

Практична робота №1

Кількість інформації та ентропія

Мета роботи: Вивчити поняття кількості інформації та ентропії, опанувати програмні засоби їх розрахунку в MATLAB.

1.1. Теоретичні відомості

Дискретне джерело інформації - це джерело, яке виробляє дискретні повідомлення, які складаються з кінцеві безлічі символів. Множина $A = \{a_1, a_2, \dots, a_m\}$ всіх можливих символів $a_1, a_2, \dots, a_m$ називається алфавітом (ансамблем) джерела, а величина m - обсягом алфавіту. Кожний символ a_i виробляється з ймовірністю $p(a_i)$, причому $\sum_{i=1}^m p(a_i) = 1$, оскільки всі символи (стани) джерела становлять повну групу випадкових подій.

Кількість інформації $I(a_i)$, що міститься у будь-якому з незалежних символів a_i , визначається виразом

$$I(a_i) = -\log_v p(a_i), \quad (1.1)$$

де v - основа логарифму.

Логарифмічна міра (1.1) має такі характеристики.

1. Кількість інформації $I(a_i)$ - величина невід'ємна і спадна зі зростанням $p(a_i)$.

2. Кількість інформації, що міститься в достовірному символі (передається з ймовірністю $p(a_i) = 1$), дорівнює нулю

$$I(a_i) = -\log_v 1 = 0.$$

3. Кількість інформації, що міститься в будь-якому числі N незалежних символів, дорівнює сумі кількостей інформації, що містяться в кожному з них

$$\begin{aligned} I(a_1, a_2, \dots, a_N) &= -\log_v p(a_1, a_2, \dots, a_N) = \\ &= -\log_v \prod_{i=1}^N p(a_i) = -\sum_{i=1}^N \log_v p(a_i) = \sum_{i=1}^N I(a_i), \end{aligned}$$

де $p(a_1, a_2, \dots, a_N) = \prod_{i=1}^N p(a_i)$ - спільна ймовірність появи незалежних символів $a_1, a_2, \dots, a_N$.

4. Міра кількості інформації (1.1) не враховує якісний зміст повідомлення, що передається (важливість для одержувача, емоційне забарвлення, можливі наслідки і т.п.), а відображає лише ступінь його невизначеності.

5. Одиниця виміру кількості інформації залежить від вибору основи логарифму v :

– при $v=10$ кількість інформації $I(a_i) = -\lg p(a_i)$ вимірюється в десяткових одиницях - дитах або Хартлі (ця одиниця зараз практично не використовується);

– при $v=e \approx 2,71$ кількість інформації $I(a_i) = -\ln p(a_i)$ вимірюється в натуральних одиницях - натах (ця одиниця зазвичай використовується при теоретичних дослідженнях);

– при $v=2$ кількість інформації $I(a_i) = -\log p(a_i)$ вимірюється у двійкових одиницях - бітах (від англ. bit - Binary digit, ця одиниця широко використовується і за умовчанням основа логарифму опускається).

6. Логарифмічна міра (1.1) з основою логарифму $v=10$ для випадку, коли всі символи рівноймовірні і $p(a_i) = \frac{1}{m}$, вперше запропонована англійським вченим Р. Л. Хартлі в 1928 році.

Ентропія дискретного джерела інформації з алфавітом $A = \{a_i\}$ – це середня кількість інформації, що припадає на один (будь-який) символ a_i , що виробляється джерелом. Ентропія $H(A)$ знаходиться як математичне сподівання дискретної випадкової величини $I(a_i)$, що визначає кількість інформації, що міститься в одному випадково вибраному символі

$$H(A) = M\{I(a_i)\}, \quad (1.2)$$

де $M\{\cdot\}$ – знак математичного сподівання, що означає усереднення випадкової величини, що стоїть у фігурних дужках, по всіх її можливих реалізаціях.

Ентропія джерела незалежних символів згідно з (1.1), (1.2) визначається за формулою

$$H(A) = -\sum_{i=1}^N p(a_i) \log p(a_i). \quad (1.3)$$

Ентропія вимірюється в одиницях: біт на символ, біт на букву, біт на повідомлення.

Властивості ентропії:

1. Ентропія є величина, обмежена нерівностями

$$0 \leq H(A) \leq \log m. \quad (1.4)$$

2. Ентропія перетворюється на нуль, якщо стан джерела повністю визначено, тобто один символ має ймовірність, рівну одиниці, а ймовірності решти символів дорівнюють нулю.

