AICC 2 - Bixio Rimoldi

Alp Ozen

Spring 2019

Contents

1 Week 2

1.1 Entropy

We begin by defining **entropy** as Shannon put it:

Definition 1. ***Entropy***
Note that this definition assumes base 2 aka. binary.

$$H(s) = -\sum_{s \in A} p(s) \log_2 p(s) \ \textit{A being our alpahabet aka. sample space}$$

and thus an equivalent definition is:

$$H(s) = E[-\log_2 p(s)]$$

And similarly, Shannon defines information as:

$$-\log_2 p(s)$$

For a random distribution we get:

Example 1.1.

$$\forall x \in A \ p(x) = \frac{1}{|A|}, \ -\log_2 p(s) = \log_2 |A|$$

Hence the entropy function $H(s) = E[\log_2 |A|] = \underbrace{\log_2 |A|}_{do\ the\ algebra}$

And now we present the **information theory inequality**

Definition 2. ***IT inequality***

$$\log_b r \leq (r-1) \log_b(e)$$

Proof. Given that

$$\ln(r) \leq (r-1)$$

and that

$$ln(r) = \frac{\log_b(r)}{\log_b(e)}$$

we are done. □

And now we present the Entropy bound theorem:

Theorem 1.1.

$$S \in A \ 0 \leq H(S) \leq \log |A|$$

Proof. We only show the RHS as the LHS is more or less trivial. Our goal is to show:

$$\begin{aligned}
\text{need to reach} H(s) - \log |A| &\leq 0 \\
E[-\log p(s)] - \log |A| & \\
&= E[\log \frac{1}{p(s)|A|}] \\
&= \sum_{s \in A} p(s)(\log \frac{1}{p(s)|A|}) \\
&\leq \underbrace{log(e) \sum [\frac{1}{|A|} - p(s)]}_{\text{using IT ineq.}} = 0
\end{aligned}$$

□

1.2 Source coding

A code is said to have a prefix if:

Definition 3. ***Prefix of a code***

For some sequence of characters $a_1 a_2 \dots a_n$ *and* $b_1 b_2 \dots b_m$ *with* $n \leq m$ *we have* $a_1 a_2 \dots a_n = b_1 b_2 \dots b_n$

A **prefix free code** also known as **instantaneouss code** is one that has no prefixes. And now we come to the important **Kraft-McMillan** result:

Theorem 1.2. *If a* $D - ary$ *code is uniquely decodable, then it satisfies:*

$$D^{-l_1} + \dots + D^{-l_m} \leq 1$$

Note that there are non-instantaneous codes that stil satisfy this inequality. By the same token, by the contrapositive, we have that if a code does not satisfy the inequality, then there exists no prefix-free version of it.

We now define the **average codeword length**

Definition 4. ***average codeword length***

$$L(S,R) = \sum_{s \in A} p_S(s) L(R(s)) \text{ where } L \text{ represents length}$$

Given this definition, another important result is:

Theorem 1.3. ***Lower bound and Upper bound for average optimal codeword length***

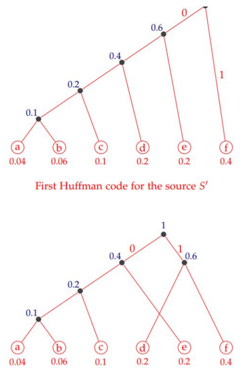
$$H_d(S) \leq L(S,R) \leq H_d(S) + 1$$

This result becomes a useful tool once we realize the similarity in the definitions as below:

$$H(S) = -\sum p(s) \log p(s)$$
$$L(S,R) = \sum p(s) L(R(s))$$

Given this, *Shannon - Fano* realized that we may define a code of length $\lceil \log_D p(s) \rceil$ This satisfies the Kraft inequality hence we now have a method of obtaining uniquely decodable code.

But as it turns out, Huffman was the first to actually find out how one finds an optimal code. We list our alphabet with probability in increasing order. Then, if say we are working in base 2, we simply continuously combine the smallest probabilities and build our branches from them. Hence, as below, we have that a Huffman code isn't always unique:



First Huffman code for the source S'

Second Huffman code for the source S'

Figure 1: Huffman codes

2 Week 3

We now define conditional entropy(the intuition for this being, we want to mesure uncertainty given knowledge of something else) as follows:

Definition 5.

$$H(X|Y) := -\sum p(x|y) \log p(x|y)$$

And the law of **total probability**

Theorem 2.1. *Imagine we take a sample space with some subset A and cut it into 3 disjoint units B_i. Now we may describe the set A as $A = (B_1 \cap A) \cup (B_2 \cap A) \cup (B_3 \cap A)$ This is equivalent to now saying:*

$$p(A) = p(B_1 \cap A) + p(B_2 \cap A) + p(B_3 \cap A)$$

Which in terms of conditional probability is

$$p(A) = p(A|B_1)p(B_1) + p(A|B_2)p(B_2) + p(A|B_3)p(B_3)$$

And another useful theorem is:

Theorem 2.2.

$$H(s_1, s_2, \ldots, s_n) \leq H(s_1) + H(s_2) + \ldots + H(s_n)$$

with equality iff the s_i are independent.

A similar result now is the chain rule of conditional entropy.

Theorem 2.3. ***Conditional entropy chain rule***

$$H(S_1, S_2, \ldots, S_n) = H(S_1) + H(S_2|S_1) + \ldots + H(S_n|S_1, \ldots, S_{n-1})$$

And we now introduce what it means for a source to be **regular**

Definition 6. ***Regular source***
A source is regular if

$$H(S) := \lim_{n\to\infty} H(S_n)$$
$$H^*(S) := \lim_{n\to\infty} H(S_n|S_1, S_2, \ldots, S_{n-1})$$

exist and are finite.

Now it should be intuitively obvious that conditioning would reduce entropy. Lets prove it.

Theorem 2.4.

$$H(X|Y) \leq H(X)$$

Proof.

$$\begin{aligned} E(\log \frac{1}{p(X|Y)}) &+ E(\log p(X)) \\ &= E(\log \frac{p(X)}{p(X|Y)}) \\ &= E(\log \frac{p(X)p(Y)}{p(X|Y)p(Y)}) \\ &\leq (\frac{p(X)p(Y)}{p(X \cap Y) - 1} \log(e) \leq 0 \end{aligned}$$

□

3 Useful links

Amazing YouTube playlist: YT Information Theory playlist