Information Theory and Predictability. Lecture 1: Introduction

1. Historical Background

The concept of entropy originated in thermodynamics in the 19th century where it was intimately related to heat flow and central to the second law of the subject. Later the concept played a central role in the physical illumination of thermodynamics by statistical mechanics. This was due to the efforts of leading physicists such as Boltzmann and Gibbs who used entropy as a measure of the disorder of the large dynamical system underlying molecular collections (see [1] and [2]).

Fisher in the 1920s (see [6]) used related concepts in the development of the foundations of theoretical statistics. This was taken further by Kullback and Leibler in the 1950s (see [5]).

In the middle of the 20th century an electrical engineer and mathematician Claude Shannon working at Bell Labs in New Jersey (see [7]) pointed out the connection between entropy and information content[1]. At that time the concept became a central part of the new discipline of information theory. Later in the 1950s the precise connection of Shannon's entropy to that from statistical mechanics was clarified in a series of elegant papers by Jaynes (see [3] and [4]). Information theory has since seen important applications in such diverse areas as complexity theory, networking analysis, financial mathematics and mathematical statistics. We will examine some of these applications in this course.

2. Entropy, an axiomatic formulation

The central idea of information theory is to measure the uncertainty associated with random variables. In general a random variable X have a certain number of outcomes x_i which have probabilities p_i of occurring. Shannon was interested in developing an intuitive idea of the uncertainty associated with X using only the p_i. He had a very practical interest in developing such a measure since a common problem in digital communications concerns the transmission of information regarding the outcome of random variables.

A concrete example is a digital picture. Each pixel location for the picture has a number of possible outcomes (color and brightness being the most important). To describe this picture in the most efficient manner is of clear practical importance. Intuitively the bigger the picture and the greater the range of colors and brightness, the larger it's digital description since one has to cover a bigger range of possible pictures. If one considers the set of all possible pictures then one can view each pixel as a random variable. The more uniform the probability of color and brightness of this pixel, the greater the choice and this implies that a larger digital description is required. This follows because the most efficient way of encoding a pixel is to

[1]When Shannon had been working on his equations for some time he happened to visit the mathematician John von Neumann who asked him how he was getting on with his theory of missing information. Shannon replied that the theory was in excellent shape except that he needed a good name for "missing information". "Why don't you call it entropy?" von Neumann suggested. "In the first place, a mathematical development very much like yours already exists in Boltzmann's statistical mechanics, and in the second place no one understands entropy very well, so in any discussion you will be in a position of advantage." Shannon took his advice.

choose a shorter code for more probable outcomes. We explore this more precisely below. There appears therefore to be a connection between the amount of effective "choice" in pictures, the size of the digital description required and the uncertainty associated with the set of all pictures. Shannon attempted to make this intuition more solid by developing a series of axioms for how this uncertainty or amount of choice might be quantified. He assumed three axioms

(1) For a random variable with uniform probabilities the entropy or uncertainty should be a monotonically increasing function of the number of outcomes for the random variable. More choice means greater uncertainty.
(2) If one splits an outcome category into subcategories then the new entropy of the extended system should be the sum of the old system entropy plus the new entropy of the split category weighted by it's probability. Creation of more choice increases uncertainty weighted according to the likelihood of the increased choice being relevant.
(3) The entropy should be a continuous function of the p_i

Let us now introduce notation. The entropy of a random variable X with n outcomes is written as $H(X) = H(p_1, p_2, \ldots, p_n)$ where the $i'th$ outcome has probability p_i. We may write the first two axioms therefore as

(1) $H(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) > H(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m})$ for $n > m$
(2) $H(q_1, \ldots, q_m, p_2, p_3, \ldots, p_n) = H(p_1, p_2, \ldots, p_n) + p_1 H(\frac{q_1}{p_1}, \frac{q_2}{p_1}, \ldots, \frac{q_m}{p_1})$ where $p_1 \equiv \sum_{i=1}^{m} q_i$

Note also we are tacitly assuming that arguments of H may be interchanged without effect.

