Лекция 15

ЭЛЕМЕНТЫ ТЕОРИИ ИНФОРМАЦИИ

План лекции

- Модель информационной системы Шеннона
- Информационная емкость сообщений для сигналов с заданным распределением частот символов
- Формулы Шеннона и Хартли
- Избыточность кодирования

Информационная модель Шеннона

- Claude Shannon “A Mathematical Theory of Communication” The Bell System Technical Journal Vol. 27, pp. 379–423, 623–656, July, October, 1948

- Дискретный или непрерывный сигнал
  - Символы – пример дискретного сигнала
- Источник (кодер)
- Приемник (декодер)
- Канал передачи сигналов
  - Канал не искажает и не теряет символы

Информационная модель Шеннона

- Три вопроса, на которые ответил Шеннон:
- Какой нужен канал, чтобы передать данный сигнал (последовательность символов) за данное время?
- За какое время можно передать данный сигнал по данному каналу?
- За какое время *нельзя* передать данный сигнал по данному каналу без потерь?

Информационная модель Шеннона

- Каким должен быть канал, чтобы передать данный сигнал за данное время?
- За какое время можно передать данный сигнал по данному каналу?

- Пропускная способность канала

- В чем измерять пропускную способность?
  - Если передача всех символов занимает одинаковое время, то в символах в секунду
  - Как быть, если передача разных символов занимает разное время?

Информационная модель Шеннона

- Пусть N(T) – число сигналов, передача которых занимает время T через данный канал
- Размер пропускной способности канала = предел log2(N(T))/T при T --> oo
- Удлинение сигнала, передаваемого через канал за время T, на X двоичных символов увеличивает пропускную способность канала на X/T
  - Соответствует интуиции

Информационная модель Шеннона

- За какое время *нельзя* передать данный сигнал по данному каналу без потерь?

- Наименьшее число двоичных символов, необходимых для записи сигнала -- объем информации в сигнале

Информационная модель Шеннона

- На практике важен *приблизительный* объем информации в сигналах *определенного* вида
  - *Точный* объём информации в *произвольном* сигнале алгоритмически невычислим
  - Пусть сигналы – это произвольные частично рекурсивные функции
  - Тогда объем информации в сигнале -- это т.н. *алгоритмическая сложность по Коломогорову* частично рекурсивной функции, соответствующей сигналу

- Шеннон предложил метод для оценки объёма информации в текстах на естественном языке

Информационная модель Шеннона

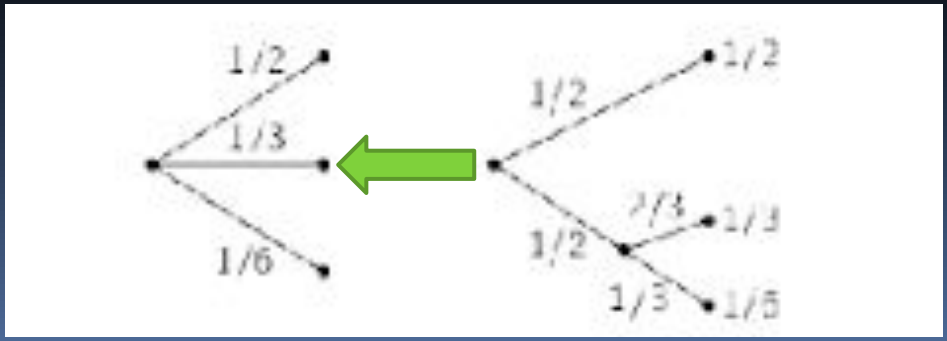
- Частота появления символа в длинном тексте не зависит от текста -- верно для всех известных языков
- Обозначим p1, p2, …, pN частоты появления символов
- Объём информации в длинном тексте на естественном языке линейно растет с длиной текста и зависит только от языка, на котором составлен текст
- Обозначим H(p1, p2, …, pN) количество информации в одном символе текста

Информационная модель Клода Шеннона

1. H должна быть непрерывна по pk

2. Значение H(1/N, 1/N, …, 1/N) должно возрастать по числу символов N

3. $H(p1, p2, \ldots, pN) = H(p1, \ldots, p_{N-1}+pN) + (p_{N-1}+pN)*H(p_{N-1}/(p_{N-1}+pN), pN/(p_{N-1}+pN))$
   - H(1/2, 1/3, 1/6) = H(1/2, 1/2) + (1/2) * H(2/3, 1/3)



Информационная модель Шеннона

- Все функции, удовлетворяющие условиям 1-3, имеют вид

$$H(p1, ..., pN) = - c \sum pk \log2(pk)$$

- Формула Шеннона для объема информации

- Формулу Шеннона для $pk=1/N$ называют формулой Хартли

Заключение

- Модель информационной системы Шеннона
- Информационная емкость сообщений для сигналов с заданным распределением частот символов
- Формулы Шеннона и Хартли
- Избыточность кодирования

Будем говорить, что источник передал приемнику некоторую

информацию о происшедшем событии, на основании

которой изменилось представление приемника о множестве

возможных исходов наблюдаемой величины.

