Information Theory and Learning

Information Theory and Games

Basic Concepts

Definition *Entropy is a measure of uncertainty of a random variable. Let X be a discrete random variable with alphabet $\mathcal{X}$.*

$$p(x) = Pr[X = x], \quad \textit{where } x \in \mathcal{X}.$$

The entropy $H(X)$ of the discrete random variable X is defined as

$$\begin{aligned} H(X) &= \mathbb{E}_p \lg \frac{1}{p(X)} \\ &= -\sum_{x \in \mathcal{X}} p(x) \lg p(x). \end{aligned}$$

Facts

1. $H(X) \geq 0$. Entropy is always nonnegative. $0 \leq p(x) \leq 1$; $-\lg p(x) \geq 0$. Hence, $E_p \lg(1/p(x)) \geq 0$.)

2. $H(X) \leq \lg |\mathcal{X}|$. Consider the uniform distribution $u(x)$. $\forall_{x \in \mathcal{X}}\ u(x) = 1/|\mathcal{X}|$. $H(u) = \sum_x (1/|\mathcal{X}|) \lg |\mathcal{X}| = \lg |\mathcal{X}|$.

3. $H(X)$ = Average number of bits required to encode the discrete random variable X.

Joint & Conditional Entropy

(X, Y) = A pair of discrete random variables with joint distribution $p(x, y)$.

Joint Entropy =

$$\begin{aligned} H(X,Y) &= \mathbb{E}_p \lg \frac{1}{p(X,Y)} \\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \lg p(x,y). \end{aligned}$$

Conditional Entropy =

$$\begin{aligned} H(Y|X) &= \mathbb{E}_p \lg \frac{1}{p(Y|X)} \\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \lg p(y|x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \lg p(y|x) \\ &= -\sum_{x\in\mathcal{X}} p(x) H(Y|x). \end{aligned}$$

Chain Rule

$$\begin{aligned} p(X,Y) &= p(X)\ p(Y|X) \quad \text{Bayes' Rule} \\ \lg p(X,Y) &= \lg p(X) + \lg p(Y|X) \\ \lg \frac{1}{p(X,Y)} &= \lg \frac{1}{p(X)} + \lg \frac{1}{p(Y|X)} \\ \mathbb{E}_p \lg \frac{1}{p(X,Y)} &= \mathbb{E}_p \lg \frac{1}{p(X)} + \mathbb{E}_p \lg \frac{1}{p(Y|X)} \quad \text{Linearity of Expectation} \\ H(X,Y) &= H(X) + H(Y|X). \end{aligned}$$

Corollary

1. $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$.
2. $H(X) + H(Y|X) = H(Y) + H(X|Y)$
   $\Rightarrow\ H(X) - H(X|Y) = H(Y) - H(Y|X)$.
3. *Note that* $H(X|Y) \neq H(Y|X)$.

Relative Entropy & Mutual Information

Definition **Relative Entropy**—*Also,* **Kullback-Liebler Distance** *between two probability mass functions* $p(x)$ *and* $q(x)$.

$$D(p||q) = \mathbb{E}_p \lg \frac{p(x)}{q(x)} = -\sum_x p(x) \lg \frac{q(x)}{p(x)}.$$

Note that $D(p||p) = 0$. If $u(x) = \frac{1}{|\mathcal{X}|}$, for all x. Then $D(p||u)$ is

$$D(p||u) = -\sum p(x) \lg \frac{1}{p(x)} + \sum p(x) \lg |\mathcal{X}| = \lg |\mathcal{X}| - H(X).$$

Definition **Mutual Information**
Let X *and* Y *be two discrete random variables with a joint probability mass function* $p(x,y)$*, and with marginal probability mass functions*

$$p(x) = \sum_{y \in \mathcal{Y}} p(x,y) \quad \& \quad p(y) = \sum_{x \in \mathcal{X}} p(x,y).$$

Mutual Information,

$$\begin{aligned} I(X;Y) &= D\Big(p(x,y) \parallel p(x)p(y)\Big) \\ &= \mathbb{E}_{p(x,y)} \lg \frac{p(x,y)}{p(x)p(y)} \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \lg \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

$$\begin{aligned} &= H(X) + H(Y) - H(X,Y) \\ &= \Big(H(X) + H(Y)\Big) - \Big(H(Y) + H(X|Y)\Big) \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y;X). \end{aligned}$$

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

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

$$I(X;Y) = I(Y;X) = H(X) + H(Y) - H(X,Y).$$

Chain Rules for Entropy, Relative Entropy and Mutual Information

$$\begin{aligned} &H(X_1, X_2, \ldots, X_n) \\ &\quad = H(X_1) + H(X_2|X_1) + \cdots \\ &\qquad + H(X_n|X_1, \ldots, X_{n-1}) \\ &\quad = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}). \end{aligned}$$

$$\begin{aligned} &I(X_1, X_2, \ldots, X_n; Y) \\ &\quad = H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n|Y) \\ &\quad = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_i) - \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_i, Y) \\ &\quad = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_i) - H(X_i|X_1, \ldots, X_i, Y) \\ &\quad = \sum_{i=1}^{n} I(X_i; Y|X_1, \ldots, X_{i-1}). \end{aligned}$$