We may now deduce the functional form of H up to a multiplicative factor: Define $A(n) \equiv H(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ and consider $n = jk$. We can group n into j groups of k equally likely outcomes and deduce from Axiom 2 that

$$A(n) = A(j) + j(\frac{1}{j})A(k)$$

$A(jk) = A(j) + A(k)$ can then be used iteratively to deduce that $A(i^j) = jA(i)$. Thus for arbitrary t, s, n and m we have

$$\begin{aligned} A(t^n) &= nA(t) \\ A(s^m) &= mA(s) \end{aligned}$$

Now for any n it is always possible to find an m such that

$$s^m \leq t^n < s^{m+1} \tag{2.1}$$

Taking logarithms of this and using the fact that this function in monotonic we get

$$m \log(s) \leq n \log(t) < (m+1) \log(s)$$

or after division by $n \log(s)$

$$\frac{m}{n} \leq \frac{\log(t)}{\log(s)} < \frac{m}{n} + \frac{1}{n}$$

which means that

$$\left| \frac{\log(t)}{\log(s)} - \frac{m}{n} \right| < \epsilon \equiv \frac{1}{n} \tag{2.2}$$

The monotonic nature of the function A also implies from (2.1) that successively

$$\begin{array}{rcl} A(s^m) & \leq A(t^n) & < A(s^{m+1}) \\ mA(s) & \leq nA(t) & < (m+1)A(s) \\ \frac{m}{n} & \leq \frac{A(t)}{A(s)} & < \frac{m}{n} + \epsilon \\ \left| \frac{A(t)}{A(s)} - \frac{m}{n} \right| & < & \epsilon \end{array}$$

Combining with (2.2) we obtain

$$\left| \frac{A(t)}{A(s)} - \frac{\log(t)}{\log(s)} \right| < 2\epsilon$$

Thus by choosing n sufficiently large we can make ϵ arbitrarily small and so to arbitrary accuracy

$$\frac{A(t)}{\log(t)} = \frac{A(s)}{\log(s)} \equiv R$$

Thus

$$A(n) = R \log(n)$$

with $R > 0$ because of Axiom 1.

Suppose now we have N possible outcomes for a random variable with rational probabilities p_i. Finding a common denominator it follows that for appropriate k_i we can write $p_i = \frac{k_i}{\sum_{i=1}^{N} k_i}$. Now divide each outcome into k_i equally likely outcomes. All subcategories will have equal probability $\frac{1}{\sum_{i=1}^{N} k_i}$ and so again by axiom 2

$$A(\sum_{i=1}^{N} k_i) = H(p_1, p_2, \ldots, p_k) + \sum_{i=1}^{N} (p_i A(k_i))$$

or

(2.3)

$$H(p_1, p_2, \ldots, p_k) = R \left(\sum_{i=1}^{N} p_i \right) \log \left(\sum_{i=1}^{N} k_i \right) - R \sum_{i=1}^{N} [p_i \log (k_i)] = -R \sum_{i=1}^{N} p_i \log p_i$$

using the quotient decomposition of the rational p_i. Axiom 3 now allows us to extend this result from rational p_i all possible real p_i.

The value chosen for R is a choice of convention. If we choose $R = 1/\log(2)$ we refer to the entropy as having units of bits while if we choose $R = 1$ then we refer to the entropy as having units of nats instead.

3. Formal definition

For the first few lectures we will deal with probability functions acting on countable sets which shall in practical applications usually be finite. Such functions are non-negative and sum to unity when the entire set upon which they are defined is included. Entropy is a mapping from the space of probability functions to the (non-negative) reals given by

$$H_x(S) \equiv -\sum_{s \in A} p(s) \log_x p(s) \tag{3.1}$$

where A is the countable set and p the probability function defined on it. The subscript x denotes the logarithm base used and, as discussed above, is usually either 2 or e. In the event that $p(s) = 0$ by convention the contributing term in equation (3.1) is also zero (as indeed it is in the limiting case $p(s) \to 0$). The capitalized S stands for a so-called random variable which has values $s \in A$ with probability $p(s)$. One can equivalently consider the entropy as acting on random variables rather than probability functions.

As mentioned previously the entropy can be interpreted as a measure of the uncertainty associated with a particular random variable and it is easy to check that it attains a minimum of zero when S has a certain value of s_0 i.e. $p(s_0) = 1$ and $p(r) = 0$ when $r \neq s_0$. It attains a maximum for the uniform probability function which evidently represents a situation of maximum uncertainty. This will be proven in the next lecture.

4. Relationship to Information

As we saw earlier Shannon introduced entropy in the context of the practical problem of communicating random variables such as pictures. We can make this a little clearer formally as follows: Suppose we have a random variable S and we wish to efficiently communicate its values $s \in A$ as they are determined. This problem is solved practically by assigning a code $c(s)$ to each possible s and then transmitting the sequence of codes as they are ascertained.