Определим *количество информации,* содержащейся в

сообщении *m,* изменяющем представление приемника о

соб

(2)

$$I(m) = -\log_2 \frac{p(S_{до})}{p(S_{после})}$$

Единицей количества информации является *бит*.

Пример 1

В семье должен родиться ребенок.

Пространство элементарных исходов данной случайной величины — *{мальчик, девочка},* — состоит из двух исходов.

Отсутствие априорной информации у приемника (родителей)

о поле малыша означает, что $S_{до}$ совпадает с этим пространством.

Сообщение источника (врача) «у вас родился мальчик» сужает

это множество предположений до множества $S_{ПоСЛЕ}$ из единственного исхода *мальчик.*

По формуле (10) $I(x) = \log_2 \frac{p(S_{до})}{p(S_{после})}$ количество $\log_2 \frac{1}{2}$ единиц информации определяется как

$log_2 2 = 1$ – ?

- 1 бит соответствует сообщению о том, что произошло одно из двух равновероятных событий;
- требуется один бит для хранения сообщений о двух равновероятных событиях.

Пример 2

Из колоды вытягивается карта. Пространство элементарных исходов — 52 карты. В отсутствие изначальной информации пространство предположений $S_{ДО\_1}$ совпадает со всем пространством.

Первое сообщение от источника «выпала трефа» сужает его до $S_{ПОСЛЕ\_1}$ из 13 возможных исходов.

Второе сообщение «выпала картинка» сужает $S_{ДО\_2} = S_{ПОСЛЕ\_1}$ до $S_{ПОСЛЕ}$ состоящего из 4 исходов.

Третье сообщение «выпала дама треф» сужает $S_{ДО\_3} = S_{ПОСЛЕ\_3}$ до $S_{ПОСЛЕ\_3}$, состоящего из единственного исхода.

Количество информации, содержащееся в первом сообщении равно $-log_2\ 13/52 = 2$ битам, во втором — $-log_2\ 4/13 = 1.5$, в третьем — $-log_2\ 1/4 = 2$ битам.

Нетрудно проверить, что суммарное количество полученной информации — 5.5 бит, совпадает с количеством информации, которое несло бы сообщение «выпала дама треф» = $-log_2\ 1/52 = 5.5$ бит.

Теорема об аддитивности информации

Теорема

Количество информации, переносимое сообщением m1 && m2 && … && mN, не зависит от порядка отдельных сообщений и равно сумме количеств информации, переносимых сообщениями m1, …, mN по отдельности.

Выберем какой-либо порядок передачи сообщений

I(W, m1) = log2(P(m1)/P(W))

I(m1, m1&&m2) = log2(P(m1&&m2)/P(m1))

I(m1 && m2 && … && m_N-1, m1 && m2 && … && mN) = log2(P(m1&&…&&mN)/P(m1&&…m_N-1))

Пример о двух источниках:

1 – p(что грань 5)=1; $\log P_{после}/P_{до} = \log 1/1 = 0$;

2 – p(что грань 5)=1/6; $\log P_{после}/P_{до} = \log 1/1/6 = \log 6 \approx 2{,}5$ бит.

Свойства информации:

Формулы Шеннона, Хартли

Предположим теперь, что источник является генератором символов из некоторого множества $\{x_1, x_2, \ldots, x_n\}$ (назовем его алфавитом источника). Эти символы могут служить для обозначения каких-то элементарных событий, происходящих в области источника, но, абстрагируясь от них, в дальнейшем будем считать, что рассматриваемым событием является поступление в канал самих символов.

Если $p(x_i)$ — вер[illegible]упления в канал символа x_i, то

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

Рассмотрим теперь модель, в которой элементарным исходом является текстовое *сообщение*. Таким образом, Ω — это множество всех цепочек символов произвольной длины.

