Лекция 3

1. Цели и задачи кодирования
2. Определения, термины и основные характеристика кодов
3. Классификация двоичных кодов
4. Способы представления кодов
5. Понятие об избыточности информации
6. Основная теорема кодирования для КБШ
7. Оптимальное (эффективное) кодирование

1. Цели и задачи кодирования

- **Кодирование широко применяется в математике и информатике (информационных технологиях)**

Примеры

1. Десятичная позиционная система счисления — способ кодирования натуральных чисел.
2. Декартовы координаты — способ кодирования геометрических объектов числами.
3. Представление данных произвольной природы в памяти компьютера*.
4. Защита информации от несанкционированного доступа.
5. Обеспечение помехоустойчивости при передаче данных по КС.
6. Сжатие информации в базах данных.
7. Составление текстов программ.

Необходимость кодирования информации, передаваемой по КС (цели кодирования)

1. Первичный алфавит ИИ достаточно объемен, а количество качественно различимых признаков сигналов мало. Поэтому информацию кодируют в некотором вторичном алфавите, который существенно меньше, чем первичный.
2. В КБШ применяют оптимальное кодирование для согласования ИИ с КС.
3. В КСШ осуществляют помехоустойчивое кодирование для повышения достоверности передачи сообщений.
4. Кодирование позволяет осуществлять первичное закрытие информации, т.е. обеспечивать скрытность передаваемых сообщений.
5. В целом, кодирование позволяет осуществлять согласование параметров КС и передаваемых сообщений, что позволяет более экономно использовать полосу частот КС, а также уменьшать стоимость передачи и хранения сообщений.

Типичная задача теории кодирования

- При заданных алфавитах A, B и множестве сообщений S найти такое кодирование* F, которое обладает определенными свойствами (т.е. удовлетворяет заданным ограничениям) и оптимально в некотором смысле.
- Критерий оптимальности обычно связан с минимизацией длин кодов.

Свойства кодирования

1. Существование декодирования*
2. Помехоустойчивость или исправление ошибок:

$$F^{-1}(\beta) = F^{-1}(\beta'),$$

   если β' в определенном смысле близко к β.
3. Заданная сложность (или простота) кодирования и декодирования**.

Множество сообщений S

- Обычно множество S является очень большим или бесконечным.
- Природа S во многом определяет оптимальное решение при одних и тех же A и B и требуемых свойствах кодирования.

Способы описания множества S

A) теоретико-множественное,

B) вероятностное,

C) логико-комбинаторное описание множества S.

2. Определения, термины, основные характеристика кодов

1. Длина кодовой комбинации (n) — количество символов в кодовом сообщении. Например,

   $$1101010 \rightarrow n=7.$$

2. Вес кодовой комбинации (w) — количество содержащихся в ней единиц. Например,

   $$11011001 \rightarrow w=5.$$

3. *Весовая характеристика W (w) — количество кодовых комбинаций с весом w. Например, код содержит комбинации: 0000, 0001, 0011, 1100, 1110. Тогда

   W (0)=1, W (1)=1, W (2)=2, W (3)=1

Определения, термины и основные характеристика кодов

4. Количество проверочных (r) и информационных символов (k) в кодовой комбинации. Если кодовая комбинация содержит n символов, то

$$n=k+r \tag{1}$$

5. Избыточность кода — отношение числа проверочных символов к общей длине кодовой комбинации

$$D=\frac{r}{n} \tag{2}$$

Определения, термины и основные характеристика кодов

6. **Скорость кода** — отношение числа информационных символов к общей длине кодовой комбинации*.

$$R = \frac{k}{n} \qquad (3)$$

7. Различают **одиночные**** ошибки и **пакеты** ошибок. Одна часть кодов обнаруживает и исправляет одиночные ошибки, другая — пакеты ошибок.

Определения, термины и основные характеристика кодов

8. Кратность одиночной ошибки — количество искаженных элементов в принятой кодовой комбинации. Например, переданной комбинации 1101 может соответствовать принятая комбинация 11<u>1</u>1 (однократная ошибка), 1<u>01</u>1 (двукратная) и т.д.
9. В результате действия ИП* ошибки в КС могут группироваться в пакеты. Пакетом или пачкой ошибок называют подряд расположенные символы, где искажен первый и последний символ. Длина пакета определяется от первого до последнего искаженного символа. Например, переданная последовательность 110<u>0</u>0<u>1</u>0<u>0</u>1001 была поражена пакетом ошибок длиной в 5 разрядов: 110100011001. Здесь поражены 4-й, 6-й и 8-й разряды.

