Entropy

let X be a discrete random variable taking values from the alphabet $\mathcal{X}$ and let

$$p(x) = P_r(X = x), \quad x \in \mathcal{X}$$

uncertainity of ~~observing~~ an outcome say $x \in \mathcal{X}$ should (intuitively) be higher the smaller $p(x)$ is, i.e., the least probable an event is the more we should be surprised by its occurance, or in other words, its observation should remove ~~the~~ more uncertainty about x. So, the uncertainty of an outcom x should be a function of $\frac{1}{p(x)}$.

The second intuitive hint is that the uncertainty should be additive and it leads us to define the measure of uncertaintly as $\log \frac{1}{p(x)}$. since for two independent outcomes x_1 and x_2, $p(x_1, x_2) = p(x_1)p(x_2)$

and $f\left(\frac{1}{p(x_1)p(x_2)}\right) = f\left(\frac{1}{p(x_1)}\right) + f\left(\frac{1}{p(x_2)}\right)$

letting $f(\cdot) = \log(\cdot)$ allows us to have the above equality.

entropy is defined as the average uncertainty, i.e., as the expected value of $\log\left(\frac{1}{p(x)}\right)$, i.e.,

$$H(X) = E\left[\log \frac{1}{p(X)}\right] = -\sum_{x \in X} p(x) \log p(x)$$

Note: to be more "mathematically correct", we should denote the entropy as $H(p)$ since it really is a function of the probability assignment on the alphabet.

Properties of $H(X)$:

$$H(X) \geq 0$$

Proof:

$0 \leq p(x) \leq 1 \Rightarrow \quad \log \frac{1}{p(x)} \geq 0$ equivalently $\log p(x) \leq 0$

$$H_b(X) = \log_b^a H_a(X)$$

$$H_b(X) = -\sum_x p(x) \log_b p(x) = -\sum_x p(x) \frac{\log_a p(x)}{\log_a b}$$

$$= -\log_b^a \sum_x p(x) \log_a p(X)$$

$$= \log_b^a H_a(X)$$

Binary Source:

Take a source with the alphabet $X=\{0,1\}$

with

$$p(x) = \begin{cases} p & x=1 \\ 1-p & x=0 \end{cases}$$

then

$$H(X) = -\sum_{X} p(x)\log p(x) = -p\log p - (1-p)\log(1-p)$$

Assume that this source generates a large number of bits, say, n. On the average, we get np ones and $n(1-p)$ zeros. The probability of this n-bit "typical" sequence is: $p^{np}(1-p)^{n(1-p)}$.

Let N_T be the total number of typical sequences it is clear that

$$N_T\, p^{np}(1-p)^{n(1-p)} < 1$$

or

$$N_T < p^{-np}(1-p)^{-n(1-p)}$$

Say, we need k bits to represent each typical sequence

$$k = \log_2 N_T < -n[p\log_2 p + (1-p)\log_2(1-p)]$$

or

$$\frac{k}{n} < H_2(p)$$

later, we show ~~the converse also~~ that the addition of a typical sequence at worst changes $\leq$ to $=$.

Joint entropy

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in \mathcal{Y}} p(x,y)\log p(x,y)$$

Conditional entropy

$$H(Y|X) = \sum_{x\in X} p(x) H(Y|X=x)$$

$$= -\sum_{x\in X} p(x) \sum_{y\in \mathcal{Y}} p(y|x) \log [p(y|x)]$$

$$= -\sum_{x\in X}\sum_{y\in \mathcal{Y}} p(x,y)\log p(y|x)$$

$$= \cancel{E[\log p(y|x)]}$$

$$= E\left[\log \frac{1}{p(y|x)}\right]$$

Chain rule:

$$H(X,Y) = H(X) + H(Y|X)$$

interpretation: the uncertainty about X and Y is the uncertainty about X plus uncertainty about Y after the uncertainty about X is resolved.

Proof:

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y)$$

$$= -\sum_x \sum_y p(x,y) \log p(x)p(y|x)$$

$$= -\sum_x \sum_y p(x,y) \log p(x) - \sum_x \sum_y p(x,y) \log p(y|x)$$

$$= -\sum_x p(x) \log p(x) - \sum_x \sum_y p(x,y) \log p(y|x)$$

$$= H(X) + H(Y|X)$$

we could also show that:

$$H(X,Y) = H(Y) + H(X|Y)$$

therefore:

channel equivocation

$$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

interpretation:

The uncertainty about X minus the uncertainty about X after knowing Y is the information Y gives about X. This is called the mutual information between X and Y, i.e., $I(X;Y)$

$I(X;Y)$ is also equal to the information X reveals about Y.

Take a channel with the input X and output Y

$$X \longrightarrow \boxed{\text{channel}} \longrightarrow Y$$

Channel Equivocation

$$I(X;Y) = H(X) - H(X|Y)$$

is the information the channel output reveals about its input.

if $H(X) = H(X|Y)$ then $I(X;Y) = 0$

that is there is no flow of information through the channel. In other words the certainty about X is the same before and after the observation of Y, i.e., Y has no relation to X and reveals nothing about X.

On the other hand if $H(X|Y) = 0$ then $I(X;Y) = H(X)$. That is after observing Y, there is no uncertainty about X.

The above two extreme cases depict total independence (0 Capacity) and total dependence.

