Lecture 1: Entropy and Data Compression

There are two fundamentally different paradigms in the study of artificial intelligence which might be characterized as the difference between the intellectual paradigm embraced by Claude Shannon versus that embraced by Noam Chomsky. Shannon based his work on probability while Chomsky based his work on possibility. In a probabilistic world view we say that one understands a subject to the extent that one can accurate judge probabilities as in "Giving the job to Mary will probably upset John". In a possibility world view one understands a subject to the extent that one can distinguish the possible from the impossible. For example, "it is not possible to go faster than the speed of light" or "angular momentum is always be conserved". Although both of these paradigms are useful, we will focus here on the Shannon paradigm of probabilistic modeling.

We want to build systems that are able to accurately assign probabilities. Perhaps the most common way of quantitatively measuring a systems ability to judge probabilistic truth is Kullback-Leibler divergence. To understand this measure we begin with the basics of information theory.

The fundamental concepts of information theory can be motivated by the problem of data compression. Suppose that we have a countable set M of messages. Suppose that we want to transmit a sequence of b messages $m_1, m_2, \ldots, m_b$ where the messages m_i are drawn IID according to P. The main theorem we consider in this section is the following (stated somewhat informally).

> **Shannon's Source Coding Theorem** [Informal Version] In the limit as the block size goes to infinity the number of bits required per message in the block is exactly the entropy $H(P)$ of P defined as follows.

$$H(P) = \mathrm{E}_{\mathrm{m}\sim\mathrm{P}}\left[\log_2(1/\mathrm{P}(\mathrm{m}))\right]$$

As a simple example suppose that P is the uniform distribution on 2^k messages. In this case we have that $H(P) = k$. As another example suppose that P assigns nonzero weight to 2^k messages but half of the weight is on a single message. In that case we can use the bit string 0 to represent the

common message and codes of length $k+1$, each starting with the the bit 1, for all other messages. In this case the average number of bits per message is $1/2 + 1/2(k+1) = 1 + k/2$. In general if the distribution is non-uniform we get greater compression by assigning fewer bits to more common messages.

We now state the main theorem more precisely. For a countable set M of messages define a code for M to be an assignment $c(m)$ of a bit string (code word) for each $m \in M$ with the property that if $m \neq m'$ then $c(m) \neq c(m')$. We will also consider coding a long sequence of messages by coding each message individually. In order for this to work one must be able to tell when one code word ends and the next begins. This can be done provided that no two distinct codes have the property that one is a prefix of the other.

Definition 1 *A code is called* prefix free *if all code words are distinct and no code word is a prefix of any other code word.*

As an example of a prefix-free code we can consider the set of all null-terminated byte (or character) strings. In this case every code word is a certain byte string and hence the length of each code is a multiple of 8.

We can now state our main theorem more precisely. We now let $|c(m)|$ be the length of the code word $c(m)$ (the number of bits used in $c(m)$). We let $\mathcal{C}(M)$ be the set of all prefix-free codes on M. We let $C^*(P)$ be defined as follows.

$$C^*(P) = \inf_{c \in \mathcal{C}(M)} \mathrm{E}_{\mathrm{m} \sim \mathrm{P}} \left[|\mathrm{c}(\mathrm{m})| \right]$$

Let b be a positive integer block size. Let M^b be the set of all tuples $\langle m_1, \ldots, m_b \rangle$ with $m_i \in M$. Let P^b be the probability distribution on M^b where each m_i is selected independently with probability distribution P. (We say that the m_i are drawn "IID" for Independently Identically Distributed). The above definition implies the following.

$$C^*(P^b) = \inf_{c \in \mathcal{C}(M^b)} \mathrm{E}_{\langle m_1, \ldots, m_b \rangle \sim \mathrm{P}^\mathrm{b}} \left[|\mathrm{c}(\langle m_1, \ldots, m_b \rangle)| \right]$$

Note that a code for M^b is coding for b messages from M so to get the

number of bits per message we should divide the length of a code word by b. We can now state the main theorem as follows.

