e-PG Pathshala

Subject : Computer Science

Paper: Machine Learning

Module: Information Theory

Module No: CS/ML/9

Quadrant I – e-text

Welcome to the e-PG Pathshala Lecture Series on Machine Learning. In this module we will discuss some fundamental aspects of Information Theory which we will later on use in the discussion of many machine learning techniques.

Learning Objectives:

The learning objectives of this module are as follows:

- To understand the principles and practice of Information Theory
- To know the interpretation of probability in real life
- To acquire knowledge about information and entropy

9.1 Information Theory -Quotes about Shannon's Concepts

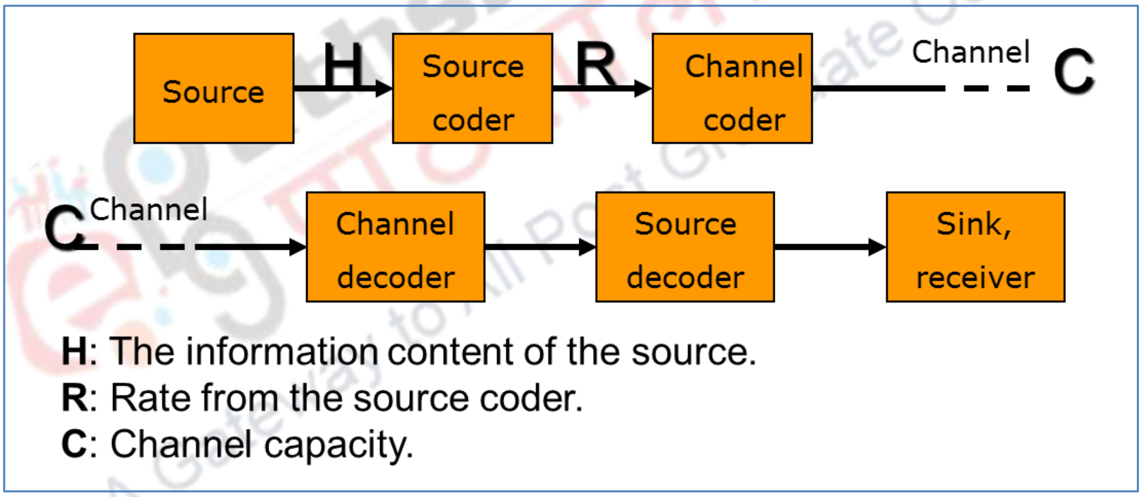
Information theory is a branch of science that deals with the analysis of a communications system for efficient and reliable transmission of information. Before we start discussing Information theory we need to know some famous quotes about Claude Shannon's ideas considered as the developer of this theory. He came up with the extraordinary idea now known as the mathematical theory of communication which was published in the Bell System Technical Journal in 1948. This landmark paper was the beginning of the branch of information theory. He proposed that the basic principles of binary or digital information can be related to fundamental physical laws. He was instrumental in shaping our digital era. Today, Shannon's theory remains the guiding foundation for the most modern, faster, more energy efficient, and more robust communication systems So

- ”What is information? Sidestepping questions about meaning, Shannon showed that it is a measurable commodity”.
- ”Today, Shannon's insight help shape virtually all systems that store, process, or transmit information in digital form, from compact discs to computers, from facsimile machines to deep space probes”.

- ”Information theory has also infilitrated fields outside communications, including linguistics, psychology, economics, biology, even the arts”.

9.2 Shannon’s Channel of Communication

The communication model considered for our purpose consists of a source that generates digital information. This information needs to be sent from the source to the receiver through a channel. The channel can be noiseless where the channel transmits symbols without causing any errors. In order to use the redundant characteristics of the information output by the source and reduce the length of the transmission, data compression also called as source coding is carried out. In this case the information needs to be decompressed at the receiver end. However sometimes the channel can be noisy which causes errors in the received symbols at the destination. To reduce the errors incurred due to noise, channel coding is carried out.



H: The information content of the source.
R: Rate from the source coder.
C: Channel capacity.

Figure 9.1 Shannon’s Channel of Communication

Figure 9.1 shows Shannon’s model of a communication channel. In this context, channel is anything that transmits or stores information such as radio link, cable, disk, CD or even a piece of paper. Considering Figure 9.1, **C** is the channel capacity, source is any source of information where **H** is considered as the information content of the source, source coder changes the information content into an efficient representation (that is data compression) where **R** is the rate of the source coder. Channel coder again changes the compressed information into an efficient representation for transmission (that is error control coding). This compressed, error coded information is then transmitted through channel C. On receiving the information it is decoded to recover from channel distortion. Now the source decoder uncompresses the information and finaaly the receiver gets the information.

