CSE533: Information Theory in Computer Science September 28, 2010

Lecture 1

Lecturer: Anup Rao *Scribe: Anup Rao*

1 Introduction

Information theory is the study of a broad variety of topics having to do with quantifying the amount of information carried by a random variable or collection of random variables, and reasoning about this information. In this course, we shall see a quick primer on basic concepts in information theory, before switching gears and getting a taste for a few domains in computer science where these concepts have been used to prove beautiful results.

2 How to measure information?

There are several ways in which one might try to measure the information carried by a random variable.

- We could measure how much space it takes to store the random variable, on average. For example, if X is uniformly random n-bit string, it takes n-bits of storage to write down the value of X. If X is such that there is a string a such that $\Pr[X = a] = 1/2$, we can save on the space by encoding X so that 0 represents a and all other strings are encoded with a leading 1. With this encoding, the expected amount of space we need to store X is at most $n/2 + 1$.
- We could measure how unpredictable X is in terms of what the probability of success is for an algorithm that tries to guess the value of X before it is sampled. If X is a uniformly random string, the best probability is 2^{-n}. If X is as in the second example, this probability is $1/2$.
- We could measure the expected length of the shortest computer program that prints out X (namely the expected Kolmogorov complexity of X). So for example, if X is random n-bit string which has only one 1, then the program needs merely needs to encode the location of the 1, which typically costs $\log n$ bits in program length.

There are just some of the options we have for measuring the information in X. Each of them sounds reasonable, and there are many other reasonable measures that make sense. Research in pseudorandomness has made gains by asking for analogous measures to those suggested above, under computational restrictions — i.e. what is the shortest encoding of X with an efficient encoding algorithm, or the best probability of predicting X using an efficient predictor algorithm. In this course, we shall focus on a few such measures, and explore applications of these ideas to problems in computer science.

3 Notation

Capital letters like X, Y, Z will be used to denote random variables, letters like S, T, U will denote sets. Calligraphic letters like $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ will be used to denote the supports of random variables (X, Y, Z). Small letters like x, y, z will be used to denote instantiations of random variables, and also elements of sets (s, t, u).

We shall use the shorthand $p(x)$ to denote the probability $\Pr[X = x]$, and $p(x, y)$ to denote $\Pr[X = x, Y = y]$. For conditional probability, we shall often use the notation $p(x|y)$ to denote $\Pr[X = x|Y = y]$.

4 The Entropy Function

The measure of information we shall focus on to start is the *entropy* function, defined as follows:

$$H(X) = \sum_x p(x) \log(1/p(x)),$$

where here we adopt the convention that $0 \log 1/0 = 0$ (which is justified by the fact that $\lim_{x \to 0} x \log(1/x) = 0$). Another way to interpret $H(X)$ is as the expected log of the probability of a sample from X,

$$H(X) = \mathbb{E}\left[\log(1/p(X))\right].$$

For $q \in [0,1]$, we shall write $H(q)$ to denote the entropy of a bit B for which $\Pr[B = 1] = q$.

Some facts are immediate from the definition:

- $H(X) \geq 0$ (since each term in the sum is non-negative). Moreover, $H(X) = 0$ if and only if $\Pr[X = a] = 1$ for some a, since otherwise, one of the terms in the sum will be strictly positive.
- $H(q) = H(1-q)$.
- $H(1/2) = 1$.

5 Axiomatic Definition

It is not clear that this is the only choice of measure of information we could have come up with, and there are others that make sense, but we shall see that this one is particularly useful. If one is so inclined, one can justify the choice by finding natural axioms that force us to pick H as the measure of entropy. Suppose we restricted ourselves to measures of information f that satisfied the following properties:

- If X is a uniformly random point from a set of size M, and Y is a uniformly random point from a set of size $M' > M$, then $f(M') > f(M)$.
- If X, Y are independent, then $f(X, Y) = f(X) + f(Y)$.
- If B_q is such that $\Pr[B_q = 1] = q$ and $\Pr[B_q = 0] = 1 - q$, then $f(B_q)$ is a continuous function of q.
- If B is a random variable taking 0/1 values, and X is another random variable, then $f(BX) = f(B) + \Pr[B = 1] \cdot f(X|B = 1) + \Pr[B = 0] \cdot f(X|B = 0)$.

Then you can show that $f(X)$ must be a scalar multiple of $H(X)$, but we leave that as an exercise (hard).

6 Some Examples

- If X is a uniformly random n-bit string, $H(X) = \sum_x p(x) \log(1/p(x)) = \sum_x 2^{-n} \log(2^n) = n$.
- If X is a uniformly random element of a set S, $H(X) = \sum_{x \in S} (1/|S|) \log(|S|) = \log(|S|)$.
- If X, Y, Z are uniformly random bits conditioned on their majority being 1, then $H(X, Y, Z) = 2$, since there are 8 3-bit strings, of which exactly 4 have majority 1. $H(X) = H(3/4)$.
- If X, Y, Z are uniformly random bits conditioned on their parity being 0 (i.e. $X + Y + Z = 0 \mod 2$), then $H(X, Y, Z) = 2$, since again, the fraction of such strings is 1/2. $H(X) = 1$, and $H(X, Y) = 2$. In this case, we see that $H(X, Y) = H(X, Y, Z)$, indicating that the first two bits already contain all information about the entire string. We shall formalize a way to measure how much information is left over in the last bit soon.

