CAS CS 591 B: Communication Complexity

Prof. Mark Bun Fall 2019

Lecture Notes 6:

Information Theory: Entropy, Mutual Information

Reading.

- Rao-Yehudayoff Chapter 6

We are starting a unit on information complexity, which is a framework for understanding communication complexity using information theory. While the scope as since broadened, Shannon introduced information theory to understand how well efficiently messages can be compressed while still being accurately transmitted. The objectives of information theory are are not quite the same as those in communication complexity, but there are nevertheless deep connections and surprisingly strong results about communication that one can prove using information.

In this lecture, we won't quite get to the relationship between information and communication, as we'll first have to get through the basic definitions and important concepts from information theory itself.

1 Entropy

In the basic information transmission problem, Alice is given a random string $A \in \{0,1\}^n$ and wishes to transmit A to Bob (who knows the distribution of A, but not its realization) using as short a message as possible. Let's look at some examples:

Example 1. If A is uniformly distributed over S for some subset $S \subset X$, then Alice can encode A using only $\log |S| \leq n$ bits.

Example 2. If A is concentrated at a single point, e.g, $A = 0^n$ deterministically, then Alice doesn't need to send anything to Bob at all.

Example 3. Suppose $A = 0^n$ with probability $1 - \varepsilon$ and A is uniformly random otherwise. We can't hope to encode A using fewer than n bits in the worst case, but on average we can do better. In particular, let us encode 0^n using the string 0 and every other $x \in \{0,1\}^n$ using the string $1x$. Then the expected length of this encoding is at most

$$(1 - \varepsilon) \cdot 1 + \varepsilon \cdot (n + 1) \approx \varepsilon n.$$

Intuitively, an efficient encoding should aim to assign shorter strings to inputs that appear with higher probability. Let p be the probability mass function of A. Let us sort the elements of X so that $p(x_1) \geq p(x_2) \geq \cdots \geq p(x_{2^n})$, and think of encoding element x_i using the integer i. We observe that for every i,

$$1 \geq p(x_1) + \cdots + p(x_i) \geq ip(x_i),$$

and hence $i \leq 1/p(x_i)$. So each x_i is encoded using about $\log i \leq \log 1/p(x_i)$ bits, and hence the expected length of an encoding is about

$$\mathbb{E}_A \left[\log \frac{1}{p(A)} \right].$$

This motivates the definition of entropy:

Definition 4. Let A be a random variable over a discrete space X with probability mass function p. Then the (Shannon) entropy of A is

$$H(A) = \mathbb{E}_A \left[\log \frac{1}{p(A)} \right] = \sum_{x \in X : p(x) > 0} p(x) \log \frac{1}{p(x)}.$$

Via Huffman codes, entropy gives a complete characterization of the expected message length needed to transmit any random variable.

Definition 5 (Huffman Coding Theorem)**.** Every random variable has an encoding with expected length at most $H(A) + 1$. Moreover, every encoding has expected length at least $H(X)$.

Huffman's Theorem (and the related Shannon Source Coding Theorem) show that entropy can be thought of as the inherent amount of information contained in a random variable. Let's record some basic mathematical properties of entropy.

- $H(A) \geq 0$ for every random variable A and $H(A) = 0$ iff A is a point mass.
- If A is uniform over a set S, then

$$H(A) = \mathbb{E}_A \left[\log |S| \right] = \log |S|.$$

- Subadditivity: If A and B are jointly distributed random variables, then $H(AB) \leq H(A) + H(B)$. (Here and elsewhere in the information-theory literature, AB denotes the joint random variable (A, B), and not the product of A and B.)

  Proof. We will use concavity of the log function to show that $H(A) + H(B) - H(AB) \geq 0$. Let $p(x, y)$ be the PMF of AB and for convenience let $p(x) = \Pr[A = x] = \sum_{y \sim B} [p(x, y)]$ and $p(y) = \Pr[B = y] = \sum_{x \sim A} [p(x, y)]$. Then

$$\begin{aligned}
H(AB) - H(A) + H(B) &= \sum_{x,y} p(x,y) \log \frac{1}{p(x,y)} - \sum_x p(x) \log \frac{1}{p(x)} - \sum_y p(y) \log \frac{1}{p(y)} \\
&= \sum_{x,y} p(x,y) \left(\log \frac{1}{p(x,y)} - \log \frac{1}{p(x)} - \log \frac{1}{p(y)} \right) \\
&= \sum_{x,y} p(x,y) \log \left(\frac{p(x)p(y)}{p(x,y)} \right) \\
&\leq \log \sum_{x,y} p(x,y) \left(\frac{p(x)p(y)}{p(x,y)} \right) \\
&= \log 1 = 0.
\end{aligned}$$

  The inequality here is an application of Jensen's Inequality: For any real-valued random variable Z and any concave function f, we have $\mathbb{E}[f(Z)] \leq f(\mathbb{E}[Z])$. □

Example 6. Suppose A, B, C are uniformly random bits conditioned on $A \oplus B \oplus C = 0$. Then $H(A) = 1$, $H(AB) = 2$, and $H(ABC) = 2$. Since C is completely determined by A and B, the random variable ABC carries no additional information over AB.

