Note of: An Introduction to Information Theory and Entropy

1 Entropy

Information Theory Measures related to how surprising or unexpected an observation or event is. This approach has been described as *information theory*.

Definition We have defined *information* strictly in terms of the probabilities of events. Therefore, let us suppose we have a set of probabilities(a probability distribution) $P = \{p1,p_2, \cdots, p_n\}$. We define the entropy of the distribution P by

$$H(P) = \sum_{i=1}^{n} p_i * \log(1/p_i)$$

We can think about this as the expected value. In other words, the entropy of a probability distribution is just the expected value of the information of the distribution.

The Gibbs Inequality First, note that the function $\ln(x)$ has derivative $1/x$. From this, we find that the tangent to $\ln(x)$ at $x = 1$ is the line $y = x - 1$. Further, since $\ln(x)$ is concave down, we have, for $x > 0$, that

$$ln(x) \leq x - 1$$

with equality only when $x = 1$. Now given two probability distributions,
$P = \{p_1, p_2, \cdots, p_n\} and$
$Q = \{q_1, q-2, \cdots, q_n\}$, where $p_i, q_i \geq 0$,
and $\sum_i p_i = \sum_i q_i = 1$, we have

$$\sum_{i=1}^{n} p_i \ln(q_i/p_i) \leq \sum_{i=1}^{n} p_i * (q_i/p_i - 1) = \sum_{i=1}^{n} (q_i - p_i) = \sum_{i=1}^{n} q_i - \sum_{i=}^{n} p_i = 1 - 1 = 0$$

with equality only when $p_i = q_i$ for all i.

Application of Gibbs Inequality We can use the Gibbs inequality to find the probability distribution which maximize the entropy function. Suppose $P = \{p_1, p_2, \cdots, p_n\}$ is a probability distribution. We have

$$\begin{aligned}
H(P) - log(n) &= \sum_{i=1}^{n} p_i * \log 1/p_i - log(n) \\
&= \sum_{i=1}^{n} p_i \log(1/p_i) - log(n) \sum_{i=1}^{n} p_i \\
&= \sum_{i=1}^{n} p_i \log(1/p_i) - \sum_{i=1}^{n} p_i \log n \\
&= \sum_{i=1}^{n} p_i \log \frac{1/n}{p_i} \\
&\leq 0
\end{aligned}$$

with equality only when $p_i = 1/n$ for all i. The last step is the application of Gibbs inequality.

That this means is that

$$0 \leq H(n) \leq \log(n)$$

That is, the maximum entropy is achieved when all the events are equally likely.

2 Shannon's Communication Theory

2.1 Introduction

In his classic 1948 papers, Claude Shannon laid the foundations for contemporary information, coding , and communication theory. He developed a general model for communication systems, and a set of theoretical tools for analyzing such systems.

His basic model consists of three parts: a sender(or source), a channel, and a receiver(or sink). His general model also includes encoding and decoding elements, and noise within the channel.

Model In Shannon's discrete model, it is assumed that the source provides a stream of symbols selected from a finite alphabet $A = \{a_1, a_2, \cdots, a_n\}$, which are then encoded. The code is sent through the channel(and possibly disturbed by noise). At the other end of the channel, the receiver will decode, and derive information from the sequence of symbols.

Given a source of symbols, and a channel with noise(in particular, a probability model for these elements), we can talk about the capacity of the channel. The general model Shannon worked with involved two sets of symbols, the input symbols and the output symbols. Let us say the two sets of symbols are

$$A = \{a_1, s_2, \cdots, a_n\} and$$
$$B = \{b-1, b_2, \cdots, b_m\}.$$

Note that we do not necessarily assume the same number of symbols in the two sets. Given the noise in the channel, when symbol b_j comes out of the channel, we can not be sure which a_i was put in. The channel is characterized by the set of probabilities $\{P(a_i|b_j)\}$.