Theorem 1 (Shannon's Source Coding Theorem)

$$\lim_{b\to\infty} \frac{1}{b} C^*(P^b) = H(P)$$

To prove this theorem we start with the following.

Lemma 2 (Kraft Inequality) *For any prefix-free code c we have the following.*

$$\sum_{m\in M} 2^{-|c(m)|} \leq 1$$

Proof: Suppose we generate a bit string by repeatedly flipping an unbiased coin and stopping as soon as the bit string we have generated is a code word. We then have that the probability of stopping with code word $c(m)$ is exactly $2^{-|c(m)|}$. The Kraft inequality then follows form the fact that probabilities sum to 1 (and there can be some nonzero probability that we miss all the code words and never stop the process). ∎

Theorem 3 *Let ℓ be an assignment of a positive integer $\ell(m)$ to each $m \in M$ satisfying the Kraft inequality:*

$$\sum_{m\in M} 2^{-\ell(m)} \leq 1$$

For any such assignment of lengths there exists a prefix-free code c with $|c(m)| = \ell(m)$.

Proof: Arrange the messages m in a sequence m_1, m_2, m_3, ... of nondecreasing length according to the assignment ℓ, i.e., such that $\ell(m_{i+1}) \geq \ell(m_i)$. Pick code words in the order given subject to the constraint the selected code word can not have as a prefix any previously selected code word. We view a code word c as having probability mass $2^{-|c|}$. When a code word $c(m)$ is selected the amount of probability mass that becomes unavailable for future assignments is $2^{-\ell(m)}$. By the Kraft inequality the amount of probability

mass remaining must be sufficient for the remainder of the code words. Furthermore, when selecting a code word for m we can consider the remaining mass to be uniformly distributed among the remaining code words of length $\ell(m)$ and hence such a code word can always be selected. ∎

Theorem 4 *For probability distribution P on a countable set of messages M there exists a code c assigning code word $c(m)$ to each $m \in M$ satisfying the following.*

$$\mathrm{E_{m\sim P}}\left[|\mathrm{c(m)}|\right] \leq \mathrm{H(P)} + 1$$

Proof: Let $\ell(m)$ be $\lceil \log_2 1/P(m) \rceil$. We have the following.

$$\begin{aligned} \sum_{m\in M} 2^{-\ell(m)} &= \sum_{m\in M} 2^{-\lceil \log_2 1/P(m)\rceil} \\ &\leq \sum_{m\in M} P(m) \\ &= 1 \end{aligned}$$

Therefore by theorem **??** there exists a code c with $|c(m)| = \ell(m)$ and hence we have the following.

$$\begin{aligned} |c(m)| &\leq \log_2(1/P(m)) + 1 \\ \mathrm{E_{m\sim P}}\left[|\mathrm{c(m)}|\right] &\leq \mathrm{E_{m\sim P}}\left[\log_2(1/\mathrm{P(m)}) + 1\right] \\ &= \mathrm{E_{m\sim P}}\left[\log_2(1/\mathrm{P(m)})\right] + \mathrm{E_{m\sim P}}\left[1\right] \\ &= H(P) + 1 \end{aligned}$$

∎

Now consider the distribution P^b on the message blocks M^b. The expected number of bits per message in a message block using a Shannon code for message blocks is the following.

$$\begin{aligned} \frac{1}{b}\mathrm{E}_{\langle m_1,\ldots,m_b\rangle\sim \mathrm{P^b}}\left[|\mathrm{c}(\langle m_1,\ldots,m_b\rangle)|\right] &\leq 1/b(H(P^b)+1) \\ &= H(P) + 1/b \end{aligned}$$

This implies that for arbitrarily long blocks the number of bits per message can be made arbitrarily close to $H(P)$. In the next lecture we will prove that no code can achieve fewer bits per code word than $H(P)$.

1 Problems

1. Let the set of messages M be the positive integers 1, 2, 3, ... Suppose we pick an integer by flipping an unbiased coin and stopping as soon as we get the first heads. We then output the number of flips. This gives $P(i) = (1/2)^i$. Give a prefix-free code $c(i)$ with $|c(i)| = \log_2(1/P(i))$. What is the entropy of this distribution?