Теория информации и кодирования

Лекция 2. Оценка количества информации. Энтропия

Лектор: Брежнев Е.В.

E-mail: e.brezhnev@csn.khai.edu

Тестовое задание

1. Какая связь между понятием сообщение, информация, сигнал?

2. Дайте определение информационной системы. Приведите пример.

3. Запишите основную задачу теории информации и кодирования.

Время 5 мин

Что понимается под источником сообщений?

Определение. Дискретный источник сообщений

Определение. Под источником информации понимают множество возможных сообщений с заданной на этом множестве вероятностной мерой

Определение. Дискретным называется источник, множество X возможных сообщений которого конечно или счетно $X=\{x_1, x_2, \ldots\}$. Подобный источник полностью описывается набором вероятностей сообщений: $p(x_i)$, $i=1,2, \ldots$. **Условие нормировки:**

$$\sum_{i=1}^{M} p(x_i) = 1 \quad \text{или} \quad \sum_{x \in X} p(x) = 1$$

Определение. Ансамбль сообщений

Ансамбль сообщений – множество возможных сообщений с их вероятностными характеристиками – *{X, p(x)}*. При этом: $X=\{x_1, x_2, \ldots, x_m\}$ – множество возможных сообщений источника; $i = 1, 2, \ldots, m$, где m – объем алфавита; $p(x_i)$ – вероятности появления сообщений, причем $p(x_i) \geq 0$ и поскольку вероятности сообщений представляют собой полную группу событий, то их суммарная вероятность равна единице

$$\sum_{i=1}^{m} p(x_i) = 1$$

Алфавит — упорядоченный набор символов, используемый для кодирования сообщений на некотором языке.

Мощность алфавита — количество символов алфавита.

Что такое количество информации?

Количество информации. Определение

Количество информации, *I(X)* - числовая величина, адекватно характеризующая актуализируемую информацию по разнообразию, сложности, структурированности (упорядоченности), определенности, ***выбору состояний отображаемой системы.***

Количество информации. Определение

Количество информации – мера неопределённости, «снятой»/устраненной при получении сообщения.

По Хартли, для того, чтобы мера информации имела практическую ценность, она должна быть такова, чтобы отражать количество информации пропорционально числу выборов.

Свойства количества информации

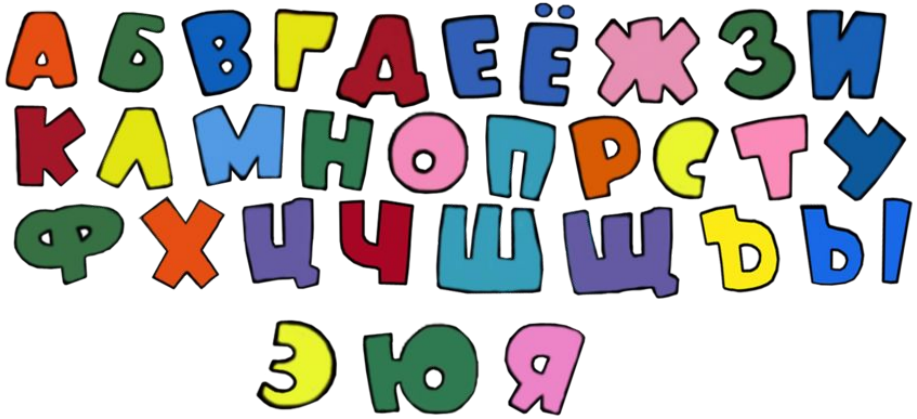
1. Количество информации в сообщении обратно – пропорционально вероятности появления данного сообщения.

2. Свойство аддитивности – суммарное количество информации двух источников равно сумме информации источников.

3. Для события с одним исходом количество информации равно нулю.

4. Количество информации в дискретном сообщении растет в зависимости от увеличения объема алфавита – m.

Мотивирующий пример (1)



Как измерить количество информации, которое может быть передано при помощи такого алфавита при условии что число знаков в сообщении равно n?

Вариант ответа: Это можно сделать, определив число N возможных сообщений, которые могут быть переданы при помощи этого алфавита.

$$N = m^n.$$

Мотивирующий пример (2)

Почему эта величина нас не устраивает:

1. При наличии алфавита, состоящего из одного символа, т.е. когда $m = 1$, возможно появление только этого символа. Следовательно, неопределенности в этом случае не существует, и появление этого символа не несет никакой информации. Между тем, значение N при $m = 1$ не обращается в нуль.

2. Для двух независимых источников сообщений (или алфавита) с N_1 и N_2 числом возможных сообщений общее число возможных сообщений $N = N_1 N_2$, в то время как количество информации, получаемое от двух независимых источников, должно быть не произведением, а суммой составляющих величин.

Количество информации. Мера Хартли (1).