По поступившему сообщению *m* можно посчитать экспериментальную *частоту* встречаемости в нем каждого символа, где *N* — общая длина сообщения, а n_i — число повторений в нем символа x_i.

$$\nu_m(x_i) = \frac{n_i}{N},$$

Понятно, что анализируя различные сообщения, мы будем получать различные экспериментальные частоты символов, но для источников, характеризующихся закономерностью выдачи символов (их называют эргодическими), оказывается, что в достаточно длинных сообщениях все частоты символов сходятся к некоторым устойчивым величинам которые можно рассматривать как *распределение вероятностей* выдачи символов данным источником.

$$p(x_i) = \lim_{N \to \infty} \frac{n_i}{N},$$

(4)

Рассмотрим сообщение m, состоящее из n_1 символов x_1, n_2 символов x_2 и т. д. в произвольном порядке, как серию элементарных событий, состоящих в выдаче одиночных символов.

Тогда вероятность появления на выходе источника сообщения m равна

$$p(m) = \frac{(n_1)^{n_1}}{N} \cdot \ldots \cdot \frac{(n_n)^{n_n}}{N} = \frac{1}{N^N} \cdot (n_1^{n_1} \cdot \ldots \cdot n_n^{n_n}).$$

Количество информации, переносимой сообщением *m* длины *N*, определяется как

$$I(m) = -\log_2 \frac{p(m)}{1} = -\log_2 \left(\left(\frac{n_1}{N} \right)^{n_1} \cdot \ldots \cdot \left(\frac{n_n}{N} \right)^{n_n} \right) = -\sum_{i=1}^{N} n_i \cdot \log_2 \left(\frac{n_i}{N} \right).$$

Количество информации, приходящейся в среднем на каждый символ в сообщении *m*, есть

$$I_0(m) = \frac{1}{N} \cdot I(m),$$

где *N* — длина сообщения *m*.

Формула **Шеннона**

Перейдем к пределу по длине всевозможных сообщений *(N —> ∞)*:

$$I_0(A) = \lim_{N\to\infty} I_0(m) = \lim_{N\to\infty} \frac{1}{N} \cdot \left(-\sum_{i=1}^{N} n_i \cdot \log_2 \left(\frac{n_i}{N} \right) \right) =$$

$$= \left(-\sum_{i=1}^{N} \lim_{N\to\infty} \left(\frac{n_i}{N} \right) \cdot \log_2 \lim_{N\to\infty} \left(\frac{n_i}{N} \right) \right).$$

По формуле (…) вспоминая, что в достаточно большом сообщении $\frac{n_i}{N}$

$p(x_i)$ = lim N->

$$I_0(A) = -\sum_{i=1}^{N} p(x_i) \cdot \log_2 p(x_i). \qquad (5)$$

Формула Хартли

Величина $I_0(A)$ характеризует среднее количество информации на

один символ из алфавита A с заданным (или экспериментально

определенным) распределением вероятностей

$$p(x_1),\ p(x_2),\ \ldots,\ p(x_N).$$

Рассмотрим случай, когда все символы в алфавите равновероятны:

$$p(x1) = p(x2) \ \ldots \ = p(x_N) = 1/N \ .$$

Среднее количество информации, приходящееся на каждый та

$$I_0(A) = -\sum_{i=1}^{N} \frac{1}{N} \cdot \log_2\left(\frac{1}{N}\right) = -N \cdot \frac{1}{N} \cdot \log_2 \frac{1}{N} = \log_2 N. \qquad (6)$$

Событие, которое может произойти или нет, называют *случайным*.

Примеры: попадание стрелка в мишень,

извлечение дамы пик из колоды карт,

выигрыш билета в розыгрыше лотереи и т. д.

На основании отдельно взятого случайного события нельзя

научно предсказать, например, какие билеты окажутся

выигрышными. Но если провести достаточно большую

последовательность испытаний, то можно выявить

определенные закономерности, позволяющие делать

количественные предсказания.

Определение

Пространство элементарных событий (исходов) Ω – множество всех различных событий, возможных при проведении эксперимента.

Элементарность исходов понимается в том смысле, что ни один из них не рассматривается как сочетание других событий.

Примеры:

1) Будем бросать монету до тех пор, пока не выпадет герб. После этого эксперимент закончим.

«Элементарный исход» этого эксперимента можно представить в виде последовательности

р, р, р, ..., р, г (где р — решка, г — герб).

Таких последовательностей бесконечно много.

Следовательно, в данном случае множество Ω бесконечно.

2) Однократное бросание игральной кости. Будем считать, что возможен только один из 6 исходов, соответствующих падению кости гранями с 1, 2,...,6 очками вверх. Каждый возможный исход удобно обозначать числом выпавших очков.

Тогда пространство элементарных событий Ω = {1,2,3,4,5,6}.

Формула $\omega \in \Omega$ означает, что элементарное событие ω является

элементом пространства Ω.

Многие события естественно описывать множествами,

составленными из элементарных исходов.

