

Entropy

Let X be a discrete random variable with alphabet \mathcal{X} and probability mass function $p(x) = \mathbb{P}\{X = x\}$, $x \in \mathcal{X}$

Definition 1. The **entropy** $H(X)$ of a discrete random variable X is defined by

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

We also write $H(p)$ for the above quantity. The \log is to the base 2 and entropy is expressed in bits. For example, the entropy of a fair coin toss is 1 bit. We will use the convention that $0 \log 0 = 0$, which is easily justified by continuity since $x \log x \rightarrow 0$ as $x \rightarrow 0$.

* *Note.* The entropy of X can also be interpreted as the expected value of $\log \frac{1}{p(X)}$, where X is drawn according to probability mass function $p(x)$. Thus

$$H(X) = \mathbb{E} \left(\log \frac{1}{p(X)} \right).$$

Lemma 1. $H(X) \geq 0$.

Proof. $0 \leq p(x) \leq 1$ implies $\log(1/p(x)) \geq 0$. □

Note that $H(X) = 0$ if and only if X is a constant.

Example 1. Let

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p, \end{cases}$$

Then

$$H(X) = -p \log p - (1 - p) \log(1 - p) = H(p).$$

In particular, $H(X) = 1$ bit when $p = 1/2$

Example 2. Let

$$X = \begin{cases} a & \text{with probability } 1/2, \\ b & \text{with probability } 1/4, \\ c & \text{with probability } 1/8, \\ d & \text{with probability } 1/8, \end{cases}$$

The entropy of X is

$$H(X) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{8} \log \frac{1}{8} - \frac{1}{8} \log \frac{1}{8} = \frac{7}{4} \text{ bits.}$$

Suppose we wish to determine the value of X with the minimum number of binary questions. An efficient first question is "Is $X = a$?" This splits the probability in half. If the answer to the first question is no, then the second question can be "Is $X = b$?" The third question can be "Is $X = c$?" The resulting expected number of binary questions required is 1.75. This turns out to be the minimum expected number of binary questions required to determine the value of X .

Joint entropy and conditional entropy

Definition 2. The **joint entropy** $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

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

which can also be expressed as

$$H(X, Y) = -\mathbb{E} \log p(X, Y).$$

Definition 3. If $(X, Y) \sim p(x, y)$ then the conditional entropy $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= - \mathbb{E}_{p(x, y)} \log p(Y|X) \end{aligned}$$

Proposition 1 (Chain Rule).

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

Proof.

$$\begin{aligned} H(X, Y) &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) p(y|x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= - \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \end{aligned}$$

□

Corollary 1.

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

* *Note.* Observe that $H(Y|X) \neq H(X|Y)$. However,

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

Relative entropy and mutual information

Definition 4. The **relative entropy** or **Kullback Leibler distance** between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$\begin{aligned} D(p\|q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= \mathbb{E}_p \log \frac{p(X)}{q(Y)} \end{aligned}$$

Definition 5. Consider two random variables X and Y with a joint probability mass function $p_{X,Y}(x, y)$ and marginal probability mass functions $p_X(x)$ and $p_Y(y)$. The **mutual information** $I(X; Y)$ is the relative entropy between the joint distribution and the product distribution $p_X(x)p_Y(y)$, i.e.

$$\begin{aligned} I(X; Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log \frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)} \\ &= D(p_{X,Y}(x, y) \| p_X(x)p_Y(y)) \\ &= \mathbb{E}_{p_{X,Y}(x,y)} \log \frac{p_{X,Y}(X, Y)}{p_X(X)p_Y(Y)} \end{aligned}$$

Relationship between entropy and mutual information

We can rewrite the definition of mutual information $I(X; Y)$ as

$$\begin{aligned} I(X; Y) &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x, y) \log \frac{p(x|y)}{p(x)} \\ &= - \sum_{x,y} p(x, y) \log p(x) + \sum_{x,y} p(x, y) \log p(x|y) \\ &= - \sum_x p(x) \log p(x) - \left(- \sum_{x,y} p(x, y) \log p(x|y) \right) \\ &= H(X) - H(X|Y) \end{aligned}$$

by symmetry, it also follows that

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

Thus X says as much about Y as Y says about X . Since $H(X, Y) = H(X) + H(Y|X)$, we have

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

Finally, we note that

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

so that entropy is also sometimes called self-information.

Proposition 2 (Mutual information and entropy).

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

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

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \quad (3)$$

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

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

Proposition 3 (Chain rule for entropy). *Let X_1, X_2, \dots, X_n be drawn according to $p(x_1, x_2, \dots, x_n)$. Then*

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

Definition 6. The **conditional mutual information** of random variables X and Y given Z is defined by

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) \\ &= \mathbb{E}_{p(x,y,z)} \log \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)} \end{aligned}$$

mutual information also satisfies a chain rule.

Proposition 4 (Chain rule for information).

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