Technical notes on information theory

Yang Xu

Information theory provides a mathematical foundation for the quantification, compression and transmission of information. This note summarizes a core set of concepts concerning *entropy* - a measure for information, or uncertainty, that is central to the work pioneered by Shannon (1948).

1. Bit

   Bit is a unit of information. 1 bit refers to the amount of information that one is uncertain about in a binary random variable that takes the value of either 0 or 1 with equal probability.

2. Surprisal

   Surprisal (s) quantifies the uncertainty in a random variable X taking a certain value x based on its probability of occurrence $p(X = x)$ or $p(x)$. Surprisal is measured in *bits* when the base of the logarithm is 2.

$$s(x) = \log_2 \frac{1}{p(x)} = -\log_2 p(x). \tag{1}$$

   Due to the logarithmic transformation, surprisal decreases monotonically as probability increases. Figure 1 illustrates this relationship. An event with zero probability would have an infinite level of surprisal, because one would be maximally uncertain about the outcome of this event (and consequentially, one would be completely surprised). An event with probability of 0.5, e.g. a fair coin toss, would have a surprisal value of 1 bit, indicated by the blue lines. A sure event would have a surprisal of 0 as indicated by the green lines, because (intuitively) the outcome would always be within one's expectation, hence entailing no surprise.

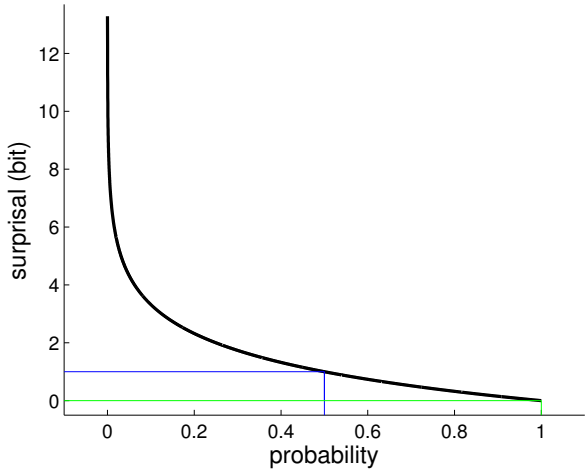

Figure 1: Surprisal *vs.* probability.

3. Entropy

   Entropy $H(\cdot)$ for a random variable X is the expected or average surprisal based on its probability distribution. Entropy is measured in *bits* when the base of the logarithm is 2.

$$H(X) = \sum_x p(x)s(x) = \sum_x p(x) \log_2 \frac{1}{p(x)} = -\sum_x p(x) \log_2 p(x). \tag{2}$$

   For example, the random variable that determines whether a coin shows up heads or tails in a given toss has the entropy:

$$H(X) \quad = \quad p(head) \log_2 \frac{1}{p(head)} + p(tail) \log_2 \frac{1}{p(tail)}. \tag{3}$$

   For a fair coin, the entropy would be $0.5 \log_2 \frac{1}{0.5} + 0.5 \log_2 \frac{1}{0.5} = 1$ bit. It is easy to show, e.g. by simulation, that entropy is maximal when the outcomes are equally probable.

4*. Source coding theorem

   It is impossible to compress an input variable (or a data source) at a rate less than its entropy without any loss of information.

5. Joint entropy

   Joint entropy of two discrete variables X and Y (swap sums with integrals for continuous variables) is the total amount of uncertainty in the outcomes of these events:

$$H(X, Y) = -\sum_{x,y} p(x, y) \log_2 p(x, y). \tag{4}$$

   Pictorically, the joint entropy is the union of areas covered by the two circles (i.e. the entropies of two individual variables) in Figure 2.

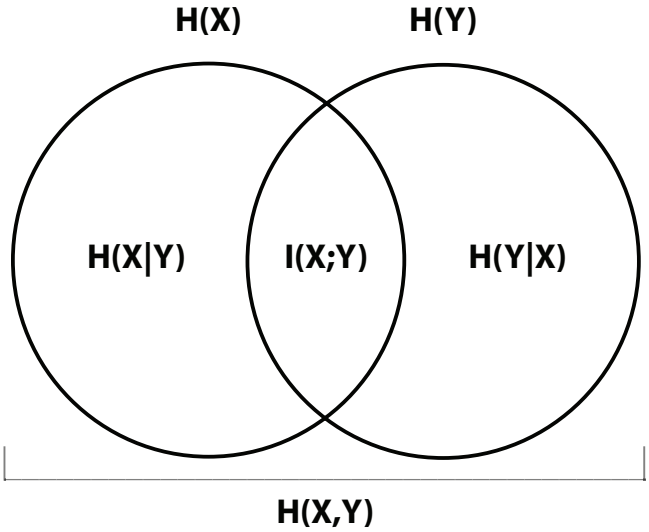

Figure 2: Illustration of entropy, joint entropy, conditional entropy, and mutual information.

