Appendix A

Information Theory

A.1 Entropy

Shannon (Shanon, 1948) developed the concept of entropy to measure the uncertainty of a discrete random variable. Suppose X is a discrete random variable that obtains values from a finite set $x_1, ..., x_n$, with probabilities $p_1, ..., p_n$. We look for a measure of how much choice is involved in the selection of the event or how certain we are of the outcome. Shannon argued that such a measure $H(p_1, ..., p_n)$ should obey the following properties

1. H should be continuous in p_i.

2. If all p_i are equal then H should be monotonically increasing in n.

3. If a choice is broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

Shannon showed that the only H that satisfies these three assumptions is of the form

$$H = -k \sum_{i=1}^{n} p_i \log p_i \tag{A.1}$$

and termed it the entropy of X, since it coincides with the notion of entropy defined in certain formulations of statistical mechanics. k is a constant that determines the units of measure, and can be absorbed in the base of the log. The current thesis adheres to the computer science literature and uses the log in base 2. To summarize, we define entropy as

Definition A.1.1 : Entropy
The *entropy* $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x} p(x) \log p(x) \tag{A.2}$$

We will also sometimes use the notation $H[\mathbf{p}]$ to denote the entropy of a random variable that has a probability distribution $\mathbf{p}$. Given several random variables we then define

Definition A.1.2 : Joint Entropy
The *joint entropy* $H(X, Y)$ of a pair of discrete random variables X and Y with a joint distribution $p(x, y)$ is defined by

$$H(X, Y) = -\sum_{x} \sum_{y} p(x, y) \log p(x, y) \tag{A.3}$$

Definition A.1.3 : Conditional entropy
Let X and Y be discrete random variables with joint distribution $p(x, y)$ and conditional distributions $p(x|y)$, then the entropy conditioned on a single symbol is defined by

$$H(X|Y = y) = -\sum_{x} p(x|y) \log p(x|y) \quad . \tag{A.4}$$

The *conditional entropy* is defined by

$$\begin{aligned} H(X|Y) &= \sum_{y} p(y) H(X|Y = y) \qquad \text{(A.5)} \\ &= -\sum_{y} p(y) \sum_{x} p(x|y) \log p(x|y) \\ &= -\sum_{x,y} p(x, y) \log p(x|y) \quad . \end{aligned}$$

Several properties of the entropy worth mentioning.

Theorem A.1.4 : Properties of H(X)
The entropies $H(X)$ of a discrete random variable X that can obtain the values $x_1, ..., x_n$, and the joint entropy $H(X, Y)$, obey the following properties

1. *Non-negativity* $H(X) \geq 0$

2. *Upper bound* $H(X) \leq \log(n)$

3. *Chain rule:* $H(X,Y) = H(X) + H(Y|X)$

4. *Conditioning reduces entropy* $H(X|Y) \leq H(X)$

5. $H(p)$ *is concave in* p

A.2 Relative entropy and Mutual information

The entropy of a variable is a measure of the uncertainty in its distribution. The relative entropy is a measure of the statistical distance between two distributions

Definition A.2.1 : Relative Entropy
The *relative entropy* or the *Kullback Leibler divergence* between to probability functions $p(x)$ and $q(x)$, is defined by

$$D_{KL}[p||q] = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{A.6}$$

The KL divergence appears in statistics as the expected value of the log likelihood ratio. It therefore determines the ability to discriminate between two states of the world, yielding sample distributions $p(x)$ and $q(x)$.

We also use sometimes a variant of D_{KL}

Definition A.2.2 : Jensen-Shannon divergence
The *Jensen-Shannon divergence* between to probability functions $p_1(x)$ and $p_2(x)$, is defined by

$$JS_\pi[p||q] = \pi_1 D_{KL}[p_1||p] + \pi_2 D_{KL}[p_2||p] \tag{A.7}$$

with $\{\pi_1, \pi_2\}$ being prior probabilities $\pi_i > 0$, $\sum_i \pi_i = 1$, and p is the weighted average $p = \pi_1 p_1 + \pi_2 p_2$.

