



Лекція 3*2. Практичні прийоми оцінювання ефективності коду.

1. Умова оптимальності коду. Теорема Фано.
2. Приклади оцінювання ефективності коду.



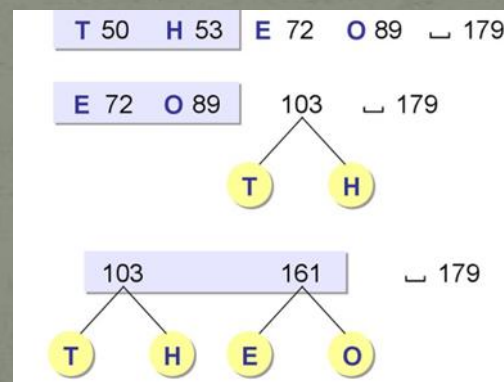
Алгоритм Шеннона-Фано



К. Шеннон и Р. Фано

	%		%		%
A	8.1	K	0.4	U	2.4
B	1.4	L	3.4	V	0.9
C	2.7	M	2.5	W	1.5
D	3.9	N	7.2	X	0.2
E	13.0	O	7.9	Y	1.9
F	2.9	P	2.0	Z	0.1
G	2.0	Q	0.2		

алгоритм стиснення даних Шеннона - Фано



Девід Хаффман

1. Умова
оптимальності
коду.
Теорема Фано.

Якщо дано джерело з обсягом алфавіту N та ентропією $H(X)$, то, застосовуючи код з основою m , можливо забезпечити середнє число двійкових символів на літеру алфавіту джерела, що лежить у межах

теорема Фано

$$\frac{H(X)}{\log_2 N} \leq n_{\text{сеп}} < \frac{H(X)}{\log_2 N} + 1.$$

Зокрема, при $N=2$ (3.2) набуває вигляду

$$H(X) \leq n_{\text{сеп}} < H(X) + 1. \quad (3.3)$$

Іншими словами, середня кількість двійкових елементів коду, що відображає символи алфавіту джерела, принаймні дорівнює ентропії джерела і може перевищувати її на одиницю. Якщо ймовірність кодових слів $p(x_k)$ задовольняє умові

$$p(x_k) = N^{-n(x_k)}.$$

можливо побудувати оптимальний код, у якого

$$n_{\text{сеп}} = \frac{H(X)}{\log_2 N}.$$

- 2. Приклади оцінювання ефективності коду.

Надмірність:

Приклад 3.1. Джерело має алфавіт із чотирьох символів А, Б, В, Г із ймовірностями $p(A)=0,5$; $p(B)=0,25$; $p(V)=p(\Gamma)=0,125$. Визначити надмірність повідомлень, складених із цього алфавіту.

Рішення. Ентропія джерела

$$\begin{aligned} H(X) &= -\sum_{i=1}^n p(x_i) \log_2 p(x_i) = \\ &= -0,5 \log_2 0,5 - 0,25 \log_2 0,25 - 0,125 \log_2 0,125 - 0,125 \log_2 0,125 = \\ &= 0,5 + 0,5 + 0,375 + 0,375 = 1,75 \text{ біт/символ} \end{aligned}$$

Максимальна ентропія $H_{\max} = \log_2 4 = 2$.

Надмірність повідомлень складе

$$D = \frac{H_{\max}(X) - H(X)}{H_{\max}(X)} = \frac{2 - 1,75}{2} = 0,125. \blacksquare$$

Приклад 3.2. Відносні частоти появи букв українського алфавіту у зв'язному тексті відповідно до [5] зведені в табл. 3.1.

