Theory of Information Measurement and Sauce Coding

Xinyao Ji

November 2, 2011

Information theory is a branch of applied mathematics which deals with the measurement and transmission of information through a channel. A typical model of communication involves three steps: (1) coding a message at its source, (2) transmitting the message through a channel, (3) decoding the message at its destination. In this talk, I will focus on the first step. I want to find a quantitative way to measure the disorder of an information sauce and a corresponding method to code it with the least uncertainty.

Let F be the set of sauce alphabets. A message is defined as a string of the elements in set A. let B be the set of code alphabets. Let G_B be the set of all strings of the elements in set B. A code is defined as a map *f* from F to G_B.

Definition1: a code is decipherable or unique decodable if every string in G_B is the image of one message.

Definition 2: A string x is a prefix in a string y if there exists a string z, such that xz=y.

A prefix-free code is decipherable because we can always find the first codeword in a message, peel it off, and continue decoding.

Theorem 1: If F is a set of sauce alphabets, B= {0,1}, f is a prefix-free code form F to G_B , for F= {X1, X2,…..Xn}, f(Xi)=Ci, Li=|Ci|, then we have the following inequality: $\sum 2^{-Li} \leqq 1$ (carft inequality)

In real information transmission, usually we don't know which sauce alphabet is under transmission. What we only know is the probability that an alphabet emerges at a certain time. Suppose F={X1, X2,…..Xn}, P(Xi)=p_i, and $\sum p_i=1$. To cut down the cost in transmission, we hope the length of a message is as short as possible. So we use the expectation E(X)= $\sum Lip_i$ as the stickyard. Our purpose is to minimize E(X) with the condition that $\sum 2^{-Li} \leqq 1$.

Let $q_i=2^{-Li}$. The question equals to maximizing $\sum p_i \log_2 q_i$, such that $\sum q_i=1$

Lemma: if $p_i \geqslant 0$, $q_i \geqslant 0$, $\sum p_i=\sum q_i=1$,then $\sum p_i \log_2(p_i/q_i) \geqq 0$

Thus, $\sum p_i \log_2(p_i) \geqslant \sum p_i \log_2(q_i)$, "=" holds iff $p_i= q_i$, $Li=\log_2(1/p_i)$

Remark: $E=-\sum p_i \log_2(p_i)$ is actually the definition of entropy in information theory. It's always non-negative. And it can be viewed as a measure of the average uncertainty associated with a random variable F.

The next question is how to use a string of 0,1 to represent Ci, such that Li= $\log_2(1/p_i)$.

The following process, called Huffman coding, provides a method to find Ci 's

F= {X1, X2,….. Xn}, P(Xi)=p_i, create n nodes corresponding to these n alphabets, probability of a alphabet is assigned to be equal to the node weight. Let set M be the collection of these n nodes.

Step1: Find and remove two nodes with smallest weights. Mark nodes A and B.

Step 2: Create a new node C. weight[C] = weight[A] + weight[B]. Create a subtree that has these

two nodes as leaves, C as root. Then insert C into the set M.
Repeat Step 1 and 2 until there is only node in the set M
For the whole binary tree, left edge is labeled 0; right edge is labeled 1. Path from root to leaf is codeword for the corresponding message
We can proof that a Huffman code is optimal.