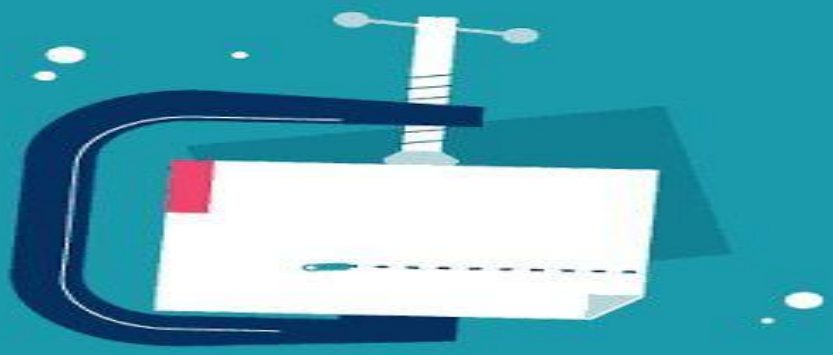
CS101 PROJECT



Project Title: Understanding Data Compression based on Shannon's Information theory

Group Members:

1.Prathistha Pandey(2022CSB1105)

2.Saaransh Sharma(2022CSB1114)

3.Sagar singh(2022CSB1115)

1. What is Data compression ?

Data compression is the process of reducing the size of our data files by using some protocols . In this we encode our data using fewer bits in order to reduce the cost of storage and increase the speed of algorithms.Compression algorithms compress the data based on its statistical properties.

There are two main types of data compression:

1.Lossless compression: This method reduces the size of data without the loss of any information.It is used to compress the files where all the detail is important.
Examples of this compression are : RAR, PNG, ZIP and GZIP.

2.Lossy compression: This method reduces the size of data by discarding some of its information.It is used for data where minor loss of quality is acceptable and achieves significant file size

reduction.These algorithms use a technique that removes all the irrelevent information.Examples:JPEG for images and MP3 for audio.

2. Shannon's Information theory

The Shannon information theory was given in the late 1940s by Claude Shannon.This theory deals with the sensible measure of information content in an event or in a message. It introduces the concept of entropy which determines the uncertainty of a possible outcome.It is a famous theory as it is applied in many fields like data compression, cryptography,telecommunication,and many more.
The important assertions of this theory are:

1.Shannon information content: It is a sensible measure of information content of an outcome and is represented by h(x).
$h(x)= (\log_2 1/P(x))$

2.Entropy: It is simply the average amount of information provided by a message or an outcome . Data compression algorithms focus on reducing the entropy of the data based on its statistical properties and hence compress it .The entropy is denoted by H(X) and is given by a simple formula given below.

$$H(X) = \Sigma P(x) \log_2 (1/P(x))$$

Here, x represents an outcome of an event X and P(x) represents the probability of the occurring of that outcome . The unit of entropy is a bit.

3.Source Coding theorem : This theorem states that the 'N' outcomes from a source X can be compressed into minimum possible bits given by 'N*H(X)' . It provides a way to reduce or compress the data to smaller size for efficient storage or transmission without losing any information.

2.1 An Intuitive Idea of Entropy

The idea of entropy can be understood more simply and intuitively by the following example.

Suppose 8 teams are playing football and every team has equal probability of winning . We want to declare the outcomes by using bits.

So for doing this we would need 3 bits to represent every possible outcome as is given by the formula.

Shannon information content for each outcome

$= \log_2 (1/1/8)) = \log_2 8 = 3$

The entropy of the event is given by

$H(X) = \Sigma P(x) \log_2 (1/P(x))$

$H(X) = 8*(⅛*3) = 3$

Total number of bits required to represent the event = N*H(X) = 8*3=24.Hence we need a total of 24 bits to represent the result .

2.2 Data Compression using Shannon's Information Theory

Say, we have a probability distribution X $\{x, A_x, P_x\}$ where;

x is a random variable.

A_x is a set of possible outcomes.

P_x is a set of probabilities of these outcomes.

