Information Theory

This is a brief tutorial on Information Theory, as formulated by Shannon [Shannon, 1948]. It is well beyond the scope of this paper to engage in a comprehensive discussion of that field; however, it is worthwhile to have a short reference of the relevant concepts. Readers interested in a deeper discussion are referred to [Cover and Thomas, 1991], where much of this material is derived.

1 Random Variables and Probabilities

A *variable* is an object, **X**, that can take on any value from a set of values Ω_x (dubbed its *domain*). These values may be discrete and finite, such as the letters of the alphabet or {0, 1}, or they may be continuous and infinite, such as any real number. A *random variabl*e is a variable whose value is unpredictable. A particular value that a random variable has taken on a some time is called a *trial*. A collection of trials is called a *sample*. A common example of a random variable is one representing the flip of a coin. The random variable may take on one of two values {H, T}. Each time we flip the coin, we have a trial. A series of coin flips is then a sample.

Associated with every random variable is a (possibly unknown) *probability distribution*, **P**(**X**). The probability of a particular value is the proportion of the number of times you expect to see that value over a very large sample. This distribution maps every possible value

of $\Omega_{\mathbf{x}}$ to a value in [0, 1]. As $\mathbf{P}(\mathbf{X})$ is a probability distribution, $\sum_{\mathbf{x} \in \Omega_{\mathbf{x}}} \mathbf{P}(\mathbf{x}) = 1$. In our coin example, we might assume a fair coin, such that $\mathbf{P}(\mathbf{X}=\text{H})=\mathbf{P}(\mathbf{X}=\text{T})=0.5$. For a two-headed coin, we might have $\mathbf{P}(\mathbf{X}=\text{H})=1$ and $\mathbf{P}(\mathbf{X}=\text{T})=0$.

For continuous random variables, there is a subtle problem. Because **X** can take on an infinite number of values, the value of any given value will almost always be zero. Instead of probability distributions, we use *probability densities* and integrate over ranges of possible values; however, the distinction is not important for the purposes of this discussion.[1]

In addition to discussing a single random variable, we have a vocabulary for discussing several at once. This is useful because random variables may be *dependent* upon one another. For example, we may define a new variable, $\mathbf{Y}=\mathbf{F}(\mathbf{X})$, where $\mathbf{F}()$ is a deterministic function. In this way, knowing **X** means that we always know the value of **Y**. On the other hand, if **X** and **Y** represent two separate coin flips then we might expect that knowing the value of one will not tell us anything about the other. If this is true, they are said to be *independent*. Of course, there are states between complete dependence and complete independence. We might have a noisy signal driving a noisy speaker. Knowing the original signal tells us something about the sound coming out, but not everything. There is still uncertainty that comes from the noisy speaker itself.

We can formalize these notions using *joint distributions*. A joint distribution, $\mathbf{P}(\mathbf{X},\mathbf{Y})$, tells us everything about the co-occurrence of events from **X** and **Y**. In fact, we can derive $\mathbf{P}(\mathbf{X})$ (and $\mathbf{P}(\mathbf{Y})$) from the joint by computing the *marginal distribution*:

$$\mathbf{P}(\mathbf{X}) = \sum_{\mathbf{y} \in \Omega_{\mathbf{y}}} \mathbf{P}(\mathbf{X}, \mathbf{Y}=\mathbf{y}).$$

Two variables are independent if and only if $\mathbf{P}(\mathbf{X},\mathbf{Y})=\mathbf{P}(\mathbf{X})\bullet\mathbf{P}(\mathbf{Y})$.

Closely related to the joint distribution is the conditional distribution:

$$\mathbf{P}(\mathbf{Y}|\mathbf{X}) = \frac{\mathbf{P}(\mathbf{X}, \mathbf{Y})}{\mathbf{P}(\mathbf{X})}$$

1. This is very important in general, however. In particular, many of the theorems that hold for discrete random variables do not hold for continuous variables. Where this is a problem, we will mention it; otherwise, when thinking of continuous random variables use integrals instead of summations.

