МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ
ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені ІВАНА ФРАНКА

Кафедра програмування

Практичне завдання № 4

СТАТИСТИЧНІ МЕТОДИ ЕКОНОМНОГО КОДУВАННЯ

з курсу "Теорія інформації"

Виконала:
студент групи ПМІ-41
Шипка Олена

Варіант **26**

Оцінка

Прийняв:
доц. Рикалюк Р.Є.
ас. Жировецький В.В.

Львів 2020

Завдання 4.1. Значення ймовірностей p_i , з якими дискретне джерело інформації генерує символи алфавіту наступні:

p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
0.32	0.09	0.05	0.15	0.02	0.22	0.08	0.07	0

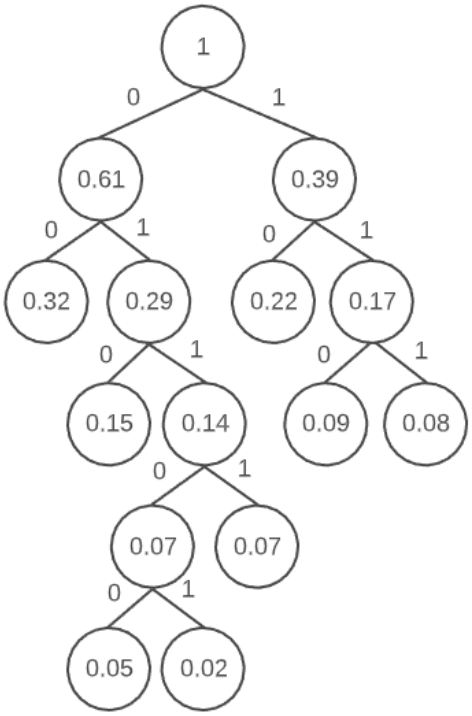
Побудувати нерівномірні ефективні коди за алгоритмами Шеннона-Фано та Хаффмена. Порівняти ефективність кодів.

Побудуємо код за алгоритмом Шеннона-Фано. Символ з ймовірністю появи 0 не кодуємо.

Ймовірність p_i	1	2	3	4	5	Код	Довжина коду l_i
0.32	0	0				00	2
0.22		1				01	2
0.15	1	0	0			100	3
0.09			1			101	3
0.08		1	0			110	3
0.07			1	0		1110	4
0.05				1	0	11110	5
0.02					1	11111	5

Обчислимо середню довжину коду $\bar{l} = \sum p_i\, l_i = 0{,}32 * 2 + 0{,}22 * 2 + 0{,}15 * 3 + 0{,}09 * 3 + 0{,}08 * 3 + 0{,}07 * 4 + 0{,}05 * 5 + 0{,}02 * 5 = 0{,}64 + 0{,}44 + 0{,}45 + 0{,}27 + 0{,}24 + 0{,}28 + 0{,}25 + 0{,}1 = 2{,}67$

Побудуємо код за алгоритмом Хаффмана



Ймовірність p_i	Код	Довжина коду l_i
0.32	00	2
0.22	10	2
0.15	010	3
0.09	110	3
0.08	111	3
0.07	0111	4
0.05	01100	5
0.02	01101	5

Обчислимо середню довжину коду $\bar{l} = 0{,}32 * 2 + 0{,}22 * 2 + 0{,}15 * 3 + 0{,}09 * 3 + 0{,}08 * 3 + 0{,}07 * 4 + 0{,}05 * 5 + 0{,}02 * 5 = 0{,}64 + 0{,}44 + 0{,}45 + 0{,}27 + 0{,}24 + 0{,}28 + 0{,}25 + 0{,}1 = 2{,}67$

Оскільки середня довжина коду при кодуванні алгоритмом Хаффмана така ж, як і при кодуванні алгоритмом Шеннона-Фано, то алгоритми є однаково ефективними.

