

Lecture 21

Instructor: Sruthi Sekar

Scribe: Geetesh Kini

1 Overview

This lecture covers some basic concepts in information theory such as **entropy**, **mutual information**, and their properties.

2 Quantifying Information bits

The first natural question that comes to mind when studying information theory is, how do we represent information mathematically?

Let's visualize this using an example:

Example 1. Consider the space of all possible 5 letter words (meaningless included). Now, if we are given the information "The word contains the letter s", it divides the space into 2 equal sets. We need to make a decision between equal sized sets (or Probability = 1/2) whether it has s or it doesn't, so we call it one bit of information.

If we are further told that "The word doesn't have t", we are further splitting these sets into sets with t and without t. So we need to make 2 decisions between equal sized sets, whether it has t or not and whether it has s or not. And the final set has a probability of 1/4.

So, can we see a relation between information bits and the probability of the final set that we get using the information?

Yes! The answer is simply

Definition 2 (Information Bits).

$$I = \log_2\left(\frac{1}{p}\right) \quad (1)$$

Where I is the number of bits of information and P is the probability of the event about which we have received information.

Notice that if an event is less likely, the occurrence of that event gives us more information compared to an event that is more likely, because it reduces the size of the final set more. So we have to make less decisions in future. And the formula says so too!

3 Shannon Entropy

Definition 3 (Shannon Entropy). We define the Shannon Entropy of a Random Variable X with pmf P(x) as the Expected value of the information bits, that is

$$H(X) = \sum_x P(x)I(x) \quad (2)$$

3.1 Huffman Encoding

Shannon entropy of a particular RV gives us the min number of bits required to encode the RV.

Example 4. Consider a RV as the colour a randomly picked ball from a bucket, we know its probability distribution as:

COLOR	PROBABILITY
Red	1/2
Green	1/4
Blue	1/8
Yellow	1/8

Figure 1: Color vs Probability Table

Now, we want to encode these colours in the most efficient possible way such that the expected value of number of bits is minimum. To do that we assign the highest probability colour, the lowest number of bits, so the encoding is:

COLOR	ENCODING SCHEME
Red	0
Green	10
Blue	110
Yellow	111

Figure 2: Huffman Encoding

Not that the number of bits assigned to each colour is same as its information bits as defined in definition 2. This makes the expected value of number of bits exactly equal to the Entropy = 1.75 and it turns out that this is the min expected value of number of bits required to represent the RV.

Here, the Probability values were powers of 0.5, as I just wanted to give an intuitive example. In general, this works too and the min number of expected value of bits is ceiling of the Entropy.

4 Binary Entropy Function

We encounter the Binary RV a lot , and hence, we study the Entropy function for it and it's properties in some detail.

Definition 5 (Binary RV). A Random Variable is Binary iff it takes the value 1 with some probability p and 0 with probability $1-p$.

Definition 6 (Binary Entropy Function). The Entropy function of a Binary RV wrt p is called the Binary Entropy function.

So it's easy to derive that:

$$H(p) = p \log\left(\frac{1}{p}\right) + (1-p) \log\left(\frac{1}{1-p}\right) \quad (3)$$

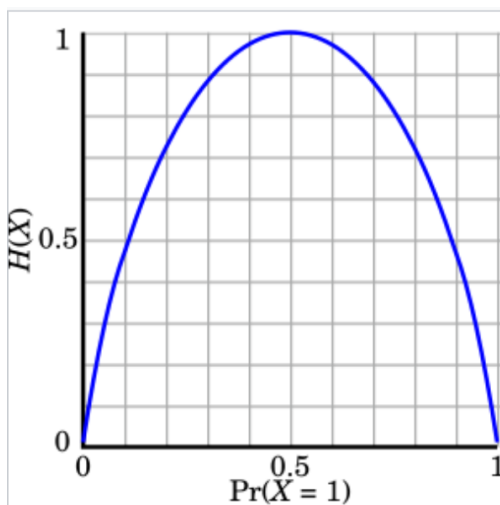


Figure 3: $H(p)$

Note that this graph is:

