Синтаксический уровень

Р. Хартли первым ввел в теорию передачи информации методологию **«измерения количества информации».**

Р. Хартли считал, что информация, это «… группа физических символов – слов, точек, тире и т. п., имеющих по общему соглашению известный смысл для корреспондирующих сторон», то есть Хартли ставил перед собой задачу ввести какую-то меру для измерения кодированной информации.

Структурный подход не учитывает содержания сообщения и связан с подсчетом числа символов в нем, то есть с его длиной

Лекция 4. Количество информации

Пусть передаётся последовательность из **n** символов $\mathbf{a_1 a_2 a_3 \ldots a_n}$, каждый из которых принадлежит алфавиту $\mathbf{A_m}$, содержащему **m** символов. Чему равно число **K** различных вариантов таких последовательностей?

Количество информации, содержащееся в такой последовательности, Хартли предложил вычислять как логарифм числа K по основанию 2:

$I = \log_2 K$, где $K = m^n$

Количество информации, содержащееся в последовательности из **n** символов из алфавита $\mathbf{A_m}$, в соответствии с формулой Хартли равно

$$I = \log_2 (m^n) = n \log_2 m$$

Лекция 4. Количество информации

Полученная мера

$$I = \log_2 (m^n) = n \log_2 m$$

Удобна !!
Ею можно оперировать как числом

При наличии нескольких источников общее количество информации

$$I(n_1, n_2, \ldots, n_n) = I(n_1) + I(n_2) + \ldots + I(n_n)$$

Другое название меры Хартли – ***аддитивная*** **мера**, поскольку слово addition с английского переводится как суммирование.

$\log_2 K$ в теории информации также называют энтропией и обозначают символом H.

Информационная энтропия *– это мера неопределённости состояния некоторой случайной величины (физической системы) с конечным или счётным числом состояний.*

Случайная величина (с.в.) – это величина, которая в результате эксперимента или наблюдения принимает числовое значение, заранее неизвестно какое.

Итак, пусть X – случайная величина, которая может принимать N различных значений x_1, x_2, … x_N; если все значения с.в. X равновероятны, то энтропия (мера неопределённости) величины X равна:

$$H(X) = \log_2 N.$$

- ***Замечание 1.*** Хартли предполагал, что все символы алфавита $\mathbf{A_m}$ могут с равной вероятностью (частотой) встретиться в любом месте сообщения.

- ***Замечание 2.*** Любое сообщение длины n в алфавите $\mathbf{A_m}$ будет содержать одинаковое количество информации.

- ***Замечание 3.*** Если случайная величина (система) может находиться только в одном состоянии (N=1), то её энтропия равна 0. Фактически это уже не случайная величина. Неопределённость системы тем выше, чем больше число её возможных равновероятных состояний.

  Энтропия и количество информации измеряются в одних и тех же единицах – в битах.

P.S.

«Осмысленное» сообщение и сообщение, полученное из него *произвольной перестановкой символов*, будут содержать одинаковое количество информации. **???**

00111, 11001 и 10101 содержат одинаковое количество информации **!!!**

С помощью символов 0 и 1 кодируется информация в компьютере и при передаче в вычислительных сетях, т.е. алфавит состоит из двух символов {0 ; 1}; один символ и в этом случае содержит

$$I = \log_2 2 = 1 \text{ бит}$$

информации, поэтому сообщение длиной **n** символов в алфавите {0 ; 1} в соответствии с формулой Хартли будет содержать **n** бит информации.

Определение. 1 бит – это энтропия системы с двумя равновероятными состояниями.

