6.441 Transmission of Information Feb 9, 2006

Lecture 2

Lecturer: Madhu Sudan *Scribe: Cy Chan*

1 Administrivia

- Questionnaire - get form from web and fill out
- Scribing - sign up on website
- Mailing List - if you haven't received an email already, tell staff
- Problem Set 1 - due 2/22/06

2 Introduction

- **Entropy** - associated with a random variable (RV) and quantifies the amount of uncertainty associated with that RV
- **Information** - associated with a pair of RVs:

$$I(X;Y) = \text{how much } Y \text{ informs us about } X$$

3 Entropy

3.1 Example

Let X, Y, and W be RVs where:

$$X = \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases}$$

$$Y = \begin{cases} 0 & \text{with probability } 7/8 \\ 1 & \text{with probability } 1/8 \end{cases}$$

$$W = \begin{cases} 0 & \text{with probability } 9/10 \\ 1 & \text{with probability } 1/20 \\ 2 & \text{with probability } 1/20 \end{cases}$$

Intuitively, X is more random than Y, but how do we make a comparison between Y and W? We need a way of quantifying the amount of randomness in each RV.

3.2 Derivation of Entropy $H(Z)$ for Bernoulli RV Z

Define Z to be a Bernoulli RV with parameter p:

$$Z = \begin{cases} 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p \end{cases}$$

How many bits are required to convey the value of Z? If we only communicate a single instance of Z, we must send at least 1 bit, but if we are sending many instances, we can batch the values as in the previous lecture and achieve an average of less than 1 bit per value.

Suppose we have a sequence $Z_1, Z_2, \ldots, Z_n$ of n independent, identically distributed (IID) RVs each with the same distribution as Z above. We prescribe the following algorithm to encode a sequence $z_1, z_2, \ldots, z_n$ drawn from this distribution:

1. Send $k = \sum_{i=1}^n z_i$, which takes k bits (takes $\log_2 n$ bits)
2. Create a table T_k that describes every sequence with k 1's and $(n-k)$ 0's
3. Send the index in the table that describes the sequence $z_1, z_2, \ldots, z_n$ (takes $\log_2 \binom{n}{k}$ bits)

We can write the expected length of the resulting encoding as

$$l = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \log_2 \binom{n}{k} + \log_2 n.$$

To simplify, we make use of the law of large numbers, from which we get

$$\Pr\left[k \notin [(p-\epsilon)n, (p+\epsilon)n]\right] \leq 2^{-\epsilon^2 n}.$$

Then,

$$l = \sum_{k=(p-\epsilon)n}^{(p+\epsilon)n} \binom{n}{k} p^k (1-p)^{n-k} \log_2 \binom{n}{k} + \sum_{k \notin [(p-\epsilon)n, (p+\epsilon)n]} \Pr[k \text{ is such}] \binom{n}{k} + \log_2 n.$$

Since $\Pr\Big[k \notin [(p-\epsilon)n, (p+\epsilon)n]\Big] \binom{n}{k} \leq 2^{-\epsilon^2 n} n$, the second term becomes vanishingly small as n gets large. Similarly, the third term $\log_2 n$ vanishes when we divide by n when taking the average encoding length per value. In the first term we note the following:

$$\sum_{k=(p-\epsilon)n}^{(p+\epsilon)n} \binom{n}{k} p^k (1-p)^{n-k} \leq \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1,$$

and for each term in the summation, $k \approx pn$, so

$$\sum_{k=(p-\epsilon)n}^{(p+\epsilon)n} \binom{n}{k} p^k (1-p)^{n-k} \log_2 \binom{n}{k} \leq 1 \cdot \log_2 \binom{n}{pn}.$$

Stirling's approximation for $n!$ implies

$$\binom{n}{pn} \approx \left(\frac{1}{p}\right)^{pn} \left(\frac{1}{1-p}\right)^{(1-p)n},$$

so for large n:

$$l \leq \log_2 \binom{n}{pn} \approx \log_2 \left[\left(\frac{1}{p}\right)^{pn} \left(\frac{1}{1-p}\right)^{(1-p)n}\right]$$

$$l \leq n \left[p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}\right].$$