9.2.1 Fundamental Theorems

Maximizing the amount of information that can be transmitted over an imperfect communication channel can be based on data compression (entropy) and transmission rate (channel capacity).

Using this basic communication model, Shannon came up with two theorems, today known as Shannon's theorems.

- **Shannon 1**: Error-free transmission is possible if $R \geq H$ and $C \geq R$.
- **Shannon 2**: Source coding and channel coding can be optimized *independently*, where *binary symbols* can be used as intermediate format assuming that the delays are arbitrarily long.

9.3 Information Source

The characteristics of any information source can be specified in terms of the number of symbols n, say S1,S2,….,Sn, and the probability of occurrence of each of these symbols, $P(S_1)$, $P(S_2)$, …, $P(S_n)$ and finally the correlation between successive symbols. A stream of symbols from the sender to the receiver is generally called a message. If a source emits signals which are independent then the source is called a memoryless source.

9.3.1 Stochastic sources

Let us assume that a source outputs symbols X1, X2, ... and that each symbol takes its value from an alphabet A = (a1, a2, …). The model: P(X1,…,XN) will be known for all combinations.

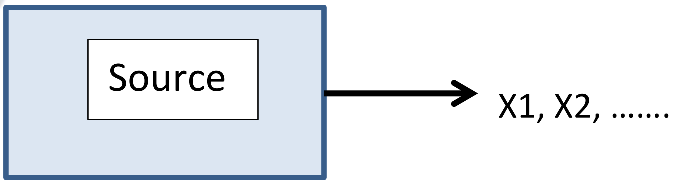


Figure 9.2 A Stochastic Source

Some examples are given below:

Example 1: A text which is a sequence of symbols each taking its value from the alphabet A = (a,.....z, A, ...Z, 1,2,3....9, !,?,....)

Example 2: A (digitized) grayscale *image* is a sequence of symbols each taking its value from the alphabet A = (0,1) or A = (0, …, 255).

For such stochastic sources there are two special cases

- The Memoryless Source where the value of each symbol is independent of the value of the previous symbols in the sequence.

  P(S1, S2, …, Sn) = P(S1) . P(S2) . … .P(Sn)

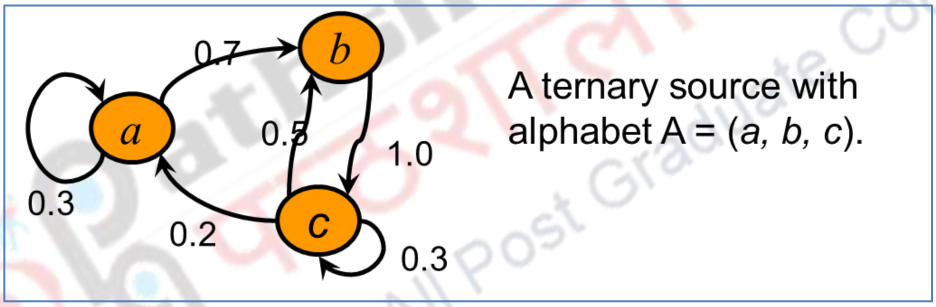
- The Markov Source where the value of each each symbol depends only on the value of the previous one in a sequence

  P(S1, S2, …, Sn) = P(S1) .P(S2|S1) .P(S3|S2) . … . P(Sn|Sn-1)

We will discuss the Markov source in detail.

9.3.2 The Markov Source

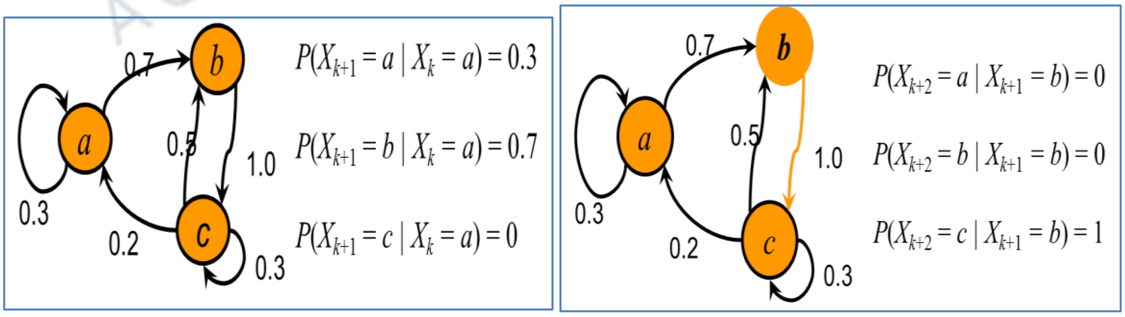
In the case of a Markov source, a symbol depends only on the previous symbol, so the source can be modelled by a state diagram. Now let us assume we have a source with an alphabet of size 3 namely A = (a,b,c).