7 Conditional Entropy

Let X, Y be two random variables. Then, expanding $H(X, Y)$ gives

$$\begin{aligned}
H(X,Y) &= \sum_{x,y} p(x,y) \log\left(\frac{1}{p(x,y)}\right) \\
&= \sum_{x,y} p(x,y) \log\left(\frac{1}{p(x)p(y|x)}\right) \\
&= \sum_{x,y} p(x,y) \log\left(\frac{1}{p(x)}\right) + \sum_{x,y} p(x,y) \log\left(\frac{1}{p(y|x)}\right) \\
&= \sum_{x} p(x) \log\left(\frac{1}{p(x)}\right)\left(\sum_y p(y|x)\right) + \sum_x p(x) \sum_y p(y|x) \log\left(\frac{1}{p(y|x)}\right) \\
&= H(X) + \sum_x p(x) H(Y|X=x)
\end{aligned}$$

We denote the second term above $H(Y|X)$. It is the expected entropy that is left in Y after fixing X. In this notation, we have just showed the *chain rule*:

Lemma 1 (Chain Rule)**.** $H(X,Y) = H(X) + H(Y|X)$

Repeated applications of this two variable chain rule give:

Lemma 2 (Chain Rule)**.** $H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_n|X_{n-1}, \ldots, X_1)$.

Revisiting our examples, we see that in the case that X, Y, Z are three random bits conditioned on the event that their parity is 0, we have that $H(X) = 1$, $H(Y|X) = 1$ and $H(Z|X,Y) = 0$, where the last equation states that for every fixing of X, Y, the value of Z is determined. On the other hand, if X, Y, Z are random bits whose majority is 1, observe that when $X = 1, Y = 0$, the last bit has some entropy. Therefore $H(Z|X,Y) > 0$, which must mean that $H(X,Y) < 2$, since $H(Z|X,Y) + H(X,Y) = H(X,Y,Z) = 2$.

8 Jensen's Inequality and Subadditivity

(After a brief discussion, we were unable to determine whether "Jensen" is pronounced "Yensen" or otherwise).

Definition 3. *We say that a function $f : (a,b) \to \mathbb{R}$ is* convex *if for every $x, y \in (a,b)$ and every $\lambda \in (0,1)$,*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

Examples of convex functions include x, e^x, x^2 and $\log(1/x)$. If $-f$ is convex, we shall say that f is *concave.*

The following is a useful inequality for dealing with the entropy function and its derivatives:

Lemma 4 (Jensen's Inequality)**.** *If f is a convex function on (a,b) and X is a random variable taking values in (a,b), then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Proof We prove the case when X takes on finitely many values. The general case follows by continuity arguments.

We prove the statement by induction on the number of elements in the support of X. If X is supported on 2 elements, the lemma immediately follows from the definition of convexity. In the general case, let us assume X is supported on $x_1, \dots, x_n$. Then,

$$\begin{aligned}
\mathbb{E}\left[f(X)\right] &= p(x_1)f(x_1) + \sum_{i=2}^{n} p(x_i)f(x_i) \\
&= p(x_1)f(x_1) + (1 - p(x_1) \sum_{i=2}^{n} p(x_i)f(x_i)/(1 - p(x_1)) \\
&\leq p(x_1)f(x_1) + (1 - p(x_1) f\left(\sum_{i=2}^{n} p(x_i)x_i/(1 - p(x_1))\right) \\
&\leq f\left(p(x_1)x_1 + (1 - p(x_1)\left(\sum_{i=2}^{n} p(x_i)x_i/(1 - p(x_1))\right)\right) \\
&= f(\mathbb{E}\left[X\right]),
\end{aligned}$$

where the first inequality follows by applying the lemma for the case when there are $n - 1$ elements in the support, and the second inequality is a direct consequence of the definition of convexity. ■

As a first application of this inequality, we show the following lemma:

Lemma 5 (Subadditivity of Entropy)**.** $H(X, Y) \leq H(X) + H(Y)$

Proof

$$\begin{aligned}
H(X, Y) - H(X) - H(Y) &= \sum_{x,y} p(x, y) \log(1/p(x, y)) - \sum_{x} p(x) \log(1/p(x)) - \sum_{y} p(y) \log(1/p(y)) \\
&= \sum_{x,y} p(x, y) \log(1/p(x, y)) - \sum_{x,y} p(x, y) \log(1/p(x)) - \sum_{x,y} p(x, y) \log(1/p(y)) \\
&= \sum_{x,y} p(x, y) \log(p(x)p(y)/p(x, y)) \\
&\leq \log\left(\sum_{x,y} p(x, y)p(x)p(y)/p(x, y)\right) \\
&= \log 1 = 0,
\end{aligned}$$

where the inequality follows from Jensen's inequality applied to the convex function $\log(1/x)$. ■

Note that the above lemma implies in particular that $H(X) + H(Y|X) \leq H(X) + H(Y)$, which means that $H(Y|X) \leq H(Y)$. In other words, conditioning can only reduce the entropy in a random variable on average.

Lemma 6. $H(Y|X) \leq H(Y)$

It is NOT true that $H(Y|X = x)$ is always smaller than $H(Y)$. Indeed, if X, Y, Z are three uniform bits conditioned on the majority being 1, we see that $H(X) = H(3/4) < 1$, yet $H(X|Y = 0, Z = 1) = 1$. However, the lemma shows that average fixings of Y, Z do reduce the entropy in X.