Parul® University

Information Theory & Coding

Prof. Pankaj Chaudhary, Assistant Professor
Information Technology

CHAPTER-2

Entropy

Why learn entropy?

- Because Entropy helps us in compressing the data. Through entropy, we got to know how many bits per symbol is enough.
- Because we have to send the message to the receiver with minimum number of bits.

Introduction of entropy

- Entropy is a measure of image information content, which is interpreted as the average uncertainty of information source.
- In Image, Entropy is defined as corresponding states of intensity level which individual pixels can adapt.
- Entropy tells us that how many bits are required for each symbol so that there is no loss of data and also the information is conveyed efficiently.

Introduction of entropy

- If the sample space is X={x1, x2, x3,…….., xN} with probability distribution P, the entropy of the probability distribution is given by

$$H(P) = -\sum_{i=1}^{N} P(x_i)\log(P(x_i))$$

Introduction of entropy

- Let us consider some examples of probability distributions and see how the Entropy is related to prediction.

❖ If $P(X1) = P(X2) = 0.5$.

So Entropy $= (-0.5)(-1) - (0.5)(-1) = +1$

In this case, X1 and X2 are equally likely to occur and the situation is as unpredictable.

❖ If $P(X1) = 1$ and $P(X2) = 0$

So Entropy$= 0$

In this case X1 will always occur because the value of entropy is zero.

Entropy of sources and their extension

Consider the information content (I) of a message of n symbols, each describing an event (E) which occurs with a probability of P(E). Then

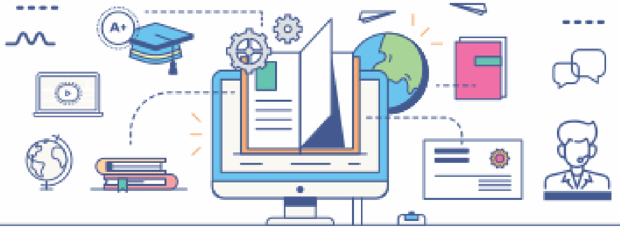
I increases as n increases

as $P(E) \rightarrow 0$, $I \rightarrow \infty$ (unexpected outcome)

If $P(E) = 1$ then $I = 0$ (outcome certain)

If two messages of n1 and n2 symbols are concatenated, the resulting information is added, i.e. $I = I1 + I2$

$I(E) = \log (1 / P(E)) = -\log P(E)$



Entropy of sources and their extension

An extension of a source is that new source which results when the emitted symbols are considered in groups.

Thus for the binary source {0,1}, the first three extensions have alphabets

{0, 1},

{00, 01, 10, 11}

{000, 001, 010, 011, 100, 101, 110, 111}.

The nth extension of a source S of q symbols is denoted by Sn. Sn has an alphabet of qn symbols, where each is a sequence of n symbols from S.

It is both obvious and easily proven that H(Sn) = nH(S). In other words the entropy of the nth extension of a source is n times the entropy of the source itself.

Loss less image compression

- Lossless compression is a class of data compression algorithms that allows the original data to be perfectly reconstructed from the compressed data.

- By contrast, lossy compression permits reconstruction only of an approximation of the original data, though usually with greatly improved compression rates

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 if 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

