Parul® University

Information Theory & Coding

Prof. Pankaj Chaudhary, Assistant Professor
Information Technology

Parul®
University

DIGITAL LEARNING CONTENT

CHAPTER-1

Basic Concept of Coding

Information Theory and Coding

- Information is the source of any communication channel whether it is an analog signal or it is a digital signal. Information can be generated in several formats, it is a sequence of letter, alphabet, symbol, binary coded symbol, images, audio, video, etc. are all considered as a form of information in a digital world.
- Information theory is a branch of probability theory which may be applied to the study of communication systems that deals with the mathematical modelling and analysis of a communication system rather than with the physical sources & physical channels.

Parul®
University

DIGITAL LEARNING CONTENT

Information Theory and Coding

- Information of an event depends only on its probability of occurrences and does not depend on its content.
- The randomness of happening of an event and the probability of its prediction as a news is known as information.
- The message associated with the least likelihood of an event contains the maximum information.

Brief of Coding Techniques

- “A Coding theory is the mathematical studies of the properties of codes and their respective suitability for specific applications”. Codes are used for data compression, cryptography, error detection and correction, data transmission and data storage.

Unique decodable codes

- Sequence of code that can be decodable in only one way.
- When a single A_X and A_Z are the source and code alphabets.
- A^+_X and A^+_Z denote sequences of one or more symbols from the source or code alphabets.
- A symbol code, C, is a mapping $A_X \rightarrow A^+_Z$. We use c (x) to denote the code-word C maps x to.
- Now define a code to be uniquely decodable if the mapping $C^+ : A^+_X \rightarrow A^+_Z$ is one-to-one.

Unique decodable codes

- A code is said to be unique decodable code if the given code can be decoded in only one way.

- E.g: Consider the codewords in the following two codes:

  (i) Code: (0, 10, 110, 111)

  (ii) Code: (0, 10, 010, 101)

Instantaneous decodable codes (IDC)

- We define a code to be instantaneously decodable if any source sequences x and x ′ in A+X for which x is not a prefix of x ′ have encodings z = C (x) and z ′ = C (x′) for which z is not a prefix of z′.
- Let us, take an example to see the working of unique decodable and IDC techniques:

Code X:
a: 10
b: 11
c: 111

Instantaneous decodable codes (IDC)

Code Y:
a: 0
b: 10
c: 110

Code X: Not uniquely decodable as both bbb and cc encode as 111111.
Code Y: Instantaneously decodable, end of each code-word marked by 0.

Instantaneous decodable codes (IDC)

Consider a code:

a1: 0
a2: 1
a3: 10
a4: 11

10 can be coded as a3 or a2a1.
So this is not UDC.

Instantaneous decodable codes (IDC)

Consider a code:

a1: 0
a2: 10
a3: 110
a4: 111

100 can be coded as a2a1.
So this is UDC.

Instantaneous decodable codes (IDC)

Insantaneous code:- A code is instantáneos if & only if no codeword is a prefix of some other codeword.

Insantaneous Decodable code: It is the code in which each codeword in any string of codewords can be decoded as soon as it is received.

E.g: (i) Code: (0, 10, 110) Instantaneous decodable

(ii) Code: (100, 101, 010, 011) Instantaneous decodable

(iii) Code: (10, 11, 111) Not instantaneous decodable because 11 is a prefix of 111.

Construction of IDC

- **Kraft's inequality:**

There is an instantaneous binary code with code-words having lengths l1, .

. . , ll if and only if:

$$\sum_{i=1}^{I} \frac{1}{2^{l_i}} \leq 1$$

Construction of IDC

- **Kraft's inequality:**

There is an instantaneous binary code with lengths 1, 2, 3, 3, since

$1/2 + 1/4 + 1/8 + 1/8 = 1$

An example of such a code is {0, 10, 110, 111}.

There is an instantaneous binary code with lengths 2, 2, 2, since

$1/4 + 1/4 + 1/4 < 1$

An example of such a code is {00, 10, 01}.

Construction of IDC

- **McMillan's Theorem:**

There is a uniquely decodable binary code with code-words having lengths l1, . . . , ll if and only if.

$$\sum_{i=1}^{I} \frac{1}{2^{l_i}} \leq 1$$

Construction of IDC

- **McMillan's Theorem:**

There is a uniquely decodable binary code with lengths 1, 2, 3, 3, since

1/2 + 1/4 + 1/8 + 1/8 = 1

An example of such a code is {0, 01, 011, 111}.

There is no uniquely decodable binary code with lengths 2, 2, 2, 2, 2, since

1/4 + 1/4 + 1/4 + 1/4 + 1/4 > 1does not satisfied the above inequality.

