Оглавление

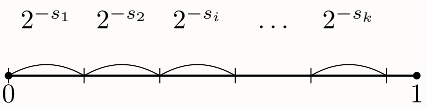
0.1 Задача о наилучших длинах кодов 2
0.2 Энтропия. 3

Лекция 7: Достаточность неравенства Крафта, задача о наилучших длинах кодов, энтропия.

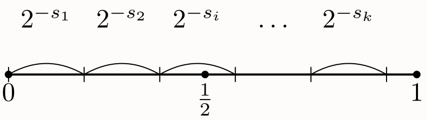
25.10.2023

Доказательство. (Достаточность)
Не умоляя общности, будем считать, что $s_1 \le s_2 \le \ldots \le s_k$. Тогда $2^{-s_1} \ge 2^{-s_2} \ge \ldots \ge 2^{-s_k}$

Теперь распложим эти значения на отрезке $[0, 1]$:



Утверждение: $\frac{1}{2}$ является границей отрезков. (не может лежать внутри какого либо 2^{-s_i})
Предположим, что это не так, тогда:



Имеем неравенства:

$$\begin{cases} 2^{-s_1} + \ldots + 2^{-s_{i-1}} \le \frac{1}{2} \\ 2^{-s_1} + \ldots + 2^{-s_{i-1}} + 2^{-s_i} > \frac{1}{2} \end{cases} \quad \text{домножим на } 2^{s_i} \Rightarrow$$

$$\Rightarrow \begin{cases} \underbrace{2^{s_i - s_1} + \ldots + 2^{s_i - s_{i-1}}}_{c} < 2^{s_i - 1} \\ \underbrace{2^{s_i - s_1} + \ldots + 2^{s_i - s_{i-1}}}_{c} + 1 > 2^{s_i - 1} \end{cases} \Rightarrow \begin{cases} c < 2^{s_i - 1} \\ c + 1 > 2^{s_i - 1} \end{cases} \quad \text{— противоречие.}$$

Таким образом какой-то отрезок из 2^{-s_i} упрется в $\frac{1}{2}$ или хотя бы не дойдет до нее. Тогда Разделим $2^{-s_1}, 2^{-s_2}, \ldots, 2^{-s_k}$ на 2 части: те, которые меньше $\frac{1}{2}$ и те, которые больше. Тем кодам, которые меньше $\frac{1}{2}$ поставим 0 в начало кода, а тем, которые больше $\frac{1}{2}$ поставим 1, тогда их длина уменьшилась на единицу. Тогда сумма тех, что слева и тех, что справа:

$$\sum 2^{-(s_i-1)} = \sum 2 \cdot 2^{-s_i} = 2 \cdot \sum 2^{-s_i} \leq 1$$

Тогда можем рекурсивно выполнять деление отрезков, т.к. если есть всего 2 символа, можем дать им коды 0 и 1 соответственно. □

0.1 Задача о наилучших длинах кодов

Пусть $A = \{c_1, \ldots, c_n\}$ — алфавит, $p_1, \ldots, p_n$ — вероятности появления символов. Пусть задан код $\varphi : A \to \{0, 1\}^*$
Фиксируем текст (большую строку) длины $N : a_1 a_2 \ldots a_N$. В этом тексте количество символов $c_i - N \cdot a_i$ (закон больших чисел). Тогда длина кодовой последовательности текста:

$$\sum_{i=1}^{n} N p_i \varphi(c_i)$$

Задача заключается в том, чтобы минимализировать длину кодовой последовательности такого текста, а т.к. N – фиксированная величина, не влияющая на коды, задача сводиться к тому, чтобы минимализировать сумму:

$$\sum_{i=1}^{n} p_i \varphi(c_i)$$

При этом для $p_i, \varphi(c_i)$ выполняются:

1. $\sum p_i = 1$
2. $0 < p_i < 1$
3. $\varphi(c_i) > 0, \varphi(c_i) \in \mathbb{Z}$
4. $\varphi(c_1), \ldots, \varphi(c_n)$ – длины кодов символов в префиксом коде, т.е. для этих длин выполняется неравенство Крафта.

Замечание. По правде говоря, из приведенного доказательства неравенства Крафта не следует, что код будет префиксным, но это можно вывести. Кто-нибудь умный скажет, как это сделать.

Теорема 1. Минимум достигается функцией:

$$H(p) = \sum_{i=1}^{n} \log_2 \frac{1}{p_i}$$

Доказательство. □

0.2 Энтропия.

Определение 1. Энтропией вероятностной схемы называется мера содержащейся в ней неопределенности. Она задается как конкретная функция $H : RS \to \mathbb{R}^+$, где RS – множество всех возможных вероятностных схем.

Функция, задающая энтропию обладает рядом свойств, и этим свойствам удовлетворяет функция 1. Это докажем позже, а сейчас рассмотрим свойства энтропии:

Свойства.

1. Мера неопределенности непрерывно зависит от вероятностей. (функция 1 этим свойством, очевидно, обладает)
2. При перестановке вероятностей мера неопределенности не меняется.
3. Необходимо ввести единицу измерения неопределенности. За единицу будем брать энтропию честной монеты: $H(\{\frac{1}{2}, \frac{1}{2}\}) = 1$ – бит.
4. Обозначим за $h(m) = H(\{\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\})$. Тогда $h(m)$ растет с ростом m. (функция 1 этим свойством тоже обладает)
5. При фиксированном m максимум энтропии достигается в случае равновероятных исходов, т.е. $h(m)$.
6. Пусть есть схемы $P_m = p_1, \ldots, p_m$ и $Q_k = q_1, \ldots, q_k$. Образуем комбинированную схему с $m-k+1$ исходами следующим образом: выбирается m-й исход в P_m и для него выбираются исходы из Q_k. Получим схему PQ с исходами:

$$1, 2, \ldots, m-1, (m, 1), (m, 2), \ldots, (m, k)$$

   Вероятность этих исходов:

$$p_1, \ldots, p_{m-1}, p_m q_1, \ldots, p_m q_k$$

   Тогда энтропия схемы $PQ : H(PQ) = H(P_k) + p_m H(Q_k)$

Доказательство. (Какие-то свойства либо уже доказаны, либо были даны без доказательства)

6

$$H(PQ) = \sum_{i=1}^{m-1} p_i \log_2 \frac{1}{p_i} + \sum_{j=1}^{k} p_m q_j \log_2 \frac{1}{p_m q_j} =$$

$$= \sum_{i=1}^{m-1} p_i \log_2 \frac{1}{p_i} + p_m \sum_{j=1}^{k} q_j \log_2 \frac{1}{p_m} + p_m \sum_{j=1}^{k} q_j \log_2 \frac{1}{q_j} =$$

$$= 1 \cdot p_m \log_2 \frac{1}{p_m} + \sum_{i=1}^{m-1} p_i \log_2 \frac{1}{p_i} + p_m \sum_{j=1}^{k} q_j \log_2 \frac{1}{q_j} = H(P_m) + p_m H(Q_k)$$

$\square$