Основы теории информации и кодирования

Александра Игоревна Кононова / **illinc@mail.ru**
+7-985-148-32-64 (телефон), +7-977-977-97-29 (WhatsApp),
gitlab.com/illinc/raspisanie

МИЭТ

10 сентября 2021 г. — актуальную версию можно найти на
https://gitlab.com/illinc/otik

Предмет теории информации. Источник информации
Измерение информации
Вероятностная мера информации
Задачи

Регламент

См. `https://gitlab.com/illinc/otik/`

Дополнительные баллы:

1. бонусные задания л/р;
2. вычитка материала — $1-4$ балла за принятое замечание, $2-8$ за принятое исправление;
3. пополнение списка литературы — $(-1)-(+8)$ баллов.

Экзамен (оценка):

5	$86-100$
4	$70-85$
3	$50-69$
2	$0-49$

Консультации — см. `gitlab.com/illinc/raspisanie`

Теория информации и связь

Теория информации — математическая теория, посвящённая измерению информации, её потока, «размеров» канала связи и т. п., особенно применительно к средствам связи:

$$x \in X \quad \sim \quad I(x)$$

x — сообщение, $X = \{x, p(x)\}$ — источник (сл. процесс/сл. величина). Дискретное x может состоять из символов или быть отдельным символом.

Информация — нематериальная сущность, при помощи которой с любой точностью можно описывать реальные (материальные), виртуальные (возможные) и понятийные сущности.

$I(x)$:

1. **Новизна** (неизмеряемость в быту).
2. **Объёмный** (длина — измерение в технике).
3. **Вероятностный** (снятая неопределённость — измерение в ТИ).

Виды источников информации

По сообщениям:

- дискретные (цифровые)/непрерывные (аналоговые);
- дискретные: качественные/количественные.

Элемент качественной информации — **символ** $a \in A$ (множество A — алфавит); конечная последовательность символов — **слово** $x \in A^+$ (строка, фраза).

Источник символов алфавита A (можно прочитать строку):

1. стационарный (вероятность символа не зависит от времени/позиции: только от контекста) / нестационарный (при сдвиге вероятности меняются);
2. **марковский** источник — вероятность символа определяется состоянием; состояние изменяется после порождения символа (новое состояние однозначно определяется предыдущим и порождённым символом);
   марковский источник порядка m — вероятность символа на i-м шаге зависит от m предыдущих символов: $i-1, i-2, \ldots, i-m$;
3. **стационарный источник без памяти** — вероятность символа $a \in A$ постоянна (равна $p(a)$);
4. **равновероятный источник** — вероятность символа $a \in A$ постоянна и одинакова для всех символов (равна $\frac{1}{|A|}$);

равновероятный $\subseteq$ стационарный без памяти $\subseteq$ марковский $\subseteq$ стационарный

Кодирование

Кодирование — преобразование дискретной информации

$$x \in X = A_1^+ \quad \rightarrow \quad code(x) \in A_2^+$$

смена алфавита, **сжатие, защита от шума,** шифрование.

x — сообщение, исходный текст, исходная строка, блок;

X — источник сообщений;

A_1 — первичный алфавит (до преобразования);

A_2 — вторичный (алфавит конечного представления).

Обычно A_1 — байты, исходные тексты x — бинарные файлы.

Единица измерения информации

Бит — количество информации в сообщении, уменьшающем неопределённость знания в два раза.

Источник с двумя равновероятными состояниями — симметричная монета

.	?	2 возможных варианта
P	Решка	1 вариант

Неопределённость уменьшилась в 2 раза: $I(\mathrm{P}) = 1$ бит

..	Две симметричные монеты	
O.	Первая — вверх орлом	2 раза ($+1$ бит)
OP	Вторая — вверх решкой	2 раза ($+1$ бит)
	4 возможных варианта	$I(\mathrm{OP}) = 2$ бита

Требования к мере информации $I(x)$

1. $I(x) \geqslant 0$.
2. Вероятностный подход: $I(x) = f(p_x)$.
3. Объёмный подход: $I(x)$ монотонно связана с затратами на передачу
   - два равновероятных сообщения — 0 и 1 (1 бит),
     четыре — 00, 01, 10, 11 (2 бита) и т. д.:
     $f\left(\frac{1}{2}\right) = 1, \quad f\left(\frac{1}{4}\right) = 2, \quad f\left(\frac{1}{8}\right) = 3, \ldots$
   - затраты на передачу независимых сообщений складываются:
     $I(x_1, \ldots, x_n) = I(x_1) + \ldots + I(x_n)$

     при этом вероятности независимых событий умножаются
     $f(p_1 \times \ldots \times p_n) = f(p_1) + \ldots + f(p_n)$.

Энтропия и информация

1865 г. — Рудольф Клаузиус ввёл в статистическую физику понятие **энтропии** — меры уравновешенности [Дж/К].