При передаче сообщений в алфавите русского языка, состоящего из 33 букв, то количество информации, содержащееся в сообщении из **n** символов, вычисленное по формуле Хартли, равно

$$I = n_* \log_2 33 = n_* 5.0444 \text{ бит}$$

Английский алфавит содержит 26 букв, один символ содержит

$$\log_2 26 = 4.7 \text{ бит}$$

Пусть система X может находиться в двух состояниях x_1 и x_2 с равной вероятностью, т.е. N = 2; тогда её энтропия

$$H(X) = \log_2 2 = 1 \text{ бит.}$$

Определение. *Ответ на вопрос любой природы (любого характера) содержит 1 бит информации, если он с равной вероятностью может быть «да» или «нет».*

Задача 1. Некто задумал натуральное число в диапазоне от 1 до 32. Какое минимальное число вопросов надо задать, чтобы гарантированно угадать задуманное (выделенное) число. Ответы могут быть только «да» или «нет».

Решение. По формуле Хартли можно вычислить количество информации, которое необходимо получить для определения выделенного элемента x из множества целых чисел {1,2,3 ……, 32}. Для этого необходимо получить $H = Log_2\ 32 = 5$ бит информации. Вопросы надо задавать так, чтобы ответы на них были равновероятны. Тогда ответ на каждый такой вопрос будет приносить 1 бит информации.

Задача 2

Имеется 27 монет, из которых 26 настоящих и одна фальшивая - она легче. Каково минимальное число взвешиваний на рычажных весах.

Решение. По формуле Хартли количество информации, которое нужно получить для определения фальшивой монеты: оно равно

$$I = Log_2 27 = Log_2 (3^3) = 3\, Log_2 3 \text{ бит.}$$

Не зная стратегии взвешивания, мы определили **I.**

Если положить на чашки весов равное количество монет, то возможны три равновероятных исхода:

левая чашка тяжелее правой (Л > П);левая чашка легче правой (Л < П);левая чашка находится в равновесии с правой (Л = П);

«Рычажные весы» могут находиться в 3 равновероятных состояниях, т.е. одно взвешивание даёт $Log_2 3$ бит информации. Всего надо получить $I = 3\, Log_2 3$ бит информации, значит надо сделать **три** взвешивания для определения фальшивой монеты.

Синтаксический уровень

Представления получателя информации о наступлении того или иного события ***недостоверны и выражаются вероятностями***, с которыми он их ожидает.

Мера неопределенности зависит от указанных вероятностей, а количество информации в сообщении определяется тем, насколько данная мера уменьшается с получением сообщения.

Статистический подход
учитывает содержание информационного сообщения

Понятия теории вероятности

Достоверное *событие* – событие, которое обязательно наступит.

$$p(\Omega) = 1$$

Невозможным называют событие, которое никогда не произойдёт.

$$p(\varnothing) = 0$$

Вероятность события определяется как отношение числа благоприятных событию исходов опыта к общему числу исходов.

Частота события – эмпирическое приближение его вероятности.

Алфавит $\mathbf{A_m}$, состоящий из **m** символов. Обозначим через $\mathbf{p_i}$ вероятность (частоту) появления i-ого символа в любой позиции передаваемого сообщения, состоящего из **n** символов. Один i – ый символ алфавита несёт количество информации равное - $\boldsymbol{\log_2 (p_i)}$**.** Перед логарифмом стоит «минус» потому, что количество информации величина неотрицательная, а $\boldsymbol{\log_2(x) < 0}$ при 0<x<1.

Количество информации, приходящееся на один символ сообщения, равно среднему значению информации по всем символам алфавита $\mathbf{A_m}$:

$$-\sum_{i=1}^{m} p_i \log_2 p_i$$

Общее количество информации, содержащееся в сообщении из **n** символов равно:

$$I = - n_* \sum_{i=1}^{m} p_i \log_2 p_i$$

равно:

$$I = - n_* \log_2 p_i$$

Если все символы алфавита $\mathbf{A_m}$ появляются с равной вероятностью, то все

$\mathbf{p_i = p}$. Так как $\mathbf{\Sigma p_i = 1}$, то $\mathbf{p = 1/m}$.

Формула (слайд 12) в случае, когда все символы алфавита равновероятны, принимает вид

$$I = n \log_2 m$$

Вывод: формула Шеннона в случае, когда все символы алфавита равновероятны, переходит в формулу Хартли (слайд 2).