Таблиця 3.1 – Відносні частоти вживання букв української алфавіту

Літера	Частота	Літера	Частота	Літера	Частота	Літера	Частота
-	0,138	Е	0,042	Я	0,019	Ю	0,008
О	0,086	С	0,037	Ь	0,016	Ж	0,007
Н	0,068	К	0,033	Б	0,013	Ш	0,005
А	0,064	М	0,029	Г	0,013	Є	0,005
И	0,055	У	0,027	Ч	0,011	Щ	0,004
В	0,046	Д	0,027	Х	0,011	Ф	0,003
Т	0,045	Л	0,027	Ї	0,010	Г	<0,001
І	0,044	П	0,025	Ц	0,010		
Р	0,043	З	0,020	Й	0,009		

Визначити ентропію та середню надмірність тексту українською мовою.

Рішення. Ентропія текстового повідомлення:

$$\begin{aligned}
 H(X) &= -\sum_{i=1}^{34} p(x_i) \log_2 p(x_i) = \\
 &= -0,138 \log_2 0,138 - 0,086 \log_2 0,086 - \dots - 0,003 \log_2 0,003 = \\
 &= 4,5356 \text{ біт/символ}
 \end{aligned}$$

Максимальна ентропія $H_{\max} = \log_2 34 = 5,0874$.

Надмірність повідомлень складе

$$D = \frac{H_{\max}(X) - H(X)}{H_{\max}(X)} = \frac{5,0874 - 4,5356}{5,0874} = 0,1085. \blacksquare$$

Усунення надмірності може досягатися використанням саме нерівномірних кодів, які враховують ймовірність появи літер у повідомленні. Літери, що мають велику ймовірність, кодуються більш короткими кодовими послідовностями, а більш довгі комбінації присвоюються рідкісними літерами.

Приклад 3.3. Розглянемо джерело з прикладу 3.1, що має алфавіт з чотирьох символів А, Б, В, Г з ймовірностями $p(A)=0,5$; $p(B)=0,25$; $p(V)=p(\Gamma)=0,125$. Для передачі каналом використовується нерівномірний код $A \rightarrow 0$, $B \rightarrow 10$, $V \rightarrow 110$, $\Gamma \rightarrow 111$.

Визначити середню кількість двійкових символів, що припадають на один символ джерела. Порівняти з ентропією джерела.

Характеристики нерівномірного коду

Символ x_i	$x_1 = A$	$x_2 = B$	$x_3 = V$	$x_4 = \Gamma$
Ймовірність p_i	0,5	0,25	0,125	0,125
Код	0	10	110	111
Число біт n_i	1	2	3	3

Середня кількість двійкових символів, що припадають на один символ джерела:

$$n_{\text{сер}} = \sum_{i=1}^4 p_i \cdot n_i = 0,5 \cdot 1 + 0,25 \cdot 2 + 0,125 \cdot 3 + 0,125 \cdot 3 = 1,75 \text{ біт}.$$

Таким чином, середня кількість двійкових розрядів, що припадають на один символ, що передається в канал, дорівнює ентропії джерела $H(X)=n_{\text{ср}}=1,75$, тобто. для даного джерела нерівномірний код виявляється більш економічним, ніж рівномірний і оптимальним. При цьому коефіцієнт стиснення

$$\mu = \frac{\log_2 N}{n_{\text{ср}}} = \frac{\log_2 4}{1,75} = \frac{2}{1,75} = 1,1429,$$

тобто швидкість передачі по каналу зв'язку при використанні нерівномірного кодування може бути в 1,1429 разів більше ніж при рівномірному кодуванні. ■

Приклад 3.4. Символи українського алфавіту закодовані відповідно до табл. 3.3. (принципи побудови кодової таблиці будуть розглянуті нижче). Визначити середню кількість двійкових символів, що припадають на один символ джерела. Порівняти з ентропією джерела.