Huffman Coding

- Huffman Coding is a technique of compressing data to reduce its size without losing any of the details. It will help to reduce the number of character from the large set of frequently occurred characters. We have to reduce the size with the help of binary tree.
- Tree can be constructed in bottom-up manner. Root node is consider as a final reduction step.
- The working of Huffman based on the Greedy approach. We have to sort the characters in ascending order.

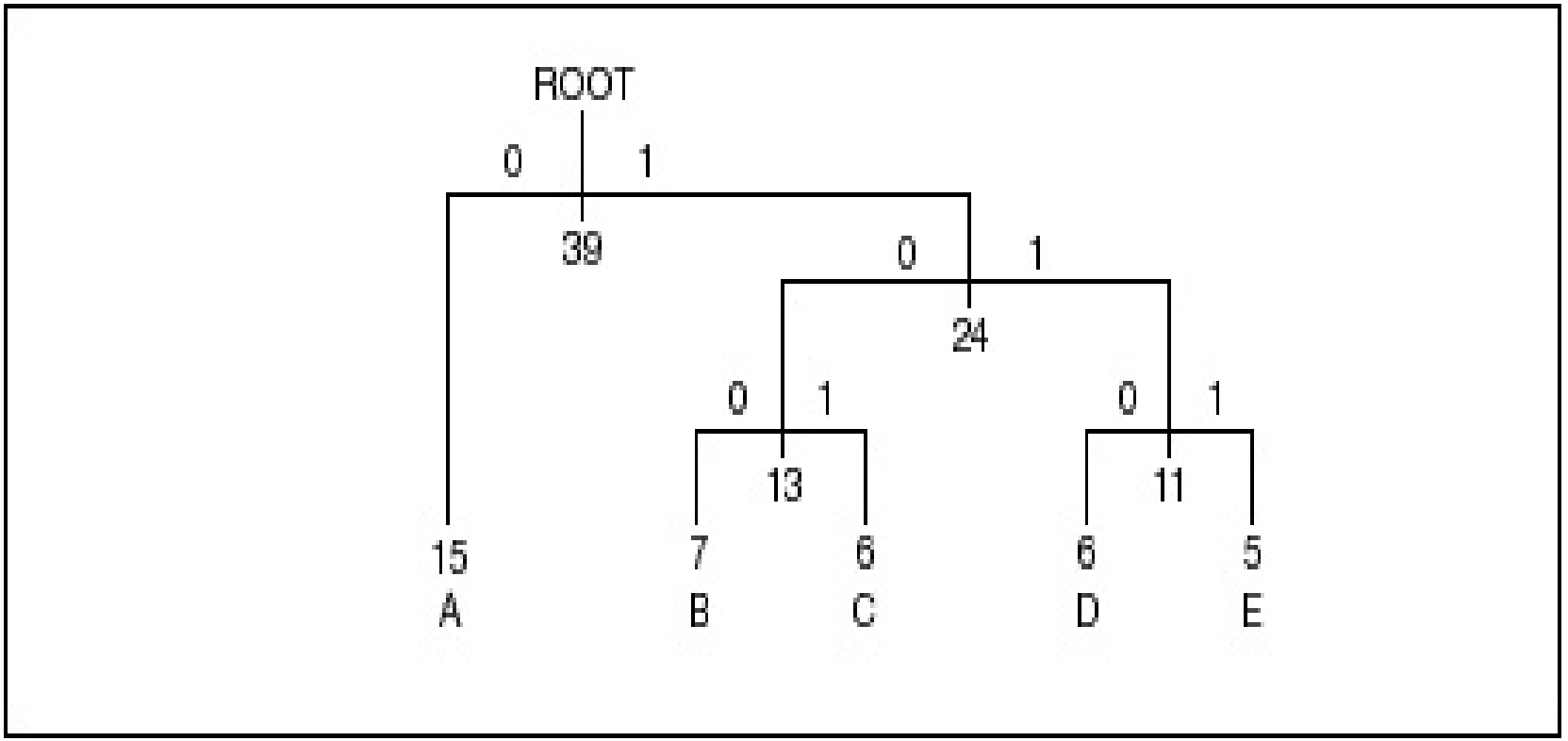
Huffman Coding

- Huffman coding will also help to remove the ambiguity from the different set of code-words. Let there be four characters p, q, r and s, and their corresponding variable length codes be 00, 01, 0 and 1. This coding leads to ambiguity because code assigned to r is the prefix of codes assigned to p and q.
- If the compressed bit stream is 0001, the de-compressed output may be “rrrs” or “rrq” or “prs” or “pq”. There is an ambiguity to select a correct code-word, this is how Huffman coding will guarantee that there will be no ambiguity after decoding the code-word.