which tells us the probability of **Y** if we already knew **X**. Note that this relationship gives us another definition for the joint; namely, that $\mathbf{P}(\mathbf{X},\mathbf{Y}) = \mathbf{P}(\mathbf{Y} \mid \mathbf{X})\bullet\mathbf{P}(\mathbf{X})$. Because joints are not order dependent, this also means that $\mathbf{P}(\mathbf{X},\mathbf{Y}) = \mathbf{P}(\mathbf{X} \mid \mathbf{Y})\bullet\mathbf{P}(\mathbf{Y})$. This observation leads us to *Bayes' rule*:

$$\mathbf{P}(\mathbf{X}|\mathbf{Y}) = \mathbf{P}(\mathbf{Y}|\mathbf{X})\frac{\mathbf{P}(\mathbf{X})}{\mathbf{P}(\mathbf{Y})}.$$

This rule turns out to be quite useful, and allows us to invert conditional probabilities.

We can construct joint and conditional distributions over three random variables, $\mathbf{P}(\mathbf{X},\mathbf{Y},\mathbf{Z})$, as well. We can also compute marginals, $\mathbf{P}(\mathbf{X}) = \sum_{\mathbf{z}\in\Omega_z,\mathbf{y}\in\Omega_y} \mathbf{P}(\mathbf{X},\mathbf{Y}=\mathbf{y},\mathbf{Z}=\mathbf{z})$. We can even define our joints in terms of conditionals: $\mathbf{P}(\mathbf{X},\mathbf{Y},\mathbf{Z})=\mathbf{P}(\mathbf{X} \mid \mathbf{Y},\mathbf{Z})\bullet\mathbf{P}(\mathbf{Y} \mid \mathbf{Z})\bullet\mathbf{P}(\mathbf{Z})$. A defintion of independence ($\mathbf{P}(\mathbf{X},\mathbf{Y},\mathbf{Z})=\mathbf{P}(\mathbf{X})\bullet\mathbf{P}(\mathbf{Y})\bullet\mathbf{P}(\mathbf{Z})$) follows naturally. Generally, we can definte joints and conditionals for any number of random variables.

2 Moments

There are several statistics we might want to use to describe the behavior of our random variables. When our random variable ranges over numbers, one of the most common statistics is the "average." We can define the *mean* or *expected value* of a random variable as:

$$\mathbf{E}_{\mathbf{X}}[\mathbf{X}] = \sum_{\mathbf{x}\in\Omega_x} \mathbf{x}\mathbf{P}(\mathbf{X}=\mathbf{x}).$$

In a common abuse of notation we will usually dismiss the subscript and refer to the expectation of **X** as simply $\mathbf{E}[\mathbf{X}]$.

What do we do in the case where we do not know the distribution of **X** and so cannot compute $\mathbf{E}[\mathbf{X}]$? If $\{\ldots x_i \ldots\}$ refers to a series of trials of **X** then we can compute a *sample mean* instead:

$$\mathbf{E}[\mathbf{X}] = \frac{1}{\mathbf{N}}\sum_{\mathbf{x}\in\Omega_x} \mathbf{x},$$

where *N* is the number of trials in the sample.

It is worth noting that the true mean is a deterministic function of the distribution of **X** while the sample mean is not. Because the samples are themselves random, we might calcuate

a different sample mean each time we pick a sample. Therefore, the sample mean is also a random variable. Luckily, the law of large numbers allows one to prove that as we take more and more trials of X, we approach an estimation of the true distribution. Thus, in the limit, the sample mean approaches the true expectation.

There are other statistics that we might compute when the mean is not enough. For example, *variance* measures the variation of values about the mean:

$$\mathbf{Var}(\mathbf{X}) = \mathbf{E}[(\mathbf{X} - \mathbf{E}[\mathbf{X}])^2] = \mathbf{E}[\mathbf{X}^2] - \mathbf{E}[\mathbf{X}]^2.$$

It is closely related to the *standard deviation*, $\sigma(X)$, which is its square root.

