Задача об оптимальном префиксном коде. Метод Хаффмена. Неравенство Крафта.

Руслан Назирович Мокаев

Математико-механический факультет,
Санкт-Петербургский государственный университет

Санкт-Петербург, 13.02.2024

Содержание лекции

- Задача об оптимальном префиксном коде
- Лемма о кратчайшем префиксе
- Лемма о соседстве самых редких символов
- Лемма об оптимальном префиксном коде для расширенного алфавита
- Метод Хаффмена
- Неравенство Крафта

Префиксный код

Λ – произвольное конечное множество (алфавит), $a \in \Lambda$ – символы. Пусть $\forall a \in \Lambda \ \exists \ l(a) \in \mathbb{N}$, $\exists \ c(a) = \{0,1\}^{l(a)}$– кодовая последовательность a.

$\forall \ a, b \in \Lambda, a \neq b \ c(a) \neq c(b)$
Это достаточное условие однозначности распознавания символа?

$\Lambda = \{a, b\}, c(a) = 10, c(b) = 100$. Расшифровать последовательность 100?

Добавим в алфавит символы $d: \ c(d) = 01$ и $e: \ c(e) = 1$.
Сообщение 1001 можно понять как ad или be.

Условие префиксности:
Определение: Код называется **префиксным**, если
$\forall \ a, b \in \Lambda \ c(a) = \omega \Rightarrow \nexists m \in \mathbb{N}_0 : c(b) = \omega\gamma$, где $\gamma \in \{0,1\}^m$.
Никакая кодовая последовательность одного символа не является началом кодовой последовательности другого символа.

Задача об оптимальном префиксном коде

Пусть $\forall\ a \in \Lambda$ соответствует вероятность $p(a)$ появления этого символа в сообщении. $\sum\limits_{a\in\Lambda} p(a) = 1$ и считаем $\forall\ a \in \Lambda\ p(a) > 0$.

Введём ДСВ l на вероятностном пространстве (Λ, p):

$$Pr\{l = x\} = Pr\big(\{a \in \Lambda \mid l(a) = x\}\big).$$

Определение: **оптимальным** называется префиксный код, минимизирующий математическое ожидание l :

$$El = \sum_{a\in\Lambda} l(a)\cdot p(a) = \sum_{x\in\mathrm{Im}(l)} x \cdot Pr\{l = x\}.$$

Чем чаще встречается символ, тем короче д.б. кодовая посл-ть.

Существование ОПК? Известно, что $El \geq 1$ (в каждой кодовой посл-ти должен быть ≥ 1 символ).
Всегда можно сделать ПК, в котором все символы имеют одинаковые длины кодовых последовательностей и эти последовательности различны ($\forall\ a \in \Lambda\ l(a) = \lceil \log_2(|\Lambda|) \rceil$).

Т.е. ПК существуют и мат. ожидание длины кодовой посл-ти ограничено.

Лемма о кратчайшем префиксе

Лемма: если в префиксном коде C существует $x \in \Lambda : \; c(x) = \omega\alpha$, где $\alpha \in \{0,1\}$ и при этом $\nexists y \in \Lambda, \; y \neq x : c(y) = \omega\gamma$, где $\gamma \in \{0,1\}^k$ (то есть, если ω не является началом никакой другой кодовой посл-ти, кроме $c(x)$), то такой код не оптимален.

▲ : рассмотрим код $C' : \; c'(x) = \omega$ и $\forall \; y \in \Lambda, y \neq x \;\; c'(y) = c(y)$. Он будет префиксным (по построению и условию леммы) и

$$El' = El - p(x) \cdot l(x) + p(x) \cdot \big(l(x) - 1\big) = El - p(x) < El.$$

Тогда код C не мог быть оптимальным $\square$.

Лемма: если в префиксном коде $C \; \exists \; a, b \in \Lambda, a \neq b$ такие, что $p(a) < p(b)$ и $l(a) < l(b)$, то такой код не оптимален.