Huffman Coding

Symbol and	Count
A	15
B	7
C	6
D	6
E	5

Huffman Coding



Huffman Coding

Assign 0 to left branches, 1 to right branches
Each encoding is a path from the root

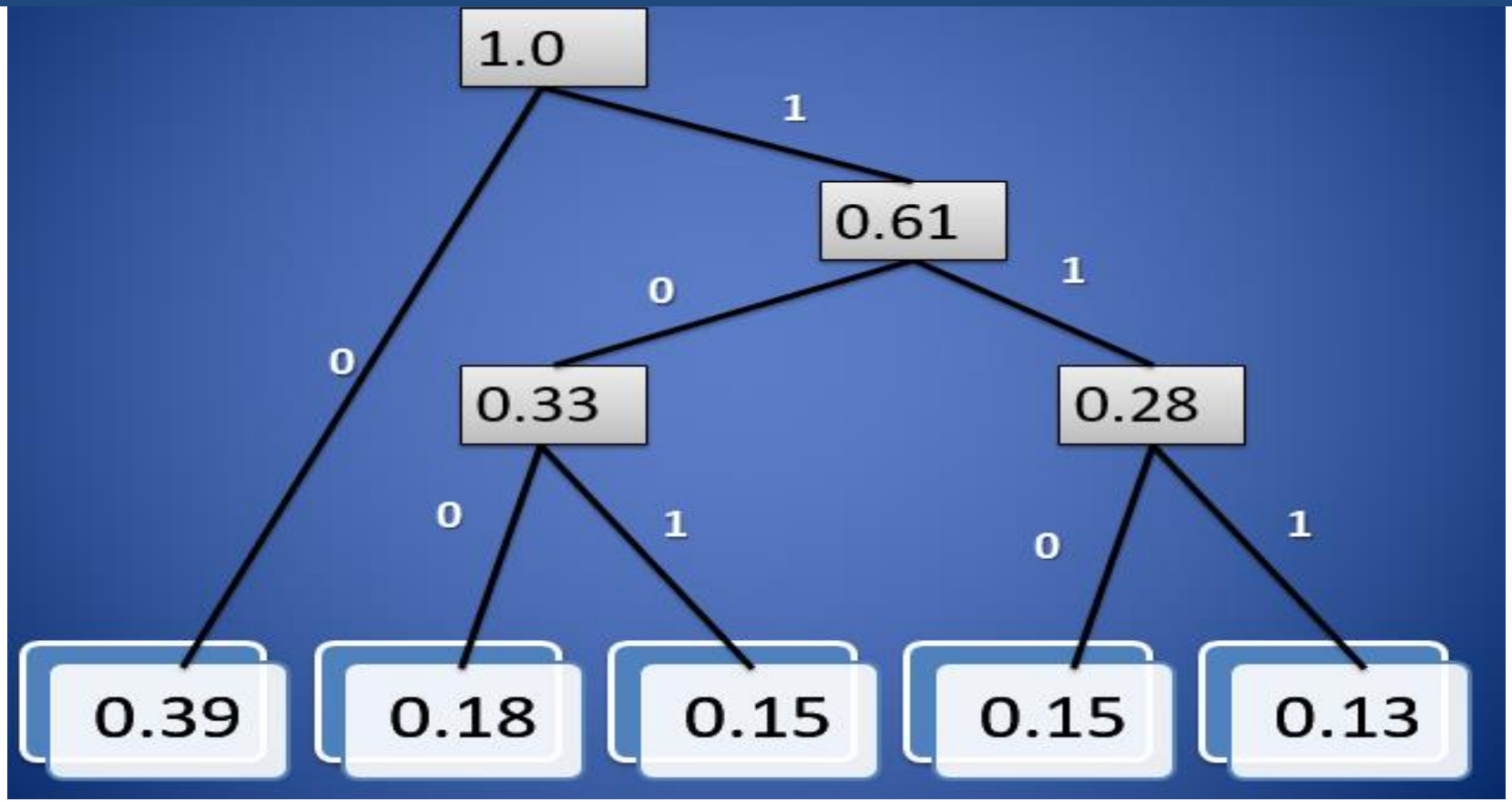
A=0
B=100
C=101
D = 110
E = 111

Each path terminates at a leaf.

Huffman Coding using probabilities

Symbol	Count	Probability
A	15	15/39 = 0.39
B	7	7/39 = 0.18
C	6	6/39 = 0.15
D	6	6/39 = 0.15
E	5	5/39 = 0.13

Huffman Coding using probabilities



Shannon-Fano Code.

- It is a technique for constructing a prefix code based on a set of symbols and their probabilities. In Shannon–Fano (SF) coding, the symbols are arranged in order from most probable to least probable, and then divided into two sets whose total probabilities are as close as possible to being equal.
- The prefix code generate by SF is not always an optimal. SF-coding is used in the IMPLODE compression method, which is part of the ZIP file format.

Shannon-Fano Code.

