СПбГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ

А.Б. Левина

Теория информации и теория кодирования

Сжатие данных

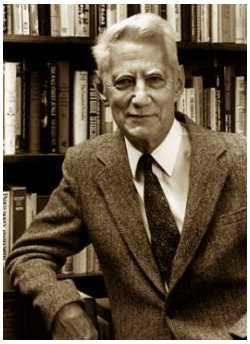
СПбГЭТУ «ЛЭТИ», 2022 г.

1 ВВЕДЕНИЕ

Мы начинаем курс «Теория информации и теория кодирования» и на данной лекции познакомимся с основами теории информации. На данной лекции будет рассмотрена история развития теории информации, применимость данного направления на практике, математические основы теории кодирования, понятие энтропии и необходимые основы теории вероятности.

2 ИСТОРИЯ РАЗВИТИЯ ТЕОРИИ ИНФОРМАЦИИ

История развития теории информации началась в 1948 году с работы Клода Шеннона «Математическая теория связи»



Клод Элвуд Шеннон (*Claude Elwood Shannon*) 30.04.1916 -24.02.2001, США — инженер, математик, криптоаналитик.

Клод Шеннон по праву считается основателем теории информации.

В 1948 году предложил использовать слово «бит» для обозначения наименьшей единицы информации.

Является автором основополагающих статей для теории информации и криптографии «Математическая теория информации» и «Теория связи в секретных системах». Клод Шеннон также сформулировал теоретические основы криптографии и внёс ключевой вклад в теорию вероятностных схем, теорию игр, теорию автоматов, теорию систем управления — области наук, входящие в понятие «кибернетика».

Введем определение, что такое теория информации.

Теория информации — раздел прикладной математики, радиотехники (теория обработки сигналов) и информатики, относящийся к измерению количества информации, ее свойств и устанавливающий предельные соотношения для систем передачи данных.

- оперирует математическими, а не реальными физическими объектами;
- Использует, главным образом, математический аппарат теории вероятности и математическую статистику.

Рассмотрим блок схему работы системы связи:

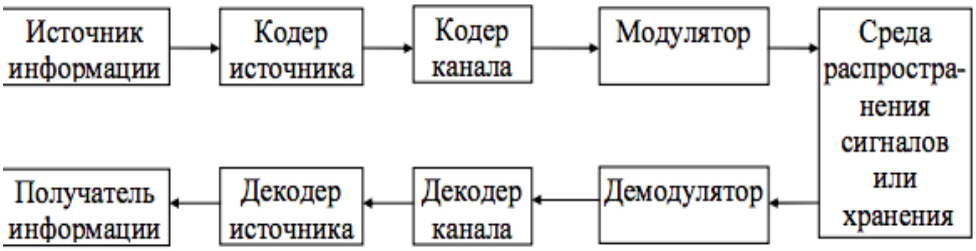


Рис. 1: Блок-схема системы связи.

Проговорим более подробно каждый объект, представленный на схеме:

1. *Источник* – любой объект порождающий сообщение, которое необходимо передать в пространстве или во времени
2. *Кодер источника* – представление информации в более компактной форме
3. *Кодер канала* – обработка информации с целью защиты при передаче
4. *Модулятор* – преобразование в сигналы, согласованные с природой канала средой накопителя

Можно выделить несколько разделов теории информации:

- *Кодирование дискретных источников* – кодирование без потерь, кодирование для каналов без шума, сжатие
- *Кодирование информации для передачи по каналу с шумом* - защита информации от помех в каналах связи
- *Кодирование с заданным критерием качества* - методы кодирования, обеспечивающие наилучший компромисс между качеством и затратами на передачу информации
- *Кодирование информации для систем со многими пользователями* - оптимальное взаимодействие абонентов, использующих общий ресурс
- *Секретная связь, системы защиты информации от несанкционированного доступа*

  Поговорим чуть более подробно о тех двух направлениях, которые будут рассматриваться в данном курсе: ч

  1. ***Кодирование источника***

     Когда говорят о кодирование источника подразумевают сжатие данных.

     Сжатие данных (*data compression*) — алгоритмическое преобразование данных, производимое с целью уменьшения занимаемого ими объёма. Применяется для

более рационального использования устройств хранения и передачи данных.

Синонимы: *упаковка данных*, *компрессия*, *сжимающее кодирование*, *кодирование источника*.

Обратная процедура называется восстановлением данных (распаковкой, декомпрессией).

2. ***Канальное (помехоустойчивое) кодирование***

У него есть две цели:

Обнаружение ошибок обеспечивается с помощью кодов, обнаруживающих ошибки

Исправление ошибок обеспечивается корректирующими кодами (коды, исправляющие ошибки, коды с коррекцией ошибок, помехоустойчивые коды).

Теория информации нераздельно связана с математикой, а более конкретно с теорией вероятности, вспомним некоторые основные понятия из теории вероятности.

3. ТЕОРИЯ ВЕРОЯТНОСТИ

Определение: Случайной величиной называется переменная X, принимающая свои значения с некоторой вероятностью.

