INTRODUCTION TO INFORMATION THEORY

KRISTOFFER P. NIMARK

These notes introduce the machinery of information theory which is a field within applied mathematics. The material can be found in most textbooks on information theory. Simple introductions are Luenberger (2006) and Pierce (1980). More advanced material can be found in Cover and Thomas (2006) and Kullback (1968).

1. Information and entropy

In the everyday meaning of the word, information means many things. To treat the topic formally, we need to be more precise.

1.1. **Defining and measuring information.**

Claude Shannon proposed to define the amount of information I contained in a message with probability p as

$$I \equiv \log (1/p) = -\log p \tag{1.1}$$

Does this definition make sense? Yes, being told that something unusual has occurred is in some sense more informative than knowing that something that often happens have occurred since it changes our beliefs about the state more. Example: The message ”It snowed in Ithaca in July” contains more information than ”It snowed in Ithaca in January.”

While the logarithm in the definition can be taken with respect to any base in (1.1), it is common in information theory to use base 2 logarithms. The amount of information is then measured in bits. Because of this convention, whenever the base of the logarithm is left unspecified, it is taken to be 2.

Date: February 13, 2020.

1.2. **Additivity of information.** Information about independent events is additive: A message about two independent events, A and B, contains

$$I_{AB} = -\log p_A p_B = -\log p_A - \log p_B = I_A + I_B \tag{1.2}$$

1.3. **Entropy.** Entropy is a measure of the expected amount of information. For two events with probability p and $1-p$, entropy is given by

$$H(p) = -p\log p - (1-p)\log(1-p) \tag{1.3}$$

1.3.1. *Example: The weather in California.* The weather in California is sunny with probability 7/8 and cloudy with probability 1/8. The average information on sunny and cloudy days is then given by

$$\begin{aligned} H &= -(7/8)\log(7/8) - (1/8)\log(1/8) \\ &= \frac{1}{8}[7\log 7 - 7\log 8 - \log 8] \\ &= \frac{1}{8}[7\times 2.81 + 7\times 3 - 3] \\ &= .54 \text{ bits.} \end{aligned}$$

Consider the general expression for a two event source

$$H(p) = -p\log p - (1-p)\log(1-p) \tag{1.4}$$

If p equals either 0 or 1, the event is completely determined and entropy is then 0. Entropy is maximized for $p = 1/2$ and $H = 1$.

1.3.2. *Example.* The game 20 questions: What questions do you start with? Which questions do you expect to allow you to reject the largest number of candidates?

1.4. **Information Sources.** An information source is defined to consist of the possible mutually exclusive outcomes (events) and their associated probabilities. The entropy of an n-event source is

$$H(p_1, p_2, ..., p_n) = -p_1 \log p_1 - p_2 \log p_2 - ... - -p_n \log p_n$$

1.4.1. *Example.* A three event source with probabilities 1/2, 1/4,1/4 have entropy

$$\begin{aligned} H\left(1/2, 1/4, 1/4\right) &= \left(\frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4\right) \\ &= \frac{3}{2} \end{aligned}$$

1.5. **Two properties of entropy.**

(1) Non-negativity:$H \geq 0$

(2) $H(p_1, p_2, ..., p_n) \leq \log n$ and with equality if $p_i = p_j \forall i, j$

1.6. **Additivity.** Entropy is additive in the same way information is. If two events, S and T are independent, then the probability of the event (T, S) is $p_T p_S$ and we then have that

$$H(S, T) = H(S) + H(T)$$

We call the (S, T) the product of two sources. If the two sources are independent, the entropy of the product is given by

$$\begin{aligned} H\left(S, T\right) &= -\sum_{s \in S, t \in T} p_s p_t \log p_s p_t \\ &= -\sum_{s \in S, t \in T} p_s p_t \left[\log p_s + \log p_t\right] \\ &= -\sum_{t \in T} p_t \left[\sum_{s \in S} p_s \log p_s\right] - \sum_{s \in S} p_s \left[\sum_{t \in T} p_t \log p_t\right] \\ &= H(S) + H(T) \end{aligned}$$

since

$$\sum_{t \in T} p_t = 1, \sum_{s \in S} p_s = 1.$$

1.7. **Mixture of sources.** Example: A source can by probability p be generated by a die and with probability $1-p$ be generated by a flip of a coin. The event space is then 1,2,3,4,5,6,heads, tails. The entropy of the source is

$$H = pH(die) + (1-p)\,H(coin) + H(p)$$

1.8. **Bits as measure.** Bits measure length of binary strings. Entropy of a string of binary numbers only coincides with the number of bits if all events are equally likely. Otherwise, the entropy is lower than the number of bits.

