2. Measure of Information

▼ 2.1 Information

an event that elicits “surprise” in the agent that observes the event

Intuitively, if an event is “surprising”, then we gained more information from the event than if the event had been something we were expecting.

Moreover, if an event is “surprising”, then presumably, this event is considered unlikely – that is, to have a low probability of occurring.

From this description, we can think of information as a positive function I, often called **self-information** that acts on probabilities:

$$I : [0, 1] \to [\infty, 0)$$

▼ 2.2 Entropy

the **entropy** of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes. Given a discrete random variable X takes exactly N values with positive probability, and is distributed according to $p : X \to [0, 1]$.

The entropy of X can be defined as:

$$H(X) = \sum_{x \in N} p(x) log(\frac{1}{p(x)})$$

also can be expressed as the expected value of the random variable $log\ 1/p(X)$,

$$H(X) = E[log(\frac{1}{p(X)})], \quad X \sim p(x)$$

Entropy satisfy:

$$0 \leq H(X) \leq log|N|$$

▼ proof

From Jensen’s Inequality, we have:

$$\begin{aligned} H(X) &= E[log(\frac{1}{p(X)})] \\ &\leq log(E[\frac{1}{p(X)}]) \\ &= log(E[\frac{1}{1/N}]) \\ &= log\ N \end{aligned}$$

Due to when every probability of X occurrence is same and equal to $\frac{1}{N}$, the function has a unique maximizer.

The entropy of an n-dimensional random vector $\mathbf{X} = (X_1, X_2, ..., X_n)$ with pmf $p(x)$ is defined as:

$$H(\mathbf{X}) = H(X_1, X_2, ..., X_n) = \sum_{\mathbf{x} \in N} p(\mathbf{x}) log(\frac{1}{p(\mathbf{x})})$$

A discrete random variable Y takes exactly M values with positive probability,

Joint Entropy: The joint entropy of random variables X and Y is simply the entropy of the vector (X, Y)

$$H(X, Y) = \sum_{x \in N} \sum_{y \in M} p(x, y) log(\frac{1}{p(x, y)})$$

Conditional Entropy: The entropy of a random variable Y conditioned on the event $\{X = x\}$ is a function of the conditional distribution $p_{Y|X}(\cdot|x)$ and is given by:

$$H(Y|X = x) = \sum_{y \in M} p(y|x) (\frac{1}{p(y|x)})$$

The conditional entropy of Y given X is a function of the joint distribution p(x, y):

$$H(Y|X) = \sum_{x \in N} p(x) H(Y|X = x) = \sum_{x,y} p(x, y) log(\frac{1}{p(y|x)})$$

Note: $H(Y|X)$ is not a random variable

Chain Rule :

$$H(X_1, X_2, .., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, .., X_1)$$

▼ 2.3 Mutual Information

Definition: mutual information is a measure of the amount of information that one random variable contains about another random variable

$$I(X; Y) = \sum_{x \in N} \sum_{y \in M} p(x, y) log(\frac{p(x, y)}{p(x)p(y)})$$

also can expressed as:

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

▼ proof

$$\begin{aligned} I(X; Y) &= \sum_{x \in N} \sum_{y \in M} p(x, y) log(\frac{p(x, y)}{p(x)p(y)}) \\ &= \sum_{x \in N} \sum_{y \in M} p(x, y) [log(\frac{1}{p(y)}) - log(\frac{1}{p(y|x)})] \end{aligned}$$

Conditional mutual information between X and Y given Z is

$$I(X;Y \mid Z) = \sum_{x,y,z} p(x,y,z) \log \left(\frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \right)$$

Chain Rule for mutual information:

$$I\left(X; Y_1, Y_2, \cdots, Y_n\right) = \sum_{i=1}^{n} I\left(X; Y_i \mid Y_1, Y_2, \cdots, Y_{k-1}\right)$$

▼ 2.4 Relative Entropy

The relative entropy between a distributions p and q is defined by:

$$D(P||Q) = \sum_{x \in N} p(x) log(\frac{p(x)}{q(x)})$$

This is also known as the Kullback-Leibler divergence.

Note: if there exists x such that p(x) > 0 and q(x) = 0, then $D(p||q) = \infty$.

