Немного кодирования

Алгоритм Шеннона

1. Упорядочить символы $a_1, a_2, \ldots, a_n$ по *возрастанию* длин кодов: $\ell_1 \le \ldots \le \ell_n$.
2. Для символа a_i вычислить:

$$r_i = \sum_{k=1}^{i-1} 2^{-\ell_1}, \quad r_1 = 0.$$

3. Взять первые l_i знаков после запятой в двоичном представлении r_i. Это и будет кодом C_i^{S} символа a_i.

Упражнение 1. Докажите что алгоритм Шеннона всегда строит беспрефиксный код, если длины $\ell_1, \ldots, \ell_n$ удовлетворяют неравенству Крафта.

Решение. Пусть даны длины кодовых слов $l_1 \le l_2 \le \ldots \le l_n$, удовлетворяющие неравенству Крафта. Рассмотрим алгоритм Шеннона. Код строится так: для символа a_i берётся двоичная запись числа $r_i = \sum_{k=1}^{i-1} 2^{-l_k}$ (где $r_1 = 0$) и в качестве кодового слова используется первые l_i бит после запятой.

Предположим, что код не является беспрефиксным. Тогда существуют индексы $i < j$ такие, что слово C_i^{S} является префиксом слова C_j^{S}. По построению C_i^{S} — это первые l_i бит числа r_i, а C_j^{S} — первые l_j бит числа r_j. Поскольку $l_i \le l_j$, то из того, что C_i^{S} — префикс C_j^{S}, следует, что двоичные представления r_i и r_j совпадают в первых l_i битах после запятой.

Заметим, что $r_j = r_i + \sum_{k=i}^{j-1} 2^{-l_k} \ge r_i + 2^{-l_i}$, так как в сумме присутствует слагаемое 2^{-l_i}. Но если два числа отличаются не менее чем на 2^{-l_i}, то их двоичные представления после запятой должны различаться хотя бы в одном из первых l_i битов. Это противоречит совпадению первых l_i битов. Следовательно, наше предположение неверно, и код является беспрефиксным.

Упражнение 2. Пусть вероятность символа a_i равна p_i. Докажите что при подходящем выборе l_i, средняя длина кода Шеннона $L(C^{\mathrm{S}})$ будет удовлетворять неравенству:

$$L_S \le \mathrm{H}_2(p_1, \ldots p_n) + 1$$

Код должен быть беспрефиксным.

Решение. Выберем длины кодовых слов $l_i = \lceil -\log_2 p_i \rceil$. Тогда $-\log_2 p_i \le l_i < -\log_2 p_i + 1$. Проверим условие Крафта:

$$\sum_{i=1}^{n} 2^{-l_i} \le \sum_{i=1}^{n} 2^{\log_2 p_i} = \sum_{i=1}^{n} p_i = 1.$$

Следовательно, для этих длин существует беспрефиксный код (по лемме Крафта, а также по предыдущему упражнению, алгоритм Шеннона его построит). Оценим среднюю длину этого кода:

$$L_S = \sum_{i=1}^{n} p_i l_i < \sum_{i=1}^{n} p_i(-\log_2 p_i + 1) = \mathrm{H}_2(p_1, \ldots, p_n) + 1.$$

Таким образом, для такого выбора l_i код Шеннона удовлетворяет требуемому неравенству.

Алгоритм Фано

1. Упорядочить символы по невозрастанию вероятностей.
2. Разделить множество символов на две части, сумма вероятностей в которых максимально близка друг к другу.
3. Всем символам в левой части присвоить в начало кода 0, в правой — 1.
4. Повторять шаги 2-3 для каждой из частей, пока в части не останется один символ.

Упражнение 3. Докажите, что алгоритм Фано всегда строит беспрефиксный код.

Решение. Алгоритм Фано строит дерево кодирования рекурсивным разбиением множества символов на две части. На каждом шаге каждому символу в одной части добавляется к префиксу 0, а в другой — 1. Таким образом, код каждого символа соответствует пути от корня дерева к листу. Поскольку разбиение производится до тех пор, пока в каждой части не останется ровно один символ, ни один код не может быть префиксом другого, так как это означало бы, что один символ находится во внутреннем узле дерева, а не в листе. Следовательно, код является беспрефиксным.

Алгоритм Хаффмана

1. Создать узлы для каждого символа с его вероятностью. Поместить все узлы в приоритетную очередь минимальная вероятность — высший приоритет).
2. Пока в очереди больше одного узла:
   (a) Извлечь два узла с *наименьшими* вероятностями.
   (b) Создать новый *внутренний* узел, вероятностью которого будет сумма вероятностей извлеченных узлов.
   (c) Сделать извлеченные узлы дочерними для нового (любому присвоить 0, другому — 1).
   (d) Поместить новый узел в очередь.
3. Оставшийся узел — корень дерева. Код символа — путь от корня к листу (0/1 на каждом ребре).

Упражнение 4. Докажите, что алгоритм Хаффмана всегда строит беспрефиксный код.