A ternary source with alphabet A = (*a*, *b*, *c*).

Figure 9.3 A Markov Source

Let us assume that initially we are in state *a*, i.e., $X_k = a$. Then we can find the probabilities for the next symbol



$$P(X_{k+1} = a \mid X_k = a) = 0.3$$
$$P(X_{k+1} = b \mid X_k = a) = 0.7$$
$$P(X_{k+1} = c \mid X_k = a) = 0$$

$$P(X_{k+2} = a \mid X_{k+1} = b) = 0$$
$$P(X_{k+2} = b \mid X_{k+1} = b) = 0$$
$$P(X_{k+2} = c \mid X_{k+1} = b) = 1$$

Figure 9.4 (a) and (b) State Transitions in an Example of a Markov Source

Figure 9.4(a) shows the probabilities of transitions from a to a, b and c respectively. Note that since there is no transition from a to c the corresponding transition probability is 0. Now let us assume we have gone from state a to b. Figure 9.4(b) shows the probabilities of transitions from b to a, b, and c. In this case, from b, the only transition possible is to c. That is, if we know that $X_{k+1} = b$, with a probabbility of 0.7, given $X_k = a$; we also know that X_{k+2} can only be c.

In a Markov source if all the states can be reached, the *stationary probabilities* for the states can be calculated from the given transition probabilities. Stationary probabilities are the probabilities $\omega_i = P(X_k = a_i)$ for any *k* when X_{k-1}, X_{k-2}, … are not given.

Markov models can be used to represent sources which have dependencies that go back more than one step by using a state diagram with several symbols in each state.

9.3.3 Stochastic Sources - Analysis and Synthesis

Stochastic models can be used for *analysing* a source. For this purpose we need to find a model that represents the real-world source in an effective manner, and then analyse the model instead of the real world. Stochastic models can also be used for *synthesizing* a source. In this case we can generate random numbers sequentially to simulate the source of the Markov model.

Measuring Information

Before we discuss entropy associated with information, let us understand how to measure the information contained in a message. In other words we need to know the amount of information a message carries from the sender to the receiver.

Example 1 -Imagine a person sitting in a room. Looking out through the window, she can clearly see that the sun is shining. If at this moment she receives a call from a neighbor saying "It is now daytime", does this message contain any information?

Example 2 -A person has bought a lottery ticket. A friend calls to tell her that she has won first prize. Does this message contain any information?

As you can conclude, in example 1 the message carries no information while the message in example 2 carries information.

Now let us assume a binary memoryless source, for example the flip of a coin. The question is the amount of information obtained when on flipping the coin we get heads.

If the coin is a fair coin, i.e., P(heads) = P (tails) = 0.5, as is normally the case the amount of information obtained is 1 bit. However if we already have the information that it is heads, i.e., P(heads) = 1, then the amount of information obtained is zero. The amount of information is greater than 0 and less than 1 if we have an unfair coin.

The information content of a message is inversely proportional to the probability of occurrence of that message. If a message is very probable, it does not contain any information. If it is very improbable, it contains a lot of information

9.4.1 Self Information

Now let us look at how Shannon viewed the information conveyed by a symbol. Assume a memoryless source with alphabet A = (a_1, …, a_n) and symbol probabilities (p_1, …, p_n). Then an important question is to find out the amount of information obtained by knowing that the next symbol is $a_{i..}$Shannon associated a concept called self information to find out amount of information conveyed by $a_{i.}$

$$I(a_i) = \log \frac{1}{p_i}$$

If we have two independent events *A* and *B*, with probabilities *P*(*A*) = p_A and *P*(*B*) = p_B. For both the events to happen, the probability is p_A . p_B. However, the amount of information should be added, not multiplied.

$$I(p_A.p_B) = I(p_A) + I(p_B)$$

Logarithms can be used to convey this information. Actually we can use any logarithm but generally if we use 2 as the base of the logarithm, and get *bits*.