Mutual information between X and Y is equal to the relative entropy between $p_{X,Y}(x,y)$and $p_X(x)\ \ p_Y(y)$,

$$I(X;Y) = D(p_{X,Y}(x,y)||p_X(x)p_Y(y))$$

Relative entropy is nonnegative. It is equal to zero if and only if $p = q$, which can be proved by **Gibbs' inequality.**

Note:

1. Mutual information is nonnegative, with equality if an only if X and Y are independent

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

2. conditioning cannot increase entropy

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

K-L divergence is convex:

$$KL[\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2] \leq \lambda KL[p_1||q_1] + (1-\lambda)KL[p_2||q_2]$$

▼ Proof

$$KL(P||Q) = \sum_{x \in N} p(x) log(\frac{p(x)}{q(x)})$$

We already know log sum inequality:

Let $a_1, ...a_n \ \ and \ \ b_1, ..., b_n$ be non-negative real numbers

$$\sum_{i=1}^{n} a_i \ log_c \frac{a_i}{b_i} \geq \sum_{i=1}^{n} a_i \ log_c \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

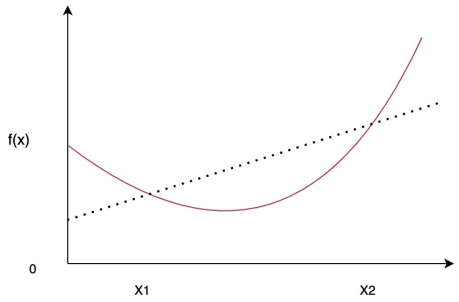
Then for K-L divergence:

$$
\begin{aligned}
KL[\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2 &= \sum_{x \in N}[\lambda p_1(x) + (1-\lambda)p_2(x)] log(\frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)}) \\
&\leq \sum_{x \in N}[\lambda p_1(x) \cdot log(\frac{\lambda p_1(x)}{\lambda q_1(x)}) + (1-\lambda)p_2(x) \cdot log(\frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)})] \\
&= \sum_{x \in N}[\lambda p_1(x) \cdot log(\frac{p_1(x)}{q_1(x)}) + (1-\lambda)p_2(x) \cdot log(\frac{p_2(x)}{q_2(x)})] \\
&= \lambda KL[p_1||q_1] + (1-\lambda)KL[p_2||q_2]
\end{aligned}
$$

▼ 2.5 Convexity & Concavity

A function $f(x)$ is convex over an interval $(a, b) \in \mathbb{R}$ if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$
f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_1)
$$



Theorem: H(X) is a concave function $p(x)$,

$$
H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2)
$$

▼ Proof

Let X be a discrete random variable with possible outcomes $\mathcal{X}$ and let $u(x)$ be the pmf of a discrete uniform distribution on $X \in \mathcal{X}$. Then, the entropy of an arbitrary pmf $p(x)$ can be rewritten as:

$$
\begin{aligned}
H[p] &= -\sum_{x \in \mathcal{X}} p(x) \cdot log\, p(x) \\
&= -\sum_{x \in \mathcal{X}} p(x) \cdot log\, \frac{p(x)}{u(x)} u(x) \\
&= -\sum_{x \in \mathcal{X}} p(x) \cdot log\, \frac{p(x)}{u(x)} - \sum_{x \in \mathcal{X}} p(x) \cdot log\, u(x) \\
&= -KL[p||u] - log\, \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} p(x) \\
&= log\, \frac{1}{|\mathcal{X}|} - KL[p||u]
\end{aligned}
$$

Then we have:

$$
log\, \mathcal{X} - H[p] = KL[p||u]
$$

Note that KL divergence is convex then :

$$\begin{aligned}\mathrm{KL}\left[\lambda p_1+(1-\lambda)p_2\|\lambda u+(1-\lambda)u\right] &\leq \lambda\mathrm{KL}\left[p_1\|u\right]+(1-\lambda)\mathrm{KL}\left[p_2\|u\right]\\ \mathrm{KL}\left[\lambda p_1+(1-\lambda)p_2\|u\right] &\leq \lambda\mathrm{KL}\left[p_1\|u\right]+(1-\lambda)\mathrm{KL}\left[p_2\|u\right]\end{aligned}$$