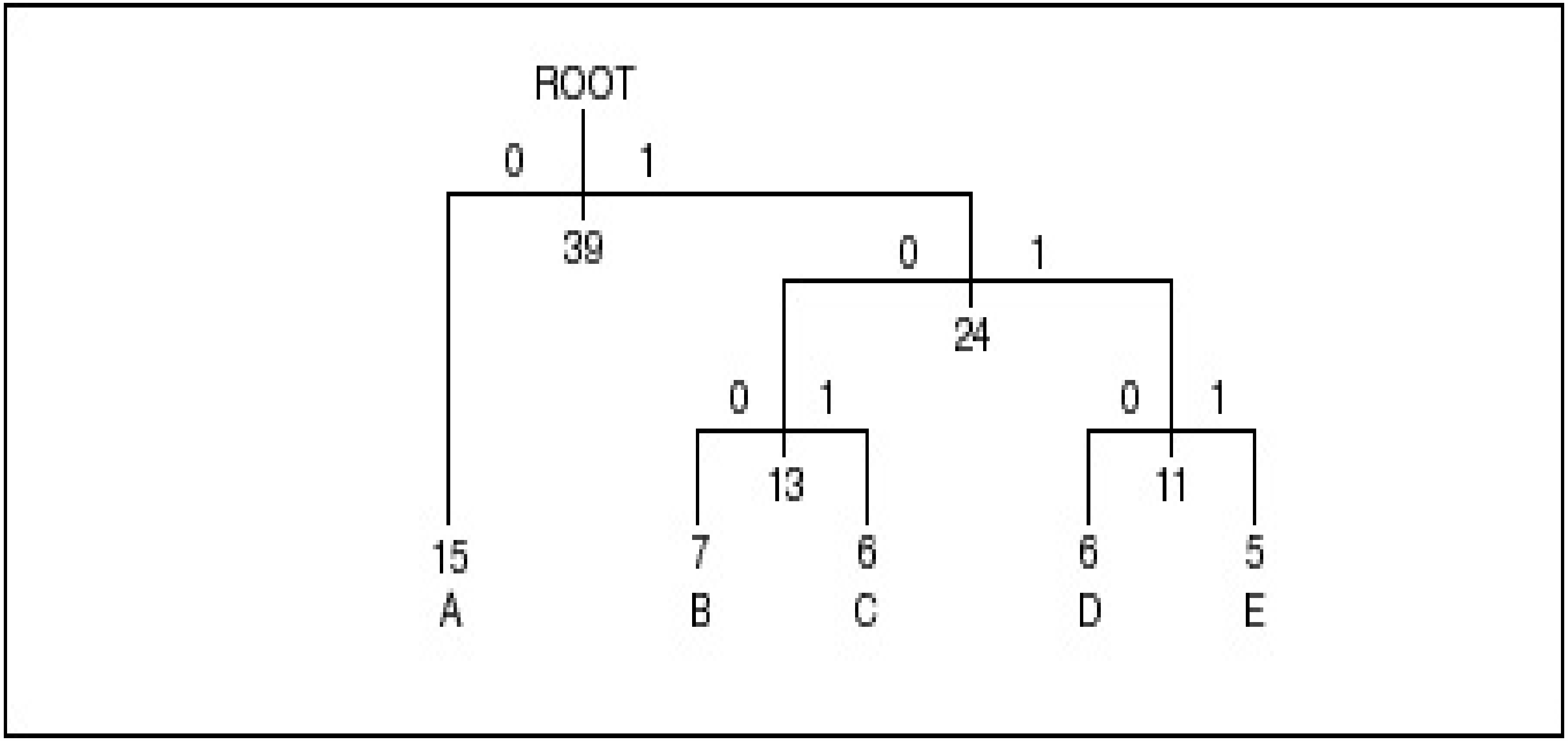
- Huffman coding will also help to remove the ambiguity from the different set of code-words. Let there be four characters p, q, r and s, and their corresponding variable length codes be 00, 01, 0 and 1. This coding leads to ambiguity because code assigned to r is the prefix of codes assigned to p and 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

