in B} p(s,t) \log \left(\frac{p(s,t)}{p(s)p(t)}\right) \tag{3.2}$$

where A and B are the sets of outcomes for X and Y. Complete independence occurs when $p(s,t) = p(s)p(t) \ \ \forall$s,t.

Intuitively mutual information measures the amount of information that two random variables have in common. It has an interesting connection to ordinary entropy and its conditional form discussed above. In particular it is quite easy to show using equations (2.2) and (3.2) that the following interesting relation holds:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{3.3}$$

which says that the mutual information is the average reduction in uncertainty of X due to knowledge of Yor symmetrically it is the reduction of uncertainty of Y due to knowledge of X.

3.2. **Basic Properties.** The concavity of the logarithm function enables us to derive important basic properties of the entropic functionals. Recall a real function $f(x)$ is *convex* on (a, b) if every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$ then

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

and a function is *concave* if $-f(x)$ is convex. Basic calculus shows that convexity of twice differentiable functions can be determined by examining the sign of their second derivative on (a, b). If it is non-negative the function is convex and obviously if it is non-positive it is concave. Since the second derivative of $\log(x)$ is $-\frac{1}{x^2}$ it is concave on its entire domain.

An important concept in probability theory is the *expectation* functional of real functions defined on the outcome space of a random variable S:

$$E(f) \equiv \sum_{s \in A} p(s)f(s)$$

which intuitively gives the average value of the function after many trials of the random variable. Note that it is easy to extend the real function f to define a function $f(S)$ of the random variable S: The outcome space of $f(S)$ is just $f(s)$ with probability $p(s)$.

Now let the outcome space be labeled by the reals then we can prove:

Theorem 1. *(Jensens Inequality) If f is a convex function and X a random variable with outcomes labeled by the reals then*

$$E(f(X)) \geq f(E(X))$$

Equality implies that X is certain with value $E(X)$.

Proof. By induction on the size of the outcome space. If this is two i.e $\{x_1, x_2\}$ then by the convexity of f we have easily that

$$f(p_1x_1 + p_2x_2) \leq p_1f(x_1) + p_2f(x_2)$$

(since $p_2 = 1 - p_1$) which is what we required. Now suppose the result holds on all outcome spaces of size $k - 1$ we show that it holds for all spaces of size k. Consider this latter situation with probability values p_i satisfying $\sum_{i=1}^{k} p_i = 1$. It is easy to see that the new set of $k - 1$ values $p_i^{'} \equiv p_i/(1 - p_k)$ is a probability set for a space of size $k - 1$ since $\sum_{i=1}^{k-1} p_i^{'} = 1$. Now we have

$$\begin{array}{rcl} \sum_{i=1}^{k} p_i f(x_i) & = & p_k f(x_k) + (1 - p_k) \sum_{i=1}^{k-1} p_i^{'} f(x_i) \\ & \geq & p_k f(x_k) + (1 - p_k) f\left(\sum_{i=1}^{k-1} p_i^{'} x_i\right) \\ & \geq & f\left(p_k x_k + (1 - p_k) \sum_{i=1}^{k-1} p_i^{'} x_i\right) \\ & = & f\left(\sum_{i=1}^{k} p_i x_i\right) \end{array}$$

where we used convexity on line 3 and the induction hypothesis on line 2. The second part is left as an exercise. □

Jensens inequality allows us to establish that relative entropy is non-negative which is a fundamental result in information theory.

Theorem 2. *$D(p||q) \geq 0$ for p and q probability functions for which relative entropy is defined. Equality holds only when $p = q$.*

Proof. We prove only the first part and the second is left as an exercise. Let the values of x for which $p(x) \neq 0$ be A and the values for which $q(x) \neq 0$ be B with $A \subset B$ for relative entropy to be defined. We have firstly easily that

$$D(p||q) = E\left(-\log\left(\frac{q}{p}\right)\right)$$

Now minus the logarithm is a convex function and also we can regard $\frac{q}{p}$ as a random variable with real valued outcomes. Thus we can use Jensens inequality to deduce that

$$D(p||q) \geq -\log\left(\sum_{x\in A} p(x)\frac{q(x)}{p(x)}\right) = -\log\left(\sum_{x\in A} q(x)\right) \geq -\log\left(\sum_{x\in B} q(x)\right) = 0$$

where we also used the fact that the logarithm is an increasing function together with the regularity assumption that $A \subset B$ in the last inequality. □

This theorem has a number of interesting consequences. Firstly since the mutual information can be expressed as a relative entropy it is also non-negative. Secondly equation (3.3) then has the consequence

$$H(X) \geq H(X|Y)$$

which means fixing a random variable never increases *on average* uncertainty of another random variable. This result can be combined with the chain rule (2.4) to give

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality only when all the random variables are independent.

In the same kind of way that we defined a conditional entropy we can also define a conditional relative entropy. We consider only the two random variables (X and Y) case but the generalization is obvious. If we fix the value of Y then both $q(X, Y)$ and $p(X, Y)$ will become $q(X|Y = y_0)$ and $p(X|Y = y_0)$ respectively and the relative entropy $D(p(X|Y = y_0)||q(X|Y = y_0))$. If this is averaged over all possible values of Y then we obtain the *conditional* relative entropy:

(3.4)

$$D(p(x|y)||q(x|y)) \equiv \sum_{y\in A} p(y) \sum_{x\in A} p(x|y) \log\left(\frac{p(x|y)}{q(x|y)}\right) = \sum_{x,y\in A} p(x, y) \log\left(\frac{p(x|y)}{q(x|y)}\right)$$

Using this definition and straightforward manipulation we can obtain an important chain rule for relative entropy:

Theorem 3. *$D(p(x, y)||q(x, y)) = D(p(y)||q(y)) + D(p(x|y)||q(x|y))$*

Proof. Exercise. □

This result and the previous Theorem will be used later to establish a generalized second law of thermodynamics.

3.3. **Fine Graining.** If we consider a certain set of n outcomes for two random variables and then we subdivide each of these categories into m_i further categories what happens to the entropy and relative entropy? Shannon's second axiom shows that the entropy increases since the amount of choice increases. It turns out that relative entropy does as well. To prove this we first establish

Theorem 4. *(Log sum inequality) If $\{a_i\}$ and $\{b_i\}$ are n positive sets of values then*

$$\sum_{i=1}^{n} a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^{n} a_i\right) \log\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right) \tag{3.5}$$

with equality only when the two sequences are constant multiples of each other.

Proof. We prove the first part and leave the second as an exercise. Consider $f(t) = t \log t$. For $t > 0$ this is clearly convex by the second derivative test and so by Jensens inequality

$$\sum p_i f(t_i) \geq f\left(\sum p_i t_i\right)$$

where the p_i are a probability set and the t_i are positive. Setting $p_i = \frac{b_i}{\sum_{j=1}^{n} b_j}$ and $t_i = \frac{a_i}{b_i}$ and using the definition of f it follows

$$\sum_i \frac{a_i}{\sum_j b_j} \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_i \frac{a_i}{\sum_j b_j}\right) \log\left(\sum_i \frac{a_i}{\sum_j b_j}\right)$$

and multiplication through by $\sum_j b_j$ gives the desired result. □

If we now recall the form of relative entropy we see that subdividing a particular outcome category into m_i new subcategories is the equivalent of replacing the right hand side of equation (3.5) with $n = m_i$ by the m_i terms of the left hand side. The effect on the relative entropy is therefore non-decreasing and usually an increase.

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry.* Translations of Mathematical Monographs, AMS, Oxford University Press, New York, 2000.