Определения, термины и основные характеристика кодов

10. Общее число кодовых комбинаций*

$$N=m^n \qquad (4)$$

11. Расстояние (d) — количество несовадающих одноименных позиций в кодовых комбинациях. Расстояние определяют как вес суммы по модулю два двух комбинаций. Например,

$$\begin{array}{r} \oplus \begin{array}{r} 11011 \\ \underline{10101} \end{array} \\ 01110 \end{array} \longleftarrow (d=w=3)$$

12. Минимальное кодовое расстояние** d_{min} — минимально возможное расстояние между любой парой кодовых комбинаций данного кода.

3.Классификация двоичных кодов

- Различают эффективные (неизбыточные) и помехоустойчивые коды

3.Классификация двоичных кодов

1. Эффективные (неизбыточные)

Классификация двоичных кодов (Пример)

Неравномерный двоичный код	Равномерный двоичный код
1	0001
10	0010
11	0011
100	0100
101	0101
110	0110
111	0111
1000	1000

Классификация двоичных кодов

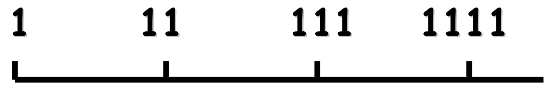
2. Помехоустойчивые

4. Способы представления кодов*

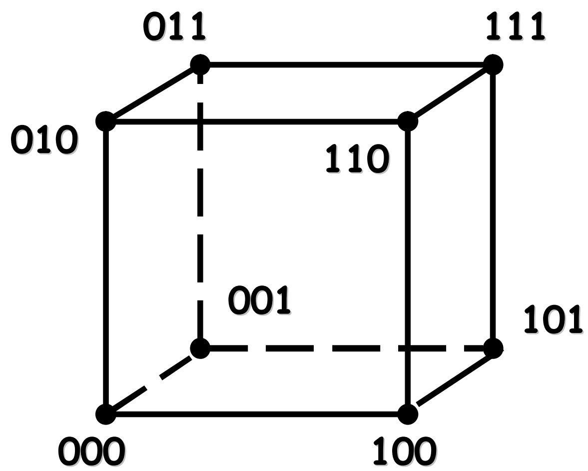
1. Геометрический

- Кодовые комбинации m-значного кода рассматривают как определенные точки в n-мерном пространстве**.

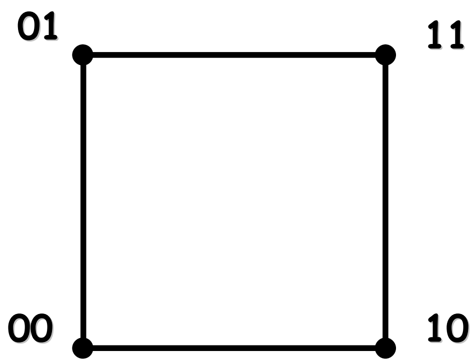
n=1



n=3



n=2



Способы представления кодов

2. Матричный*

- Пусть дана матрица **G** размерности $k \times n$, состоящая из элементов g_{ij} , где i — номер строки, j — номер столбца.
- Элементы g_{ij} принимают значения 0 или 1.
- Кодирование реализуется операцией

$$\vec{b} = \vec{a}G \qquad (1)$$

или

$$b_j = a_1 g_{1j} + a_2 g_{2j} + \ldots + a_k g_{kj}$$

где кодовые слова рассматриваются как векторы, т.е как матрицы–строки размера $1 \times n$.

Способы представления кодов

Пример

Рассмотрим матрицу 3×6

$$G=\begin{bmatrix}1&0&0&1&1&0\\0&1&0&0&1&1\\0&0&1&1&1&1\end{bmatrix}$$

Тогда кодирование задано отображениями*

000→000000; 001→001111; 010→010011;
011→011100; 100→100110; 101→101001;
110→110101; 111→111010.

Способы представления кодов

Выводы

1. Рассмотренный пример показывает **преимущества** матричного кодирования перед табличным: достаточно запомнить k кодовых слов вместо 2^k слов.
2. Кодирование не должно приписывать одно и то же кодовое слово разным исходным сообщениям. Простой способ добиться этого состоит в том, чтобы k первых столбцов G образовывали единичную подматрицу*.

