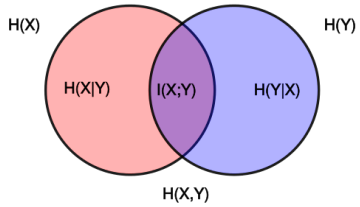
WIKIPEDIA

Conditional entropy

In information theory, the **conditional entropy** (or **equivocation**) quantifies the amount of information needed to describe the outcome of a random variable Y given that the value of another random variable X is known. Here, information is measured in shannons, nats, or hartleys. The *entropy of Y conditioned on X* is written as $\mathrm{H}(Y|X)$.



Venn diagram showing additive and subtractive relationships various information measures associated with correlated variables X and Y. The area contained by both circles is the joint entropy $\mathrm{H}(X,Y)$. The circle on the left (red and violet) is the individual entropy $\mathrm{H}(X)$, with the red being the conditional entropy $\mathrm{H}(X|Y)$. The circle on the right (blue and violet) is $\mathrm{H}(Y)$, with the blue being $\mathrm{H}(Y|X)$. The violet is the mutual information $\mathrm{I}(X;Y)$.

Contents

Definition

The conditional entropy of Y given X is defined as

$$\mathrm{H}(Y|X) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)} \qquad \text{(Eq.1)}$$

where $\mathcal{X}$ and $\mathcal{Y}$ denote the support sets of X and Y.

Note: It is conventioned that the expressions $0\log 0$ and $0\log c/0$ for fixed $c>0$ should be treated as being equal to zero. This is because $\lim_{\theta\to 0^+}\theta\log c/\theta = 0$ and $\lim_{\theta\to 0^+}\theta\log\theta = 0$[1]

Motivation

Let $\mathrm{H}(Y|X=x)$ be the entropy of the discrete random variable Y conditioned on the discrete random variable X taking a certain value x. Denote the support sets of X and Y by $\mathcal{X}$ and $\mathcal{Y}$. Let Y have probability mass function $p_Y(y)$. The unconditional entropy of Y is calculated as $\mathrm{H}(Y) := \mathbb{E}[\mathrm{I}(Y)]$, i.e.

$$\mathrm{H}(Y) = \sum_{y\in\mathcal{Y}}\Pr(Y=y)\,\mathrm{I}(y) = -\sum_{y\in\mathcal{Y}} p_Y(y)\log_2 p_Y(y),$$

where $\mathrm{I}(y_i)$ is the information content of the outcome of Y taking the value y_i. The entropy of Y conditioned on X taking the value x is defined analogously by conditional expectation:

$$\mathrm{H}(Y|X=x) = -\sum_{y\in\mathcal{Y}}\Pr(Y=y|X=x)\log_2\Pr(Y=y|X=x).$$

$\mathrm{H}(Y|X)$ is the result of averaging $\mathrm{H}(Y|X=x)$ over all possible values x that X may take.

Given discrete random variables X with image $\mathcal{X}$ and Y with image $\mathcal{Y}$, the conditional entropy of Y given X is defined as the weighted sum of $\mathrm{H}(Y|X=x)$ for each possible value of x, using $p(x)$ as the weights:[2]:15

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

Properties

Conditional entropy equals zero

$\mathrm{H}(Y|X)=0$ if and only if the value of Y is completely determined by the value of X.

Conditional entropy of independent random variables

Conversely, $\mathrm{H}(Y|X)=\mathrm{H}(Y)$ if and only if Y and X are independent random variables.

