CS174 Fall 98: Lecture Note 8

Michael Jordan, March 2002.

In this lecture we discuss an important consequence of the law of large numbers—the ability to *compress* data sequences. We discuss the relationship between the law of large numbers, entropy and compression.

Entropy

Definition: Given a discrete random variable X with probability mass function $p(x)$, define the *entropy* of X as follows:

$$H(X) \triangleq -\mathrm{E}\left(\log p(X)\right) = -\sum_x p(x)\log p(x).$$

Note that $p(x)$ plays two roles in the definition of entropy; it is the quantity with respect to which we are averaging, and its logarithm is the quantity being averaged.

Entropy has an appealing concrete interpretation: *entropy is a lower bound on the expected code length.*

What do we mean by the "expected code length"? Let the random variable X take its values in a set $\mathcal{X}$, assumed finite for simplicity. A *code* is an invertible mapping $C : \mathcal{X} \to \mathcal{D}^*$, where $\mathcal{D}^*$ is the set of strings from an alphabet $\mathcal{D}$. Typically, $\mathcal{D} = \{0, 1\}$, in which case a code assigns a sequence of zeros and ones to each value x in the range of the random variable X.

Let $l(x)$ denote the length of the codeword $C(x)$. We can now define the *expected code length*:

$$L(C) \triangleq \mathrm{E}\left(l(X)\right) = \sum_x p(x)l(x).$$

Moreover, given this definition, we can state the following theorem:

Theorem 1 *$L(C) \geq H(X)$, for any code C.*

That is, however we choose to map observations of X to codewords, on average the length of our codewords will be no smaller than the lower bound given by the entropy. (Any particular codeword may of course be shorter than the entropy bound. The theorem is a statement about *average* code lengths).

We will not prove this theorem here, but instead we will describe a general procedure that allows us to design codes whose expected code lengths are arbitrarily close to entropy. Understanding this design procedure will provide some insight into the reason that entropy plays the role that it does as a gold standard for coding.

Entropy and the law of large numbers

Recall the (weak) law of large numbers, which we proved in Note 6:

Theorem 2 *Given a set of n iid random variables $\{X_i\}$ with finite second moment, we have:*

$$\frac{1}{n}\sum_{i=1}^{n} x_i \to \mathrm{E}\left(X\right) \quad \textit{in probability.}$$

That is, for any $\epsilon > 0$, we have:

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right\} \to 0,$$

as $n \to \infty$.

Now note a trivial but important fact: if X_i is a random variable, then $\log p(X_i)$ is also a random variable, and moreover if the set of random variables $\{X_i\}$ are iid, then the set of random variables $\{\log p(X_i)\}$ are also iid. Thus we can apply the weak law of large numbers to the set $\{\log p(X_i)\}$ as follows:

$$\begin{aligned}
-\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) &= -\frac{1}{n}\log\left(\prod_{i=1}^{n} p(x_i)\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i) \\
&\to -\mathrm{E}\left(\log p(X)\right) \quad \text{in probability} \\
&= H(X).
\end{aligned}$$

We see that the average of the negative log probability of a set of n iid observations approaches the entropy as n goes to infinity.

Typical sets

The relationship between the law of large numbers and the entropy is often described in terms of "typical sets." Given fixed values of ϵ and n, define the *typical set* $A_\epsilon^{(n)}$ as the set of all sequences $(x_1, x_2, \ldots, x_n)$ whose negative log probability is near the entropy:

$$H(X) - \epsilon \leq -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon.$$

Note that this is different from the set of the most probable sequences (e.g., for a Bernoulli random variable with $0 < p < 1$, the sequence of all ones is not in the typical set for large n).

The derivation that we presented above shows that the probability of the typical set goes to one as n goes to infinity.

Moreover, the probability of each member of the typical set is essentially the same (within ϵ of entropy). Thus the typical set is a set that has most of the probability and has a nearly uniform distribution on its members.

From these facts follows another important fact—the typical set is *small*:

$$|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}.$$

To prove this, we compute:

$$1 = \sum_{(x_1,\ldots,x_n)} p(x_1, \ldots, x_n)$$

$$
\begin{aligned}
&\geq \sum_{(x_1,\ldots,x_n)\in A_\epsilon^{(n)}} p(x_1,\ldots,x_n) \\
&\geq \sum_{(x_1,\ldots,x_n)\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \\
&= 2^{-n(H(X)+\epsilon)} A_\epsilon^{(n)}.
\end{aligned}
$$

Thus, entropy can be viewed as a measure of volume—it measures the size of a typical set.

Compression

The properties of the typical set imply that we can compress data.

To show this, let us consider the following (simple) coding mechanism:

- Order all sequences in the typical set, and give each sequence an index. Since the number of sequences is bounded above by $2^{n(H(X)+\epsilon)}$, we need no more than $n(H(X)+\epsilon)+1$ bits to do this.
- Order and index all sequences not in the typical set. We need no more than $n\log|\mathcal{X}|+1$ bits to do this.
- Add an initial flag bit of 1 to the sequences in the typical set and an initial flag bit of 0 to the sequences not in the typical set.

We now compute the average code length of our code. Let x^n denote the sequence $(x_1, x_2, \ldots, x_n)$. We have:

$$
\begin{aligned}
\mathrm{E}\left(l(C)\right) &= \sum_{x^n} p(x^n)l(x^n) \\
&= \sum_{x^n\in A_\epsilon^{(n)}} p(x^n)l(x^n) + \sum_{x^n\in A_\epsilon^{(n)c}} p(x^n)l(x^n) \\
&\leq \sum_{x^n\in A_\epsilon^{(n)}} p(x^n)[n(H+\epsilon)+2] + \sum_{x^n\in A_\epsilon^{(n)c}} p(x^n)(n\log|\mathcal{X}|+2) \\
&= \Pr\{A_\epsilon^{(n)}\}[n(H+\epsilon)+2] + \Pr\{A_\epsilon^{(n)c}\}(n\log|\mathcal{X}|+2) \\
&\leq n(H+\epsilon) + \epsilon n\log|\mathcal{X}| + 2 \\
&= n(H+\epsilon'),
\end{aligned}
$$

where $\epsilon' = \epsilon + \epsilon\log|\mathcal{X}| + \frac{2}{n}$ can be made arbitrarily small (choosing ϵ first and then n).

Thus we have proven the following theorem:

Theorem 3 *Let X^n be a sequence of n iid random variables having probability mass function $p(x)$. Let $\epsilon > 0$. Then there exists a code from observations x^n to binary strings such that:*

$$
\mathrm{E}\left(\left[\frac{1}{n}l(X^n)\right]\right) \leq H(X)+\epsilon,
$$

for n sufficently large.

Although this theorem establishes the possibility of coding arbitrarily close to entropy, it does not yield a practical code. (Why?) To obtain a practical code we turn block coding algorithms, in particular to Huffman codes and arithmetic codes. The former are described on the accompanying handout.

Reference:

Cover, T., and Thomas, J. (1991). *Elements of Information Theory.* New York: John Wiley and Sons.