Розв'язок

Таблиця 3.3 – Нерівномірне кодування букв українського алфавіту

Літера	Ймовірність p_i	Код	Число біт n_i	Літера	Ймовірність p_i	Код	Число біт n_i
-	0,138	101	3	З	0,020	110111	6
О	0,086	1111	4	Я	0,019	110100	6
Н	0,068	1001	4	Ь	0,016	011110	6
А	0,064	1000	4	Б	0,013	010100	6
И	0,055	0110	4	Г	0,013	001001	6
В	0,046	0011	4	Ч	0,011	001000	6
Т	0,045	0001	4	Х	0,011	1101101	7
І	0,044	0000	4	Ї	0,010	1101100	7
Р	0,043	11101	5	Ц	0,010	1101011	7
Е	0,042	11100	5	Й	0,009	0111111	7
С	0,037	11001	5	Ю	0,008	0111110	7
К	0,033	11000	5	Ж	0,007	0101011	7
М	0,029	01110	5	Ш	0,005	11010101	8
У	0,027	01001	5	Є	0,005	11010100	8
Д	0,027	01000	5	Щ	0,004	01010101	8
Л	0,027	01011	5	Ф	0,003	010101001	9
П	0,025	00101	5	Г	0,000	010101000	9

Рішення. Середня кількість двійкових символів, що припадають на один символ джерела

$$n_{сер} = \sum_{i=1}^{34} p_i \cdot n_i = 0,138 \cdot 3 + 0,086 \cdot 4 + \dots + 0,002 \cdot 9 + 0,000 \cdot 9 = 4,5720.$$

Таким чином, середня кількість двійкових розрядів на один символ, що передається в канал, $n_{сер}=4,572$, що дещо більше, ніж визначена в прикладі 3.2. ентропії джерела $H(X)=4,5356$. Однак $n_{сер}$ менше максимальної ентропії $H_{max}=\log_2 34=5,0874$, тобто для зазначеного джерела нерівномірний код виявляється більш економічним, ніж рівномірний. При цьому коефіцієнт стиснення

$$\mu = \frac{\log_2 N}{n_{сер}} = \frac{\log_2 34}{4,572} = \frac{5,0875}{4,572} = 1,1128,$$

тобто швидкість передачі по каналу зв'язку при використанні нерівномірного кодування може бути в 1,1128 разів більше ніж при рівномірному кодуванні.

Приклад 3.5. Нехай ймовірності появи кожного з дев'яти повідомлень зведено в табл. 3.5. Побудувати код Шеннона-Фано, визначити максимальну ентропію H_{max} , ентропію джерела $H(X)$, середнє число біт на символ $n_{сер}$, коефіцієнт стиснення μ .

Рішення. Побудову коду зведемо у табл. 3.5.

СИМВОЛ x_i	$p(x_i)$	Розбиття повідомлень на групи					Код	$n(x_i)$	$n(x_k) \cdot p(x_i)$
x_1	0,35	1	1				11	2	0,7
x_2	0,15	1	0				10	2	0,3
x_3	0,13	0	0	1			001	3	0,39
x_4	0,09	0	0	0			000	3	0,27
x_5	0,09	0	1	1	1		0111	4	0,36
x_6	0,08	0	1	1	0		0110	4	0,32
x_7	0,05	0	1	0	0		0100	4	0,2
x_8	0,04	0	1	0	1	1	01011	5	0,2
x_9	0,02	0	1	0	1	0	01010	5	0,1

Доведемо, що знайдений код близький до оптимального. Ентропія джерела повідомлення:

$$H(X) = - \sum_{i=1}^9 p(x_i) \log_2 p(x_i) = - (0,35 \log_2 0,35 + 0,15 \log_2 0,15 + 0,13 \log_2 0,13 + \\ + 0,09 \log_2 0,09 + 0,09 \log_2 0,09 + 0,08 \log_2 0,08 + 0,05 \log_2 0,05 + \\ + 0,04 \log_2 0,04 + 0,02 \log_2 0,02) = 2,75 \text{ біт / повідомл.}$$

Середня довжина кодового слова або середня кількість біт на символ

$$n_{cp} = \sum_{i=1}^9 p(x_i) n(x_i) = 0,7 + 0,3 + 0,39 + 0,27 + 0,36 + 0,32 + 0,2 + 0,2 + 0,1 = 2,84 \text{ біт / сим.}$$

$H(X) \approx n_{cp}$, отже, отриманий код задовольняє умові Фано (3.3) і дуже близький до оптимального, проте не досягає лівої межі (3.3).