3. Ентропія максимальна та дорівнює

$$H_0(A) = \log m. \quad (1.5)$$

якщо всі символи джерела рівноймовірні, тобто для всіх значень $i = 1, 2, \dots, m$ ймовірності становлять $p(a_i) = 1/m$.

4. Ентропія двійкового джерела з числом символів $m = 2$ при $p(a_1) = p$ і $p(a_2) = 1 - p$ згідно (1.3) визначається виразом

$$H(A) = -p \log p - (1 - p) \log(1 - p). \quad (1.6)$$

З формули (1.6) випливає, що ентропія двійкового джерела змінюється від нуля до одиниці, досягаючи максимуму за значення аргументу $p = 1 - p = 0,5$. Таким чином, 1 біт – це кількість інформації, що міститься в одному із двох рівноймовірних незалежних символів.

5. Ентропія об'єднання (або спільна ентропія) кількох статистично незалежних джерел з алфавітами $A_1, A_2, \dots, A_N$ дорівнює сумі ентропій цих джерел

$$H(A_1, A_2, \dots, A_N) = \sum_{k=1}^N H(A_k). \quad (1.7)$$

7. Ентропія характеризує середню невизначеність вибору одного символу з алфавіту джерела. При її визначенні використовують лише ймовірності символів, повністю ігноруючи їх змістовну сторону.

2. Ентропія як міра невизначеності узгоджується з експериментальними даними, отриманими щодо психологічних реакцій людини і, зокрема, реакції вибору. Так, встановлено, що час безпомилкової реакції на послідовність рівноймовірних подразників, що хаотично чергуються (наприклад, лампочки, що загоряються) зростає зі збільшенням їх кількості так само, як ентропія.

3. Термін «ентропія» в теорії інформації та вираз (1.3) для її обчислення вперше були введені в 1948 американським ученим К. Є. Шенноном. Він висловив твердження, а математиком Я.А. Хінчіним суворо доведено, що формула (1.3) є єдиним функціоналом, що задовольняє сформульованим властивостям.

1.2. Розрахункові завдання

Завдання 1.1. Побудувати графіки залежності кількості інформації від ймовірності символу за формулою (1.1) за різних значень v .

```
%Script_1_1
%Побудова графіків залежності кількості інформації
від ймовірності символу
clear all; close all; clc
%Значення ймовірності
p=0:.001:1;
%Значення кількості інформації
I1=-log2(p);
I2=-log(p);
I3=-log10(p);
%Побудова графіків
plot(p,I1,p,I2,p,I3)
```

```

grid
xlabel('Ймовірність символу p(a)')
ylabel('Кількість інформації I(a), біт')
title('Залежність кількості інформації від
ймовірності символу')
legend('\nu = 2', '\nu = e', '\nu = 10')
ylim([0 8])

```

Завдання 1.2. Побудувати графік залежності ентропії двійкового джерела від ймовірності символу за формулою (1.6).

```

%Script_1_2
%Побудова графіка ентропії двійкового джерела в
залежності від ймовірності символу
clear all; close all; clc
^Значення ймовірності
p=0:.001:1;
%Значення ентропії
H=-p.*log2(p)-(1-p).*log2(1-p);
%Побудова графіка
plot(p,H)
grid
xlabel('Вірогідність символу p')
ylabel('Ентропія джерела H(A), біт/символ')
title('Ентропія двійкового джерела в залежності від
ймовірності символу')

```

Завдання 1.3. Дискретне джерело має обсяг алфавіту $m = 3$. Визначити ентропію джерела, якщо:

а) символи алфавіту рівноймовірні;

б) символи виробляються із ймовірностями $p(a_1) = 0,25$; $p(a_2) = 0,3$; $p(a_3) = 0,45$.

На яку величину зменшується ентропія у другому випадку?

```

%Script_1_3
%Розрахунок ентропії дискретного джерела без
пам'яті
clear all; close all; clc
%Завдання анонімної функції
%-----
fun=@(x) -x.*log2(x+(x==0));
%-----
%Завдання ймовірностей символів

```

```

p=[.25 .3 .45];
disp('Ймовірності символів:')
disp(p)
%Сума ймовірностей символів
S=sum(p);
disp('Сума ймовірностей:')
disp(s)
%Визначення обсягу алфавіту (довжини вектора)
m=length(p);
disp('Обсяг алфавіту:')
disp(m)
%Визначення максимальної ентропії
H0=log2(m);
disp('Максимальна ентропія, біт/символ:')
disp(H)
%Обчислення ентропії джерела
H=sum(fun(p));
disp('Ентропія джерела, біт/символ:')
disp(H)
%Визначення різниці ентропій
dH=H0-H;
disp('Різниця ентропій, біт/символ:')
disp(dH)

```

Завдання 1.4. Дискретне джерело має обсяг алфавіту $m = 4$. Визначити ентропію джерела, якщо:

- а) символи виробляються з однаковими ймовірностями;
- б) ймовірності символів дорівнюють $p(a_1) = 0,1$; $p(a_2) = 0,2$; $p(a_3) = 0,3$, $p(a_4) = 0,4$.