- Convex function of p
- $H(p) = 0$ only when $p = 0$ or 1
- Max when $p = 0.5$, $H(0.5) = 1$

5 Properties of $H(X)$

- $H(X) \geq 0$
- $|range(X)| \leq N$ implies $H(X) \leq \log_2 N$
 - a) $H(X) = 0$ iff X is constant
 - b) $H(X) = \log_2 N$ if X is uniform
- If X, Y are independent RVs, then $H(XY) = H(X) + H(Y)$
Proof: The joint entropy of X and Y is defined as:

$$H(XY) = - \sum_x \sum_y P(x, y) \log P(x, y)$$

If X and Y are independent, then:

$$P(x, y) = P(x)P(y)$$

Substituting this into the definition of joint entropy:

$$\begin{aligned}
H(XY) &= - \sum_x \sum_y P(x, y) \log P(x, y) \\
&= - \sum_x \sum_y P(x)P(y) \log (P(x)P(y)) \\
&= - \sum_x \sum_y P(x)P(y) [\log P(x) + \log P(y)] \\
&= - \sum_x \sum_y P(x)P(y) \log P(x) - \sum_x \sum_y P(x)P(y) \log P(y)
\end{aligned}$$

We now use the fact that $\sum_y P(y) = 1$ and $\sum_x P(x) = 1$:

$$\begin{aligned}
H(XY) &= - \sum_x P(x) \log P(x) \sum_y P(y) - \sum_y P(y) \log P(y) \sum_x P(x) \\
&= - \sum_x P(x) \log P(x) - \sum_y P(y) \log P(y) \\
&= H(X) + H(Y)
\end{aligned}$$

Therefore, for independent random variables X and Y :

$$H(XY) = H(X) + H(Y)$$

6 Mutual Information

Definition 7 (Mutual Information). We define Mutual information between 2 RVs X and Y as

$$I(X; Y) = H(X) + H(Y) - H(XY) \quad (4)$$

Expanding the equation, we get

$$I(X; Y) = \sum_x \sum_y P(x, y) \log \left(\frac{P(x, y)}{P(x)P(y)} \right)$$

This measures how much knowing one variable reduces the uncertainty about the other.

Example 8. Consider the following example to understand this better:

$Y \backslash X$	♥	♦	♠
♥	$\frac{1}{2}$	0	0
1	0	$\frac{1}{16}$	$\frac{1}{16}$
2	0	$\frac{1}{16}$	$\frac{1}{16}$
3	0	$\frac{1}{16}$	$\frac{1}{16}$
4	0	$\frac{1}{16}$	$\frac{1}{16}$

Their Marginal Distributions are as follows

X	Value	$P(X)$
♥		$\frac{1}{2}$
♦		$\frac{1}{4}$
♠		$\frac{1}{4}$

Y	Value	$P(Y)$
	♥	$\frac{1}{2}$
	1	$\frac{1}{8}$
	2	$\frac{1}{8}$
	3	$\frac{1}{8}$
	4	$\frac{1}{8}$

Hence, calculating the entropies, we get $H(X) = 1.5$, $H(Y) = 2$, $H(XY) = 2.5$,

This gives $I(X;Y) = 1$

Now, clearly, the mutual dependence of X and Y is like: If X is hearts, Y is hearts and vice versa, and if they're not hearts, they're independent and they're hearts with probability 0.5. Hence, it's 1 bit of mutual information.

6.1 Properties of Mutual Information

- $I(X;Y) = 0$ iff X and Y are independent.
- $I(X;Y) \geq 0$

Exercise 9. Prove that $I(X;Y) \geq 0$

Proof. Consider the negative mutual information:

$$-I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x,y)} = \mathbb{E}_{p(x,y)} \left[\log \frac{p(x)p(y)}{p(x,y)} \right]$$

Since $\log$ is a concave function, by Jensen's inequality:

$$\mathbb{E}[\log Z] \leq \log \mathbb{E}[Z]$$

for any random variable Z . Let $Z = \frac{p(x)p(y)}{p(x,y)}$, then:

$$\mathbb{E}_{p(x,y)} \left[\log \frac{p(x)p(y)}{p(x,y)} \right] \leq \log \left(\mathbb{E}_{p(x,y)} \left[\frac{p(x)p(y)}{p(x,y)} \right] \right)$$

