Information theory: Prefix codes

Yuval Filmus

November 19, 2017

1. A *prefix code* is a mapping C from a finite non-empty set S to $\{0,1\}^*$ such that for $x \neq y$, $C(x)$ is not a prefix of $C(y)$. Prefix codes can be described as trees in which the edges are labeled with 0 or 1.

   If every internal node has exactly two children, we say that the code is *complete*.

   Our first goal is to prove Kraft's inequality: A prefix code with codeword lengths $\ell_1, \ldots, \ell_n$ exists if and only if

   $$\sum_{i=1}^{n} 2^{-\ell_i} \leq 1.$$

   In fact, a stronger result holds: a complete prefix code with these codeword lengths exists if and only if

   $$\sum_{i=1}^{n} 2^{-\ell_i} = 1.$$

   (a) Suppose that $x_1, x_2, \ldots$ is an infinite stream of independent uniformly random bits. Show that the probability that a word w is a prefix of the stream is $2^{-|w|}$.

   (b) Deduce that if $w_1, \ldots, w_n$ is a prefix code, then

   $$\sum_{i=1}^{n} 2^{-|w_i|} \leq 1.$$

   (c) Show that equality holds iff the code is complete.

   (d) Suppose that $\sum_{i=1}^{n} 2^{-\ell_i} = 1$. Show that if $n > 1$ then there exist $i < j$ such that $\ell_i = \ell_j$.

   (e) Assuming $\ell_i = \ell_j$, show how to construct a prefix code with codeword lengths $\ell_1, \ldots, \ell_n$ given a prefix code with codeword lengths $\ell_1, \ldots, \widehat{\ell_i}, \ldots, \widehat{\ell_j}, \ldots, \ell_n, \ell_i - 1$, where the hats denote that ℓ_i, ℓ_j are removed.

   (f) Show that if $\sum_{i=1}^{n} 2^{-\ell_i} = 1$ then there exists a prefix code with codeword lengths ℓ_i.

   (g) Show that if $\sum_{i=1}^{n} 2^{-\ell_i} < 1$ then there exists an index j such that $\sum_{i \neq j} 2^{-\ell_i} + 2^{-(\ell_j - 1)} \leq 1$.

   (h) Show how to construct a prefix code with codeword lengths $\ell_1, \ldots, \ell_n$ given a prefix code with codeword lengths $\ell_1, \ldots, \ell_{j-1}, \ell_j - 1, \ell_{j+1}, \ldots, \ell_n$.

   (i) Complete the proof of Kraft's inequality.

   Kraft's inequality can be generalized to the countably infinite case.

2. We now introduce entropy and describe one of its interpretations.
   The entropy of a discrete random variable X taking values in a finite (or countable) set R is
   $$H(X) = \sum_{i \in R} \Pr[X = i] \log \frac{1}{\Pr[X = i]}.$$
   The base of the logarithm depends on the context — we will use base 2.

   Let $T(X)$ be the minimum value of $\mathbb{E}[|C(X)|]$, where C goes over all prefix codes. (The optimal code C can be found in time $O(|R| \log |R|)$ using *Huffman's algorithm.*)

   (a) Consider the optimization problem "minimize $\sum_i p_i \ell_i$ subject to $\sum_i 2^{-\ell_i} = 1$", where $\ell_i \in \mathbb{R}$. Use Lagrange multipliers (or any other method) to show that the optimal solution is $\ell_i = \log(1/p_i)$, and that the optimal value is $H(X)$, where X is defined by $\Pr[X = i] = p_i$.

   (b) Deduce that $T(X) \geq H(X)$.

   (c) Considering $\ell_i = \lceil \log(1/p_i) \rceil$, show that $T(X) < H(X) + 1$. (This is known as *Shannon–Fano encoding.*)

   (d) Show that if $c < 1$ then the inequality "$T(X) < H(X) + c$" doesn't always hold.

   (e) Let $X_1, X_2, \ldots$ be independent copies of X. Show that
   $$H(X_1, \ldots, X_n) \leq T(X_1, \ldots, X_n) < H(X_1, \ldots, X_n) + 1,$$
   treating $X_1, \ldots, X_n$ as a single random variable ranging over R^n.

   (f) Show that if X, Y are independent then $H(X, Y) = H(X) + H(Y)$.

   (g) Deduce that
   $$H(X) = \lim_{n \to \infty} \frac{T(X_1, \ldots, X_n)}{n}.$$

3. The entropy of a Bernoulli p random variable is denoted by $h(p)$ (the "entropy function").

   (a) Give a formula for $h(p)$.

   (b) Show that $h(p) = h(1 - p)$.

   (c) Show that $h(p)$ is increasing for $p \leq 1/2$ and decreasing for $p \geq 1/2$.

   (d) Show that $h(p)$ is concave.

   (e) Estimate $h(1/2 + \epsilon)$.

   (f) Estimate $h(\epsilon)$.

Exercise Suppose that X is a random variable attaining at most n different values.

(a) Show that $T(X) \leq \lceil \log n \rceil$, and deduce that $H(X) < \log n + 1$.

(b) Use the formula $H(X) = \lim_{m \to \infty} T(X_1, \ldots, X_m)/m$ to show that $H(X) \leq \log n$.

(c) Can $H(X)$ equal $\log n$?

Challenge A *uniquely decodable code* is a function $C \colon S \to \{0,1\}^*$ whose extension $C^* \colon S^* \to \{0,1\}^*$ (defined by $C^*(s_1 \ldots s_m) = C(s_1) \ldots C(s_m)$) is one-to-one. Prove *McMillan's inequality*: a uniquely decodable code satisfies Kraft's inequality.