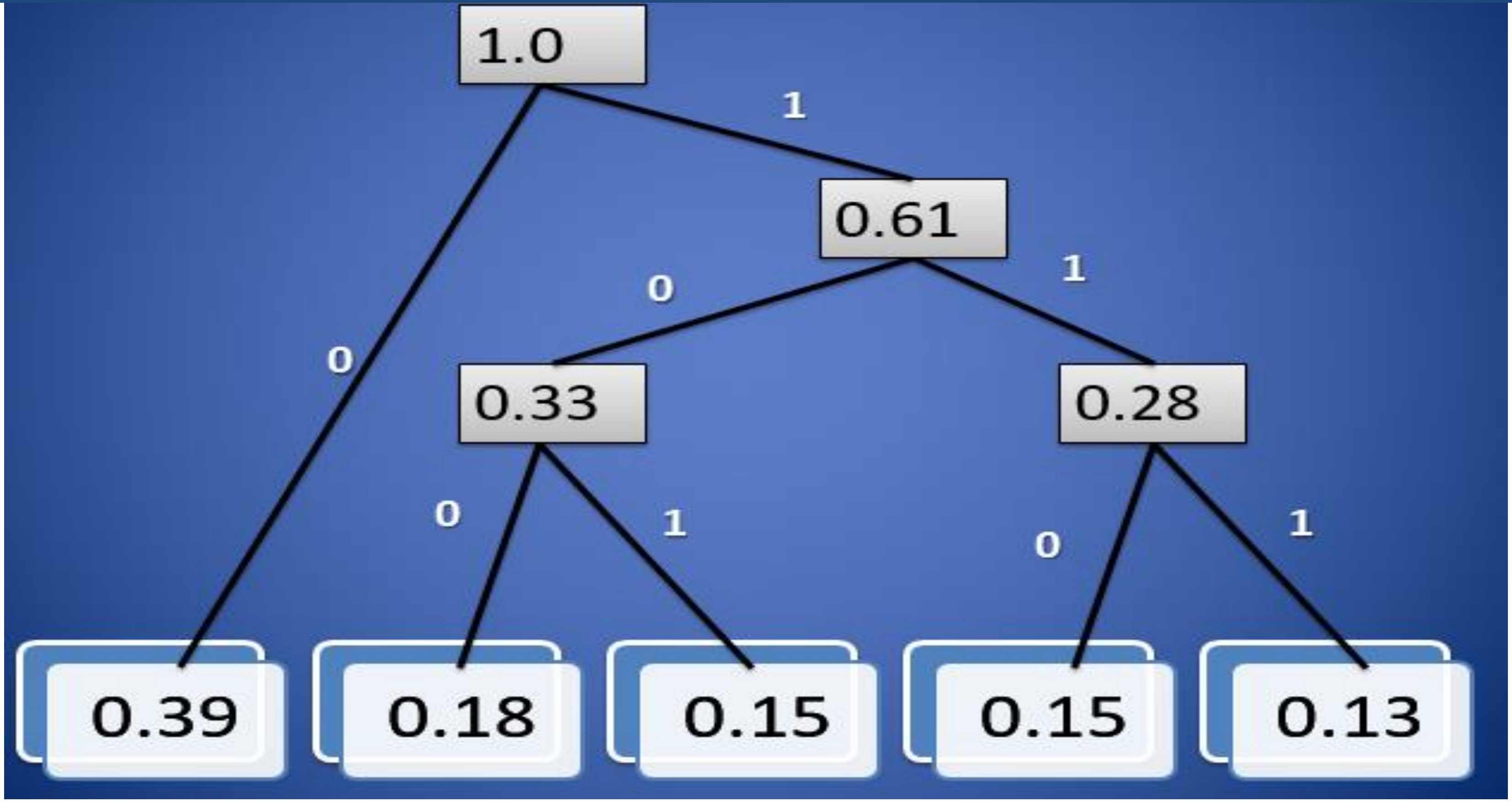
A=0
B=100
C=101
D = 110
E = 111

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



References

1. N. Abramson, Information and Coding, McGraw Hill, 1963.
2. http://www.stat.yale.edu/~yw562/teaching/itlectures.pdf
3. Gray, R. M., Entropy and Information Theory, Springer (2011).
4. http://poseidon.csd.auth.gr/LAB_PUBLICATIONS/Books/dip_material/chapter_4/chap4en.pdf

DIGITAL LEARNING CONTENT

Parul® University

www.paruluniversity.ac.in