Немного кодирования

Лемма Крафта

Рассмотрим источник, который генерирует сообщения из алфавита

$$\mathcal{A} = \{a_1, a_2, \ldots, a_n\}.$$

Кодер преобразует символы в последовательности кодовых знаков из алфавита

$$J = \{0, 1, \ldots, q-1\}.$$

Код — это отображение

$$f\colon \mathcal{A} \longrightarrow J^*,$$

Определение. *Код называется* ***однозначно декодируемым****, если продолжение f^* кода f на $\mathcal{A}^* \to J^*$, определяемое правилом:*

$$f^*(u_1, \ldots, u_n) = f(u_1) \ldots f(u_n)$$

является инъективным отображением[1]*.*

Далее под $l(a)$, где $a \in \mathcal{A}^*$ будем понимать длину кодового слова $|f^*(u)|$.

Определение. *Код называется* ***беспрефиксным****, если ни одно кодовое слово не является началом другого кодового слова, т.е. для любых $a_1, a_2 \in \mathcal{A}$ из условия $a_2 \neq a_2$ следует, что $f(a_1)$ не является префиксом $f(a_2)$ и наоборот.*

Беспрефиксные коды декодируются мгновенно: при чтении входного потока достаточно сразу определить конец текущего кодового слова без заглядывания вперёд.

Упражнение 1. Приведите пример кода, который однозначно декодируется, но не является беспрефиксным.

Решение. Возьмём алфавит $I = \{a, b, c\}$ и зададим код

$$f(a) = 1, \qquad f(b) = 101, \qquad f(c) = 11011.$$

Нетрудно проверить что такой код однозначно декодируется.

Лемма (Лемма Крафта)**.** *Для алфавита источника $\mathcal{A}$ размера n и алфавита кодировщика J размера q, однозначно-декодируемый код $f\colon \mathcal{A} \to J$ с длинами слов $l_1, \ldots, l_n$ существует тогда и только тогда, когда*

$$\sum_{i=1}^{n} 2^{-l_i} \leqslant 1$$

В дальнейшем будем полагать $q = 2$. Следующие четыре задачи доказывают необходимость условия Крафта.

Упражнение 2. Пусть $f\colon \mathcal{A} \to J^*$ — произвольный, $\ell = \max_{a \in \mathcal{A}} l(a)$. Докажите, что для любого $k \in \mathbb{N}$

$$\left(\sum_{a \in \mathcal{A}} 2^{-l(a)}\right)^k = \sum_{i=1}^{k\ell} \#\{u \in \mathcal{A}^k \mid l(u) = i\} \cdot 2^{-i}.$$

Решение. Распишем левую часть:

$$\left(\sum_{a \in \mathcal{A}} 2^{-l(a)}\right)^k = \left(\sum_{u_1 \in \mathcal{A}} 2^{-l(u_1)}\right) \cdots \left(\sum_{u_k \in \mathcal{A}} 2^{-l(u_k)}\right)$$

После раскрытия произведения получаем сумму по всем k-кратным упорядоченным наборам символов

$$\sum_{(u_1, \ldots, u_k) \in \mathcal{A}^k} 2^{-l(u_1)} \cdots 2^{-l(u_k)} = \sum_{(u_1, \ldots, u_k) \in \mathcal{A}^k} 2^{-\left(l(u_1) + \cdots + l(u_k)\right)} = \sum_{u \in \mathcal{A}^k} 2^{-l(u)}$$

[1] Для двух слов a и b запись ab обозначает их конкатенацию

Группируя слагаемые по одинаковым значения длины $i = l(u)$, получаем

$$\sum_{u \in \mathcal{A}^k} 2^{-l(u)} = \sum_{i=1}^{k\ell} \#\{u \in \mathcal{A}^k \mid l(u) = i\} \cdot 2^{-i}.$$

Что и требовалось доказать.

Упражнение 3. Пусть теперь f — однозначно-декодируемый код (необязательно беспрефиксный). Докажите, что для любого $i \in \mathbb{N}$ количество слов $u \in \mathcal{A}^*$, таких, что $l(u) = i$, не превышает 2^i:

$$\#\{u \in \mathcal{A}^* \mid l(u) = i\} \leq 2^i.$$

Решение. Каждое кодовое слово длиной i является некоторой битовой строкой из множества $J^i = \{0,1\}^i$. Поскольку в множестве J^i ровно 2^i различных строк, количество различных кодовых слов любой длины не может превышать это число, иначе нарушится инъективность f^*.