Завдання 4.2. Алфавіт дискретного джерела інформації складається з чотирьох символів $X = \{A, B, C, D\}$. Значення ймовірностей виникнення символів наступні:

$p(A)$	$p(B)$	$p(C)$	$p(D)$
0.42	0.37	0.18	0.03

Побудувати нерівномірні ефективні коди за алгоритмами Шеннона-Фано або Хаффмена для кодування поодиноких символів джерела та слів довжиною у два символи. Оцінити та порівняти ефективність отриманих кодів. Побудованими кодами закодувати фрагмент повідомлення довжиною у 30 символів, що був згенерований джерелом. Фрагмент повідомлення: $BBBCBABBBAACCBCCAAACABBABADAAA$

Побудуємо код за алгоритмом Шеннона-Фано для кодування поодиноких символів:

Символ	Ймовірність p_i	1	2	3	Код	Довжина коду l_i
A	0.42	0			0	1
B	0.37	1	0		10	2
C	0.18		1	0	110	3
D	0.03			1	111	3

Обчислимо середню довжину коду $\bar{l} = 0{,}42 * 1 + 0{,}37 * 2 + 0{,}18 * 3 + 0{,}03 * 3 = 0{,}42 + 0{,}74 + 0{,}54 + 0{,}09 = 1{,}79$

Ентропія джерела становить $H(X) = -(0{,}42log_2 0{,}42 + 0{,}37log_2 0{,}37 + 0{,}18log_2 0{,}18 + 0{,}03log_2 0{,}03) = 0{,}5256 + 0{,}5307 + 0{,}4453 + 0{,}1518 = 1{,}6534$

Відносна різниця між $\bar{l}$ та H(X) становить: $\frac{\bar{l}-H(X)}{H(X)} * 100\% = \left(\frac{1{,}79-1{,}653}{1{,}653}\right) * 100\% = 8{,}259\%$

Тепер побудуємо код для кодування слів довжиною у два символи:

$$p(AA) = 0{,}42 * 0{,}42 = 0{,}1764$$
$$p(AB) = 0{,}42 * 0{,}37 = 0{,}1554$$
$$p(AC) = 0{,}42 * 0{,}18 = 0{,}0756$$
$$p(AD) = 0{,}42 * 0{,}03 = 0{,}0126$$
$$p(BA) = 0{,}37 * 0{,}42 = 0{,}1554$$
$$p(BB) = 0{,}37 * 0{,}37 = 0{,}1369$$
$$p(BC) = 0{,}37 * 0{,}18 = 0{,}0666$$
$$p(BD) = 0{,}37 * 0{,}03 = 0{,}0111$$
$$p(CA) = 0{,}18 * 0{,}42 = 0{,}0756$$
$$p(CB) = 0{,}18 * 0{,}37 = 0{,}0666$$
$$p(CC) = 0{,}18 * 0{,}18 = 0{,}0324$$
$$p(CD) = 0{,}18 * 0{,}03 = 0{,}0054$$
$$p(DA) = 0{,}03 * 0{,}42 = 0{,}0126$$
$$p(DB) = 0{,}03 * 0{,}37 = 0{,}0111$$
$$p(DC) = 0{,}03 * 0{,}18 = 0{,}0054$$
$$p(DD) = 0{,}03 * 0{,}03 = 0{,}0009$$

Символ	Ймовірність p_i	1	2	3	4	5	6	7	8	Код	Довжина коду l_i
AA	0.1764	0	0							00	2
AB	0.1554		1	0						010	3
BA	0.1554			1						011	3
BB	0.1369	1	0	0						100	3
AC	0.0756			1	0					1010	4
CA	0.0756				1					1011	4
BC	0.0666		1	0	0					1100	4
CB	0.0666				1					1101	4
CC	0.0324			1	0	0				11100	5
AD	0.0126					1				11101	5
DA	0.0126				1	0	0			111100	6
BD	0.0111						1			111101	6
DB	0.0111					1	0			111110	6
CD	0.0054						1	0		1111110	7
DC	0.0054							1	0	11111110	8
DD	0.0009								1	11111111	8

Обчислимо середню довжину коду $\bar{l} = (0{,}1764 * 2 + 0{,}1554 * 3 + 0{,}1554 * 3 + 0{,}1369 * 3 + 0{,}0756 * 4 + 0{,}0756 * 4 + 0{,}0666 * 4 + 0{,}0666 * 4 + 0{,}0324 * 5 + 0{,}0126 * 5 + 0{,}0126 * 6 + 0{,}0111 * 6 + 0{,}0111 * 6 + 0{,}0054 * 7 + 0{,}0054 * 8 + 0{,}0009 * 8)/2 =$

$= (0{,}3528 + 0{,}4662 + 0{,}4662 + 0{,}4107 + 0{,}3024 + 0{,}3024 + 0{,}2664 + 0{,}2664 + 0{,}162 + 0{,}063 + 0{,}0756 + 0{,}0666 + 0{,}0666 + 0{,}0378 + 0{,}0432 + 0{,}0072)/2 =$