Mutual Information We consider information we get from observing a symbol b_j. Given a probability model of the source, we have an priori estimate $P(a_i)$ that symbol a_i will be sent next. Upon observing b_j, we can revise our estimate to $P(a_i|b_j)$. The change in our information(the mutual information) will be given by:

$$\begin{aligned} I(a_i; b_j) &= log(\frac{1}{P(a_i)} - log(\frac{1}{P(a_i|b_j)}) \\ &= log(\frac{P(a_i|b_j)}{P(a_i)} \end{aligned}$$

We have the properties:

$$\begin{aligned} I(a_i; b_j) &= I(b_j; a_i) \\ I(a_i; b_j) &= log(P(a_i|b_j)) + I(a_i) \\ I(a_i, b_j) &\leq I(a_i) \end{aligned}$$

If a_i and b_j are independent (i.e., if $P(a_i, b_j) = P(a_i) * P(b_j)$), then

$$I(a_i, b_j) = 0$$

Average the mutual information over all the symbols

$$I(A; b_j) = \sum_i P(a_i|b_j) * I(a_i; b_j)$$
$$= \sum_i P(a_i|b_j) * log(\frac{P(a_i|b_j)}{P(a_i)})$$

Channel Capacity C We define the Channel Capacity to be

$$C = \max_{P(a)} I(A; B)$$

Property We have the nice property that if we are using the channel at its capacity, then for each of the a_i

$$I(a_i; B) = C$$

and thus, we can maximize channel use by maximizing the use for each symbol independently.

Shannon Main Theorem For any channel, there exist ways of encoding input symbols such that we can simultaneously utilize the channel as closely as we wish to the capacity, and at the same time have an error rate as close to zero as we wish.

This is actually quite a remarkable theorem. We might naively guess that in order to minimize the error rate, we would have to use more o the channel capacity for error detection/correction, and less for actual transmission of information. Shannon showed that it is possible to keep error rates low and still use the channel for information transmission at(or near) its capacity.

3 Some other measurements

Moment There have been various approaches to expanding on the idea of entropy as a measure of complexity. One useful generalization of entropy was developed by the Hungarian methematician A. Renyi. His method involves looking at the moments of order q of a probability distribution $\{p_i\}$

$$S_q = \frac{1}{q-1} log \sum_i (p_i)^q$$

how?

If we take the limit as $q \to 1$, we get:

$$S_1 = \sum_i p_i log(1/p_i)$$

the entropy we have previously defined. We can then think of S_q as a generalized entropy for any real number q.

Dimension Expanding on these generalized entropies, we can then define a generalized dimension associated with a data set. If we imaging the data set to be distributed among bins of diameter r, we can let p_i be the probability that a data item falls in the i-th bin(estimated by counting the data elements in the bin, and dividing by the total number of items). We can then, for each q, define a dimension

why make such definition?

$$D_q = \lim_{r \to 0} \frac{1}{q-1} \frac{log \sum_i (p_i)^q}{log(r)}$$

why do we call this a generalized dimension? Consider D_0. First, we sill adopt the convention that $(p_i)^0$ when $p_i = 0$. Also, let N_r be the number of non-empty bins(i.e., the number of bins of diameter r it takes to cover the data set).

Then we have:

$$D_0 = \lim_{r \to 0} \frac{log \sum_i (p_i)^0}{log(1/r)} = \lim_{r \to 0} \frac{log(N_r)}{log(1/r)}$$

Thus, D_0 is the hausdorff dimension D, which is frequently in the literature called the fractal dimension of the set.

Need Explain

Example 1 Consider the unit interval $[0, 1]$. Let $r_k = 1/(2)^k$. Then $N_{rk} = 2^{2k}$, and

$$D_0 = \lim k \to \infty \frac{log(2^k)}{log(2^k)} = 1$$

Example 2 Consider the unit square $[0, 1] * [0, 1]$. Again, let $r_k = 1/2^k$. Then $N_{rk} = 2^{2k}$, and

$$D_0 = \lim_{k \to \infty} \frac{log(2^{2k})}{log(2^k)} = 2$$

Example 3 Consider the Cantor Set

4 Examples using Bayes' Theorem