Note of: An Introduction to Information Theory and Entropy

Notes for Materials "An introduction to information theory and entropy", by Tom Carter.

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

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

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:

$$I(a_i; b_j) = I(b_j; a_i)$$
$$I(a_i; b_j) = log(P(a_i|b_j)) + I(a_i)$$
$$I(a_i, b_j) \leq I(a_i)$$

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

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

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 Analog Channels

Problem Model Suppose we have a signalling system using band-limited signals(i.e., the frequencies of the transmissions are restricted to lie within some specified range). Let us call the bandwidth W. Let us further assume we are transmitting signals of duration T. In order to reconstruct a given signal, we will need $2WT$why?? samples of the signal. Thus, if we are sending continuous signals, each signal can be represented by $2WT$ numbers x_i, taken at equal intervals.

why define energy in such form?? We are associated each signal an energy, given by

$$E = \frac{1}{2W} \sum_{i=1}^{2WT} (x_i)^2$$

The distance of the signal(from the original) will be

$$r = (\sum (x_i)^2)^{1/2} = (2WE)^{1/2}$$

We can define the signal power to be the average energy

$$S = \frac{E}{T}$$

Then the radius of the sphere of transmitted signals will be

$$r = (2WST)^{1/2}$$

Each signal will be disturbed by the noise in the channel. If we measure the power of the noise N added by the channel, the disturbed signal will lie in a sphere around the original signal of radius $(2WNT)^{1/2}$.

Thus the original sphere must be enlarged to a larger radius to enclose the disturbed signals. The new radius will be

$$r = (2WT(S+N))^{1/2}$$

In order to use the channel effectively and minimize error(misreading of signals.), we will want to put the signals in the sphere, and separate them as much as possible(and have the distance between the signals at least twice what the noise contributes...). We thus want to divide the sphere up into sub-spheres of radius $= (2WNT)^{1/2}$. From this, we can get an upper bound on the number M of possible messages that we can reliably distinguish. We can use the formula for the volume of an n-dimensional sphere

$$V(r,n) = \frac{\pi^{n/2} r^n}{\Gamma(n/2+1)}$$

We have the bound

$$M \leq \frac{\pi^{WT}(2WT(S+N))^{WT}}{\Gamma(WT+1)} \frac{\Gamma(WT+1)}{\pi^{TW}(2WTN)^{WT}}$$

The information sent is the log of the number of messages sent(assuming they are equally likely), and hence

$$I = log(M) = WT * log(1 + S/N)$$

We thus have the usual signal/noise formula for channel capacity.

5 A Maximum Entropy Principle

Model Suppose we have a system for which we can measure certain macroscopic characteristics. Suppose further that the system is made up of many microscopic elements, and the system is free to vary among various states.

Conclusion With probability essentially equal to 1, the system will be observed in states with maximum entropy.

We will then sometimes be able to gain understanding of the system by applying a maximum information entropy principle(MEP), and using Lagrange multipliers, derive formulae for aspects of the system.

Suppose we have a set of macroscopic measurable characteristics $f_k, k = 1, 2, \cdots, M$(which we can think of as constraints on the system), which we assume are related to microscopic characteristics via

$$\sum_i p_i * (f_i)^{(k)} = f_k$$

Of course, we also have the constraints

$$p_i \geq 0$$
$$\sum_i p_i = 1$$

(how to use Lagrange multipliers?? We want to maximize the entropy $\sum_i p_i log(1/p_i)$, subject to these constraints. Using Lagrange multipliers λ_k(one for each constraint), we have he general solution:

$$p_i = exp(-\lambda - \sum_k \lambda_k (f_i)^{(k)}$$

If we define Z, called the partition function, by

$$Z(\lambda_1, \cdots, \lambda_M) = \sum_i exp(-\sum_k \lambda_k (f_i)^{(k)})$$

then we have $e^\lambda = Z$, or $\lambda = ln(z)$

6 Application: A Boltzmann Economy

Model Suppose there is a fixed amount of money(M dollars), and a fixed number of agents(N) in the economy. Suppose that during each time step, each agent randomly selects another agent and transfers one dollar to the selected agent. An agent having no money doesn't go in debt. What will the long term(stable) distribution of money be? We can image that every agent starts with approximately the same amount of money, although in the long run, the starting distribution shouldn't matter.