Naturally each code has a certain length $l(c(s))$ and an obvious question is: How can the expected length of the transmission per random realization be minimized and what is this minimum expected length? In other words how do we minimize

$$L \equiv \sum_{s \in A} p(s) l(c(s))$$

It would clearly make sense if we want a shorter transmission of codes to choose l smaller when p is larger. An example helps make this concrete and hopefully clearer. Let us use bits to do the coding (i.e. a 1 or a 0)

s	$p(s)$	Code 1	Code2	Code 3	Code 4
a	0.5	0	0	10	0
b	0.25	0	010	00	10
c	0.125	0	01	11	110
d	0.125	0	10	110	111

Code 1 has $L = 1\,bit$ and so is very short but is obviously not useful because it cannot be decoded at the other end.

Code 2 has $L = 1.75\,bits$ and so is longer but it does have a unique code for each outcome/letter. However consider the transmission 010. This could be either b or it could be ca so it cannot be uniquely decoded either.

Code 3 has $L = 2.125\,bits$ is still longer but this time can be uniquely decoded. One must however wait to the end of the message before decoding so it is not *instantaneously* uniquely decodeable. Consider 11000 whose first four symbols could be cb but the last 0 tells us that instead the answer is db.

Code 4 has $L = 1.75\,bits$ which is the same length as as Code 2 but is now *instantaneously* uniquely decodeable (no code is a prefix of another, whereas for code 3, c is a prefix of d). Notice that we have chosen short codes for the high probabilities and long codes for the low probabilities.

Obviously the first two codes are not useful and the third somewhat impractical. Shannon proved the following result which we will discuss at length in Lecture 4:

Theorem. *There exists a uniquely decodeable Code of minimum length which is instantaneously decodeable. Moreover this minimal length is given by the inequality*

$$H_2(S) \leq L \leq H_2(S) + 1$$

Note that for the example above $H_2(S) = 1.75\, bits$ so Code 4 is an optimal minimal length useful code. The entropy is sometimes not achieved as the minimal Code length because each code member is required to be of integral length while probabilities can be arbitrary.

In summary then the entropy $H(X)$ is the amount of information required on average to describe the random variable X for communication purposes.

5. **Overview of applications**

In this course we will focus on several applications connected primarily with statistical physics although we shall examine data compression since this is where information theory began. This field is remarkably interdisciplinary and we provide here a very brief summary of this diversity.

5.1. **Communication Theory.** We have briefly touched on this above. The study of the rate, flow and distortion of information on a network uses information theory in a central way. In addition the subject of optimal compression of data is an important topic in computer science and was as we have seen the original motivation of Shannon. We shall study this further in Lecture 4.

5.2. **Mathematical Statistics.** A central concept here in significance testing is the comparison of different probability distributions. A generalization of entropy to relative entropy allows for a natural measure of the distance between distributions and this can be used in likelihood and hypothesis testing. Additionally it is useful in fitting distributions to data to assume minimum knowledge about random variables (least biased estimation). Maximizing the entropy of the fitted distribution is a very natural way to achieve this. We shall examine the so called "maxent" techniques further in Lecture 6.

5.3. **Complexity theory.** This was introduced by Kolmogorov who was interested in the most efficient algorithmic (computer program) description of data and the further use of such algorithms as the simplest possible prediction tools (Ochams Razor). This has deep links with entropy/information and is viewed by many as a more fundamental field.

5.4. **Statistical Mechanics.** The concept of entropy originated here however information theory provides a more general mathematical framework. Jaynes was able to reformulate this field in terms of a maximum information theoretic entropy principle as we shall see in Lecture 6. There are also interesting links to irreversible processes or non-equilibrium statistical mechanics as we shall see in Lecture 3 and the second half of the course.

Other important applications can be found in probability (large deviation theory) and in financial market investment theory.

References

[1] L. Boltzmann. *Lectures on Gas Theory*. Dover, March 1995.

[2] J. Gibbs. *Elementary Principles in Statistical Mechanics*. Dover, 1960.

[3] Jaynes E. T. Information theory and statistical mechanics. *Phys. Rev.*, 106:620, 1957.

[4] Jaynes E. T. Information theory and statistical mechanics II. *Phys. Rev.*, 108:171, 1957.

[5] S. Kullback and R.A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.

[6] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London A.*, 222:309–368, 1922.

[7] Shannon C. E. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423,623–656, 1948.