Максимальна ентропія $H_{\max} = \log_2 N = \log_2 9 = 3,1699$.

$$\text{Коефіцієнт стиснення } \mu = \frac{\log_2 N}{n_{cp}} = \frac{\log_2 9}{2,84} = \frac{3,1699}{2,84} = 1,1162. \blacksquare$$

Приклад 3.6. Нехай алфавіт джерела містить вісім елементів: А, Б, В, Р, Д, Е, Ж, З, що з'являються з ймовірностями: $p(A)=0,16$; $p(B)=0,11$; $p(V)=0,09$; $p(\Gamma)=0,03$; $p(\Gamma)=0,39$; $p(D)=0,13$; $p(E)=0,03$; $p(\epsilon)=0,06$. Побудувати код Шеннона-Фано і Гаффмана, визначити максимальну ентропію $H_{\max}$, ентропію джерела $H(X)$, середнє число біт на символ $n_{\text{сер}}$, коефіцієнт стиснення μ . Порівняти ефективність кодів Шеннона-Фано та Гаффмана.

Рішення. Розташуємо літери алфавіту повідомлень у порядку зменшення ймовірностей і надамо їм для зручності позначення $x_1, \dots, x_8$. x_1 відповідає літері Г, $x_2 \sim A$, $x_3 \sim D$, $x_4 \sim B$, $x_5 \sim V$, $x_6 \sim \epsilon$, $x_7 \sim E$, $x_8 \sim \Gamma$.

Побудову коду Шеннона-Фано зведемо у табл. 3.6.

x_i	$p(x_i)$	Розбиття повідомлень на підгрупи					Код	$n(x_i)$	$n(x_k) \cdot p(x_i)$
x_1	0,39	1	1				1 1	2	0,78
x_2	0,16	1	0				1 0	2	0,32
x_3	0,13	0	1	1			0 1 1	3	0,39
x_4	0,11	0	1	0			0 1 0	3	0,33
x_5	0,09	0	0	0			0 0 0	3	0,27
x_6	0,06	0	0	1	1		0 0 1 1	4	0,24
x_7	0,03	0	0	1	0	1	0 0 1 0 1	5	0,15
x_8	0,03	0	0	1	0	0	0 0 1 0 0	5	0,15

Ентропія повідомлень

$$H(X) = -\sum_{i=1}^8 p(x_i) \log_2 p(x_i) = -(0,39 \log_2 0,39 + 0,16 \log_2 0,16 + 0,13 \log_2 0,13 + 0,11 \log_2 0,11 + 0,09 \log_2 0,09 + 0,06 \log_2 0,06 + 0,03 \log_2 0,03 + 0,03 \log_2 0,03 = 2,5455 \text{ біт / повідомл.}$$