Analysis For this model, we are interested in looking at the distribution of money in the economy, so we are looking at the probabilities $\{p_i\}$ that an agent has the amount of money i. We are hoping to develop a model for the collection $\{p_i\}$.

If we let n_i be the number of agents who have i dollars, we have two constraints

$$\sum_i n_i * i = M$$
$$\sum_i n_i = N$$

Phrased differently(using $p_i = n_i/N$), this says

$$\sum_i p_i * i = \frac{M}{N}$$
$$\sum_i p_i = 1$$

We now apply Lagrange multipliers how??:

$$L = \sum_i p_i * ln(1/p_i) - \lambda[\sum_i p_i * i - M/N] - \mu[\sum_i p_i - 1]$$

from which we get

$$\frac{\partial_L}{\partial_{p_i}} = -[1 + \ln(p_i)] - \lambda i - \mu = 0$$

We can solve this for p_i:

$$ln(p_i) = -\lambda * i + (1 + \mu)$$
$$p_i = e^{-\lambda_0} e^{-\lambda i}$$

where we set $1 + \mu \equiv \lambda_0$.

Putting in constraints, we have

$$\begin{aligned}1 &= \sum_i p_i \\ &= \sum_i e^{-\lambda_0} e^{-\lambda i} \\ &= e^{-\lambda_0} \sum_{i=0}^{M} e^{-\lambda i}\end{aligned}$$

and

$$\begin{aligned}\frac{M}{N} &= \sum_i p_i \\ &= \sum_i e^{-\lambda_0} e^{-\lambda i} * i \\ &= e^{-\lambda_0} \sum_{i=0}^{M} e^{\lambda i} * i\end{aligned}$$

We can approximate(for large M)

$$\sum_{i=0}^{M} e^{\lambda i} \approx \int_0^M e^{\lambda x} dx \approx \frac{1}{\lambda}$$
$$\sum_{i=1}^{M} e^{\lambda i} * i \approx \int_0^M x e^{\lambda x} dx = \frac{1}{\lambda^2}$$

From these, we have(approximately)

$$e^{\lambda_0} = \frac{1}{\lambda}$$
$$e^{\lambda_0} \frac{M}{N} = \frac{1}{\lambda^2}$$

From this, we get

$$\lambda = \frac{N}{M} = e^{\lambda_0}$$

and thus(letting $T = \frac{M}{N}$) we have:

$$p_i = e^{-\lambda_0} e^{-\lambda i} = \frac{1}{T} e^{-\frac{i}{T}}$$

This is a Boltzmann-Gibbs distribution, where we can think of T(the average amount of money per agent) as the "temperature", and thus we have a "Boltzmann economy"...

Note: this distribution also solves the functional equation why???

$$p(m_1)p(m_2) = p(m_1 + m_2)$$

7 Application: A Power Law

Model Suppose that a (simple) economy is made up of many agents a, each with wealth at time t in the amount of $w(a, t)$. We are interested in looking at the distribution of wealth in the economy, so we will assume there is some collection $\{w_i\}$ of possible values for the wealth an agent can have, and associated probabilities $\{p_i\}$ that an agent has wealth $\{w_i\}$. We are hoping to develop a model of the collection $\{w_i\}$.

Analysis In order to apply the maximum entropy principle, we want to look at global(aggregate/macro) observables of the system that reflect(or are made up of) characteristics of(micro) elements of the system.