Theorem A.2.3 : Properties of D_{KL}
Let $p(x)$ *and* $q(x)$ *be two probability distributions, Then*

1. $D_{KL}[p||q] \geq 0$ *with equality iff* $p(x) = q(x) \forall x$.

2. *$D_{KL}[p||q]$ is convex w.r.t the pair (p, q).*

Definition A.2.4 : Mutual Information

The *mutual information* $I(X; Y)$ of two random variables X and Y is the KL divergence between their joint distribution and the product of their marginals

$$I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad . \tag{A.8}$$

By this definition the mutual information provides some measure of the dependence between the variables. From the non negativity of the D_{KL} we obtain

Theorem A.2.5 : Non negativity of I(X;Y)

Let X and Y be two discrete random variables, then

$$I(X; Y) \geq 0 \tag{A.9}$$

and equality iff X and Y are independent.

Theorem A.2.6 : Properties of the mutual information

Let X and Y be two discrete random variables, then their mutual information $I(X; Y)$ obeys

1. *Symmetry $I(X; Y) = I(Y; X)$.*

2. *$I(X; Y) = H(X) - H(X|Y)$.*

3. *$I(X; X) = H(X)$.*

4. *Chain rule: $I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, ..., X_{i-1})$.*

5. *Data processing inequality: if (X, Y, Z) form a Markov chain, then $I(X; Y) \geq I(X; Z)$. As a consequence, $I(X; Y) \geq I(X; f(Y))$ for any function f of Y.*

A.3 Extensions

While the above notions were defined for discrete variables, entropy and mutual information can be extended to continuous variables (Shanon, 1948; Cover & Thomas, 1991). This issue is beyond of the scope of the current manuscript. Also, the notion of information can be extended to more than two variables using the view that information measure the KL distance from independence

Definition A.3.1 : Multi Information

The *multi information* $I(X_1; \ldots ; X_n)$ of n random variables is the KL divergence between their joint distribution and the product of their marginals

$$I(X_1; \ldots ; X_n) = \sum_{x_1,\ldots,x_n} p(x_1, ..., x_n) \log \frac{p(x_1, ..., x_n)}{\prod_i p(x_i)} \quad . \tag{A.10}$$

By this definition the multi information provides some measure of the dependence between all the variables. From the non negativity of the D_{KL} we obtain that the multi information is non negative. The properties of the multi information measure are further discussed in (Studenty & Vejnarova, 1998).

Appendix B

Table of symbols

AI	Auditory cortex
BF	Best Frequency
$D_{KL}[p\|\|q]$	The Kullback Liebler divergence (Definition A.2.1)
$H(X)$	The entropy of a discrete variable X (Definition A.1.1)
$I(X;Y)$	The Mutual information of two variables X and Y (A.2.4)
$I[p]$	The mutual information of variables with a joint distribution p
IC	inferior colliculus
$JS[p\|\|q]$	The Jensen-Shannon divergence (A.2.2)
MGB	Medial Geniculate body of the thalamus
MI	Mutual information
n	Sample size
N	Number of variables
p	Probability distribution. $p(X,Y)$ is the joint distribution of X and Y
$\hat{p}$	Probability distribution that is estimated from empirical data
R	Neural responses (a random variable)
S	Stimulus (a random variable)
STRF	Spectro-Temporal Receptive Field
$T(R)$	A statistic of the responses

References

Cover, T., & Thomas, J. (1991). *The elements of information theory.* New York: Plenum Press.

Shanon, C. (1948). A mathematical theory of communication. *The Bell systems technical journal*, *27*, 379-423,623-656.

Studenty, M., & Vejnarova, J. (1998). The multi-information function as a tool for measuring stochastic dependence. In M. Jordan (Ed.), *Learning in graphical models* (p. 261-297). Cambridge, MA: MIT Press.