Kulback Leibler Distance

Kulback-Leibler distance or relative entropy is defined as

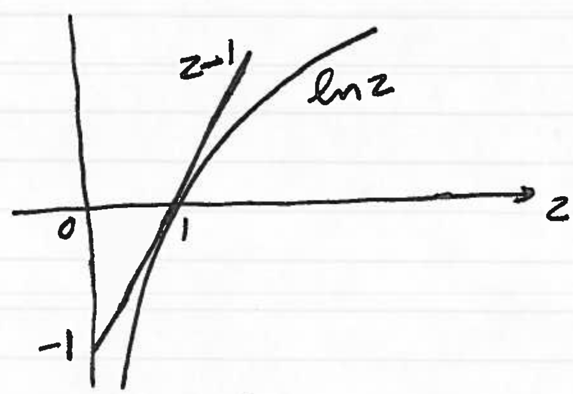
$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$D(p\|q)$ is a measure of the degree of closeness of two probability distributions $\{p(x)\}$ and $\{q(x)\}$ defined on X.

Theorem: $D(p\|q) \geqslant 0$ with equality if $p(x) = q(x)$ for all x.

Proof:

We use the inequality $\ln z \leq z-1$ with equality if $z=1$



$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \log e \sum_x p(x) \ln \frac{p(x)}{q(x)}$$

$$= -\log e \sum_x p(x) \ln \frac{q(x)}{p(x)} \overset{(a)}{\geqslant} -\log e \sum_x p(x) \left[\frac{q(x)}{p(x)} - 1\right]$$

$$= -\log e \left[\sum_x q(x) - \sum_x p(x)\right] = 0$$

(a) holds with equality if $\frac{q(x)}{p(x)} = 1$. That is $D(p\|p) = 0$

Now, let's look at a channel with input $X \sim p(x)$ output $Y \sim p(y)$ and transition probabity $p(y|x)$.

If the input and output were independent, i.e., zero capacity $p(x,y) = p(x)p(y)$. Else $p(x,y) = p(x)p(y)$ with $p(y|x) \neq p(y)$.

We would like to see how different the joint distribut of the (x,y) is from $p(x)p(y)$. To do so, we ~~define~~ compute

$$D(p(x)p(y)\|p(x,y)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

We then show that

$$D(p(x)p(y)\|p(x,y)) = I(X;Y)$$

<u>Proof:</u>

$$D(p(x,y)\|p(x)p(y)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{p(x)p(y)} = \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{p(y)}$$

$$= \sum_x \sum_y p(x,y) \log \frac{p(x|y)}{p(x)}$$

$$= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y)$$

$$= H(X) - H(X|Y) = I(X;Y)$$

$$= \sum_{x,y} p(x,y) \log p(y|x) - \sum_{x,y} p(x,y) \log p(y)$$

$$= H(Y) - H(Y|X) = I(Y;X)$$

$$= I(X;Y)$$

Summary of the properties of $I(X;Y)$

1) $I(X;Y) = H(X) - H(X|Y)$

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

3) $H(X,Y) = H(X) + H(Y|X) \Rightarrow \boxed{I(X;Y) = H(X) + H(Y) - H(X,Y)}$

4) $I(X;Y) = I(Y;X)$

5) $I(X;X) = H(X)$

6) $I(X;Y) \geq 0$

Chain rule for entropy

$$
\begin{aligned}
H(X_1, X_2, \dots X_n) &= H(X_1) + H(X_2, \dots X_n | X_1) \\
&= H(X_1) + H(X_2|X_1) + H(X_3, X_4 \cdots X_n | X_1, X_2) \\
&\vdots \\
&= H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1 \cdots X_{n-1}) \\
&= \sum_{i=1}^{n} H(X_i | X_{i-1} \cdots X_1)
\end{aligned}
$$

Conditional Mutual information:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

Chain rule for information:

$$I(X_1,\cdots X_n;Y) = H(X_1,\cdots X_n) - H(X_1,\cdots X_n|Y)$$

$$= \sum_{i=1}^{n} H(X_i|X_{i-1}\cdots X_1) - \sum_{i=1}^{n} H(X_i|X_{i-1},\cdots X_1,Y)$$

$$= \sum_{i=1}^{n} I(X_i;Y|X_1,\cdots X_{i-1})$$

Theorem: $H(X) \leq \log|\mathcal{X}|$ were $|\mathcal{X}|$ is the number of elements in the range of X. The equality holds if X is distributed uniformly over $\mathcal{X}$.

Proof: let $u(x) = \frac{1}{|\mathcal{X}|}$ be the uniform distribution over $\mathcal{X}$. Then:

$$\sum_x p(x) \log \frac{p(x)}{u(x)} = \log e \sum_x p(x) \log \frac{p(x)}{u(x)}$$

$$= -\log e \sum_x p(x) \log \frac{u(x)}{p(x)}$$

$$\geq -\log e \sum_x p(x)\left[\frac{u(x)}{p(x)} - 1\right] = 0$$

but

$$\sum_x p(x) \log \frac{p(x)}{u(x)} = \sum_x p(x) \log p(x) - \sum_x p(x) \log u(x)$$

$$= -H(X) + \sum p(x) \log |X| \geq 0$$

$$\Rightarrow \quad H(X) \leq \log |X|$$

Note: $D(p\|u) = -H(X) + \log |X|$

$\Rightarrow H(X) = \log |X| - D(p\|u)$

interpretation: the farther p is from u, the smaller $H(X)$

Theorem: Conditioning reduces entropy

$$H(X|Y) \leq H(X)$$

with equality iff X and Y are independent.

Proof:

$$I(X;Y) = H(X) - H(X|Y) \geq 0$$

$$\Rightarrow \quad H(X|Y) \leq H(X)$$