В 1928 г. американский инженер Р. Хартли предложил научный подход к оценке сообщений. Предложенная им формула имела следующий вид:

$$I(N) = \log N,$$

где N - количество равновероятных сообщений (событий).

Эта мера удовлетворяет свойствам количества информации.

Если же все множество возможных сообщений состоит из одного ($N = m = 1$), то $I(N) = \log 1 = 0$, что соответствует отсутствию информации в этом случае.

При наличии независимых источников информации с N_1 и N_2 числом возможных сообщений

$$I(N) = \log N = \log N_1 N_2 = \log N_1 + \log N_2$$

Количество информации. Мера Хартли (2).

Если возможность появления любого символа алфавита равновероятна, то эта вероятность $p = 1/m$. Полагая, что $N = m$,

$$I = \log N = \log m = \log (1/p) = -\log p.$$

Аксиомы количества информации (требования к универсальной информационной мере)

1. Количество информации в сообщении x зависит только от его вероятности

$$I(x) = f(p(x)),\ \forall x \in X.$$

2. Неотрицательность количества информации:

$$I(x) \geq 0,\ \forall x \in X,$$

причем $I(x)=0$ только для достоверного события ($p(x)=1$).

3. Аддитивность количества информации для независимых сообщений:

$$I(x, y) = I(x) + I(y).$$

В каких единицах измеряется информация?

Единицы измерения информации (1)

В простейшем случае, когда m=2

$$I = -\log_2 p = -\log_2 1/2 = \log_2 2 = 1.$$

Полученная единица количества информации, представляющая собой выбор из двух равновероятных событий, получила название **двоичной единицы, или бита.**

Единицы измерения информации (2)

В зависимости от основания логарифма используют следующие единицы информации:

2 – [бит] (*bynary digit* – двоичная единица), используется при анализе информационных процессов в ЭВМ и др. устройствах, функционирующих на основе двоичной системы счисления;

e – [нит] (*natural digit* – натуральная единица), используется в математических методах теории связи;

10 – [дит] (*decimal digit* – десятичная единица), используется при анализе процессов в приборах работающих с десятичной системой счисления.

Недостатки меры Хартли

Формула Хартли позволяет определить количество информации в сообщении только для случая, когда появление символов равновероятно и они статистически независимы.

На практике эти условия выполняются редко. При определении количества информации необходимо учитывать не только количество разнообразных сообщений, которые можно получить от источника, но и вероятность их получения.

Решение задач (1)

Пример 1

Какое количество информации Вы получили, если узнали на какое поле шахматной доски какая фигура и какого цвета поставлена? (Цветов 2; полей 64; разных фигур 6.)

Решение

Число различных исходов: $N = 2 \cdot 64 \cdot 6$.

Количество информации: $I = \log N = 1 + 6 + 2.58 = 9.58$[бит].

Ответ: число в диапазоне 9.5÷9.6.

Решение задач (2)

Пример 2.

Необходимо определить, какое количество информации в битах содержит книга, написанная на русском языке, содержащая 200 страниц (каждая страница содержит приблизительно 2000 символов).

Для решения воспользуемся мерой Хартли

$$I = n \cdot \log_2 m = 200 \cdot 2000 \cdot \log_2 32 = 2000000 \text{ бит.}$$

Здесь n - длина сообщения в знаках:

$$n = 200 \text{ страниц} \cdot 2000 \text{ знаков/страница} = 400000 \text{ знаков.}$$

Ответ: книга содержит около 2-х мегабит (Мбит) информации

Решение задач (3)

Пример 3

Какое количество информации содержится во фразе:

«Кому на Руси жить хорошо?»

(n = 25 (включая пропуски и знак вопроса), M = 32 (русские буквы) + пропуск + (.; -; !; ?) = 37 знаков.)

Решение: $I = \log N = \log M^{n} = 25 \cdot \log 37 = 25 \cdot 5.21 = 130.25$.

Ответ: число в диапазоне 130.2÷130.3.

Решение задач (4)

Пример 3. Имеются 192 монеты. Известно, что одна из них - фальшивая, например, более легкая по весу. Определить, сколько взвешиваний нужно произвести, чтобы выявить ее.

Если положить на весы равное количество монет, то получим 3 независимые возможности: а) левая чашка ниже; б) правая чашка ниже; в) чашки уравновешены. Таким образом, каждое взвешивание дает количество информации $\log_2 3$

следовательно, для определения фальшивой монеты нужно сделать не менее k взвешиваний, где наименьшее
k удовлетворяет условию $\log_2 3^k \geq \log_2 192$

Ответ - необходимо сделать не менее 5 взвешиваний (достаточно 5).