2 Conditional Entropy

Let us revisit the information transmission problem, but suppose Alice is given A and Bob is given B, where A and B are possibly correlated random variables. Can Alice use the fact that Bob knows B to save on the cost of transmitting A? If A and B are independent, then the answer is no. But revisiting the above example, if A and B were, say, uniform bits conditioned on $A \oplus B = 0$, then B completely determines A, and hence Alice doesn't need to send anything. (Even though A itself has positive entropy.)

The notion of conditional entropy allows us to formulate how much information the random variable A still contains after Bob has observed B:

Definition 7. Let A and B be jointly distributed random variables. Then the conditional entropy of A given B is

$$H(A|B) = \mathbb{E}_{y \sim B} H(A|B = y).$$

Theorem 8 (Chain Rule)**.** *For every A, B, we have*

$$H(AB) = H(B) + H(A|B).$$

Proof. We compute using Bayes' rule

$$\begin{aligned} H(AB) &= \sum_{x,y} p(x,y) \log \left(\frac{1}{p(x,y)} \right) \\ &= \sum_{x,y} p(x,y) \log \left(\frac{1}{p(y) \cdot p(x|y)} \right) \\ &= \sum_{y} p(y) \log \frac{1}{p(y)} + \sum_{y} p(y) \sum_{x} p(x|y) \log \frac{1}{p(x|y)} \\ &= H(B) + H(A|B). \end{aligned}$$

□

The Chain Rule plus subadditivity of entropy implies that conditioning can only reduce entropy:

$$H(A|B) = H(AB) - H(B) \leq H(A) + H(B) - H(B) \leq H(A).$$

However, note that conditioning on a specific realization of B does not necessarily reduce entropy, i.e., $H(A|B = y)$ could be larger than $H(A)$. (Can you find a counterexample?)

3 Mutual Information

Conditional entropy tells us how much information is left in A once B is revealed. We can also ask how much information is *learned* about A when B is revealed. This quantity is captured by the notion of mutual information.

Definition 9. The mutual information between two random variables A, B is defined by

$$\begin{aligned} I(A;B) = I(B;A) &= H(A) - H(A|B) \\ &= H(B) - H(B|A) \\ &= H(A) + H(B) - H(AB). \end{aligned}$$

Let's record some properties of mutual information:

- $I(A;B) \geq 0$, since conditioning always reduces entropy.
- A and B are independent iff $I(A;B) = 0$.
- If $A = B$, then $I(A;B) = H(A) = H(B)$.
- Unlike entropy, mutual information can be either subadditive or superadditive. If A, B, C are uniform bits conditioned on being equal, then $I(AB;C) = 1 < 2 = I(A;C) + I(B;C)$. On the other hand, if A, B, C are uniform conditioned on $A \oplus B \oplus C = 0$, then $I(AB;C) = 1 > 0 = I(A;C) + I(B;C)$.

We can also define conditional mutual information:

Definition 10. The mutual information between two random variables A, B conditioned on a third random variable C is defined by

$$\begin{aligned} I(A;B|C) &= \mathbb{E}_{z \sim C} I(A;B|C=z) \\ &= H(A|C) - H(A|BC) \\ &= H(A|C) + H(B|C) - H(AB|C). \end{aligned}$$

Example 11. Unlike with entropy, mutual information can increase under conditioning. Returning to our example of A, B, C being uniform conditioned on $A \oplus B \oplus C = 0$, we have $I(A;B) = 0$ but $I(A;B|C) = 1$.

Mutual information also satisfies a chain rule:

Theorem 12. $I(AB;C) = I(B;C) + I(A;C|B)$

Proof.

$$\begin{aligned} I(AB;C) &= H(C) - H(C|AB) \\ &= H(C) - H(C|B) + H(C|B) - H(C|AB) \\ &= I(B;C) + I(A;C|B). \end{aligned}$$

□