▲ : проверим, что для кода C', в котором $c'(a) = c(b), \;\; c'(b) = c(a)$ и $\forall \; x \in \Lambda : \;\; x \neq a, x \neq b \; c'(x) = c(x)$ верно $El - El' > 0$.

$$El - El' = p(a)l(a) + p(b)l(b) - p(a)l(b) - p(b)l(a) =$$
$$= \big(p(a) - p(b)\big)\big(l(a) - l(b)\big) > 0 \; \square$$

Лемма о соседстве самых редких символом

Лемма: Пусть $a, b \in \Lambda,\ a \neq b$ – символы с наименьшими вероятностями ($\forall\ x \in \Lambda\ p(x) \geq p(b) \geq p(a)$). Тогда $\exists$ ОПК: $c(a) = \omega 0, c(b) = \omega 1$, где $\exists\ k \in \mathbb{N}_0 :\ \omega \in \{0,1\}^k$, и это самые длинные кодовые последовательности.

$\blacktriangle$: пусть C' – ОПК. По лемме о кратчайшем префиксе a и b имеют самые длинные кодовые последовательности в C' :
$\forall\ x \in \Lambda, x \neq a, x \neq b\ l'(a) \geq l'(b) \geq l'(x)$

Если $c(a) = \overline{\omega}\gamma,\ \overline{\omega} \in \{0,1\}^{l'(b)}, \gamma \in \{0,1\}^{l'(a)-l'(b)}$, то $\overline{\omega}$ не является началом никакой кодовой посл-ти (т.к. остальные кодовые посл-ти не длиннее $\overline{\omega}$ и $\nexists$ символа с кодовой посл-тью $\overline{\omega}$ в силу префиксности C') $\Rightarrow$ можно сократить кодовую посл-ть a, создав более оптимальный код (?!)

$\Rightarrow$ из оптимальности C' следует $l'(a) = l'(b)$. Пусть $c'(b) = \omega 1$, тогда, если $\exists\ x \in \Lambda :\ c'(x) = \omega 0$, то построим ОПК C :
$c(a) = c'(x), c(x) = c'(a),\ \forall\ z \in \Lambda, z \neq a, z \neq x\ c(z) = c'(z)$.

если $\nexists\ x \in \Lambda :\ c'(x) = \omega 0$, то по лемме с прошлого слайда C' – не оптимален (?!) $\square$.

Лемма об ОПК для расширенного алфавита

Лемма: Пусть $a, b \in \Lambda$, $a \neq b$ – символы с наименьшими вероятностями. $\Lambda' = \Lambda \setminus \{a, b\} \cup \{\underbrace{ab}\}$, где $\underbrace{ab} \notin \Lambda$, $p(\underbrace{ab}) = p(a) + p(b)$.

Пусть C' – ОПК для Λ', $c'(\underbrace{ab}) = \omega$. Тогда для Λ код $C:$ $c(a) = \omega 0$, $c(b) = \omega 1$, $\forall\ x \in \Lambda$, $x \neq a, x \neq b$ $c(x) = c'(x)$ будет ОПК.

$\blacktriangle : l(a)p(a) + l(b)p(b) = (l'(\underbrace{ab}) + 1)(p(a) + p(b)) = l'(\underbrace{ab})p(\underbrace{ab}) + p(\underbrace{ab})$

Тогда $El = El' + p(\underbrace{ab})$.

Пусть $\overline{C}$ – ОПК для Λ и $E\bar{l} < El$. Л. о соседстве: $\bar{c}(a) = \gamma 0$, $\bar{c}(b) = \gamma 1$.

Построим $\overline{C}'$ для Λ' : $\bar{c}'(\underbrace{ab}) = \gamma$ и $\forall x \in \Lambda$, $x \neq a, x \neq b$ $\bar{c}'(x) = \bar{c}(x)$

$\overline{C}'$ – префиксный? По Лемме о кратчайшем префиксе $\nexists$ символа с кодовой посл-тью длины $> \bar{l}(a)$. Никакой символ не мог иметь кодовую посл-ть γ, т.к. $\overline{C}$ префиксный. Единственные две посл-ти длины $\bar{l}(a)$, начинающиеся на γ, – это коды a и b. Но их нет в Λ'. При этом $E\bar{l} = E\bar{l}' + p(\underbrace{ab})$.

По предположению $El' + p(\underbrace{ab}) = El > E\bar{l} = E\bar{l}' + p(\underbrace{ab})$ (?!) опт-ти C'

$\Rightarrow E\bar{l} \geq El$, но т.к. $\overline{C}$ – ОПК $\Rightarrow E\bar{l} = El$ и C – ОПК. $\square$

Алгоритм Хаффмана построения ОПК

Задача: нужно построить ОПК на алфавите Λ, $|\Lambda| = M$. По лемме об ОПК для расширенного алфавита задачу построения ОПК можно свести к такой же задаче, но с исходным алфавитом с числом букв на единицу меньше, и с набором вероятностей, получающимся из первоначального сложением двух наименьших вероятностей.