$$
\begin{aligned}
& D\Big(\, p(x,y) \;||\; q(x,y) \,\Big) \\
&= \sum_x \sum_y p(x,y) \lg \frac{p(x,y)}{q(x,y)} \\
&= \sum_x \sum_y p(x,y) \lg \frac{p(x)\; p(y|x)}{q(x)\; q(y|x)} \\
&= \sum_x \sum_y p(x,y) \lg \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \lg \frac{p(y|x)}{q(y|x)} \\
&= \sum_x p(x) \lg \frac{p(x)}{q(x)} + \sum_y p(y|x) \lg \frac{p(y|x)}{q(y|x)} \\
&= D\Big(\, p(x) \;||\; q(x) \,\Big) + D\Big(\, p(y|x) \;||\; q(y|x) \,\Big).
\end{aligned}
$$

Information Inequality

$$
\begin{aligned}
-D(p||q) &= \sum_x p(x) \lg \frac{q(x)}{p(x)} \qquad \lg \text{ is a concave function} \\
&\leq \lg \sum_x p(x) \frac{q(x)}{p(x)} \;\leq\; \lg \sum_x q(x) = \lg 1 = 0.
\end{aligned}
$$

Theorem $D(p||q) \geq 0$ *(with equality iff* $p(x) = q(x)$ *for all* x*.)*

Corollary

$$
I(X;Y) = D\Big(\, p(x,y) \;||\; p(x)p(y) \,\Big) \;\geq\; 0,
$$

(with equality iff X *and* Y *are independent, i.e.,* $p(x,y) = p(x)p(y)$ *for all* x *and* y*.)*

Let $u(x) = \frac{1}{|\mathcal{X}|}$.

$$
D(p \;||\; u) = \lg |\mathcal{X}| - H(X) \geq 0.
$$

Hence,

$$H(X) \leq \lg |\mathcal{X}|,$$

(with equality iff X has a uniform distribution over $\mathcal{X}$.)

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

Theorem

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

Conditioning reduces entropy.

$$\begin{aligned} H(X_1, \ldots, X_n) &= \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}) \\ &\leq \sum_{i=1}^{n} H(X_i) \end{aligned}$$

Corollary

$$H(X_1, \ldots, X_n) \ \leq \ \sum_{i=1}^{n} H(X_i),$$

with equality iff X_i's are independent.

Stationary Markov Process

- *Markovian*

$$Pr[X_n|X_1, \ldots, X_i] = Pr[X_n|X_i], \quad i \leq n.$$

- *Stationary*

$$Pr[X_n|X_1, \ldots, X_i] = Pr[X_{n+1}|X_2, \ldots, X_{i+1}].$$

$$\begin{aligned} H(X_n|X_1) &\geq H(X_n|X_1, X_2) \quad \text{conditioning reduces entropy} \\ &= H(X_n|X_2) \quad \text{Markov} \\ &= H(X_{n-1}|X_1) \quad \text{Stationary .} \end{aligned}$$

2nd Law of Thermodynamics

Theorem *Conditional entropy $H(X_n|X_1)$ increases with time n for a stationary Markov process.*

Relative entropy $D(\pi_n\|\pi'_n)$ decreases with time n.

Let π_n and π'_n be two postulated probability distributions on the state space of a Markov Process. At time $n+1$, the distribution changes to π_{n+1} and π'_{n+1}, governed by the transition probabilities $r(x_n, x_{n+1})$.

Thus

$$\begin{aligned} p(x_n, x_{n+1}) &= p(x_n)r(x_n, x_{n+1}) \\ &= p(x_n)p(x_{n+1}|x_n) \end{aligned}$$

similarly,

$$\begin{aligned} q(x_n, x_{n+1}) &= q(x_n)r(x_n, x_{n+1}) \\ &= q(x_n)q(x_{n+1}|x_n) \end{aligned}$$

Thus, we have

$$\begin{aligned} & D\Big(p(x_n, x_{n+1}) \parallel q(x_n, x_{n+1})\Big) \\ &= D\Big(p(x_n) \parallel q(x_n)\Big) + D\Big(p(x_{n+1}|x_n) \parallel q(x_{n+1}|x_n)\Big) \\ &= D\Big(p(x_n) \parallel q(x_n)\Big). \end{aligned}$$

And

$$D\Big(p(x_n, x_{n+1}) \parallel q(x_n, x_{n+1})\Big)$$

$$
\begin{aligned}
&= D\Big(p(x_{n+1}) \parallel q(x_{n+1})\Big) + D\Big(p(x_n|x_{n+1}) \parallel q(x_n|x_{n+1})\Big) \\
&\geq D\Big(p(x_{n+1}) \parallel q(x_{n+1})\Big).
\end{aligned}
$$

We conclude that

$$
D\Big(p(x_n) \parallel q(x_n)\Big) \geq D\Big(p(x_{n+1}) \parallel q(x_{n+1})\Big).
$$

Thus the relative entropy for this system must decrease:

$$
\begin{aligned}
D(\pi_1 \| \pi_1') &\geq D(\pi_2 \| \pi_2') \geq \cdots \\
&\geq D(\pi_n \| \pi_n') \geq D(\pi_{n+1} \| \pi_{n+1}') \geq \cdots\cdots \to 0.
\end{aligned}
$$