Дискретная математика 2 семестр ПИ, Лекции

Собрано 19 февраля 2022 г. в 13:39

Содержание

Раздел #1: Кодирование информации

1.1. Задача об оптимальном префиксном коде

Пусть 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, где $l(a)$ – длина.
Очевидно, условие $\forall a, b \in \Lambda \to (a \neq b \Rightarrow c(a) \neq c(b))$ не является достаточным для однозначного распознавания символов.

Def 1.1.1. *Код называется префиксным, если* $\forall a, b \in \Lambda\ c(a) = \omega \Rightarrow \nexists\ m \in \mathbb{N}_0 : c(b) = \omega\gamma$, *где* $\gamma \in \{0,1\}^m$

Пусть $\forall a \in \Lambda$ соответствует вероятность $p(a)$ появления этого символа в сообщении. $\sum_{a\in\Lambda} p(a) = 1$ и считаем $\forall a \in \Lambda\ p(a) > 0$.
Введем дискретную случайную величину $l : \forall a \in \Lambda\ Pr\{l = l(a)\} = p(a)$ – длина кодовой последовательности символа в сообщении.

Def 1.1.2. *Оптимальным называется префиксный код, минимизирующий математическое ожидание* $l : \mathbb{E}l = \sum_{a\in\Lambda} l(a)p(a)$

Чем чаще встречается символ, тем короче должна быть кодовая последовательность.
Почему вообще ОПК существует? Известно, что $\mathbb{E}l \geqslant 1$ (в каждой кодовой последовательности должен быть хотя бы один символ). Всегда можно сделать префиксный код, 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. префиксный код существует и матожидание длины кодовой последовательности ограничено.

Lm 1.1.3. Если в некотором 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 последовательности, кроме $c(x)$), то код $C' : c'(x) = \omega, \forall y \in \Lambda, y \neq x\ c' = c(y)$ будет префиксным (по построению и условию леммы) и $\mathbb{E}l' = \mathbb{E}l - p(x)l(x) + p(x)(l(x)-1) = \mathbb{E}l - p(x) < \mathbb{E}l$. Тогда код C точно не мог быть оптимальным.

Lm 1.1.4 (Лемма о кратчайшем префиксе)**.** Если в префиксном коде C $\exists a, b \in \Lambda, a \neq b : p(a) < p(b), l(a) < l(b)$, то такой код не оптимален.

Доказательство. 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 $\mathbb{E}l - \mathbb{E}l' > 0$.

$$\mathbb{E}l - \mathbb{E}l' = p(a)l(a) + p(b)l(b) - p(a)l(b) - p(b)l(a) = (p(a) - p(b))(l(a) - l(b)) > 0$$

■

Lm 1.1.5 (Лемма о соседстве самых редких символов)**.** Пусть $a, b \in \Lambda, a \neq b$ – символы с наименьшими вероятностями ($\forall x \in \Lambda\ p(x) \geqslant p(b) \geqslant p(a)$). 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.

Доказательство. Пусть C' – ОПК. По лемме о кратчайшем префиксе a и b имеют самые длинные кодовые последовательности в $C' : \forall x \in \Lambda, x \neq a, x \neq b \; l'(a) \geqslant l'(b) \geqslant l'(x)$
Если $c(a) = \omega\gamma, \omega \in \{0,1\}^{l'(b)}, \gamma \in \{0,1\}^{l'(a)-l'(b)}$ и ω не является началом никакой кодовой последовательности (т.к. остальные кодовые последовательности не длиннее ω и $\nexists$ символа с кодовой последовательностью ω в силу префиксности C') $\Rightarrow$ можно сократить кодовую последовательность 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 построим ОПК $C : c(a) = \omega 0, \forall z \in \Lambda, z \neq a \; c(z) = c'(z)$. ■

Lm 1.1.6 (Лемма об ОПК для расширенного алфавита)**.** Пусть $a, b \in \Lambda, a \neq b$ – символы с намиеньшими 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 $\Lambda', c'(\underbrace{ab}) = \omega$. 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.

Доказательство. $l(a)p(a)+l(b)p(b) = (l'(\underbrace{ab})+1)(p(a)+p(b)) = l'(\underbrace{ab})p(\underbrace{ab})+p(\underbrace{ab})$. Тогда $\mathbb{E}l = \mathbb{E}l' + p(\underbrace{ab})$.

