Статистика ФИВТ ПМИ

Прикладной поток

Лекция 14

Кодирование

Алфавит: {**A**, **B**, **C**}

Как его закодировать с помощью 0 и 1?

Правило кодирования:

$$\begin{array}{lcl} \mathbf{A} & \longrightarrow & \mathbf{00} \\ \mathbf{B} & \longrightarrow & \mathbf{01} \\ \mathbf{C} & \longrightarrow & \mathbf{10} \end{array}$$

Дано сообщение:

AAAAAACAABAAACAAAAACABAAA
AAAAAABABCBABAAABBAABAAAA

Закодированное сообщение:

00 00 00 00 00 00 10 00 00 01 00 00 00 10 00 00 00 00 00 10 00 01 00 00 00
00 00 00 00 00 00 01 00 01 10 01 00 01 00 00 00 01 01 00 00 01 00 00 00 00

Длина символа: 2

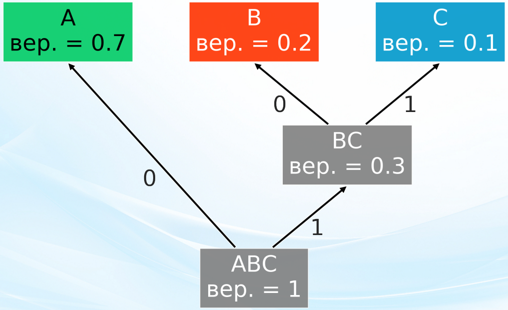
Код Хаффмана

Алфавит: {**A**, **B**, **C**}

Известны вероятности появления символов: {**0.7**, **0.2**, **0.1**}

Хочется уменьшить длину закодированного сообщения.

Метод построения оптимального кода (Хаффман):



Правило кодирования:

$\mathbf{A} \longrightarrow \mathbf{0}$
$\mathbf{B} \longrightarrow \mathbf{10}$
$\mathbf{C} \longrightarrow \mathbf{11}$

Декод-ние однозначно, т.к. ни один код не является префиксом другого.

Средняя длина символа: $0.7 \cdot 1 + 0.2 \cdot 2 + 0.1 \cdot 2 = 1.3$

Пример

Исходное кодирование

Закодированное сообщение:

00 00 00 00 00 00 10 00 00 01 00 00 00 10 00 00 00 00 00 10 00 01 00 00 00
00 00 00 00 00 00 01 00 01 10 01 00 01 00 00 00 01 01 00 00 01 00 00 00 00

Количество символов: 100; Длина символа: 2

Код Хаффмана

Закодированное сообщение:

0 0 0 0 0 0 11 0 0 10 0 0 0 11 0 0 0 0 0 11 0 10 0 0 0 0 0 0 0 0 0 10 0 10 11 10
0 10 0 0 0 10 10 0 0 10 0 0 0 0

Количество символов: 63; Средняя длина символа: 1.3

Средняя длина символа

Утверждение:

Для кодирования символа, встречающегося с вероятностью p_j в "идеале" нужно $\log_2 \frac{1}{p_j}$ бит. Приближение — коды Хаффмана.

Пример:

Символы равновероятны $\Rightarrow$ для каждого символа нужно $\lceil \log_2 k \rceil$ бит.

Пусть символы $a_1, \ldots, a_k$ встречаются с вер-тями $p_1, \ldots, p_k$.

$H(\mathrm{P}) = -\sum_{j=1}^{k} p_j \log_2 p_j$ — **энтропия**

Энтропия — средняя длина символа при оптимальном кодировании.

В нашем случае для вероятностей $\{\mathbf{0.7}, \mathbf{0.2}, \mathbf{0.1}\}$

$H(\mathrm{P}) = -0.7 \log_2 0.7 - 0.2 \log_2 0.2 - 0.1 \log_2 0.1 \approx 1.157$

А мы построили код со средней длинной символа 1.3.

Кодирование с помощью другого распределения

Что будет, если будем кодировать кодом, построенным по распр.

$Q = \{q_1, ..., q_k\}$, если истинное распр. $P = \{p_1, ..., p_k\}$?

Алфавит: {**A**, **B**, **C**}

Истинные вероятности появления символов: P ={**0.7**, **0.2**, **0.1**}

Предполагаемые вер-ти появления символов: Q ={**0.4**, **0.5**, **0.1**}

Правило кодирования для Q:

$$\begin{aligned} \mathbf{A} &\longrightarrow \mathbf{10} \\ \mathbf{B} &\longrightarrow \mathbf{0} \\ \mathbf{C} &\longrightarrow \mathbf{11} \end{aligned}$$

Закодированное сообщение:

10 10 10 10 10 10 11 10 10 0 10 10 10 11 10 10 10 10 10 11 10 0 10 10 10 10 10 10 10 10 10 0 10 0 11 0 10 0 10 10 10 0 0 10 10 0 10 10 10 10

Количество символов: 91

Средняя длина символа: $0.7 \cdot 2 + 0.2 \cdot 1 + 0.1 \cdot 2 = 1.8$

Кодирование с помощью другого распределения

Что будет, если будем кодировать кодом, построенным по распр. $\mathrm{Q} = \{q_1, \ldots, q_k\}$, если истинное распр. $\mathrm{P} = \{p_1, \ldots, p_k\}$?

$H(\mathrm{P}, \mathrm{Q}) = -\sum\limits_{j=1}^{k} p_j \log_2 q_j$ — **кросс-энтропия**

Кросс-энтропия — средняя длина символа при кодировании алфавита вероятностями появления символов Q, если на самом деле они появляются с вероятностями P.

$KL(\mathrm{P}, \mathrm{Q}) = H(\mathrm{P}, \mathrm{Q}) - H(\mathrm{P})$ — **дивергенция Кульбака-Лейблера**

$KL(\mathrm{P}, \mathrm{Q}) = \sum\limits_{j=1}^{k} p_j \log_2 \frac{p_j}{q_j}$

Дивергенция Кульбака-Лейблера — избыточная длина символа при кодировании алфавита вероятностями появления символов Q, если на самом деле они появляются с вероятностями P.

Кодирование с помощью другого распределения

Алфавит: {**A**, **B**, **C**}

Истинные вероятности появления символов: P ={**0.7**, **0.2**, **0.1**}

Предполагаемые вер-ти появления символов: Q ={**0.4**, **0.5**, **0.1**}

$H(\mathrm{P}) = -0.7\log_2 0.7 - 0.2\log_2 0.2 - 0.1\log_2 0.1 \approx 1.157$

$H(\mathrm{P},\mathrm{Q}) = -0.7\log_2 0.4 - 0.2\log_2 0.5 - 0.1\log_2 0.1 \approx 1.458$

$KL(\mathrm{P},\mathrm{Q}) = H(\mathrm{P},\mathrm{Q}) - H(\mathrm{P}) \approx 1.458 - 1.157 = 0.301$

В теории мы тратим лишние 0.3 бита на символ.

Для приближающих кодов Хаффмана:

Средняя длина символа при кодировании по P: 1.3

Средняя длина символа при кодировании по Q: 1.8

Избыточная длина символа: 0.5

Хочется назвать $KL(\mathsf{P}, \mathsf{Q})$ расстоянием

от истинного распределения P

до предполагаемого распределения Q.

Далее решать задачу: $KL(\mathsf{P}, \mathsf{Q}) \to \min\limits_{\mathsf{Q}}$


ВСЁ!