На яку величину зменшується ентропія у другому випадку?

Вказівка: використовувати Script_1_3.

Завдання 1.5. Дискретне джерело має обсяг алфавіту $m = 5$. Визначити ентропію джерела для наступних випадків:

- а) символи виробляються з однаковими ймовірностями;
- б) ймовірності символів $p(a_1) = 0,8$; $p(a_2) = 0,15$; $p(a_3) = 0,03$, $p(a_4) = 0,014$, $p(a_5) = 0,01$.

На яку величину зменшується ентропія у другому випадку?

Вказівка: використовувати Script_1_3.

Завдання 1.6. Побудувати графік функції $f[p(a_i)] = -p(a_i) \log p(a_i)$ і дослідити вплив ймовірності $p(a_i)$ появи окремого символу a_i на величину ентропії джерела в цілому. Дослідити функцію $f[p(a_i)]$ на екстремум.

```
%Script_1_6
%Побудова графіка одного доданку в залежності
від ймовірності символу
clear all; close all; clc
%Завдання анонімної функції
%-----
fun=@(x) -x.*log2(x+(x==0));
%-----
%Завдання значень ймовірності
p=0:001:1;
%Розрахунок значень ентропії
y=fun(p);
%Побудова графіка
plot(p,y)
grid
xlabel('Ймовірність символу p')
ylabel('Функція f(p), біт')
title('Графік функції f(p)')
%Визначення максимуму функції за графіком
[fm,im]=max(y);
disp('Визначення максимуму функції за графіком')
disp('Максимум функції, біт:')
disp(fm)
disp('Значення аргументу:')
disp(p(im))
%Чисельне визначення максимуму функції
[pm,fm]=fminbnd(@(x) x.*log2(x+(x==0)),.2,.7);
disp('Чисельне визначення максимуму функції')
disp('Максимум функції, біт:')
disp(-fm)
disp('Значення аргументу:')
disp(pm)
%Точне визначення максимуму функції
pm=1/exp(1);
fm=-pm*log2(pm);
disp('Точне визначення максимуму функції')
disp('Максимум функції:')
disp(fm)
```

```

disp('Значення аргументу:')
disp(pm)
%Позначення максимуму на графіку
line([0 pm],[fm fm],'LineStyle','--')
line([pm pm],[fm 0],'LineStyle','--')
hold on
plot(pm, fm, 'r. ')

```

1.3. Звіт та контрольні питання

Звіт з лабораторної роботи оформляється у вигляді документа MS Word і повинен містити:

1. Назва лабораторної роботи (шрифт Times New Roman).
2. Результати виконання всіх розрахункових завдань, включаючи лістинги m-файлів, що копіюються з вікна **Editor**, результати їх виконання, що копіюються з вікна **Command Window** (шрифт Courier New), а також отримані графіки, що копіюються з вікна **Figure** за командою **Edit | Copy Figure**.

Захист лабораторної роботи проводиться на основі поданого звіту та відповідей на контрольні питання з наступного списку:

1. Що таке дискретне джерело інформації?
2. Дайте визначення логарифмічного заходу кількості інформації $I(a)$.
3. Перерахуйте властивості логарифмічної міри кількості інформації.
4. При однаковому значенні ймовірності символу яке значення кількості інформації буде найбільшим, а яке буде найменшим - виражене в дитах, натах або в бітах?
5. Дайте визначення ентропії дискретного джерела з алфавітом $A = \{a_1, a_2, \dots, a_m\}$ як математичного сподівання випадкової дискретної величини $I(a_i)$.
6. Запишіть вираз для ентропії дискретного джерела статистично незалежних символів.
7. Перерахуйте властивості ентропії дискретного джерела статистично незалежних символів.
8. Запишіть вираз для ентропії дискретного джерела з об'ємом алфавіту m і незалежними рівноймовірними символами.
9. Запишіть вираз ентропії двійкового джерела незалежних символів.
10. При яких значеннях ймовірності появи i -го символу досягаються найбільш суттєві значення доданків виду $-p(a_i)\log p(a_i)$?