$$\begin{aligned}\log|\mathcal{X}|-\mathrm{H}\left[\lambda p_1+(1-\lambda)p_2\right] &\leq \lambda\left(\log|\mathcal{X}|-\mathrm{H}\left[p_1\right]\right)+(1-\lambda)\left(\log|\mathcal{X}|-\mathrm{H}\left[p_2\right]\right)\\ \log|\mathcal{X}|-\mathrm{H}\left[\lambda p_1+(1-\lambda)p_2\right] &\leq \log|\mathcal{X}|-\lambda\mathrm{H}\left[p_1\right]-(1-\lambda)\mathrm{H}\left[p_2\right]\\ -\mathrm{H}\left[\lambda p_1+(1-\lambda)p_2\right] &\leq -\lambda\mathrm{H}\left[p_1\right]-(1-\lambda)\mathrm{H}\left[p_2\right]\\ \mathrm{H}\left[\lambda p_1+(1-\lambda)p_2\right] &\geq \lambda\mathrm{H}\left[p_1\right]+(1-\lambda)\mathrm{H}\left[p_2\right]\end{aligned}$$

Jesen's Inequality:

If f is a convex function over an interval $\mathcal{I}$ and X is a random variable with support $\mathcal{X}\subset\mathcal{I}$ then

$$E[f(X)]\geq f(E[X])$$

▼ 2.6 Data Processing Inequality

Markov Chain: Random variables X, Y, Z from a Markov chain, denoted $X\to Y\to Z$

if X, Z are independent conditioned on Y.

$$p(x,z|y)=p(x|y)p(z|y)$$

alternatively

$$p(x,y,z)=p(x)p(y|x)p(z|y)$$

Data Processing Inequality:

If $X\to Y\to Z$, then

$$I(X;Y)\geq I(X;Z)$$

▼ Proof

$$\begin{aligned}I(X;Z) &= H(X)-H(X|Z)\\ &\leq H(X)-H(X|Z,Y)\\ &= H(X)-H(X|Y)\\ &= I(X;Y)\end{aligned}$$

▼ 2.7 Fano's Inequality

Suppose we want to estimate a random variable X from an observation Y .
The probability of error for an estimator $\hat{X} = \psi(Y)$ is

$$P(e) = \mathbb{P}[\hat{X} \neq X]$$

Theorem:

For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$,

$$H_b(P(e)) + P(e)(log(|\mathcal{X}|) - 1) \geq H(X|Y)$$

▼ Proof

Define a indicator function E which indicates the event that our estimate $\hat{X} = \psi(Y)$is in error.

$$E := \begin{cases} 1, & if \ \ \hat{X} \neq X \\ 0, & if \ \ \hat{X} = X \end{cases}$$

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) \\ &= H(E|\hat{X}) + H(X|E, \hat{X}) \end{aligned}$$

Here $H(E|X, \hat{X}) = 0$ (we already know the information of $X, \hat{X}$, then the whether is an error is definite)

Then $H(X|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X})$

Expending $H(X|E, \hat{X})$

$$\begin{aligned} H(X \mid E, \hat{X}) &= \underbrace{H(X \mid E = 0, \hat{X})}_{=0} \cdot P(E = 0) + H(X \mid E = 1, \hat{X}) \cdot \underbrace{P(E = 1)}_{=P(e)} \\ &= H(X \mid E = 1, \hat{X}) \cdot P(e) \end{aligned}$$

Since $E = 0$ means $X = \hat{X}$; being given the value of $\hat{X}$ allows us to know the value of X with certainty. This makes the term $H(X \mid E = 0, \hat{X}) = 0$. On the other hand, $E = 1$ means that $\hat{X} \neq X$, hence given the value of $\hat{X}$, we can narrow down X to one of $|\mathcal{X}| - 1$ different values, allowing us to upper bound the conditional entropy $H(X \mid E = 1, \hat{X}) \leq \log(|\mathcal{X}| - 1)$.

Hence

$$H(X \mid E, \hat{X}) \leq \log(|\mathcal{X}| - 1) \cdot P(e)$$

The other term, $H(E \mid \hat{X}) \leq H(E)$, because conditioning reduces entropy. Because of the way E is defined, $H(E) = H_b(e)$, meaning that $H(E \mid \hat{X}) \leq H_b(e)$. Putting it all together,

$$H(X \mid \hat{X}) \leq H_b(e) + P(e) \log(|\mathcal{X}| - 1)$$

Because $X \to Y \to \hat{X}$ is a Markov chain, we have $I(X; \hat{X}) \leq I(X; Y)$ by the data processing inequality, and hence $H(X \mid \hat{X}) \geq H(X \mid Y)$, giving us

$$H(X \mid Y) \leq H_b(e) + P(e) \log(|\mathcal{X}| - 1)$$