For example, we can look at the growth rate of the economy. A reasonable way to think about this is to let $R_i = w_i(t_1)/w_i(t_0)$ and $R = W(t_1)/W(t_0)$(where t_0 and t_1 represent time steps of the economy). The growth rate will then be $\ln(R)$. We then have the two constraints on the p_i how the first constraint???

$$\sum_i p_i * ln(R_i) = ln(R)$$

$$\sum_i p_i = 1$$

We now apply Lagrange multipliers

$$L = \sum_i p_i \ln(1/p_i) - \lambda[\sum_i p_i ln(R_i) - ln(R)] - \mu[\sum_i p_i - 1]$$

from which we get

$$\frac{\partial L}{\partial p_i} = -[1 + \ln(p_i)] - \lambda ln(R_i) - \mu = 0$$

We cab solve this for p_i

$$p_i = e^{-\lambda_0} e^{-\lambda ln(R_i)} = e^{-\lambda_0} (R_i)^{-\lambda}$$

(where we have set $1 + \mu \equiv \lambda_0$.)

Solving, we get $\lambda_0 = ln(Z(\lambda))$ (by using $\sum_i p_i = 1$), where $Z(\lambda) \equiv \sum_i (R_i)^{-\lambda}$(the partition function) normalizes the probability distribution to sum to 1. from this we see the power law(for $\lambda > 1$):

$$p_i = \frac{(R_i)^{-\lambda}}{Z(\lambda)}$$

Continuous Version We let $R = w(T)/w(0)$ be the relative wealth at time T. We ant to find the probability density function $f(R)$, that is i.e., find f(R) that maximize the entropy

$$max_{\{f\}} H(f) = -\int_1^\infty f(R) ln(f(R)) dR$$

subject to

$$\int_1^\infty f(R) dR = 1$$

$$\int_1^\infty f(R) ln(R) dR = C ln(R)$$

where C is the average number of transactions per time step. (why????)

We need to apply the calculus of variations to maximize over a class of functions.

When we are solving an extremal problem of the form

$$F[x, f(x), f'(x)]dx$$

we work to solve

$$\frac{\partial F}{\partial f(x)} - \frac{d}{dx}(\frac{\partial F}{\partial f'(x)}) = 0$$

Our Lagrangian is of the form

$$L \equiv -\int_1^\infty f(R) ln(f(R)) dr - \mu(\int_1^\infty f(R) dR - 1) - \lambda(\int_1^\infty f(R) ln(R) dR - C * ln(R))$$

neglect following analysis...

8 Application to Physics(laser)

Model For a laser, we will be interested in the intensity of the light emitted, and the coherence property of the light will be observed in the second moment of the intensity. The electric field strength of such a laser will have the form

$$E(x,t) = E(t)sin(kx)$$

and $E(t)$ can be decomposed in the form

$$E(t) = B * e^{-iwt} + B * e^{iwt}$$

If we measure the intensity of the light over time intervals long compared to the frequency, but small compared to fluctuations of $B(t)$, the output will be proportional to $BB*$ and to the loss rate sk, of the laser:

$$I = 2kBB*$$

The intensity squared will be

$$I^2 = 4k^2B^2B*^2$$

Analysis If we assume that B and B^* are continuous random variables associated with a stationary process, then the information entropy of the system will be

$$H = \int p(B,B*)log(\frac{1}{p(B,B*)})d^2B$$

The two constraints why??? on the system will be the averages of the intensity and the square of the intensity:

$$f_1 =< 2kBB* >,$$
$$f_2 =< 4k^2B^2B*^2 >$$

Then, of course, we will let

$$f^{(1)}_{B,B*} = 2kBB*$$
$$f^{(2)}_{B,B*} = 4k^2B^2B*^2$$

We can now use the method outlined above, finding the maximum entropy general solution derived via Lagrange multipliers for this system.

Applying the general solution, we get

$$P(B,B^*) = exp[-\lambda - \lambda_1 2kBB* - \lambda_2 4k^2B^2b*^2]$$

or in other notation:

$$p(B,B^*) = N * exp(-\alpha|B|^2 - \beta|B|^4)$$

This function in laser physics is typically derived by solving the Fokker-Planck equation belonging to the Langevin equation for the system.