Например, событие, состоящее в появлении четного

числа очков, описывается множеством $S = \{2,4,6\}$.

Формула $S \subseteq \Omega$ означает, что событие S является

подмножеством пространства Ω.

- *Случайная величина —> переменная*
- *Элементарный исход —> значение переменной*
- *Пространство элементарных исходов —> область значений*
- *Событие —> подмножество области значений*

Определим формально *меру события μ,* как отображение из

пространства Ω в M, обладающее следующими свойствами:

1) $\mu(\theta) = 0,$ где θ - пустое множество, т.е. множество, не содержащее ни одного элемента;

2) $S_1 \subseteq S_2 \Rightarrow \mu(S_1) \leq \mu(S_2),$

3) mu(S1 U S2)=mu(S1)+mu(S2)-mu(S1/\S2)

Введем функцию *p(S) вероятности события* как численного выражения возможности события *S* на заданном пространстве элементарных исходов Ω следующим образом:

$$p(S) = \frac{\mu(S)}{\mu(\Omega)} = \frac{\text{Число желательных исходов}}{\text{Число всех возможных исходов}} \quad (1)$$

«Желательные» исходы - элементарные исходы, образующие событие S.

$0 \le p(S) \le 1$ $p(0) = 0$, $p(\Omega) = 1$.

Событие с вероятностью 1 содержит все элементарные исходы и,

следовательно, происходит наверняка.

Событие с вероятностью 0 не содержит ни одного исхода,

следовательно, не происходит никогда.

Говорят, что заданы вероятности элементарных событий,

если на Ω задана неотрицательная числовая функция *p* такая,

что:

$$\sum_{\omega \in \Omega} p(w) = 1.$$

Вероятность того, что при бросании кости выпадет единица,

равна $$\frac{\mu(\{1\})}{\mu(\{1,2,3,4,5,6\})} = \frac{1}{6}.$$

Вероятность появления четного числа очков равна

$$\frac{\mu(\{2,4,6\})}{\mu(\{1,2,3,4,5,6\})} = \frac{3}{6} = \frac{1}{2}.$$

Паскаль в письмах к Ферма в 1654 г. писал:
«Как велика вероятность, что когда я проснусь ночью и посмотрю на часы, то большая стрелка будет стоять между 15 и 20 минутами?»
И в этом же письме приводит рассуждения о том, что вероятность того, что стрелка часов будет находиться в этом промежутке, равна 5/60=1/12.

Теорема о сложении вероятностей

Если пересечение событий *A* и *B* непусто, то

$$p(A \cup B) = p(A) + p(B) - p(A \cap B).$$

(Это следует из аксиомы 3 для меры.)

Пример. Найдем вероятность того, что вытащенная из полной колоды карта окажется пикой или картинкой.

Пусть событию *A* соответствует извлечение из колоды карт пики, событию *B* — картинки.

Для каждой карты из колоды вероятность вытащить ее равна 1/52.

Число пик в полной колоде равно 13. Следовательно, вероятность

события *A* равна 13/52=1/4. Число картинок равно 16, вероятность

события *B* равна 16/52 = 4/13.

События *A* и *B* имеют непустое пересечение. Множество $A \cap B$ состоит

из четырех элементов,следовательно, $p(A \cap B) = 4/52 = 1/13$.

$$p(A \cup B) = p(A) + p(B) - p(A \cap B = 1/4+4/13-1/13=25/52.$$

Вероятность того, что вытащенная из полной колоды карта окажется

Теорема об умножении вероятностей

Рассмотрим теперь серию экспериментов, в которой некоторая

случайная величина наблюдается последовательно несколько

раз. Последовательные события называются *независимыми*,

если наступление каждого из них не связано ни с каким из

других.

Например, исходы при бросании кости являются

независимыми событиями, а последовательные вытягивания

карт из одной и той же колоды без возврата — нет.

Теорема. Вероятность того, что независимые события S1, S2

произойдут в одной серии испытаний, равна

Определим формально *меру события μ*, как отображение из

пространства Ω в *N*, обладающее следующими свойствами:

1) $\mu(\emptyset) = 0$, где $\emptyset$ — пустое множество, т. е. множество, не содержащее ни одного элемента;

2) $S_1 \subseteq S_2 \Rightarrow \mu(S_1) \leq \mu(S_2)$, где $S_1 \subseteq \Omega, S_2 \subseteq \Omega$;

3) $\mu(S_1 \cup S_2) = \mu(S_1) + \mu(S_2) - \mu(S_1 \cap S_2)$.