The mean is the first *moment* of the random variable **X**. In general, there are **k** moments, each denoted by $\mathbf{E}[\mathbf{X}^k]$. When you subtract the mean from **X** before taking the expectation, $\mathbf{E}[(\mathbf{X} - \mathbf{E}[\mathbf{X}])^k]$, you have a *central moment*. The variance is therefore the second central moment of **X**. Often, in order to control for scale, we compute a *normalized central moment:*

$$\frac{\mathbf{E}[(\mathbf{X} - \mathbf{E}[\mathbf{X}])^k]}{\sigma(\mathbf{X})^k}.$$

Each increasing moment can be used to further classify the behavior of a random variable. In the work for which this tutorial was first written we often used kurtosis—the fourth normalized central moment—as a convenient measure of the peakedness of a distribution.

3 Entropy

Although it is in principle a very old concept, entropy is generally credited to Shannon because it is the fundamental measure in information theory. Entropy is often defined as an expectation:

$$\mathbf{H}(\mathbf{X}) = -\mathbf{E}[\log \mathbf{P}(\mathbf{X})] = -\sum_{\mathbf{x} \in \Omega_x} \mathbf{P}(\mathbf{X}=\mathbf{x}) \log \mathbf{P}(\mathbf{X}=\mathbf{x})$$

where $0 \log(0) = 0$. The base of the logarithm is generally 2. When this is the case, the units of entropy are *bits*.

Entropy captures the amount of randomness or uncertainty in a variable. This, in turn, is a measure of the average length of a message that would have to be sent to describe a sam-

ple. Recall our fair coin from § 1. It's entropy is: $-(0.5\log 0.5 + 0.5\log 0.5) = 1$; that is, there is one bit of information in the random variable.This means that on average we need to send one bit per trial to describe a sample. This should fit your intuitions: if I flip a coin 100 times, I'll need 100 numbers to describe those flips, if order matters. By contrast, our two-headed coin has entropy $-(1\log 1 + 0\log 0) = 0$. Even if I flip this coin 100 times, it doesn't matter because the outcome is always heads. I don't need to send any message to describe a sample.

There are other possibilities besides being completely random and completely determined. Imagine a weighted coin, such that heads occurr 75% of the time. The entropy would be: $-(0.75\log 0.75 + 0.25\log 0.25) = 0.8113$. After 100 trials, I'd only need a message of about 82 bits on average to describe the sample. Shannon showed that there exists a coder that can construct messages of length $\mathbf{H}(\mathbf{X})+1$, nearly matching this ideal rate.

Just as with probabilities, we can compute joint and conditional entropies. Joint entropy is the randomness contained in two variables, while conditional entropy is a measure of the randomness of one variable given knowledge of another. Joint entropy is defined as:

$$\mathbf{H}(\mathbf{X}, \mathbf{Y}) = -\mathbf{E}_{\mathbf{X}}[\mathbf{E}_{\mathbf{Y}}[\log \mathbf{P}(\mathbf{X}, \mathbf{Y})]] = -\sum_{\mathbf{x} \in \Omega_z, \mathbf{y} \in \Omega_y} \mathbf{P}(\mathbf{X}=\mathbf{x}, \mathbf{Y}=\mathbf{y}) \log \mathbf{P}(\mathbf{X}=\mathbf{x}, \mathbf{Y}=\mathbf{y})$$

while the condtional entropy is:

$$\mathbf{H}(\mathbf{Y}|\mathbf{X}) = -\mathbf{E}_{\mathbf{X}}[\mathbf{E}_{\mathbf{Y}}[\log \mathbf{P}(\mathbf{Y}|\mathbf{X})]] = -\sum_{\mathbf{x} \in \Omega_z, \mathbf{y} \in \Omega_y} \mathbf{P}(\mathbf{X}=\mathbf{x}, \mathbf{Y}=\mathbf{y}) \log \mathbf{P}(\mathbf{Y}=\mathbf{y}|\mathbf{X}=\mathbf{x}).$$

There are several interesting facts that follow from these definitions. For example, two random variables, **X** and **Y**, are considered independent if and only if $\mathbf{H}(\mathbf{Y}|\mathbf{X}) = \mathbf{H}(\mathbf{Y})$ or $\mathbf{H}(\mathbf{X}, \mathbf{Y}) = \mathbf{H}(\mathbf{X}) + \mathbf{H}(\mathbf{Y})$. It is also the case that $\mathbf{H}(\mathbf{Y}|\mathbf{X}) \leq \mathbf{H}(\mathbf{Y})$ (knowing more information can never increase our uncertainty). Similarly, $\mathbf{H}(\mathbf{X}, \mathbf{Y}) \leq \mathbf{H}(\mathbf{X}) + \mathbf{H}(\mathbf{Y})$. It is also the case that $\mathbf{H}(\mathbf{X}, \mathbf{Y}) = \mathbf{H}(\mathbf{Y}|\mathbf{X}) + \mathbf{H}(\mathbf{X}) = \mathbf{H}(\mathbf{X}|\mathbf{Y}) + \mathbf{H}(\mathbf{Y})$.These relations hold in the general case of more than two variables.