$$I(A) = \log(p_A)?$$

Since the information conveyed increases with decreasing probabilities information is given by taking the negative of the logarithm.

$$I(A) = -\log(p_A) = \log \frac{1}{p_A}$$

Example 1 $\quad \mathrm{p_i} = 0.5 \Rightarrow \mathrm{I}(0.5) = \log_2 \frac{1}{0.5} = 1[\mathrm{bit}]$

Example 2 $\quad \mathrm{p_i} = 1 \Rightarrow \mathrm{I}(1) = \log \frac{1}{1} = 0$

9.4.2 Entropy

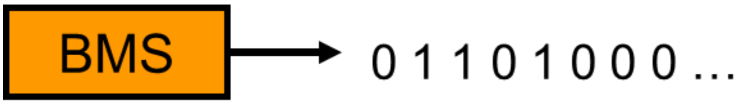
In order to find information about the symbols in a sequence we can *average over the probability over all the symbols*, given as:

$$\boxed{H = \sum_{1}^{N} p_i a_i}$$

This value H(*X*) is called the first order *entropy* of the source. This can be regarded as the degree of *uncertainty* about the following symbol.

The ***minimum*** average number of binary digits needed to specify a source output (message) uniquely is called source entropy. Entropy is the average length of the message needed to transmit an outcome using the optimal code.

Example: Binary Memoryless Source

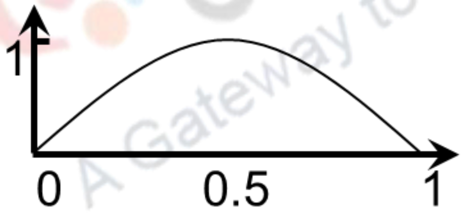


$$p = P(X_k = 1), q = P(X_k = 0) = 1 - p$$

Then

$$\boxed{H = p \log \frac{1}{p} + (1 - p) \log \frac{1}{1 - p}}$$

where H is often denoted as h(p). The uncertainty (information) is greatest when p=q= ½



9.4.2.1 Entropy: Three properties

Entropy is a measurement of information. It has three basic properties. It can be shown that $0 \leq H \leq \log N$. Maximum entropy ($H = \log N$) is reached when all symbols are equiprobable, i.e., $p_i = 1/N$. The difference between maximum entropy and actual entropy $\log N - H$ is called the *redundancy* of the source.

9.5 Entropy for Memory and Markov Sources

Entropy for Memory Source

Assume a block of source symbols ($X_1, \ldots, X_n$) , we can define the *block entropy*:

$$H(X_1,\ldots, X_n) = \sum_1^{Nn} P(X_1,\ldots, X_n)\log\frac{1}{P(X_1,\ldots, X_n)}$$

In this case summation is carried over all possible combinations of *n* symbols. If we let the block length go towards infintity we can divide by *n* to get the number of bits / symbol. Thus the entropy for a memory source is defined as:

$$H_\infty = \lim_{n\to\infty}\frac{1}{n}H(X_1,\ldots, X_n)$$

Entropy for a Markov Source

The entropy for a state S_k can be expressed as

$$H(Sk) = \sum_{l=1}^{r} P_{kl}\log\frac{1}{P_{kl}}$$

In this case P_{kl} is the transition probability from state k to state l. Averaging over all states, we get the entropy for the Markov source as

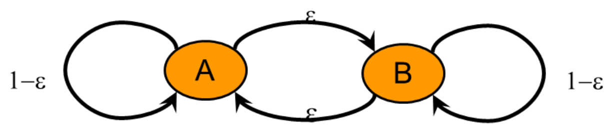
$$H_M = \sum_{k=1}^{r} P(S_k)H(S_k)$$

where summation lower index k = 1.

The Run-length Source

Certain sources generate long *runs* or *bursts* of equal symbols. Let us look at an example

Example:



Probability for a burst of length *r*: $P(r) = (1\text{-}\varepsilon)^{r-1}.\varepsilon$ and the Entropy: $H_R = -\sum_{r=1}^{1} P(r) \log P(r)$.
If the average run length is taken to be ρ, then $H_R/\rho = H_M$.

9.6 The Source Coding Theorem

As we have already discussed entropy is the smallest number of bits allowing error-free representation of the source. Shannon shows that source coding algorithms exist that have a unique average representation length that approaches the entropy of the source and we cannot reduce beyond this length. To understand this aspect let us consider typical sequences.

Typical Sequences

We assume a *long* sequence from a binary memoryless source with *P*(1) = *p*. Among n bits, there will be approximately *w* = *n* .*p* ones. Thus, there is *M* = (*n choose w*) such *typical sequences*!