1877 г. — Людвиг Больцман установил связь энтропии с вероятностью.

1901 г. — Макс Планк определил энтропию как $H = k \cdot \ln(\Omega)$, где k — коэффициент Больцмана [Дж/К].

1921 г. — Роналд Фишер ввёл термин «информация» (информация, которую можно извлечь из имеющихся данных, **имеет предел**).

1928 г. — Ральф Хартли — логарифмическая мера информации для **равновероятных** событий.

1948 г. — Клод Шеннон — вычисление количества информация и энтропии.

Основное соотношение между энтропией и информацией:

$$I + \frac{\log_2 e}{k} H = const \quad \text{[бит]} \qquad \left(\frac{dI}{dt} = -\frac{\log_2 e}{k} \frac{dH}{dt} \quad \text{[бит/с]} \right).$$

Теорема Шеннона для сжатия

Первая теорема Шеннона (для сжатия): $|code(X)| \geqslant I(X)$

ℕB: **усреднение по источнику X!**

При отсутствии помех средняя длина кода может быть сколь угодно близкой к средней информации сообщения.

Следствия:

1. не существует архиватора, который любой файл сжимает до 8 байт;
2. не существует архиватора, который любой блок из 9 байт сжимает до 8 байт.
3. не существует и такого архиватора, который любой блок из $N+1$ бит сжимает ровно до N бит, ни при каком N.

Формула Хартли для равновероятных событий

Источник X порождает N **равновероятных** сообщений x ($\forall x \in X: \; p(x) = p = \frac{1}{N}$).

$$I(x) = I(X) = I = \log_2 N = -\log_2(p) \quad \text{или } 2^I = N$$

где $I(x)$ — количество информации в сообщении x;
$I(X)$ — **среднее** кол-во информации в одном сообщении источника X.

Если $N = 2$, то $I = 1$ бит.

Подбрасывание монеты

..	4 варианта	2 бита

Угадывание слов по словарю

.....	175 слов	7,5 бит
.а.и.а	122 слова	6,9 бит
р.б.т.	4 слова	2 бита

Формула Шеннона для неравновероятных событий

Количество информации I в сообщении с вероятностью $p(x)$:

$$I(x) = -\log_2 p(x)$$

Свойства:

1. Неотрицательность: $I(x) \geqslant 0, x \in X$.
2. Монотонность: $x_1, x_2 \in X, p(x_1) \geqslant p(x_2) \rightarrow I(x_1) \leqslant I(x_2)$.
3. Аддитивность: для независимых сообщений $x_1, \ldots, x_n$
   $I(x_1, \ldots, x_n) = \sum_{i=1}^{n} I(x_i)$
4. Для равновероятных событий соответствует формуле Хартли.

Среднее количество информации дискретного источника $X = \{x, p(x)\}$:

$$I(X) = \sum_{x_i \in X} \Big(p(x_i) \cdot I(x_i) \Big) = - \sum_{x_i \in X} \Big(p(x_i) \cdot \log_2 p(x_i) \Big)$$

Задачи (равновероятный источник)

1. Найти количество информации в событии «три симметричные монеты выпали все вверх решкой».
2. Найти количество информации в источнике «три разные симметричные монеты».

Задачи (стационарный источник без памяти)

1. Найти количество информации в событии «две из трёх неразличимых симметричных монет выпали вверх решкой, третья — орлом».
2. Найти количество информации в источнике «три неразличимые симметричные монеты».

3. Найти количество информации в событии «из урны с 3 белыми и 5 чёрными шарами извлекли чёрный шар».
4. Найти количество информации в событии «из урны с 3 белыми и 5 чёрными шарами извлекли белый шар».
5. Найти количество информации в источнике «урна с 3 белыми и 5 чёрными шарами».

Задачи (стационарный источник с памятью)

1. Источник X генерирует последовательность подстрок «хрю» и «мяу» (с равной вероятностью), не разделяя их (например, «хрюхрюхрюмяухрюмяумяухрюмяумяу...»).
   Из случайного места последовательности (не обязательно с начала подстроки) читается три символа подряд (сообщение x).

   Найти количество информации в событии «$x =$ рюх».

2. Источник X аналогично генерирует посл-ть «ку» и «кукареку» (например, «кукукукукарекукукукукарекукукукарекукукукукарекукукукареку...»).
   Из случайного места посл-ти читается два (три) символа подряд (x).

   Найти количество информации в событиях:

   - $x =$ ка;
   - $x =$ кар;
   - $x =$ ку;
   - $x =$ ук;
   - $x =$ ек.

   Подсказка: основная проблема в том, что часть символов — одинаковые. Пусть они разные...

Продолжаем?

МИЭТ

www.miet.ru

Александра Игоревна Кононова / **illinc@mail.ru**
+7-985-148-32-64 (телефон), +7-977-977-97-29 (WhatsApp),
gitlab.com/illinc/raspisanie
https://gitlab.com/illinc/otik/