The entropy $H(Z)$ is the average encoding length per value:

$$H(Z) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}.$$

3.3 Extensions to Non-Bernoulli discrete RVs

What if we have a RV that takes N values? Consider a RV Z that takes values in $\{1, 2, \ldots, N\}$, where $p_i = \Pr[Z = i]$. We define two new RVs, Z_1 and Z_2, where

$$Z_1 = \begin{cases} 0 & \text{if } Z = 1 \\ 1 & \text{otherwise} \end{cases}$$

$$\Pr[Z_1 = 0] = p_1, \text{ and } \Pr[Z_1 = 1] = 1 - p_1,$$

$$Z_2 = Z | \{Z_1 = 1\}$$

$$\Pr[Z_2 = i] = \frac{p_i}{1 - p_1}, \text{ for } i \in \{2, 3, \ldots, N\}.$$

We can show that

$$H(Z) = H(Z_1) + \Pr[Z_1 = 1] H(Z_2),$$

and by induction that

$$H(Z) = \sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i}.$$

Note that we would have gotten the same answer no matter how we partition the sequence and assign new random variables.

3.4 Properties of Entropy

The entropy function satisfies the following three properties:

1. $H(p_1, p_2, \ldots, p_N)$ is symmetric in its arguments
2. $H(p_1, p_2, \ldots, p_N) = H(p_1, 1 - p_1) + (1 - p_1) H(\frac{p_2}{1-p_1}, \frac{p_3}{1-p_1}, \ldots, \frac{p_N}{1-p_1})$
3. $H(p_1, p_2, \ldots, p_N) \leq \log_2 N$

In property 3, the inequality is strict unless $p_i = \frac{1}{N}$ for all i. In other words, maximum entropy occurs when the probability mass is evenly distributed. For probability functions with unbounded support, it is possible to have unbounded entropy. For example, over all densities on the real line, the density that maximizes entropy (indeed the differential entropy) for a given variance is a gaussian distribution. For densities over positive reals with a given mean, the entropy maximizing density is the exponential distribution. This is because the square of a zero mean Gaussian random variable is exponentially distribution with the mean equal to its variance.

Other functions may satisfy the above three requirements, but if we change property 3 to

3′. $H(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}) = \log_2 N$,

then properties 1, 2, and 3′ imply our specific entropy function $H(Z) = \sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i}$.

3.5 Joint and Conditional Entropy

We can extend our definition of entropy to include joint distributions of RVs. If we have a pair of RVs (X, Y) with density $P(X, Y)$ over $\Omega_x \times \Omega_y$, we define the joint entropy as:

$$H(X,Y) = \sum_{x \in \Omega_x, y \in \Omega_y} P(X = x, Y = y) \log_2 \frac{1}{P(X = x, Y = y)}.$$

We define conditional entropy $H(X|Y)$ as the average (over Y) entropy of X given Y:

$$H(X,Y) = \sum_{y \in \Omega_y} P_y(y)\ H(X|Y = y).$$

Intuitively, we sense that $H(X)$ should be no smaller than $H(X|Y)$, which we will prove next lecture. In the satellite example in the previous lecture, if X is the satellite transmission and Y is what Earth received, $H(X|Y)$ is the number of bits necessary to fix the errors.

4 Information

How much information does Y give about X (and vice versa)? First, we state the **chain rule of entropy**:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

Rearranging the terms, we define the quantity

$$I(X;Y) \triangleq H(X) - H(X|Y) = H(Y) - H(Y|X)$$

as the **mutual information** between X and Y. Since $H(X|Y) \leq H(X)$, the mutual information is always non-negative. As an example, consider tossing 10 coins and letting X be the values of the first 7 coins and Y be the value of the last 5 coins. Then

$$H(X) = 7 \text{ and } H(Y) = 5$$
$$H(X|Y) = 5 \text{ and } H(Y|X) = 3$$
$$I(X;Y) = I(Y;X) = 2.$$