Количество энтропии **H** произвольной системы **X** (случайной величины), которая может находиться в **m** различных состояниях $x_1, x_2, \ldots x_m$ с вероятностями $p_1, p_2, p_3 \ldots p_m$, вычисленное по формуле Шеннона, равно

$$H(X) = -\sum_{i=1}^{m} p_i * \log_2 p_i$$

Шеннона, равно

$$H(X) = - * \log_2 p_i$$

При $p_1+p_2+p_3+\ldots+p_m = 1$. Если все p_i одинаковы, то все состояния системы X равновероятны; в этом случае $p_i = 1/m$, и эта формула переходит в формулу Хартли (слайд 4):

$$H(X) = \log_2 m$$

- ***Замечание.*** Количество энтропии системы (случайной величины) X не зависит от того, в каких конкретно состояниях $x_1, x_2, \ldots x_m$ может находиться система, но зависит от числа **m** этих состояний и от вероятностей $p_1, p_2, p_3 \ldots p_m$, с которыми система может находиться в этих состояниях. Это означает, что две системы, у которых число состояний одинаково, а вероятности этих состояний $p_1, p_2, p_3 \ldots p_m$ (с точностью до порядка перечисления), имеют равные энтропии.

Синтаксический уровень

Частотные вероятности русских букв

i	Символ	p_i	i	Символ	p_i	i	Символ	p_i
1	Пробел	0,175	13	К	0,028	25	Ч	0,012
2	О	0,090	14	М	0,026	26	Й	0,010
3	Е	0,072	15	Д	0,025	27	Х	0,009
4	Ё	0,072	16	П	0,023	28	Ж	0,007
5	А	0,062	17	У	0,021	29	Ю	0,006
6	И	0,062	18	Я	0,018	30	Ш	0,006
7	Т	0,053	19	Ы	0,016	31	Ц	0,004
8	Н	0,053	20	З	0,016	32	Щ	0,003
9	С	0,045	21	Ь	0,014	33	Э	0,003
10	Р	0,040	22	Ъ	0,014	34	Ф	0,002
11	В	0,038	23	Б	0,014			
12	Л	0,035	24	Г	0,013			

Количество информации по формуле Хартли

$I = \log_2 34 \approx 5$ **бит.**

Количество информации по формуле Шеннона

$I = H \approx 4{,}72$ ***бит.***

Оценки количества информации, полученные при *структурном* и *статистическом* подходах, не совпадают.

Теорема

Максимум энтропии H(X) достигается в том случае, когда все состояния системы ***равновероятны.***
Это означает, что

$$-\sum_{i=1}^{m} p_i \log_2 p_i \leq \log_2 m$$

- ≤

Среди всех систем с двумя состояниями наибольшая энтропия будет у системы с равновероятными состояниями, т.е. когда $p_1=p_2=1/2$.

Количество энтропии такой системы равно
$H(X) = -(1/2*\log_2(1/2) + 1/2*\log_2(1/2)) = -\log_2(1/2) = \log_2(2) = 1$

Это количество принимается за ***единицу измерения энтропии (информации)*** и называется 1 бит (1 bit

Пример. Вы хотите угадать количество очков, которое выпадет на игральном кубике. Вы получили сообщение, что выпало чётное число очков. Какое количество информации содержит это сообщение?

Решение. Энтропия системы «игральный кубик» H_1 равна $\log_2 6$, т.к. кубик может случайным образом принять шесть равновозможных состояний {1, 2, 3, 4, 5, 6}. Полученное сообщение уменьшает число возможных состояний до трёх: {2, 4, 6}, т.е. энтропия системы теперь равна $H_2 = \log_2 3$. Приращение энтропии равно количеству полученной информации $\mathbf{I = H_1 - H_2 = \log_2 6 - \log_2 3 = \log_2 2 = 1}$ **bit.** На примере разобранной задачи можно пояснить одно из распространённых определений единицы измерения – 1 бит:

1 бит - количество информации, которое уменьшает неопределённость состояния системы в два раза.