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.