Chain rule

Assume that the combined system determined by two random variables X and Y has joint entropy $\mathrm{H}(X,Y)$, that is, we need $\mathrm{H}(X,Y)$ bits of information on average to describe its exact state. Now if we first learn the value of X, we have gained $\mathrm{H}(X)$ bits of information. Once X is known, we only need $\mathrm{H}(X,Y)-\mathrm{H}(X)$ bits to describe the state of the whole system. This quantity is exactly $\mathrm{H}(Y|X)$, which gives the *chain rule* of conditional entropy:

$$\mathrm{H}(Y|X)\ =\ \mathrm{H}(X,Y)-\mathrm{H}(X).$$[2]:17

The chain rule follows from the above definition of conditional entropy:

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

In general, a chain rule for multiple random variables holds:

$$\mathrm{H}(X_1,X_2,\ldots,X_n)=\sum_{i=1}^{n}\mathrm{H}(X_i|X_1,\ldots,X_{i-1})$$[2]:22

It has a similar form to chain rule in probability theory, except that addition instead of multiplication is used.

Bayes' rule

Bayes' rule for conditional entropy states

$$\mathrm{H}(Y|X)\ =\ \mathrm{H}(X|Y)-\mathrm{H}(X)+\mathrm{H}(Y).$$

Proof. $\mathrm{H}(Y|X)=\mathrm{H}(X,Y)-\mathrm{H}(X)$ and $\mathrm{H}(X|Y)=\mathrm{H}(Y,X)-\mathrm{H}(Y)$. Symmetry entails $\mathrm{H}(X,Y)=\mathrm{H}(Y,X)$. Subtracting the two equations implies Bayes' rule.

If Y is conditionally independent of Z given X we have:

$$\mathrm{H}(Y|X,Z)\ =\ \mathrm{H}(Y|X).$$

Other properties

For any X and Y:

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

where $\mathrm{I}(X;Y)$ is the mutual information between X and Y.

For independent X and Y:

$$\mathrm{H}(Y|X) = \mathrm{H}(Y) \text{ and } \mathrm{H}(X|Y) = \mathrm{H}(X)$$

Although the specific-conditional entropy $\mathrm{H}(X|Y=y)$ can be either less or greater than $\mathrm{H}(X)$ for a given random variate y of Y, $\mathrm{H}(X|Y)$ can never exceed $\mathrm{H}(X)$.

Conditional differential entropy

Definition

The above definition is for discrete random variables. The continuous version of discrete conditional entropy is called *conditional differential (or continuous) entropy*. Let X and Y be a continuous random variables with a joint probability density function $f(x,y)$. The differential conditional entropy $h(X|Y)$ is defined as[2]:249

$$h(X|Y) = -\int_{\mathcal{X},\mathcal{Y}} f(x,y)\log f(x|y)\,dx dy \qquad \text{(Eq.2)}$$

Properties

In contrast to the conditional entropy for discrete random variables, the conditional differential entropy may be negative.

As in the discrete case there is a chain rule for differential entropy:

$$h(Y|X) \,=\, h(X,Y) - h(X)$$[2]:253

Notice however that this rule may not be true if the involved differential entropies do not exist or are infinite.

Joint differential entropy is also used in the definition of the mutual information between continuous random variables:

$$\mathrm{I}(X,Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$$

$h(X|Y) \leq h(X)$ with equality if and only if X and Y are independent.[2]:253

Relation to estimator error

The conditional differential entropy yields a lower bound on the expected squared error of an estimator. For any random variable X, observation Y and estimator $\widehat{X}$ the following holds:[2]:255

$$\mathbb{E}\left[\left(X - \widehat{X}(Y)\right)^2\right] \geq \frac{1}{2\pi e} e^{2h(X|Y)}$$

This is related to the uncertainty principle from quantum mechanics.

Generalization to quantum theory

In quantum information theory, the conditional entropy is generalized to the conditional quantum entropy. The latter can take negative values, unlike its classical counterpart.

See also

- Entropy (information theory)
- Mutual information
- Conditional quantum entropy

- Variation of information
- Entropy power inequality
- Likelihood function

References

1. "David MacKay: Information Theory, Pattern Recognition and Neural Networks: The Book" (http://www.inference.org.uk/mackay/itprnn/book.html). *www.inference.org.uk*. Retrieved 2019-10-25.
2. T. Cover; J. Thomas (1991). *Elements of Information Theory* (https://archive.org/details/elementsofinform0000cove). ISBN 0-471-06259-6.