Пример:

Переменная X принимает значение s с вероятностью 0,3 p(X=s)=0,3

Определение: X - дискретная случайная величина, то множество вероятностей всех ее значений называется *распределением вероятностей,* а функция p(X=x) сопоставляющая значению функции какую-то вероятность - *плотностью распределения вероятности.*

1. $p(X=x)\geq 0$
2. $\sum_x p(X=x)=1$

Пример:

Рассмотрим колоду из 52 карт:

V – случайная величина появление карты определенного достоинства

S – масть

C – цвет

$$p(C=красный)=\frac{1}{2},$$

$$p(V=туз\ треф)=\frac{1}{52},$$

$$p(S=трефа)=\frac{1}{4}.$$

Определение: X и Y случайные величины с распределение вероятности p(X=x) и p(Y=y). Вероятность одновременного равенства X=x и Y=y называется *совместной вероятностью* P(X=x, Y=y).

$$p(C = красный, S = трефа) = 0, \quad p(C = красный, S = бубна) = \frac{1}{4},$$
$$p(C = красный, S = черва) = \frac{1}{4}, \quad p(C = красный, S = пика) = 0,$$
$$p(C = черный, S = трефа) = \frac{1}{4}, \quad p(C = черный, S = бубна) = 0,$$
$$p(C = черный, S = черва) = 0, \quad p(C = черный, S = пика) = \frac{1}{4}.$$

Определение: Две случайные величины X иY называются *независимыми*, если для всех возможных значений x и y имеет место равенство:

$$p(X = x, Y = y) = p(X = x) \cdot p(Y = y).$$

Определение: Условной вероятностью p(X=x|Y=y) случайных величин X и Y называется вероятность того, что переменная X принимает значение x при условии что Y=y.

$$p(S = пика | C = красный) = 0$$

$$p(V = туз\ пик | C = черный) = \frac{1}{26}.$$

Теорема Байеса:

Если p (Y=y) > 0, то

$$p(X = x|Y = y) = \frac{p(X = x) \cdot p(Y = y|X = x)}{p(Y = y)} = \\ = \frac{p(X = x, Y = y)}{p(Y = y)}.$$

Рассмотрим наш пример с картами:

$$p(S = пика | C = красный) = \frac{p(S = пика, C = красный)}{p(C = красный)} = \\ = 0 \cdot \left(\frac{1}{4}\right)^{-1} = 0.$$

$$p(V = туз\ пик | C = черный) = \frac{p(V = туз\ пик, C = черный)}{p(C = черный)} =$$
$$= \frac{1}{52} \cdot \left(\frac{1}{2}\right)^{-1} = \frac{2}{52} = \frac{1}{26}.$$

Формула полной вероятности позволяет вычислить вероятность интересующего события через условные вероятности этого события в предположении неких гипотез, а также вероятностей этих гипотез.

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

Введем еще два очень важных понятия.

Определение: Математическое ожидание— среднее значение случайной величины (распределение вероятностей стационарной случайной величины) при стремлении количества выборок или количества измерений (иногда говорят — количества испытаний) её к бесконечности.

Если X -дискретная случайная величина, имеющая распределение $\mathbb{P}(X = x_i) = p_i$, $\sum_{i=1}^{\infty} p_i = 1$, то $M[X] = \sum_{i=1}^{\infty} x_i\, p_i$.

Определение: Дисперсия случайной величины — мера разброса значений случайной величины относительно её математического ожидания.

Дисперсией случайной величины называют математическое ожидание квадрата отклонения случайной величины от её математического ожидания.

Пусть X — случайная величина, определённая на некотором вероятностном пространстве, тогда дисперсией называется $D[X] = M\left[\left(X - M[X]\right)^2\right]$

где M обозначает математическое ожидание.

Рассмотри несколько задач:

1. По каналу связи с помехами передается одна из двух команд управления в виде 11111 и 00000, вероятности передачи этих команд соответственно равны 0,7 и 0,3. Вероятность правильного приема каждого из символов 0 и 1 равна 0,6. Символы искажаются помехами независимо друг от друга. На выходе канала

имеем кодовую комбинацию 10110.

Определить какая комбинация была передана.

2. По двоичному каналу связи с помехами передаются цифры 1 и 0 с вероятностями $p1=p2=0.5$. Вероятность перехода единицы в единицу и нуля в ноль соответственно равны $p(1/1)=p$, $p(0/0)=q$.

Определить закон распределения вероятностей случайной величины X – однозначного числа, получаемого на приемной стороне.

3. Производится прием символов 0 и 1 до первого появления символа 1. Вероятность появления 1 при приеме $p=0,4$. Принимается не более четырех символов.

Вычислить $M(X)$, $D(X)$.

4. ИЗМЕРЕНИЕ ИНФОРМАЦИИ

Возникает вопрос – какой может быть мера информации?

Ответом является - мера связанная с затратами на передачу сообщения.