$= \frac{3{,}3555}{2} = 1.677775$

Відносна різниця між $\bar{l}$ та H(X) становить: $\frac{\bar{l} - H(X)}{H(X)} * 100\% = \left(\frac{1{,}67 - 1{,}653}{1{,}653}\right) * 100\% = 1.028\%$

Закодуємо фрагмент повідомлення $BBBCBABBBAACCBCCAAACABBABADAAA$

Кодом для кодування поодиноких символів:

$B\ \ B\ \ B\ \ C\ \ B\ A\ B\ B\ \ B\ A\ A\ \ C\ \ \ C\ \ \ B\ \ C\ \ \ C\ A\ A\ A\ C\ \ A\ B\ \ B\ A\ B\ A\ D\ \ A\ A\ A$

10 10 10 110 10 0 10 10 10 0 0 110 110 10 110 110 0 0 0 110 0 10 10 0 10 0 111 0 0 0

Довжина коду – 55 символів.

Кодом для кодування слів довжиною у два символи:

$BB\ \ BC\ \ \ BA\ \ BB\ \ BA\ \ \ AC\ \ \ \ CB\ \ \ CC\ \ \ AA\ \ AC\ \ AB\ \ BA\ \ BA\ \ \ \ DA\ \ \ AA$

100 1100 011 100 011 1010 1101 11100 00 1010 010 011 011 111100 00

Довжина коду – 52 символи.

Ефективнішим є кодування кодом для слів довжиною у два символи.

Завдання 4.3. Алфавіт марковського дискретного джерела інформації, що має глибину пам'яті $h = 1$, складається з трьох символів: $X = \{A, B, C\}$. Значення умовних ймовірностей виникнення символів наступні:

$$\begin{pmatrix} 0.01 & 0.75 & 0.24 \\ 0.33 & 0.22 & 0.45 \\ 0.01 & 0.83 & 0.16 \end{pmatrix}$$

1. Побудувати нерівномірні ефективні коди за алгоритмом Шеннона-Фано або Хаффмена для кодування поодиноких символів джерела та слів довжиною у два символи.
2. Побудувати марковський алгоритм для кодування символів джерела.

3. Оцінити та порівняти ефективність отриманих кодів та марковського алгоритму.
4. Побудованими кодами закодувати фрагмент повідомлення $ACBBABCBBABCCBABCBCB$ довжиною у 20 символів, що був згенерований джерелом.

Оскільки глибина пам'яті h = 1, то кількість S станів джерела дорівнює потужності його алфавіту, тобто S = 3. Для кожного стану, який визначений попереднім символом на виході джерела будуємо нерівномірний код. Застосовуємо методику Шеннона-Фано. Отримаємо коди:

Після A

Символ	Ймовірність	1	2	Код
B	0.75	0		0
C	0.24	1	0	10
A	0.01		1	11

Середня довжина коду $\overline{l_A} = 0{,}75 * 1 + 0{,}24 * 2 + 0{,}01 * 2 = 0{,}75 + 0{,}48 + 0{,}02 = 1{,}25$

Після B

Символ	Ймовірність	1	2	Код
C	0.45	0		0
A	0.33	1	0	10
B	0.22		1	11

Середня довжина коду $\overline{l_B} = 0{,}45 * 1 + 0{,}33 * 2 + 0{,}22 * 2 = 0{,}45 + 0{,}66 + 0{,}44 = 1{,}55$

Після C

Символ	Ймовірність	1	2	Код
B	0.83	0		0
C	0.16	1	0	10
A	0.01		1	11

Середня довжина коду $\overline{l_C} = 0{,}83 * 1 + 0{,}16 * 2 + 0{,}01 * 2 = 0{,}83 + 0{,}32 + 0{,}02 = 1{,}17$

Середню довжину коду $\bar{l}$ можна знайти за формулою

$$\bar{l} = \overline{l_A} * p(A) + \overline{l_B} * p(B) + \overline{l_C} * p(C)$$

де $p(A), p(B), p(C)$ – безумовні ймовірності появи символів A, B, C на виході джерела.