Уменьшаем пока не получится алфавит из двух букв. ОПК для алфавита из 2-х букв – $\{0, 1\}$.

Строже: $\Lambda_0 := \Lambda$. $\forall k \in 0 : (M-3)$ берём $a_k, b_k \in \Lambda_k : \forall\ x \in \Lambda_k, x \neq a_k$, $x \neq b_k$ $p(a_k) \leq p(b_k) \leq p(x)$ и построим $\Lambda_{k+1} = \Lambda_k \setminus \{a_k, b_k\} \cup \{\underbrace{a_k b_k}\} \ldots$

Для $\Lambda_{M-2} = \{a_{M-2}, b_{M-2}\}$ оптимальным будет код $C_{M-2}:\ c_{M-2}(a_{M-2}) = 0, c_{M-2}(b_{M-2}) = 1$, т.к. для него $El_{M-2} = 1$.

Теперь для $k \in 1 : (M-2)$ есть ОПК C_k для Λ_k. По Лемме об ОПК для расширенного алфавита строится ОПК C_{k-1} для Λ_{k-1} такой, что $c_{k-1}(a_{k-1}) = c_k(a_{k-1}b_{k-1})0$, $c_{k-1}(b_{k-1}) = c_k(a_{k-1}b_{k-1})1$, $\forall x \in \Lambda_k, x \neq \underbrace{a_{k-1}b_{k-1}}\quad c_{k-1}(x) = c_k(x)$

Выполняем пока не получится C_0 – ОПК для $\Lambda_0 = \Lambda$.

$\Lambda_0 = \{a, b, c, d, e, f, g\}$, $p(a) = 0.13, p(b) = 0.08, p(c) = 0.25,$
$p(d) = 0.18, p(e) = 0.03, p(f) = 0.12, p(g) = 0.21.$

$a_0 = e, b_0 = b$, $\Lambda_1 = \{a, \underbrace{eb}, c, d, f, g\}$, $p(a) = 0.13,$
$p(\underbrace{eb}) = 0.11, p(c) = 0.25$, $p(d) = 0.18, p(f) = 0.12, p(g) = 0.21.$

$a_1 = \underbrace{eb}, b_1 = f$, $\Lambda_2 = \{a, \underbrace{ebf}, c, d, g\}$,
$p(a) = 0.13$, $p(\underbrace{ebf}) = 0.23, p(c) = 0.25$, $p(d) = 0.18, p(g) = 0.21.$

$a_2 = a, b_2 = d$, $\Lambda_3 = \{\underbrace{ad}, \underbrace{ebf}, c, g\}$,
$p(\underbrace{ad}) = 0.31, p(\underbrace{ebf}) = 0.23, p(c) = 0.25, p(g) = 0.21.$

$a_3 = g, b_3 = \underbrace{ebf}$, $\Lambda_4 = \{\underbrace{ad}, \underbrace{gebf}, c\}$,
$p(\underbrace{ad}) = 0.31, p(\underbrace{gebf}) = 0.44, p(c) = 0.25.$

$a_4 = c, b_4 = \underbrace{ad}$, $\Lambda_5 = \{\underbrace{cad}, \underbrace{gebf}\}$, $p(\underbrace{cad}) = 0.56, p(\underbrace{gebf}) = 0.44.$

Тогда $c_5(\underbrace{gebf}) = 0, c_5(\underbrace{cad}) = 1.$

Теперь раскрываем алфавит обратно:
$c_4(\underbrace{gebf}) = 0, c_4(c) = 10, c_4(\underbrace{ad}) = 11.$

$c_3(g) = 00, c_3(\underbrace{ebf}) = 01, c_3(c) = 10, c_3(\underbrace{ad}) = 11.$