КОНЕЦ ЛЕКЦИИ

Избыточность кодирования

Оказывается, что величина $I_0(A)$ определяет предел сжимаемости кода:

никакой двоичный код не может иметь среднюю длину меньшую, чем I_0,

в противном случае можно было бы передать некоторое количество информации меньшим числом битов, что невозможно.

Таким образом, любой код может быть лишь в большей или меньшей степени *избыточным*.

Относительная избыточность кода характеризуется как отношение числа

«избыточных» битов в коде к общей длине кода,

то избыточное число битов есть $L - N * I_0(A)$,

(сообщение из N символов алфавита A с информационной емкостью $I_0(A)$,

код длины L б

$$\frac{L - N \cdot I_0(A)}{L} = 1 - \frac{N}{L} \cdot I_0(A). \quad (7)$$

ть каждого символа кода:

Заметив, что $\lim_{N\to\infty}$ L/N - есть средняя длина кодового слова $K_0(A)$, получим независимое от сообщения соотношение для избыточности кода:

$$Z(K) = 1 - I_0(A)/K_0(A).$$

Оптимальный код с нулевой избыточностью является код со средней длиной кодового слова $K_0 = I_0(A)$ битов или наиболее близкий к нему.

Резюме. $I_0(A)$ показывает, какое в среднем количество двоичных символов

нужно для записи всех кодовых слов алфавита А при произвольном кодировании «символ —> слово».

Для алфавитов с равновероятными символами формула Хартли определяет

минимальную необходимую длину кодового слова, например для алфавита

ASCII: $I_0(ASCII) = log_2 256 = 8$ бит.

Таким образом, любой 8-битный код для *ASCII* будет оптимальным.

Посчитаем информационную емкость кода: длина исходного

сообщения $N = 18$, длина кода $L = 39$ битов.

Удельная информационная емкость алфавита A с распределением

P есть

$$I_0(A) = \frac{8}{18}\cdot\log_2\frac{18}{4} + \frac{1}{18}\cdot\log_2\frac{18}{1} + \frac{7}{18}\cdot\log_2\frac{18}{7} + \frac{2}{18}\cdot\log_2\frac{18}{2} = 2.1 \quad .$$

Избыточность кода

$$Z = 1 - \frac{N}{L}\cdot I_0(A) = 1 - \frac{18}{39}\cdot 2.1 = 0.03,$$

Реализация проекта

рхиватор должен вызываться из командной строки,

ормат вызова:

```
arc.exe –[axdlt] arc[.ext]  file_1 file_2 … file_n
```

оддерживаемые операции:

- *a*- поместить файл(ы) в архив;
- *x* - извлечь файл(ы) из архива;
- *d* - удалить файл(ы) из архива;
- *l* - вывести информацию о файлах, хранящихся в архиве;
- *t* - проверить целостность архива.

Проверка целостности архива

```
_stat, _wstat, _stati64, _wstati64
int _stat(const char* path, struct _stat *buffer);

#include <sys/stat.h>
```

CRC32 – проверка контрольных сумм

Построение дерева Хаффмана

Вход:

A – исходный набор символов $<a_1,...,a_N>$,

$P=<p_1,p_2,...,p_N>$ - распределение их частот;

– $W_0 = \{<a_1,p_1>,...,<a_N,p_N>\}$ (начальный набор свободных узлов соответствует встречающимся символам);

– цикл по i от 0 до N-1

W_i = Шаг_построения(W_{i-1});

Выход:

Дерево Хаффмана, построенное в цикле с корневым узлом, содержащимся в W_N.

Код Хаффмана

Алгоритм:

1. Определить алфавит A = { c_1, c_2 , ... , c_n } сообщения S и подсчитать число вхождений p_1, p_2, ... , p_n в S
2. Построить дерево оптимального префиксного двоичного кода для S используя свойства 1-8 оптимального кода – полученный префиксный двоичный код называется кодом Хаффмана (1951, David A. Huffman, Massachusetts Institute of Technology)
3. Закодировать сообщение S используя код Хаффмана

Критерии качества кодирования:

- минимальная длина кода;
- однозначное декодирование.

Информационная модель Клода Шеннона

- Пусть в области источника происходит наблюдение за некоторой случайной величиной.
- Приемник может иметь некоторое априорное представление о множестве $S_{до}$ возможных исходов этой величины до того, как произошло наблюдение.
- Когда ничего не известно заранее, $S_{до}$ принимается за все пространство возможных исходов Ω.
- Источник передает приемнику сообщение о произошедшем наблюдении, после получения которого множество предположительных исходов у приемника сужается до $S_{ПоСЛЕ}$.
- Это представление будем называть апостериорным.