There are several facts about discrete entropy, **H**(), that do not hold for continuous or *differential entropy*, **h**(). The most important is that while $\mathbf{H}(\mathbf{X}) \geq 0$, **h**() can actually be negative. Worse, even a distribution with an entropy of $-\infty$ can still have uncertainty. Luckily, for

us, even though differential entropy cannot provide us with an absolute measure of randomness, it is still that case that if $\mathbf{h}(\mathbf{X}) \geq \mathbf{h}(\mathbf{Y})$ then **X** has more randomness than **Y**.

4 Mutual Information

Although conditional entropy can tell us when two variables are completely independent, it is not an adequarte measure of dependence. A small value for $\mathbf{H}(\mathbf{Y}|\mathbf{X})$ may imply that **X** tells us a great deal about **Y** or that **H(Y)** is small to begin with. Thus, we measure dependence using *mutual information*:

$$\mathbf{I}(\mathbf{X}, \mathbf{Y}) = \mathbf{H}(\mathbf{Y}) - \mathbf{H}(\mathbf{Y}|\mathbf{X}).$$

Mutual information is a measure of the reduction of randomness of a variable given knowledge of another variable. Using properties of logarithms, we can derive several equivalent definitions:

$$\begin{aligned} \mathbf{I}(\mathbf{X}, \mathbf{Y}) &= \mathbf{H}(\mathbf{Y}) - \mathbf{H}(\mathbf{Y}|\mathbf{X}) \\ &= \mathbf{H}(\mathbf{X}) - \mathbf{H}(\mathbf{X}|\mathbf{Y}) \\ &= \mathbf{H}(\mathbf{X}) + \mathbf{H}(\mathbf{Y}) - \mathbf{H}(\mathbf{X}, \mathbf{Y}) \\ &= \mathbf{I}(\mathbf{Y}, \mathbf{X}) \end{aligned}$$

In addition to the definitions above, it is useful to realize that mutual information is a particular case of the *Kullback-Leibler divergence*. The KL divergence is defined as:

$$\mathbf{D}(\mathbf{p}||\mathbf{q}) = \int \mathbf{p}(\mathbf{x}) \log \frac{\mathbf{p}(\mathbf{x})}{\mathbf{q}(\mathbf{x})}.$$

KL divergence measures the difference between two distributions. It is sometimes called the relative entropy. It is always non-negative and zero only when **p**=**q**; however, it is not a distance because it is not symmetic.

In terms of KL divergence, mutual information is:

$$\mathbf{D}(\mathbf{P}(\mathbf{X}, \mathbf{Y})||\mathbf{P}(\mathbf{X})\mathbf{P}(\mathbf{Y}))) = \int \mathbf{P}(\mathbf{X}, \mathbf{Y}) \log \frac{\mathbf{P}(\mathbf{X}, \mathbf{Y})}{\mathbf{P}(\mathbf{X})\mathbf{P}(\mathbf{Y})}.$$

In other words, mutual information is a measure of the difference between the joint probability and product of the individual probabilities. These two distributions are equivalent only when **X** and **Y** are independent, and diverge as **X** and **Y** become more dependent.

References

Bell, A. and Sejnowski, T. (1995). An Information-Maximization Approach to Blind Source Separation and Blind Deconvolution. *Neural Computation*, 7, 1129-1159.

Cover, T. and Thomas, J. (1991). Elements of Information Theory. John Wiley and Sons.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell Systems Technical Journal*, 27.

Viola, P. (1995). *Aligment by Maximization of Mutual Information.* PhD Thesis. AI Technical Report No. 1548. Massachusetts Institute of Technology.