Середня кількість біт на символ

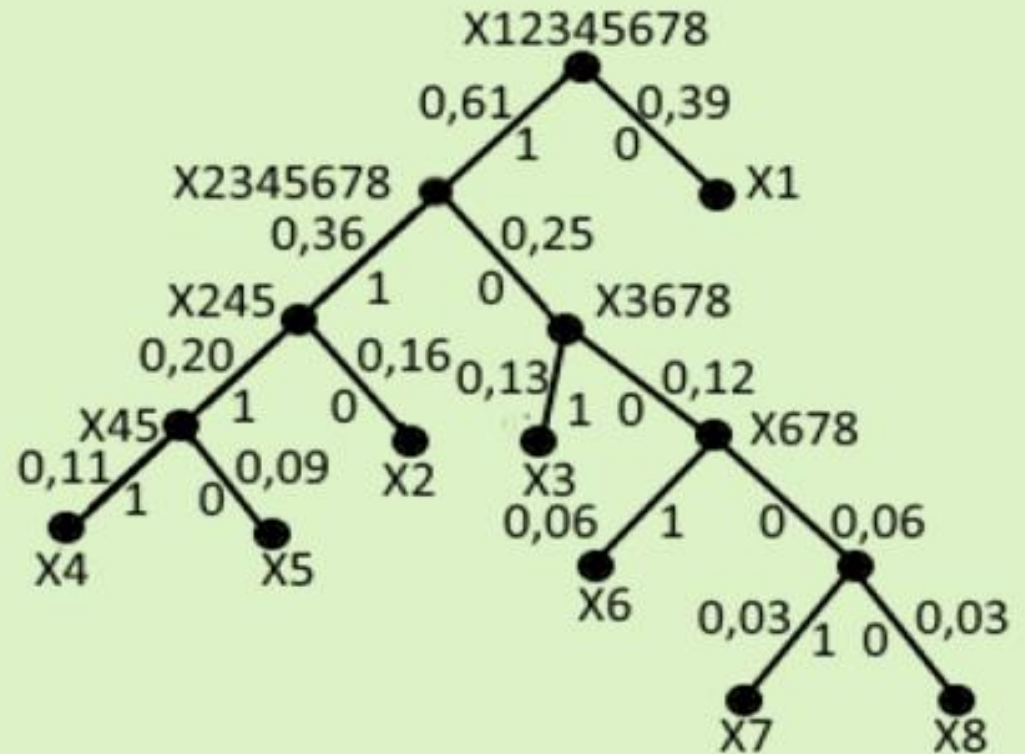
$$n_{\text{сери}} = \sum_{i=1}^8 p(x_i) n(x_i) = 0,78 + 0,32 + 0,39 + 0,33 + 0,27 + 0,15 + 0,15 = 2,63 \text{ біт/симв.}$$

Максимальна ентропія $H_{\max} = \log_2 N = \log_2 8 = 3$.

Коефіцієнт стиснення $\mu = \frac{\log_2 N}{n_{\text{ср}}} = \frac{\log_2 8}{2,63} = \frac{3}{2,63} = 1,1407$.

Побудову коду Гаффмана зведемо в табл. 3.7 і збудуємо кодове дерево

Крок 1		2		3		4		5		6		7	
x_i	$p(x_i)$	x_i	$p(x_i)$	x_i	$p(x_i)$	x_i	$p(x_i)$	x_i	$p(x_i)$	x_i	$p(x_i)$	x_i	$p(x_i)$
x_1	0,39	x_1	0,39	x_1	0,39	x_1	0,39	x_1	0,39	x_1	0,39	$x_{2453678}$	0,61
x_2	0,16	x_2	0,16	x_2	0,16	x_{45}	0,20	x_{3678}	0,25	x_{245}	0,36	x_1	0,39
x_3	0,13	x_3	0,13	x_3	0,13	x_2	0,16	x_{45}	0,20	x_{3678}	0,25		
x_4	0,11	x_4	0,11	x_{678}	0,12	x_3	0,13	x_2	0,16				
x_5	0,09	x_5	0,09	x_4	0,11	x_{678}	0,12						
x_6	0,06	x_6	0,06	x_5	0,09								
x_7	0,03	x_{78}	0,06										
x_8	0,03												



Кодові символи, одержані з табл. 3.7 та рисунку кодового дерева мають такий вигляд:

Літера	Г	А	Д	Б	В	Є	Е	Г
x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Код	0	110	101	1111	1110	1001	10001	10000

Для коду Гаффмана середня кількість біт на символ:

$$n_{серх} = \sum_{i=1}^8 p(x_i) n(x_i) = 0,39 + 0,48 + 0,39 + 0,44 + 0,36 + 0,24 + 0,15 + 0,15 = 2,60 \text{ біт/сим.}$$

$$\text{Коефіцієнт стиснення } \mu = \frac{\log_2 N}{n_{cp}} = \frac{\log_2 8}{2,60} = \frac{3}{2,60} = 1,1538.$$

Аналіз отриманих результатів показує, що відоме з теорії співвідношення $H(X) \leq n_{серх} \leq n_{серш} \leq H_{\max}$ виконується.