Only these sequences are interesting. All other sequences will appear with smaller probability as *n becomes larger*. Now we need to find out how many sequences are actually typical sequences.

$$\boxed{\mathrm{M} = (\mathrm{nCw}) = \frac{\mathrm{n!}}{\mathrm{w!}(\mathrm{n}-\mathrm{w})!}}$$

(Stirling: n! ≈ $\frac{n^n}{e^n}\sqrt{2\pi n}$) bits/symbol

$$\boxed{\frac{1}{n}\log(\frac{n^{n}}{w^{w}(n-w)n-w}\cdot\frac{const}{\sqrt{n}})}$$

Equation 1

Enumeration needs log *M* bits, i.e,

$$\boxed{\frac{1}{n}\log M}$$

bits per symbol!

Deriving from equation 1 by substituting w=n.p we get

-p log p – (1-p) log (1-p) = h(p) =H(X).

Thus we can conclude that we need *H*(*X*) bits per symbol to code any typical sequence.

Example –Sequence

Consider the following sequence:

1 2 1 2 4 4 1 2 4 4 4 4 4 4 1 2 4 4 4 4 4 4

Obtaining the probability from the sequence

1 four times (4/22), 2 four times (4/22), and 4 fourteen times (14/22)

$$a_i = 1 \Rightarrow I(1) = \log_2 \frac{1}{0.1818} = 2.4595$$

$$a_i = 2 \Rightarrow I(2) = \log_2 \frac{1}{0.1818} = 2.4595$$

$$a_i = 4 \Rightarrow I(4) = \log_2 \frac{1}{0.6363} = 0.6522$$

- On *average over all the symbols*, we get:

$$\boxed{H = \sum_1^N p_i I(a_i)}$$

The entropy H = 0.447 + 0.447 + 0.415 = 1.309 bits

There are 22 symbols, we need 22 * 1.309 = 28.798 (29) bits to transmit

Now we can consider the symbols as blocks of 12, 44 (now total symbols-11) -

12 appears 4/11 and 44 appears 7/11

$$a_i = 12 \Rightarrow I(12) = \log_2 \frac{1}{0.3636} = 1.4595$$

$$a_i = 44 \Rightarrow I(1) = \log_2 \frac{1}{0.6363} = 0.6522$$

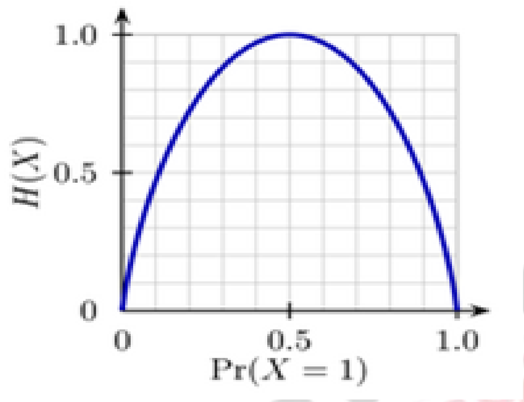
$$H = \sum_1^N p_i I(a_i) = 0.530 + 0.415 = 0.945 \ bits$$

There are 11 symbols, we need 11 * 0.945 = 10.395 (11) bits to transmit information – that is we might be able to find patterns with less entropy

9.7 Using information theory

Entropy measures the amount of uncertainty in a probability distribution:

Consider tossing a biased coin. If you toss the coin VERY often, the frequency of heads is, say, p, and hence the frequency of tails is 1-p. (fair coin p=0.5). Uncertainty in any actual outcome is given by entropy: Uncertainty is zero if p=0 / 1 and maximal if we have p=0.5.



- p examples which are true (positive) and n examples which are false (negative). -Our best estimate of true or false is given by:

$$P(true) \approx p / p + n$$
$$p(false) \approx n / p + n$$

$$Entropy\left(\frac{p}{p+n}, \frac{n}{p+n}\right) \approx -\frac{p}{p+n}\log\frac{p}{p+n} - \frac{n}{p+n}\log\frac{n}{p+n}$$

Information theory can be used to determine how much information we gain if we disclose the value of some attribute? (for example in a decision tree). Entropy can be used as a measure of the quality of our models and as a measure of how different two probability distributions are. We will discuss these aspects later.

Summary

- Explained the model of communication
- Outlined the sources of information
- Explained entropy

- Discussed source and Markov source
- Described typical sequences