Решение. Доказательство беспрефиксности повторяет доказательство беспрефиксности кода Фано mutatis mutandis.

Кодируем...

Упражнение 5. Для алфавита $\{A, B, C, D\}$ с вероятностями $\mathbb{P}(A) = 0.45$, $\mathbb{P}(B) = 0.25$, $\mathbb{P}(C) = 0.2$, $\mathbb{P}(D) = 0.1$ выполните:

1. Постройте код Шеннона. Вычислите среднюю длину L_{S}.
2. Постройте код Фано. Вычислите среднюю длину L_{F}.
3. Постройте код Хаффмана. Вычислите среднюю длину L_{H}.
4. Вычислите энтропию источника $\mathrm{H}_2(X) = -\sum p_i \log_2 p_i$.
5. Сравните избыточности кодов: $R = L_{\text{code}} - H(X)$.

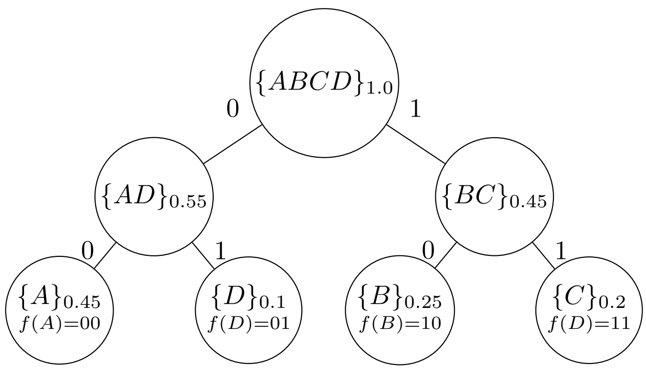
Решение. Дано: $\mathbb{P}(A) = 0.45$, $\mathbb{P}(B) = 0.25$, $\mathbb{P}(C) = 0.2$, $\mathbb{P}(D) = 0.1$.
1. **Код Шеннона.** Вычислим длины: $l_i = \lceil -\log_2 p_i \rceil$.

$$\begin{aligned}
-\log_2 0.45 &\approx 1.152, \quad l_A = 2;\\
-\log_2 0.25 &= 2.000, \quad l_B = 2;\\
-\log_2 0.2 &\approx 2.322, \quad l_C = 3;\\
-\log_2 0.1 &\approx 3.322, \quad l_D = 4.
\end{aligned}$$

Символы уже упорядочены по возрастанию длин. Вычислим r_i:

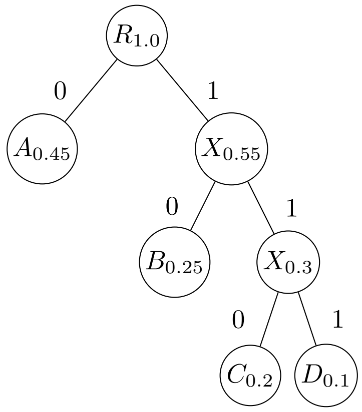
$$\begin{aligned}
&r_A = 0 = 0.0000_2, \quad \text{код: } 00;\\
&r_B = 2^{-2} = 0.25 = 0.0100_2, \quad \text{код: } 01;\\
&r_C = 2^{-2} + 2^{-2} = 0.5 = 0.1000_2, \quad \text{код: } 100;\\
&r_D = 2^{-2} + 2^{-2} + 2^{-3} = 0.5 + 0.125 = 0.625 = 0.1010_2, \quad \text{код: } 1010.
\end{aligned}$$

Средняя длина: $L_S = 0.45 \cdot 2 + 0.25 \cdot 2 + 0.2 \cdot 3 + 0.1 \cdot 4 = 0.9 + 0.5 + 0.6 + 0.4 = 2.4$.
2. **Код Фано.** Алгоритм строит дерево сверху вниз. Графически процесс можно изобразить так:



Средняя длина: $L_F = 2$.

3. **Код Хаффмана.** Имеем:



Коды:
A : 0 (путь: 0)
B : 10 (путь: 1→0)
C : 110 (путь: 1→1→0)
D : 111 (путь: 1→1→1)

Средняя длина: $L_H = 1.85$.

Энтропия и избыточности. Пусть X обозначает случайный символ. Имеем:

$$\begin{aligned}
\mathrm{H}_2(X) &= -\sum p_i \log_2 p_i \\
&\approx 0.45 \cdot 1.152 + 0.25 \cdot 2 + 0.2 \cdot 2.322 + 0.1 \cdot 3.322 \\
&\approx 0.5184 + 0.5 + 0.4644 + 0.3322 = 1.815.
\end{aligned}$$

Тогда:

$$\begin{aligned}
R_S &= L_S - \mathrm{H}_2(X) \approx 2.4 - 1.815 = 0.585; \\
R_F &= L_{SF} - \mathrm{H}_2(X) \approx 2 - 1.815 = 0.185; \\
R_H &= L_H - \mathrm{H}_2(X) \approx 1.85 - 1.815 = 0.035.
\end{aligned}$$