Compute the expectation inside the logarithm:

$$\mathbb{E}_{p(x,y)} \left[\frac{p(x)p(y)}{p(x,y)} \right] = \sum_{x,y} p(x,y) \cdot \frac{p(x)p(y)}{p(x,y)} = \sum_{x,y} p(x)p(y) = 1$$

Thus:

$$-I(X;Y) \leq \log(1) = 0$$

which implies:

$$I(X;Y) \geq 0$$

Equality occurs when $\frac{p(x)p(y)}{p(x,y)} = c$ almost surely for some constant c . Since:

$$\sum_{x,y} p(x,y) = \sum_{x,y} p(x)p(y) = 1$$

we must have $c = 1$, meaning $p(x,y) = p(x)p(y)$ for all x, y . This is precisely the definition of X and Y being independent. $\square$

- $I(X;X) = H(X)$
- $I(X;Y) = I(Y;X)$
- Conditional Entropy

Definition 10 (Conditional Entropy).

$$H(Y|X) = \mathbb{E}_{x \in X} [H(Y|X = x)] \quad (5)$$

- a) $H(Y|X) = H(Y) - I(X;Y) \geq 0$
- b) $H(X|Y) = H(X) - I(X;Y) \geq 0$

- Chain Rule:

- a) $H(XY) = H(X) + H(Y|X)$
- b) $H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2|X_1) + \dots H(X_n|X_1, X_2, X_3, \dots, X_{n-1})$

The following venn diagram helps us in remembering these formulas:

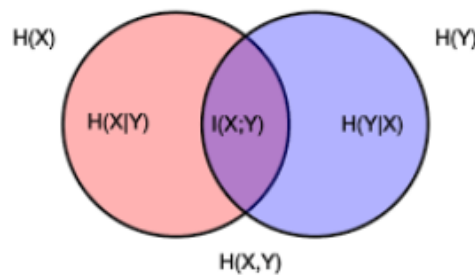


Figure 4: Venn Diagram

- Conditional mutual information

Definition 11 (Conditional mutual information).

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(XY|Z) = \mathbb{E}_{z \in Z} [I(X|Z = z, Y|Z = z)] \quad (6)$$

Example 12. X, Y are independent random bits

$$Z = X \oplus Y$$

$$I(X;Y) = 0$$

$I(X;Y|Z) = 1$ since, given Z , we know if $X = Y$ or not

7 Application in Information Complexity

Recall the Distributional Communication Complexity setting: Alice and Bob have received n bit inputs $x \in X$ and $y \in Y$ respectively according to a given distribution $\mu_{X,Y}$ and they're supposed to find $f(x,y)$ by communicating the least number of bits possible.

T_μ is the transcript they use for communication

$D_{\mu_{X,Y}}^\varepsilon$: min communication cost over all deterministic protocols π such that $err_\mu(\pi) = Pr_{x,y \in \mu_{X,Y}} [\pi(x,y) \neq f(x,y)] \leq \varepsilon$

Now one possible solution is Alice sends complete information of X to Bob, this will take at least $H(X)$ bits and can be done within $H(X)+1$ bits on an average.

Definition 13 (Information Complexity of π under μ).

$$IC_{\mu}(\pi) = I(T_{\pi}; Y|X) + I(T_{\pi}; X|Y) \quad (7)$$

where $I(T_{\pi}; Y|X)$ represents the amount of info the conversation teaches Alice about Bob's input and vice versa

Definition 14 (Information Complexity of f under μ).

$$IC_{\varepsilon}^{\mu}(f) = \min_{\pi} \{IC_{\mu}(\pi) : err_{\mu}(\pi) \leq \varepsilon\} \quad (8)$$

where $I(T_{\pi}; Y|X)$ represents the amount of info the conversation teaches Alice about Bob's input and vice versa

Fact:

$$D_{\varepsilon}^{\mu} \geq IC_{\varepsilon}^{\mu}(f)$$

... to be continued