Пусть $\overline{C}$ – ОПК для Λ и $\mathbb{E}\bar{l} < \mathbb{E}l$. По лемме о соседстве: $\bar{c}(a) = \gamma 0, \bar{c}(b) = \gamma 1$. Построим $\overline{C}'$ для $\Lambda' : \bar{c}'(\underbrace{ab}) = \gamma$ 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 последовательностью длины $> \bar{l}(a)$. Никакой символ не мог иметь кодовую последовательность γ, т.к. $\overline{C}$ префиксный. Единственные две последовательности длины $\bar{l}(a)$, начинающиеся на γ – это коды a и b. Но их нет в Λ'. При этом $\mathbb{E}\bar{l} = \mathbb{E}\bar{l}' = p(\underbrace{ab})$. По предположению $\mathbb{E}l' + p(\underbrace{ab}) = \mathbb{E}l > \mathbb{E}\bar{l} = \mathbb{E}\bar{l}' + p(\underbrace{ab})$ (?!) оптимальности $C' \Rightarrow \mathbb{E}\bar{l} \geqslant \mathbb{E}l$, но т.к. $\overline{C}$ – ОПК $\Rightarrow \mathbb{E}\bar{l} = \mathbb{E}l$ и C – ОПК. ■

Задача: нужно построить ОПК на алфавите $\Lambda, |\Lambda| = M$. 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, получющимся из первоначального сложением двух наименьших вероятностей.
Уменьшаем пока не получится алфавит из двух букв. ОПК для алфавита из 2-х букв – $\{0,1\}$.
Строже: $\Lambda_0 := \Lambda$. $\forall k \in 0...(M-3)$ берем $a_k, b_k \in \Lambda_k : \forall x \in \Lambda_k, x \neq a_k, x \neq b_k \; p(a_k) \leqslant p(b_k) \leqslant p(x)$ и построим $\Lambda_{k+1} = \Lambda_k \setminus \{a_k, b_k\} \cup \{\underbrace{a_k b_k}\}$...

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 $\mathbb{E}l_{M-2} = 1$.
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 $c_{k-1}(a_{k-1}) = c_k(a_{k-1}b_{k-1})0, c_{k-1}(b_{k-1}) = c_k(a_{k-1}b_{k-1})1, \forall x \in \Lambda_k, x \neq \underbrace{a_{k-1}b_{k-1}} \; c_{k-1}(x) = c_k(x)$.

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

Пример 1.1.7. $\Lambda_0 = \{a,b,c,d,e,f,g\}, p(a) = 0.13, p(b) = 0.08, p(c) = 0.25, p(d) = 0.18, p(e) = 0.03, p(f) = 0.12, p(g) = 0.21$.
$a_0, = e, b_0 = b, \Lambda_1 = \{a, \underbrace{e,b}, c, d, f, g\}, p(a) = 0.13, p(\underbrace{eb}) = 0.11, p(c) = 0.25, p(d) = 0.18, p(f) =$

$0.12, p(g) = 0.21$.
$a_1 = \underbrace{eb}, b_1 = f, \Lambda_2 = \{a, \underbrace{ebf}, c, d, g\}, p(a) = 0.13, p(\underbrace{ebf}) = 0.23, p(c) = 0.25, p(d) = 0.18, p(g) = 0.21$.
$a_2 = a, b_2 = d, \Lambda_3 = \{\underbrace{ad}, \underbrace{ebf}, c, g\}, p(\underbrace{ad}) = 0.31, p(\underbrace{ebf}) = 0.23, p(c) = 0.25, p(g) = 0.21$.
$a_3 = g, b_3 = \underbrace{ebf}, \Lambda_4 = \{\underbrace{ad}, \underbrace{gebf}, c\}, p(\underbrace{ad}) = 0.31, p(\underbrace{gebf}) = 0.44, p(c) = 0.25$. $a_4 = c, b_4 = \underbrace{ad}, \Lambda_5\{\underbrace{cad}, \underbrace{gebf}\}, p(\underbrace{cad}) = 0.56, p(\underbrace{gebf}) = 0.44$. Тогда $c_5(\underbrace{gebf}) = 0, c(\underbrace{cad}) = 1$.
Теперь раскрываем алфавит обратно:
$c_4(\underbrace{gebf}) = 0, c_4(c) = 10, c_4(\underbrace{ad}) = 11$.
$c_3(g) = 00, c_3(\underbrace{ebf}) = 01, c_3(c) = 10, c_3(\underbrace{ad}) = 11$.
$c_1(g) = 00, c_1(\underbrace{eb}) = 010, c_1(f) = 011, c_1(c) = 10, c_1(a) = 110, c_1(d) = 111$.
$c_0(g) = 00, c_0(e) = 0100, c_0(b) = 0101, c_0(f) = 011, c_0(c) = 10, c_0(a) = 110, c_0(d) = 111$.

1.2. Неравенство Крафта

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

Теорема 1.2.1. Для того, чтобы набор длин $l_1, ..., l_m$ мог быть набором длин кодовых последовательностей некоторого ПК для алфавита из m символов необходимо и достаточно, чтобы $\sum_{i=1}^{m} 2^{-l_i} \leqslant 1$.