- For a given list of symbols, develop a corresponding list of probabilities or frequency counts.
- So that each symbol's relative frequency of occurrence is known.
- Sort the lists of symbols according to frequency with the most frequently occurring symbols at the top.
- Divide the list into two parts, with the total frequency counts of the upper half being as close to the total of the bottom half as possible.

Shannon-Fano Code.

- The upper half of the list is assigned the binary digit 0, and the lower half is assigned the digit 1.
- This means that the codes for the symbols in the first half will all start with 0, and the codes in the second half will all start with 1.
- Recursively apply the steps 3 and 4 to each of the two halves, subdividing groups and adding bits to the codes until each symbol has become a corresponding code leaf on the tree.

Parul® University

DIGITAL LEARNING CONTENT

Shannon-Fano Code.

Consider six symbol alphabet where the probability of each symbol is given below:

A 0.30

B 0.25

C 0.20

D 0.12

E 0.08

F 0.05

Shannon-Fano Code.

Consider six symbol alphabet where the probability of each symbol is given below:

		stage 1	stage 2	stage 3	stage4	code
A	0.30	0	0	-	-	00
B	0.25	0	1	-	-	01
C	0.20	1	0	-	-	10
D	0.12	1	1	0	-	110
E	0.08	1	1	1	0	1110
F	0.05	1	1	1	1	1111

Shannon-Fano Code.

- Example: The source of information A generates the symbols {A, B, C, D and E} with corresponding probabilities {0.4, 0.3, 0.15, 0.1 and 0.05}. Encoding the source symbols using binary encoder and Shannon-Fano encoder gives:

- Binary Code: {000,001,010,011,100}. Average length=3

- Shannon-Fano: {0, 10, 110, 1110, 1111} = Average length=2.05

$$Entropy\ (H) = -\sum_{i=0}^{n} Prob_i log(Prob_i)$$

Shannon-Fano Code.

After substituting the Probability value in the above formula we get, Entropy=2.0087.

Now we have to calculate the efficiency,

$$Efficiency\ (E) = \frac{Entropy\ (H)}{Average\ codeword\ Length\ (L)}$$

Efficiency of the binary code= 67%.

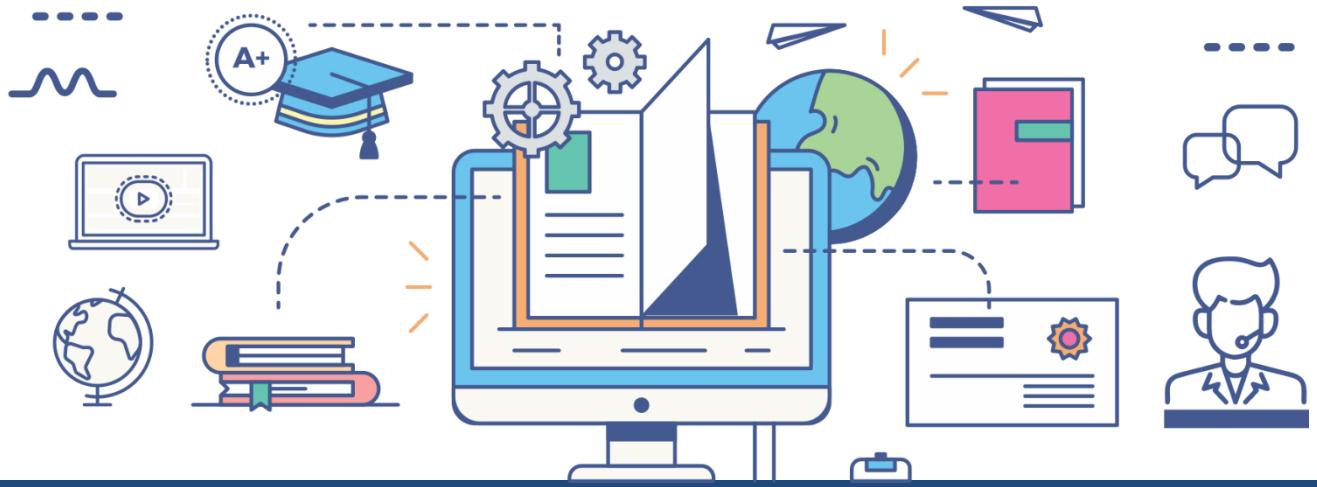
And Efficiency of the Shannon-Fano code= 98%.

Thus we can conclude from the above example that the efficiency of the Shannon-Fano encoder is much higher than that of the binary encoder.

References

1. https://www.cl.cam.ac.uk/teaching/0809/InfoTheory/InfoTheoryLectures.pdf
2. https://cs.nyu.edu/~roweis/csc310-2005/notes/lec2x.pdf.
3. http://www.cs.toronto.edu/~radford/csc310/week2.pdf

DIGITAL LEARNING CONTENT



Parul® University

www.paruluniversity.ac.in