2. Differential entropy

Entropy is only defined for discrete random variables. The corresponding concept for continuously distributed random variables is called *differential entropy.* Not much changes in terms of results and interpretations, apart from that integrals replace summations.

Definition 1. *The differential entropy of X when X is a continuous random variable is defined as*

$$h(X) = -\int_S f(x)\log f(x)dx$$

where S is the support of f

Example 1. *Uniform distribution $X \sim (0,a)$ so that $f(x) = \frac{1}{a}$*

$$\begin{aligned} h(X) &= -\int_0^a \frac{1}{a}\log\frac{1}{a}dx \\ &= \log a \end{aligned}$$

Entropy is thus increasing in length of interval, which is intuitive.

Example 2. *Normal distribution* $X \sim (0, \sigma^2)$ *so that* $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/\left(2\sigma^2\right)}$

$$\begin{aligned} h(X) &= -\int f(x) \log \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/\left(2\sigma^2\right)} \right] dx \\ &= \frac{1}{2} \log 2\pi e \sigma^2 \end{aligned}$$

The entropy of a normally distributed variable is thus increasing in its variance, which is also intuitive since entropy is also a natural measure of uncertainty.

3. Channels

An information channel transmits information about the random variable X (input) to the random variable Y (output). A discrete channel is defined by transition probabilities between states of the input X and the output Y, or equivalently, the conditional distribution of outputs (signals), given the inputs (states).

Example 3. *Binary symmetric channel* $X, Y \in \{0, 1\}$

$$\begin{aligned} p(Y &= X) = p \\ p(Y &\neq X) = 1 - p \end{aligned}$$

Example 4. *Binary asymmetric channel* $X, Y \in \{0, 1\}$

$$\begin{aligned} p(Y &= 1 \mid X = 1) = p \\ p(Y &= 0 \mid X = 0) = q \end{aligned}$$

and $p \neq q$.

3.1. **Conditional and joint entropies.**

Conditional entropy is denoted $H(Y \mid X)$.It can be derived by starting with the entropy of Y conditional on a particular realization of X

$$H\left(Y \mid x_i\right) = \sum_{y_j} p\left(y_j \mid x_i\right) \log \frac{1}{p\left(y_j \mid x_i\right)}.$$

Definition 2. *Conditional entropy $H(Y \mid X)$*

$$H(Y \mid X) = \sum_{x_i} H\left(Y \mid x_i\right) p\left(x_i\right).$$

Example 5. *The toss of a die. Let Y be outcome of a six sided die and let X be the events High (5 or 6) and Not High (1,2,3, or 4). The conditional entropy is*

$$\begin{aligned} H(Y &\mid X) = \frac{1}{3}\log 2 + \frac{2}{3}\log 4 \\ &= \frac{5}{3} \\ &< 2.56 \\ &= H(Y) \end{aligned}$$

This is a general result:

Lemma 1. *Conditioning on a random variable cannot increase entropy*

$$0 \leq H(Y \mid X) \leq H(Y)$$

However, it is not generally true that $H(Y \mid x_i) \leq H(Y)$.

Lemma 2. *Joint entropy*

$$\begin{aligned} H(X,Y) &= H(Y \mid X) + H(X) \\ &= H(X \mid Y) + H(Y) \end{aligned}$$

3.2. **Mutual Information.**

Definition 3. *The mutual information of a random variable X given a random variable Y is*

$$I(X;Y) = H(X) - H(X \mid Y)$$

Interpretations: Mutual information is the reduction in uncertainty about X that occurs when Y is observed

Properties of mutual information

(1) $I(X;Y) = H(X,Y) - H(X \mid Y) - H(Y \mid X)$

(2) $I(X;Y) = H(X) - H(Y \mid Y)$

(3) $I(X;Y) = H(Y) - H(Y \mid X)$

(4)
$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

(5) $I(X;Y) \geq 0$

(6) $I(X;X) = H(X)$

References

[1] Cover, T.M. and J.A. Thomas, *Elements of Information Theory*, 2006, Wiley.

[2] Kullback S. *Information theory and statistics.* New York: Dover, 2nd ed.. 1968.

[3] Luenberger, D.G., *Information Science*, 2006, Princeton University Press.

[4] Pierce, *An Introduction to Information Theory: Symbols, Signals and Noise,* Dover books, 1980.