Мера Шеннона (1)

Количество информации, в сообщении, состоящем из *n* не равновероятных его элементов равно (эта мера предложена в 1948 г. К. Шенноном)

$$I(x) = -n\sum_{i=1}^{m} p_i \log p_i$$

Мера Шеннона (2)

Среднее количество информации для всей совокупности сообщений можно получить путем усреднения по всем независимым событиям:

$$I(x) = -\sum_{i=1}^{m} p_i \log p_i$$

Связь меры Хартли и Шеннона

Формула, предложенная Хартли, представляет собой частный случай более общей формулы Шеннона. Если в формуле Шеннона принять, что

$p_1 = p_2 = \ldots = p_i = \ldots = p_N = 1/N$, то

$$I = -\sum_{i=1}^{N} \frac{1}{N} \log \frac{1}{N} = -\log \frac{1}{N} = \log N .$$

Недостатки меры Шеннона

Не всегда представляется возможным заранее установить перечень всех состояний системы и вычислить их вероятности.

Рассматривается только формальная сторона сообщения, не учёт содержания, ценности информации.

Возможно ли измерить неопределенность?

Понятие энтропии (1)

Впервые, сущность энтропии и ее меру открыл в 1871 году великий физик Людвиг Больцман. Он рассматривал количество неопределенности в ансамбле молекул газа и решил проблему физического смысла энтропии как меру хаоса в ансамбле молекул (некотором объеме газа).

$$S = k \cdot \ln W,$$

Понятие энтропии (2)

Энтропия - мера неопределенности информации.

Энтропия - математическое ожидание *H(x)* случайной величины *I(x)* определенной на ансамбле *{X, p(x)}*, т.е. она характеризует среднее значение количества информации, приходящееся на один символ.

$$H(X) = M[I(X)] = \frac{I(X)}{n} = -\sum_{i=1}^{m} p(x_i) \cdot log\ p(x_i)$$

$\sum_{i=1}^{n}$

Понятие энтропии (3)

Если система A имеет n возможных состояний

$$A_1, A_2, \ldots, A_n$$

причем вероятности этих состояний равны соответственно

$$p_1, p_2, \ldots, p_n;$$
$$p_1 + p_2 + \ldots + p_n = 1,$$

то энтропией системы A называется величина:

$$\mathbf{H(A) = -(p_1 \cdot \log_2 p_1 + p_2 \cdot \log p_2 + \ldots + p_n \cdot \log p_n),}$$

$$H(A) = -\sum_{i=1}^{n} p_i \bullet \log_2 p_i$$

Понятие энтропии (4)

Единицы измерения энтропии

Один бит - это энтропия простейшей физической системы, которая может быть только в одном из двух состояний, причем эти состояния равновероятны.

Пример

Пусть система A обладает двумя состояниями A_1 и A_2 с вероятностями $p_1 = 0,5$ и $p_2 = 0,5$. Тогда энтропия такой системы равна

$$H(A) = - (0,5 \cdot \log_2 0,5 + 0,5 \cdot \log_2 0,5) = 1$$

Свойства энтропии

1) энтропия всегда неотрицательна, так как значения вероятностей выражаются величинами, не превосходящими единицу, а их логарифмы - отрицательными числами или нулем;

2) если $p_i = 1$ (а все остальные $p_j = 0$, $j = 1, ..., (n-1)$), то $H(A) = 0$. Это тот случай, когда об опыте или величине все известно заранее и результат не дает новую информацию;

3) $H(A) = H_{max}$ при $p_1 = p_2 = ... = p_n = 1 / n$

4) энтропия системы, состоящей из двух подсистем A и B (состояния системы образуются совместной реализацией объектов A и B), то есть:

H(AB) = H(A) + H(B). (доказать)

Избыточность сообщений

Одной из информационных характеристик источника дискретных сообщений является избыточность, которая определяет, какая доля максимально-возможной энтропии не используется источником

$$R = \frac{H_{max} - H(x)}{H_{max}} = 1 - \frac{H(x)}{H_{max}} = 1 - \mu$$

где μ – коэффициент сжатия.

“+” - избыточность необходима для обеспечения достоверности передаваемых данных, т.е. надежности СПД, повышения помехоустойчивости. Чем ближе энтропия источника к максимальной, тем рациональнее работает источник.

“-” - Избыточность приводит к увеличению времени передачи сообщений, уменьшению скорости передачи информации, излишней загрузки канала,

Связь между энтропией и количеством информации

Вопрос 1: Что означает равенство I = H?

Количество информации только тогда равно энтропии, когда неопределенность ситуации снимается полностью

Вопрос 2: Что означает $H_{apr} - H_{aposter}$?

Вопрос 3: Что априорно, а что апостериорно? Информация или неопределенность?

Решение задачи

Пример 1. Определить энтропию СВ X, заданной распределением

X	1	2	3	4	5	6	7	8
p	0.1	0.2	0.1	0.05	0.1	0.05	0.3	0.1.

Спасибо за внимание!