Способы представления кодов

3. Алгебраический

- При практическом построении помехоустойчивых кодов применяют математический аппарат алгебраической теории кодирования. В частности, кодовое слово представляют в виде многочлена с фиктивной переменной x:

$$\boxed{F(x) = a_0 + a_1x + a_2x^2 + \ldots + a_{n-1}x^{n-1} = \sum_{i=0}^{n-1} a_i x^i} \qquad (2)$$

где n-длина кодового слова;

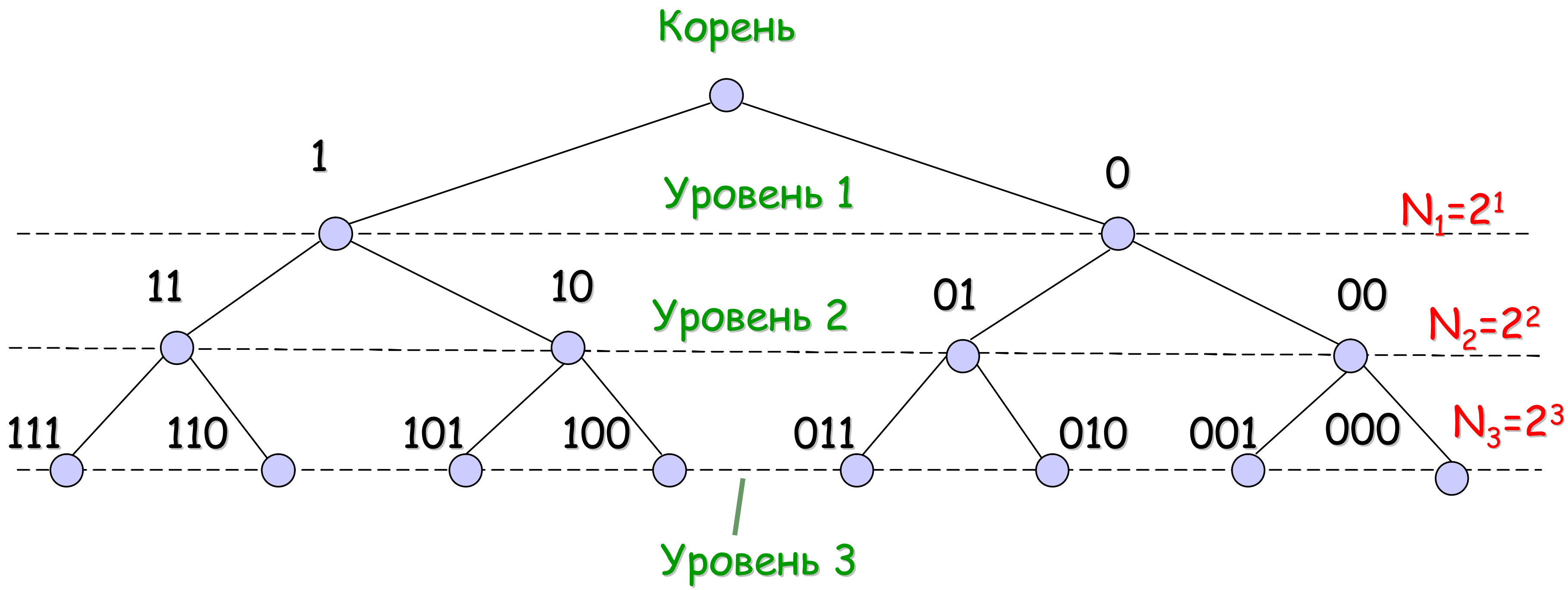
$a_i \in \{0;1\}$**

Способы представления кодов

4. С помощью кодовых деревьев

В общем виде кодовое дерево представляет собой граф, состоящий из узлов и ветвей, соединяющих узлы, расположенные на разных уровнях. Истоком графа является корень. Каждый уровень содержит m^n узлов, где n — номер уровня, а m — значность кода.

Способы представления кодов



Способы представления кодов

Выводы

- При помощи кодовых деревьев наглядно представляют коды, обладающих свойством **префикса***.
- Префиксные коды можно получить последовательным вычеркиванием последнего знака кодовой комбинации.
- Например, префиксами комбинации **A=1101101001** являются: 1; 11; 110; 1101; 11011; 110110; 1101101; 11011010; 110110100; 1101101001.