$Ax = \{a_1, a_2, a_3, \ldots\ldots, a_i\}$,

$Px = \{p_1, p_2, p_3, \ldots\ldots, p_i\}$,

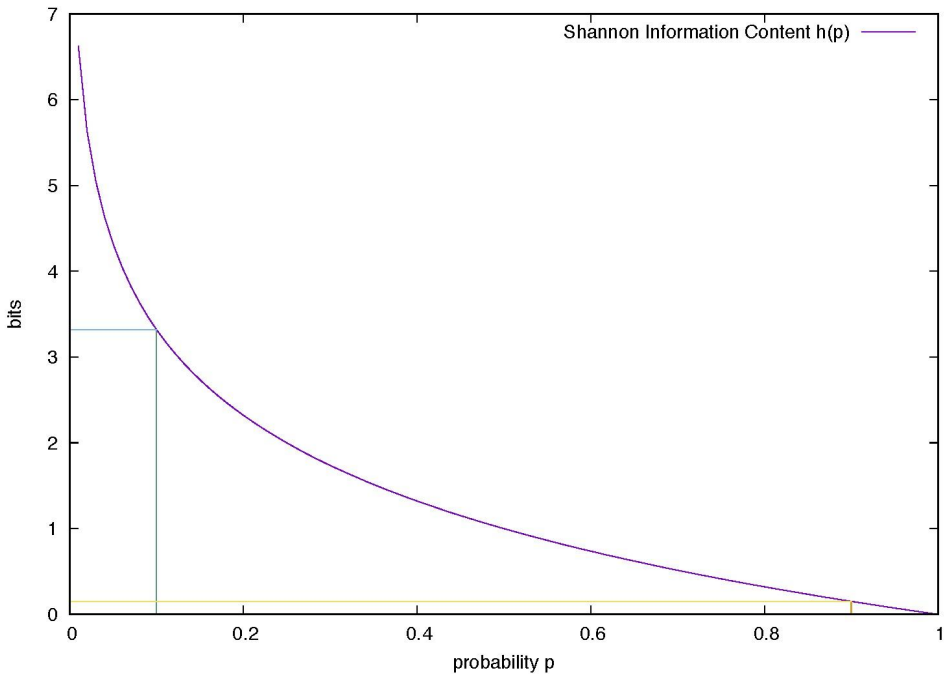
such that, $P(x=a_i)=p_i$;

$\sum p_i = 1$;

Shannon's Idea:

Shannon's information content {h(x)} of an outcome:

$$h(x) = (\log_2 1/P(x)) \text{ bits ;}$$



According to Shannon the least probable events provide us the most information content while it is vice-versa for events with a higher probability.

Shannon claimed that h(x) is the compressed file length we should aim for.

2.3 Shannon Fano Coding Algorithm

This encoding algorithm is used for lossless compression and was developed by Claude Shannon and Robert M. Fano.This algorithm provides a unique variable length code to the symbols based on the probability of their occurrence.It gives shorter length codes to more frequently occurring symbols and longer length codes to less frequently occurring symbols.

The steps followed are:

1.Find the probability of occurrence of each element and then sort them.
2.Build a binary tree by recursively dividing the elements into two approximately equal parts until all the elements are divided and a binary tree is obtained.

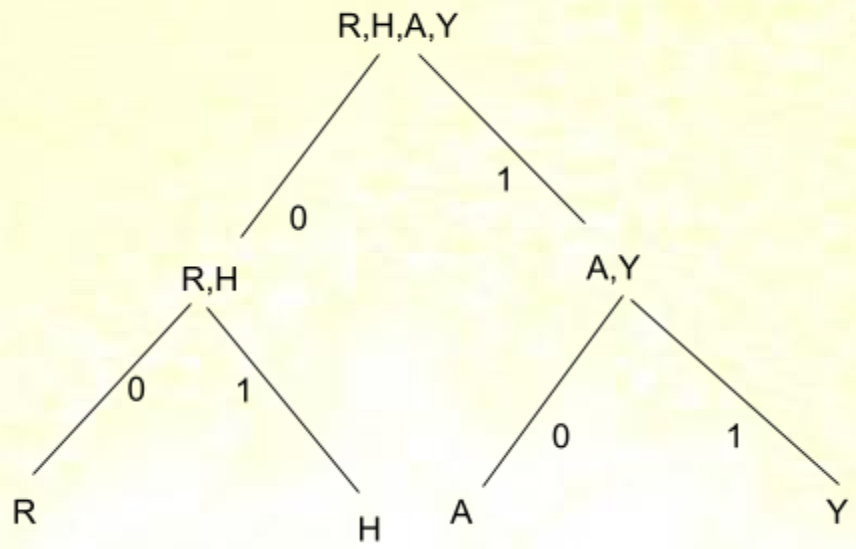
An example is provided below to understand this algorithm completely.

Coding of word “HARRY”
The frequency count is :

symbol	H	A	R	Y

frequency	1	1	2	1
probability	1/5	1/5	2/5	1/5

Form the binary tree:



SYMBOL	COUNT	CODE	NO. of bits used	h(x)
H	1	01	2	2.32
A	1	10	2	2.32
R	2	00	2	1.32
Y	1	11	2	2.32

Total no. of bits used = (2+2+2*2+2)=10

Assume each symbol requires 8 bits,total bits=8*5=40

Percentage of compression = (40-10)*100/40 = 75%

3. Conclusion

We have seen how data compression plays a very vital role in transferring data and information, when the data sizes become absurd it is an absolute necessity to compress its size into smaller bits. Claude Shannon's information theory has helped achieve this very efficiently. His concept of information content and entropy is pretty accurate and can be applied to practical problems where we need to attain information and data. The idea behind data compression using Shannon's theory is that we try compressing the bit size as much as possible while retaining the maximum amount of information, this is achieved by the Shannon-Fano coding algorithm which is used for lossless compression. There are many practical applications of this technique in the domains of telecommunication,

image compression (eg. JPEG), video compression (eg. MPEG), audio compression (eg. MP3).