COMPSCI 650 Applied Information Theory Jan 19, 2016

Lecture 1

Instructor: Arya Mazumdar *Scribe: Arya Mazumdar*

1 What is news

First, news is generally considered to be something especially unusual. The journalism truism is that dog bites man is not a story, but man bites dog is. Thats not a judgment on whether dog bites matter; its a judgment about whats surprising[1].

2 Transmission of Information

When I speak English and make a few speling mistakes, you can stil understand what I mean!

Consider sending 1 bit of information over a channel that flips a bit with probability $p < \frac{1}{2}$. Such a channel is called a *binary symmetric channel (BSC)* and is depicted below.

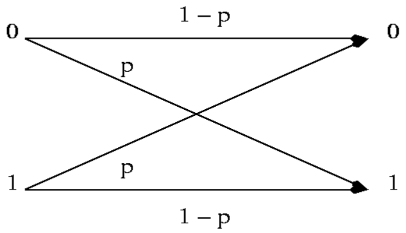

Figure 1: BSC

Now instead of sending this message uncoded, we add some redundancy. Let us repeat this bit three times. At the receiver, we choose the majority among the three bits. What is the probability of error?

If you think, erroneous reception happens only when two or three bits are received in error. That has probability

$$P_3 = p^3 + 3p^2(1-p).$$

We are interested in the case, when $P_3 < p$. That happens when:

$$p^3 + 3p^2(1-p) < p$$

or

$$p^2 + 3p - 3p^2 < 1$$

or

$$2p^2 - 3p + 1 > 0$$

[1] Washington Post, Nov 16, 2015.

or

$$(2p-1)(p-1) > 0$$

or

$$p < \frac{1}{2}.$$

Interesting. This is true by assumption. So we have reduced the probability of error by repeating a bit thrice. Is that a free lunch?

No, the reduction in the probability of error came at the cost of reducing the rate of transmission. The rate of transmission is $\frac{1}{3}$. By repeating n (an odd number) times, the probability of error will be,

$$P_n = \sum_{i=\frac{n+1}{2}}^{n} \binom{n}{i} p^i (1-p)^{n-i} \to 0,$$

as $n \to \infty$. Why? Because by repeating enough number of time, this number can be made to be smaller than any positive number. But rate of information transmission is $\frac{1}{n}$, also going to 0 as $n \to \infty$.

Shannon (1948) showed that there exist more clever way of adding redundancy to messages, such that probability of error goes to zero, but rate of information goes to a finite positive number called *capacity* of the communication channel. For BSC the capacity is $1 - h(p)$ where

$$h(p) = -p \log_2 p - (1-p) \log_2 (1-p),$$

is called the *binary entropy* function.

What? How did this happen?

3 Compression of music, movies and images

We sometime compress a file for storage. Everyone here used zip, compress or some other command (or have you not)? Before the invention of tape-recorders etc. people used *Shorthand.* How much can you compress data? You cannot surely compress it to zero length!

4 Intuitions about measure of information

Suppose an event A has probability $p(A)$. Occurrence of A may convey certain information to us. However, if $H(A)$ (abusing notation), is a measure that information, the first thing we notice is that $H(A)$ must be a function of $p(A)$ and not A itself.

Think about this. Also, if $p(A)$ is small then $H(A)$ should be large, is not it?

Moreover, consider an independent event B. What is the conveyed information if A and B both happens? Indeed, $H(A \wedge B)$ should equal to $H(A) + H(B)$. Indeed, since $H(A) \triangleq H(p(A))$ we must have that,

$$H(p(A) \cdot p(B)) = H(p(A)) + H(p(B)).$$

Can you recall a function that converts product to summation?

5 Entropy

Let us restrict ourselves to discrete random variable only. Let X be a random variable, that takes value in the alphabet $\mathcal{X}$. Let $p(x) \equiv P(X = x)$. The *entropy* of the random variable X is defined to be,

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x).$$

The unit of entropy depends on the base of the logarithm. If the base is 2, the entropy is measured in bits.

Note that $H(X) \geq 0$ because $0 \leq p(x) \leq 1$.

Also notice that for a Bernoulli(p) random variable, entropy is equal to the binary entropy function.

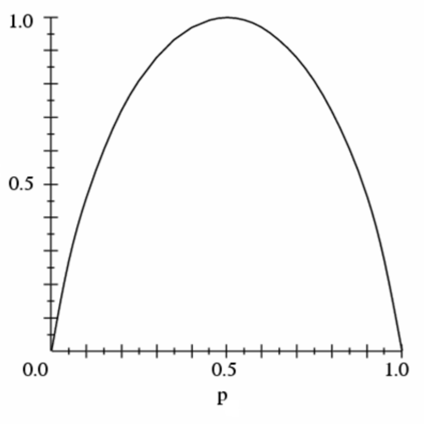

Figure 2: Binary Entropy Function

Example: Say $\mathcal{X} = \{1, 2, 3\}$ and $p(1) = p(2) = \frac{1}{4}$, $p(3) = \frac{1}{2}$. Then,

$$H(X) = 2 \cdot \frac{1}{4} \log_2 4 + \frac{1}{2} \log_2 2 = 1 + \frac{1}{2} = \frac{3}{2}.$$

6 Joint entropy and conditional entropy

The joint entropy $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as,

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y).$$

The conditional entropy of Y given $X = x$ is

$$H(Y|X = x) = -\sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x).$$

And the conditional entropy of Y given X is,

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x).$$

Chain Rule

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x)p(y|x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)[\log p(x) + \log p(y|x)] \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) + H(Y|X) \\
&= H(X) + H(Y|X).
\end{aligned}$$

7 Relative entropy

The relative entropy or KullbackLeibler distance between two probability mass functions $p(x)$ and $q(x)$ is defined as,

$$D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}.$$

You know what!

$$D(p\|q) \geq 0.$$

Always.