5. Понятие об избыточности информации

- Известно, что максимальное количество информации на символ сообщения можно получить только в случае равновероятных и независимых символов.
- Реальные коды обычно не удовлетворяют этому условию в полной мере.
- Поэтому вводят понятие информационной избыточности сообщения.
- Избыточность характеризует информационную недогруженность символа

Понятие об избыточности информации

- Различают естественную и искусственную избыточность.
- Естественная избыточность* относится к первичным алфавитам (либо заложена в структуре сообщения), а искусственная – к вторичным.
- Естественная избыточность бывает семантической и статистической.

Понятие об избыточности информации

- Семантическая избыточность связана с тем, что мысль, высказанная в сообщении м.б сформулирована более кратко. Устраняют семантическую избыточность в первичном алфавите (аббревиатуры, условные сокращения часто повторяющихся слов).

Понятие об избыточности информации

- Статистическая избыточность связана с тем, что символы ИИ имеют разную вероятность появления в сообщении. Некоторые ИИ могут иметь взаимозависимые вероятности появления символов (источники с памятью).

Пример

- Максимальная энтропия английского языка $H_{max}=\log_2 26=4{,}7$ бит. Если учесть взаимозависимость между символами и статистику следования слов в английских текстах, то энтропия английского языка не превысит 2 бит.

Понятие об избыточности информации

Выводы

Сообщения на естественных языках, используемые для передачи информации, можно значительно сжать.

- Для оценки сжатия сообщений (данных) применяют коэффициент сжатия

$$\mu = K_{сж} = \frac{H}{H_{max}} \qquad (1)$$

где H—текущая энтропия;

$$H_{max} = \log_2 m \qquad (2)$$

— максимальная энтропия, определяемая при равновероятном появлении символов в сообщении.

Понятие об избыточности информации

Полную статистическую избыточность определяют по формуле

$$D = 1 - \frac{H}{H_{max}} = 1 - \mu \qquad (3)$$

Ее составляющими являются:

- D_s — избыточность, вызванная статистической связью* между символами сообщения;
- D_p — избыточность, связанная с неравномерным распределением символов в сообщении.

Понятие об избыточности информации

$$D_s = 1 - \frac{H}{H'} \qquad (4)$$

где

$$H = H(X/Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i)p(y_j/x_i)\cdot \log_2 p(y_j/x_i), \qquad (5)$$

$$H' = -\sum_i p_i \log_2 p_i \qquad (6)$$

Понятие об избыточности информации

- Избыточность D_p характеризует информационный резерв сообщений с равновероятными символами относительно сообщений, символы которых неравновероятны*

$$D_p = 1 - \frac{H'}{H_{max}} \qquad (7)$$

Полная избыточность рассчитывается по формуле

$$D = D_S + D_P - D_P \times D_S \qquad (8)$$

Понятие об избыточности информации

Избыточность округления*

- Эта составляющая избыточности может проявляться при передаче десятичных цифр двоичным кодом.

Пример

- Длину кодовой комбинации вычисляют по формуле

$$L \geq \frac{\log_2 m_1}{\log_2 m_2} = \varphi \qquad (9)$$

- где m_1 и m_2 — количество качественных признаков первичного и вторичного алфавитов

Понятие об избыточности информации

- Определим избыточность округления при передаче пяти сообщений двоичным блочным кодом.
- Из (8) следует, что

$$L \geq \frac{\log_2 5}{\log_2 2} \approx 2{,}32 \text{ бит}$$

- Однако, количество разрядов должно быть **целым**, поэтому результат округляем до трех.
- Тогда*

$$D_o = \frac{k - \varphi}{k} = \frac{3 - 2{,}32}{3} \approx 0{,}23$$

Понятие об избыточности информации

Выводы

- Избыточность округления не возникает, если общее количество сообщений равно целочисленной степени двойки.
- Искусственная избыточность связана с введением r дополнительных проверочных символов помимо k информационных ($D=r/k$).

6. Основная теорема кодирования для КБШ (Средняя длина кодового слова)

- В КБШ потери информации отсутствуют. Однако, для построения эффективной информационной системы, обеспечивающей передачу больших объемов информации при минимальных временных и материальных затратах, а также для однозначного декодирования принятых сообщений необходимо решить ряд технических задач.