6. Conditional entropy

Conditional entropy $H(X|Y)$ is the amount of uncertainty in X given what one knows about Y, and *vice versa* for $H(Y|X)$:

$$H(X|Y) = -\sum_{x,y} p(x,y) \log_2 p(x|y); \tag{5}$$

$$H(Y|X) = -\sum_{x,y} p(x,y) \log_2 p(y|x). \tag{6}$$

Pictorially, $H(X|Y)$ is equivalent to the area occupied by $H(X,Y)$ (i.e. joint entropy of X and Y) with the area under $H(Y)$ excluded in Figure 2, and similarly, $H(Y|X)$ is equivalent to the area under $H(X,Y)$ with the area under $H(X)$ subtracted. Thus, the conditional entropies can be also formulated in terms of the joint and individual entropies:

$$H(X|Y) = H(X,Y) - H(Y); \tag{7}$$

$$H(Y|X) = H(X,Y) - H(X). \tag{8}$$

7. Mutual information

Mutual information (MI) $I(\cdot;\cdot)$ between two variables X and Y quantifies the amount of information that is "shared" between the variables, or the degree of dependence between two variables. Pictorially, MI is the area where $H(X)$ and $H(Y)$ overlap in Figure 2. Thus MI can be calculated in multiple ways:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) = H(X,Y) - H(X|Y) - H(Y|X) \\ &= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}. \end{aligned} \tag{9}$$

When two variables are independent, their mutual information is 0 since $p(x,y) = p(x)p(y)$.

8. Kullback-Leibler divergence

Kullback-Leibler divergence, or KL divergence, measures the relative entropy or divergence between two probability distributions:

$$KL(P_x||P_y) = \sum_i p_x(i) \log \frac{p_x(i)}{p_y(i)}. \tag{10}$$

It is an asymmetric measure such that $KL(P_x||P_y) \neq KL(P_y||P_x)$. It is also not hard to show, by comparing Equations 9 and 10, that mutual information between X and Y is equivalent to the KL divergence (when the base of the logarithm is 2): $I(X;Y) = KL(p(x,y)||p(x)p(y))$. This equivalence provides the interpretation that mutual information measures the divergence between the joint probability and the product of the marginal probablities of two variables.

9*. Channel capacity (noisy-channel coding theorem)

The channel capacity C is defined as the maximal amount of information that can be transmitted between an input X and an output Y, namely the supremum of mutual information between the two variables considering all possible values of the input:

$$C = sup_{p(x)} I(X;Y). \tag{11}$$

10*. Maximum entropy principle

The maximum entropy principle (Jaynes, 1957) provides a bridge between information theory and probability theory. It states that given certain *a priori* knowledge, the distribution that best represents the state of knowledge is the one with maximal entropy. As such, this principle explains why certain probability distributions take the forms they do. Below are two examples.

Case 1: Uniform distribution (discrete)
It can be shown that a uniform distribution maximizes the entropy of a probability distribution $P(X)$ subject to no more prior knowledge than that the probability masses need to sum to 1. This can be formulated in terms of a Langrange function $\mathcal{L}(\cdot)$ as follows:

$$\mathcal{L}(X, \lambda) = H(X) + \lambda(\sum_x P(X = x) - 1) = -\sum_x p(x) \log p(x) + \lambda(\sum_x p(x) - 1). \quad (12)$$

Maximizing this function involves setting derivatives with respect to probability of each value of $x = x'$, i.e. $p(x')$, to 0, and similarly with respect to the Langrange multiplier λ:

$$\frac{\partial \mathcal{L}}{\partial p(x')} = -\log p(x') - 1 + \lambda = 0 \ , \forall x'; \quad (13)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_x p(x) - 1 = 0. \quad (14)$$

Equation 13 suggests that $p(x') = e^{\lambda - 1}$, $\forall x'$, hence $p(x) = \frac{1}{N}$ (N is the total number of possible values of x) is a probability distribution that satisfies the *a priori* constraint and maximizes the uncertainty under that constraint.

Case 2: Gaussian distribution (continuous)
It can be shown that a Gaussian distribution maximizes the entropy of a probability distribution $f(x)$ subject to the prior knowledge that 1) the probability distribution sums to 1; 2) the mean of the distribution is μ; 3) the variance of the distribution is σ^2. These constraints can be formulated with Langrange multiples λ_0, λ_1, and λ_2:

$$\mathcal{L}(X, \lambda_0, \lambda_1, \lambda_2) = H(X) + \lambda_0(\int_x f(x)dx - 1) + \lambda_1(\int_x xf(x)dx) + \lambda_2(\int_x x^2 f(x)dx - \sigma^2). \quad (15)$$

Maximizing this Langrange function (by setting the partial derivatives to 0) would yield a Gaussian distribution, although we omit the details here because the derivation is beyond the scope of the course. The constraints on mean and variance are a special case of constraints on N orders of moments (M) of a distribution, where the generalized Langrangian is:

$$\mathcal{L}(X, \{\lambda_0, \lambda_1, ..., \lambda_N\}) = H(X) + \sum_{i=0}^{N} \lambda_i(\int_x g_i(x) f(x)dx - M_i). \quad (16)$$

Here $g_i(x)$ is a polynomial function of order i, e.g. $g_2(x) = x^2$. It can be shown that a general solution to the maximum entropy distribution is ($Z(\lambda)$ is the normalizing constant):

$$f(x) = Z(\lambda) e^{\sum_{i=1}^{N} \lambda_i g_i(x)}. \quad (17)$$