Ці ймовірності знаходимо розв’язавши систему

$$\begin{cases} p(A) = p(A)p(A|A) + p(B)p(A|B) + p(C)p(A|C) \\ p(B) = p(A)p(B|A) + p(B)p(B|B) + p(C)p(B|C) \\ 1 = p(A) + p(B) + p(C) \end{cases}$$

Розв’язком цієї системи є $p(A) = 0.159, p(B) = 0.466, p(C) = 0.375$

Тоді $\bar{l} = 1{,}25 * 0{,}159 + 1{,}55 * 0{,}466 + 1{,}17 * 0{,}375 = 0{,}199 + 0{,}722 + 0{,}438 = 1{,}36$

Ентропію марківського джерела з глибиною пам’яті $h = 1$ обчислюється за формулою

$$H_{mem}^{h=1}(X) = -\sum_i p(x_i) \sum_j p(x_j|x_i) log_2 p(x_j|x_i)$$

$$\begin{aligned} H(X|x_0) &= 0{,}01log_2(0{,}01) + 0{,}75log_2(0{,}75) + 0{,}24log_2(0{,}24) + \\ &= -0{,}066 + -0{,}311 + -0{,}494 += -0{,}872 \\ H(X|x_1) &= 0{,}33log_2(0{,}33) + 0{,}22log_2(0{,}22) + 0{,}45log_2(0{,}45) + \\ &= -0{,}528 + -0{,}481 + -0{,}518 += -1{,}527 \\ H(X|x_2) &= 0{,}01log_2(0{,}01) + 0{,}83log_2(0{,}83) + 0{,}16log_2(0{,}16) + \\ &= -0{,}066 + -0{,}223 + -0{,}423 += -0{,}713 \\ H_{mem}^{h=1}(X) &= -(0{,}159 * -0{,}8720{,}466 * -1{,}5270{,}375 * -0{,}713) \\ &= -(-0{,}139 + -0{,}712 + -0{,}267+) = 1{,}117 \end{aligned}$$

Відносна різниця між $\bar{l}$ та $H_{mem}^{h=1}$(X) становить: $\frac{\bar{l} - H_{mem}^{h=1}(X)}{H_{mem}^{h=1}(X)} * 100\% = \left(\frac{1{,}36 - 1{,}117}{1{,}117}\right) * 100\% = 21{,}694\%$

Також знайдемо безумовні коди, для кодування першого символу в повідомленні

Символ	Ймовірність	1	2	Код
B	0.466	0		0
C	0.375	1	0	10
A	0.159		1	11

Та побудуємо код для кодування слів довжиною у два символи:

$$\begin{aligned} p(AA) &= 0{,}159 * 0{,}159 = 0{,}025 \\ p(AB) &= 0{,}159 * 0{,}466 = 0{,}074 \\ p(AC) &= 0{,}159 * 0{,}375 = 0{,}06 \\ p(BA) &= 0{,}466 * 0{,}159 = 0{,}074 \\ p(BB) &= 0{,}466 * 0{,}466 = 0{,}217 \\ p(BC) &= 0{,}466 * 0{,}375 = 0{,}175 \\ p(CA) &= 0{,}375 * 0{,}159 = 0{,}06 \\ p(CB) &= 0{,}375 * 0{,}466 = 0{,}175 \\ p(CC) &= 0{,}375 * 0{,}375 = 0{,}14 \end{aligned}$$

Символ	Ймовірність p_i	1	2	3	4	Код	Довжина коду l_i
BB	0.217	0	0			00	2
BC	0.175		1	0		010	3
CB	0.175			1		011	3
CC	0.14	1	0	0		100	3
BA	0.074			1		101	3
AB	0.074		1	0	0	1100	4
CA	0.06				1	1101	4
AC	0.06			1	0	1110	4
AA	0.0254				1	1111	4

Застосовуючи отримані коди, закодуємо фрагмент повідомлення $ACBBABCBBABCCBABCBCB$

Безумовним кодом для кодування поодиноких символів:

$$A\ C\ B\ B\ A\ B\ C\ B\ B\ A\ B\ C\ C\ B\ A\ B\ C\ B\ C\ B$$

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

Довжина коду: 30

Кодом для кодування слів довжиною у два символи:

$$AC\ \ BB\ \ AB\ \ CB\ \ BA\ \ BC\ \ CB\ \ AB\ CB\ CB$$

1110 00 1100 011 101 010 011 1100 011 011

Довжина коду: 32

Із застосуванням марківського алгоритму

$$A\ \ C\ B\ \ B\ \ A\ \ B\ C\ B\ B\ \ A\ \ B\ C\ \ C\ B\ \ A\ \ B\ C\ B\ C\ B$$

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

Довжина коду: 28

В цьому випадку марківський алгоритм виявився найефективнішим.