CS909 Topics In Algorithms	October 29,2003
Lecture 25	
Lecturer: Amit Kumar	*Scribe: Sharad Jain*

1 Introduction

We continue our discussion of information theory in this lecture. Recall that we derived an expression for the entropy of a random variable and proved two key lemmas in the previous lecture. In today's lecture we shall continue our discussion of the concept of entropy.

2 Information Theory

Theorem 25.1 The entropy of random variable X with a set of possible outcomes $\{x_1, x_2, \ldots, x_n\}$, having probabilities $\{p_1, p_2, \ldots, p_n\}$ is maximum when $p_1 = p_2 = \cdots = p_n = 1/n$.

Proof: Let A be $n \times n$ doubly stocastic matrix(i.e., $\sum_i A_{ij} = 1$ and $\sum_j A_{ij} = 1$) with $A_{i,j} = 1/n$ and B be the $n \times 1$ matrix with $B_i = p_i$, $i = 1, \ldots, n$. Notice that $A \cdot B = C$ where C is the $n \times 1$ matrix with $C_i = 1/n$, $i = 1 \ldots n$. From Lemma 2 in the previous lecture, we have,

$$-\sum_{i=1}^{n} C_i \ln C_i = \geq -\sum_{i=1}^{n} B_i \ln B_i \geq -\sum_{i=1}^{n} p_i \ln p_i$$

But $-\sum_{i=1}^{n} p_i \ln p_i$ is $H(X)$. So the maximum value of $H(X)$ is $-\sum_{i=1}^{n} C_i \ln C_i$, which holds when B is some re-ordering of C. This proves the theorem. $\square$

2.1 Joint Entropy of Random Variables

Let X and Y be two random variables where X takes values in the set $\{x_1, x_2, \ldots, x_n\}$ and Y takes values in the set $\{y_1, \ldots, y_n\}$. Let $p(x_i, y_j)$ be the joint probability distribution of X and Y. Let $p(x_i)$ and $p(y_j)$ be the probability distributions of X and Y respectively. The joint entropy of random variables X and Y, $H(X, Y)$ is defined as

$$H(X, Y) = -\sum_{x_i \epsilon X, y_j \epsilon Y} p(x_i, y_j) \ln p(x_i, y_j).$$

Theorem 25.2 $H(X) + H(Y) \geq H(X, Y)$ with equality holding iff X and Y are independent.

Proof:

$$\begin{aligned}
H(X)+H(Y) &= -\sum_{i=1}^{n} p(x_i)\ln p(x_i) - \sum_{j=1}^{m} p(y_j)\ln p(y_j) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\ln p(x_i) - \sum_{j=1}^{m}\sum_{i=1}^{n} p(x_i,y_j)\ln p(y_j) \\
&= \sum_{i,j} p(x_i,y_j)(\ln p(x_i)+\ln p(y_j)) \\
&= -\sum_{i,j} p(x_i,y_j)\ln p(x_i)p(y_j) \\
&\geq -\sum_{i,j} p(x_i,y_j)\ln p(x_i,y_j)
\end{aligned}$$

where the last inequality follows from lemma 1 of the previous lecture. But $-\sum_{i,j} p(x_i,y_j)\ln p(x_i,y_j)$ is $H(X,Y)$. This implies the theorem. □

Corollary 25.3 $H(X_1,X_2,\ldots,X_n) \leq H(X_1)+H(X_2)+\cdots+H(X_n)$

2.2 Conditional Entropy of Random Variables

Let X and Y be random variables as in the previous section. We want to define the notion of $H(X|Y)$, the conditional entropy of X given the random variable Y. Let us see how to define it intuitively. Fix a value y_j of Y. Suppose $Y=y_j$. Given this condition, the random variable X takes the value x_i with probability $p(X=x_i|Y=y_j)$. So it is natural to define

$$H(X|y_j) = -\sum_i p(X=x_i|Y=y_j)\ln p(X=x_i|Y=y_j).$$

Since Y takes the value y_j with probability $p(y_j)$ we can define $H(X|Y)$ as follows

$$\begin{aligned}
H(X|Y) &= \sum_j p(y_j)H(X|y_j) \\
&= -\sum_j p(y_j)\sum_i p(x_i|y_j)\ln p(x_i|y_j) \\
&= -\sum_{i,j} p(x_i,y_j)\ln p(x_i|y_j).
\end{aligned}$$

Lemma 25.4

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

Proof:

$$\begin{aligned}
H(X)+H(Y)-H(X,Y) &= H(X)-\sum_j p(y_j)\ln p(y_j) + \sum_i\sum_j p(x_i,y_j)\ln p(x_i,y_j) \\
&= H(X)-\sum_j\sum_i p(x_i,y_j)\ln p(y_j) + \sum_i\sum_j p(x_i,y_j)\ln p(x_i,y_j)
\end{aligned}$$

$$
\begin{aligned}
&= H(X) - \sum_j \sum_i p(x_i, y_j)(\ln p(y_j) - \ln p(x_i, y_j)) \\
&= H(X) - \sum_j \sum_i p(x_i, y_j)(\ln p(y_j)/p(x_i, y_j)) \\
&= H(X) + \sum_j \sum_i p(x_i|y_j)p(y_j)(\ln p(x_i|y_j)) \\
&= H(X) + \sum_j p(y_j) \sum_i p(x_i|y_j)(\ln p(x_i|y_j)) \\
&= H(X) + \sum_j p(y_j) H(X|y_j) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

□

Corollary 25.5

$$H(X|Y) \leq H(X)$$

Proof: Proof follows directly from the previous lemma and the previous theorem. □

2.3 The Noiseless Coding Theorem

Suppose we have a random variable X which can take n different values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$. Once the random variable takes a value, we want to convey this value by transmitting it to some other place. What are the number of bits needed to convey this information. Clearly, $\log n$ bits are enough because we can always encode the n possibilities in n bits. Can we do better sometimes ? For example, suppose there are 4 possibilities 0,1,2,3 with probabilities 0.7,0.2.0.05,0.05 respectively. Suppose we encode 0 by 0, 1 by 100, 2 by 101, and 3 by 111. We would require only 1.6 bits /symbol as against 2 bits. So what is the maximum compression possible? We will continue it in next lecture.