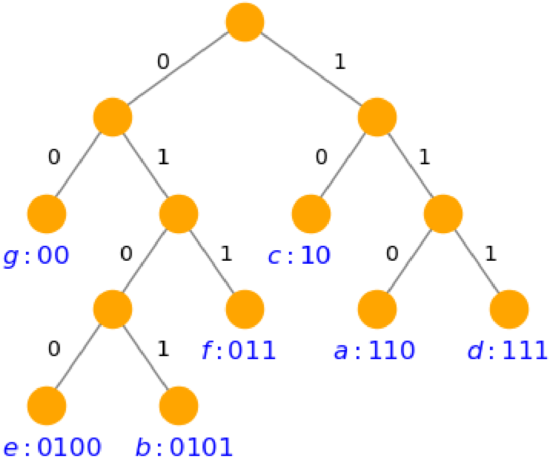
$c_2(g) = 00, c_2(\underbrace{ebf}) = 01, c_2(c) = 10, c_2(a) = 110, c_2(d) = 111.$

$c_1(g) = 00, c_1(\underbrace{eb}) = 010, c_1(f) = 011, c_1(c) = 10,$
$c_1(a) = 110, c_1(d) = 111.$

$c_0(g) = 00, c_0(e) = 0100, c_0(b) = 0101, c_0(f) = 011,$
$c_0(c) = 10, c_0(a) = 110, c_0(d) = 111.$

Параллельно с построением кода можно строить соответствующее ему двоичное дерево.

Код – это набор путей из корня в произвольную вершину в произвольном двоичном дереве, префиксный код – набор путей из корня в листья в произвольном двоичном дереве.


0
1
0
1
0
1
g : 00
0
1
c : 10
0
1
0
1
f : 011
a : 110
d : 111
e : 0100
b : 0101

Неравенство Крафта

Задача: Пусть задан набор длин $l_1, \dots, l_m$, не все обязательно различны. Может ли такой набор оказаться набором длин некоторого префиксного кода?

Теорема: Для того, чтобы набор длин $l_1, \dots, l_m$ мог быть набором длин кодовых посл-тей некоторого ПК для алфавита из m символов необходимо и достаточно, чтобы $\sum\limits_{i=1}^{m} 2^{-l_i} \leq 1$.

$\blacktriangle$: $\Rightarrow$) $\exists$ ПК для алфавита с кодовыми посл-тями с длинами $l_1, \dots, l_m$.
Множ-во кодовых посл-тей – набор всех путей на двоичном дереве от корня к листьям.
Корень – нулевой уровень. Далее последовательно увеличиваем номер по мере удаления от корня.

Каждой вершине v на уровне t сопоставим число $a(v) = 2^{-t}$.

Пусть вершина v на уровне t – не лист. Т.е. на уровне $t+1$ есть ≥ 1 вершина, получившаяся из v. Обозначим его $N(v)$. Тогда $a(v) \geq \sum\limits_{u \in N(v)} a(u)$.

Просуммируем нер-ва для всех не листов: $\sum\limits_{v \text{ не лист}} a(v) \geq \sum\limits_{u \text{ не корень}} a(u) \Rightarrow$

$\Rightarrow 2^0 \geq \sum\limits_{u \text{ листья}} a(u)$. Необходимость доказана.

$\Leftarrow$) выполнено нер-во и пусть $l_1 \leq \ldots \leq l_m$.
n_j – число листьев на уровне j: $n_j = \Big|\{i : l_i = j, i \in 1 : m\}\Big|$.

$\sum\limits_{i \in 1:m} 2^{-l_i} \leq 1 \Rightarrow \sum\limits_{j \in 1:l_m} 2^{-j} n_j \leq 1$. Тогда для $\forall j \in 1 : l_m$ справедливо
$n_j \leq 2^j - (2^{j-1} n_1 + \ldots + 2n_{j-1})$

Пусть $m \neq 1$. Выделим на первом уровне вершин $n_1 \leq 2$, на втором уровне останется $2(2 - n_1)$. Известно, что $n_2 \leq 2^2 - 2n_1 \Rightarrow$ осталось не меньше, чем требуется для второго уровня.

$(j-1)$-уровень: было свободно $2^{j-1} - (2^{j-2} n_1 + \ldots + 2n_{j-2})$ и n_{j-1} не больше этой величины. Выделим n_{j-1} узлов, останется
$2^{j-1} - (2^{j-2} n_1 + \ldots + 2n_{j-2}) - n_{j-1}$. Значит на j-м уровне будет
$2 \cdot (\ldots) = 2^j - (2^{j-1} n_1 + \ldots + 2n_{j-1})$. $\square$