Когда мы производим измерение информации мы должны помнить некоторые свойства:

1. Сообщение случайное событие
2. X дискретное множество, x€X с вероятностью p(x)
3. μ(x) – мера информации ансамбля X= {x, p(x)}

Для меры информации выполнено:

1. μ(x) ≥ 0
2. Мера **однозначно** определяет вероятность сообщение: μ(x) пишем μ(p(x))
3. $\mu(p(x)^m)=m\mu(p(x))$

4. $\mu(x_1, \ldots, x_n) = \mu(x_1) + \ldots, + \mu(x_n)$ для независимых $x_1, \ldots, x_n$

Определение: Собственной информацией I(x) сообщения x из дискретного ансамбля X={x, p(x)}, называется величина, вычисляемая по формуле: $I(x) = -\log p(x)$.

Пример:

1. X= {a,b,c,d}, p(a)=1/2, p(b)=1/4, p(c)=p(d)=1/8

?. Найти собственную информацию для каждой буквы.

2. X= {a,b}, p(a)=0,05, p(b) =0,95

?. Найти собственную информацию для каждой буквы.

Свойства собственной информации:

1. Неотрицательность: $I(x) \geq 0$, при x из X
2. Монотонность: если x_1, x_2 из X, $p(x_1) \geq p(x_2)$, то $I(x_1) \geq I(x_2)$
3. Аддитивность: Для независимых сообщений $x_1, \ldots, x_n$ имеет место равенство

$$I(x_1, \ldots, x_n) = \sum_{i=1}^{n} I(x_i).$$

5. ЭНТРОПИЯ

Введем еще одно понятие, которое будет с нами на протяжении всего курса.

Определение: Энтропией дискретного ансамбля X= {x, p(x)} называется

$$H(X) = \mathrm{M}\left[-\log p(x)\right] = -\sum_{x \in X} p(x) \log p(x)$$

Энтропия – количественная мера неопределенности о том, какое из сообщений будет порождено источником.

Свойства энтропии

1. $H(X) \geq 0.$

2. $H(X) \leq \log |X|$ равенство возможно, когда все элементы X равновероятны.

$$
\begin{aligned}
H(X) - \log |X| &\overset{(a)}{=} -\sum_{x \in X} p(x) \log p(x) - \sum_{x \in X} p(x) \log |X| = \\
&\overset{(b)}{=} \sum_{x \in X} p(x) \log \frac{1}{p(x)|X|} \leq \\
&\overset{(c)}{\leq} \log e \left[\sum_{x \in X} p(x) \left(\frac{1}{p(x)|X|} - 1 \right) \right] = \\
&= \log e \left(\sum_{x \in X} \frac{1}{|X|} - \sum_{x \in X} p(x) \right) = 0 \quad .
\end{aligned}
$$

3. X = {x; p(x)} и Y = {y = f (x), p(y)}, тогда H(Y) ≤H(X), равенство выполняется если f(x) отличается от p(x) только порядком следования элементов

4. Если X и Y независимы, то H(XY)=H(X)+H(Y)

5. Энтропия - выпуклая функция распределения вероятностей на элементах X

6. Пусть X={x, p(x)} и A входит в X введем X’ ={x, p’(x)} p’(x):

$$
p'(x) = \begin{cases} \frac{P(A)}{|A|}, x \in A, \\ p(x), x \notin A. \end{cases}
$$

тогда H(X’) ≥ H(X)

7. Задан X и на множестве его элементов определена функция g(x), введем

Y={y=g(x)}. Тогда H(Y) ≤ H(X). Равенство имеет место когда g(x) обратима.

6. УСЛОВАНЯ ЭНТРОПИЯ.

Введем так же понятие условной энтропии.

Для любого фиксированного y из Y можно построить условное распределение вероятностей p(x|y) на множестве X и для каждого x принадлежащего X подсчитать собственную информацию,

$$I(x|y) = -\log p(x|y),$$

которую называют *условной собственной информацией* сообщения x при фиксированном y.

Определение: Условной энтропией X при фиксированном y принадлежащем Y называется:

$$H(X|y) = -\sum_{x \in X} p(x|y) \log p(x|y),$$

Определение: Условной энтропией X при фиксированном У является

$$H(X|Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x|y)$$

Запишем свойства условной энтропии:

<u>Свойства</u>

1. $H(X|Y) \geq 0$
2. $H(X|Y) \leq H(X)$
3. $H(X)=H(X)+H(X|Y)=H(Y)+H(X|Y)$
4. $H(X_1,\ldots X_n)=H(X_1)+H(X_2|X_1)+\ldots.+H(X_n|X_1\ldots X_{n-1})$

5. H(X|YZ)≤H(X|Y)

$$H(X_1 ... X_n) \leq \sum_{i=1}^{n} |H(X_i),$$

7. ЗАКЛЮЧЕНИЕ

На данной лекции были рассмотрены: история развития теории информации, применимость данного направления на практике, математические основы теории кодирования, понятие энтропии и необходимые основы теории вероятности.