Упражнение 4. Пусть теперь f — однозначно-декодируемый код (необязательно беспрефиксный). Докажите, что для любого $i \in \mathbb{N}$ количество слов $u \in \mathcal{A}^*$, таких, что $l(u) = i$, не превышает 2^i:

$$\left(\sum_{a \in \mathcal{A}} 2^{-l(a)}\right)^k \leq kl$$

Решение. По предыдущим упражнениям:

$$\left(\sum_{a \in \mathcal{A}} 2^{-l(a)}\right)^k = \sum_{i=1}^{k\ell} \underbrace{\#\{u \in \mathcal{A}^k \mid l(u) = i\}}_{\leq 2^i} \cdot 2^{-i} \leq \sum_{i=1}^{k\ell} 2^i \cdot 2^{-i} = k\ell$$

Упражнение 5. Пусть f — однозначно-декодируемый код. Докажите, что

$$\sum_{a \in \mathcal{A}} 2^{-l(a)} \leq 1.$$

Решение. Возьмём произвольное $k \in \mathbb{N}$ и применим предыдущее упражнение:

$$\left(\sum_{a \in \mathcal{A}} 2^{-l(a)}\right)^k \leq k\ell$$

Отсюда

$$\sum_{a \in \mathcal{A}} 2^{-l(a)} \leq (k\ell)^{1/k}.$$

Поскольку ℓ фиксировано, предел правой части при $k \to \infty$ равен 1:

$$\lim_{k \to \infty} (k\ell)^{1/k} = 1.$$

Следовательно,

$$\sum_{a \in \mathcal{A}} 2^{-l(a)} \leq 1.$$

Средняя длина и энтропия

Будем теперь считать, что источник генерирует символ a_i с вероятностью p_i, (все $p_i > 0, \sum p_i = 1$).

Определение. ***Средняя длина*** *кода f с длинами $l_i = l(a_i)$ определяется как*

$$L(C^f) = \sum_{i=1}^{n} p_i l_i.$$

Теорема (Теорема Шенонна)**.** *Пусть f — однозначно-декодируемый код, имеющий наименьшую среднюю длину среди всех однозначно-декодируемых кодов. Тогда*

$$\mathrm{H}_2(p_1, \ldots, p_n) \leq L(C^f) \leq \mathrm{H}_2(p_1, \ldots, p_n) + 1$$

Упражнение 6. Докажите, что для любого однозначно-декодируемого кода выполняется неравенство:

$$L(C^f) \geq \mathrm{H}_2(p_1, \ldots, p_n),$$

Решение. Чтобы оценить нижнее значение, найдем минимум функции $L(C^f)$. Рассмотрим задачу минимизации средней длины при фиксированных вероятностях p_i и выполнении условия Крафта

$$\sum_{i=1}^{n} p_i l_i \to \min$$
$$\sum_{i=1}^{n} 2^{-l_i} \leq 1, \qquad l_i \in \mathbb{N}.$$

Два замечания:

- Условие целочисленности l_i можно ослабить — это только ухудшит оценку минимума.
- Точка минимума находится на границе множества $\sum_{i=1}^{n} 2^{-l_i} \leq 1$, т.е. когда

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

Поскольку, если бы точка минимума находилась внутри, то можно было бы уменьшить одно из l_i на достаточно малое число — это привело бы к тому что неравенство Крафта все еще сохранялось, но средняя длина уменьшилась.

Итак, свелись к следующей задаче:

$$\sum_{i=1}^{n} p_i l_i \to \min$$
$$\sum_{i=1}^{n} 2^{-l_i} = 1, \qquad l_i \in \mathbb{R}.$$

Запишем функцию Лагранжа:

$$\mathcal{L}(l_1, \ldots, l_n, \lambda) = \sum_{i=1}^{n} p_i l_i + \lambda \cdot \left(\sum_{i=1}^{n} 2^{-l_i} - 1 \right).$$

Найдём стационарные точки, дифференцируя по l_i:

$$\frac{\partial \mathcal{L}}{\partial l_i} = p_i - \lambda \cdot 2^{-l_i} \cdot \ln 2 = 0 \Longrightarrow 2^{-l_i} = \frac{p_i}{\lambda \ln 2}.$$

Подставляя в условие Крафта:

$$\sum_{i=1}^{n} \frac{p_i}{\lambda \ln 2} = 1 \Longrightarrow \lambda = \frac{1}{\ln 2}.$$

Отсюда

$$2^{-l_i} = p_i \quad \Longrightarrow \quad l_i = -\log_2 p_i.$$

Подставляя найденные l_i в целевую функцию, получаем

$$L(C^f) \geq \mathrm{H}_2(p_1, \ldots, p_n)$$