A. Разные символы первичного алфавита, из которого составлены сообщения, должны иметь различные кодовые комбинации.
B. Код д.б. построен так, чтобы можно было четко отделить начало и конец букв первичного алфавита.
C. Код д.б. максимально кратким: чем меньшее число элементарных символов требуется для передачи данного сообщения, тем ближе скорость передачи информации к пропускной способности КС.

- Требование **А)** очевидно, т.к. при одинаковых кодовых обозначениях букв алфавитов их нельзя будет различить при декодировании.
- Требование **В)** можно удовлетворить различными способами:
  - Введением в код дополнительного разделительного символа (паузы), что значительно удлиняет коды, а следовательно, и время передачи сообщения;
  - применяя префиксные коды;
  - применяя комбинации **равномерного*** кода, в котором все буквы передаются комбинациями равной длины.

- Чтобы выполнить требование C), применяют теорему кодирования

1. При кодировании множества сигналов с энтропией H в алфавите, насчитывающем m* символов, при условии отсутствия шумов, средняя длина кодового слова не м.б. меньше, чем H/(log m).
2. Если вероятности символов не являются отрицательными степенями числа m, то точное достижение границы H/(log m) невозможно, но при кодировании достаточно длинными блоками к этой границе можно сколь угодно приблизиться.

Основная теорема кодирования для КБШ

- Обозначим количество независимых букв в блоке **M**, а среднюю длину кодового слова данного алфавита **L**. Тогда

$$\frac{H}{\log m} \le L < \frac{H}{\log m} + \frac{1}{M} \qquad (1)$$

- Из (1) видно, что если количество независимых букв в блоке велико ($M \to \infty$), то среднее число элементарных символов, затрачиваемых на передачу одной буквы, неограниченно приближается к величине **H/(logm)**

Основная теорема кодирования для КБШ

- Если коды двоичные (**m=2**), то основную теорему кодирования формулируют следующим образом:

  > при кодировании сообщений в двоичном алфавите с ростом количества кодовых слов **среднее число двоичных знаков** на букву сообщений **приближается** к **энтропии** источника сообщений.

Основная теорема кодирования для КБШ

Среднее число двоичных знаков на букву в точности равно энтропии источника сообщений, если:

1. кодируемый алфавит равновероятный и $m=2^i$ ($i = 1, 2, 3,\dots$);
2. вероятности появления сигналов являются целочисленными отрицательными степенями двойки ($p_i=2^{-i}$).

Основная теорема кодирования для КБШ

Выводы

1. Чем длиннее первичное кодовое слово, тем точнее величина $H/(\log m)$ характеризует среднюю длину кодового слова.
2. Чем больше длина блока, тем меньше разность между верхней и нижней границами, определяющими среднее число элементарных символов на букву сообщения.
3. Из какого бы числа букв не состоял алфавит, целесообразно кодировать сообщения не побуквенно, а поблочно.
4. Энтропия первичного алфавита может характеризовать возможный предел сокращения кодового слова во вторичном алфавите.

7. Оптимальное (эффективное) кодирование

- Оптимальное кодирование предполагает, что помехи в КС отсутствуют или минимальны. Поэтому оптимальные коды имеют кодовое расстояние равное единице, т.е. они лишены искусственной избыточности и, следовательно, не обнаруживают и не исправляют ошибок. Преимущество оптимальных кодов состоит в том, что они позволяют увеличивать скорость передачи и уменьшают расход памяти.
- Все оптимальные коды делятся на равномерные и неравномерные*.

Оптимальное (эффективное) кодирование

Определения

1. Оптимальным кодированием называется преобразование символов первичного алфавита m_1 в кодовые слова во вторичном алфавите m_2, при котором средняя длина сообщений во вторичном алфавите имеет минимально возможную для данного m_2 длину.
2. Оптимальными называются коды, представляющие кодируемые символы кодовыми словами минимальной средней длины.

Оптимальное (эффективное) кодирование

Свойства оптимальных кодов

1. Минимальная средняя длина кодового слова оптимального кода обеспечивается в том случае, когда избыточность каждого кодового символа сведена к минимуму (в идеальном случае — к нулю).
2. Кодовые слова оптимального кода должны строиться из равновероятных и взаимонезависимых символов.