

Conditional Entropy and Mutual Information

STATISTICS

An intuitive introduction to multivariate and conditional entropy, showing how mutual information measures the overlap between two random variables.

PUBLISHED

August 25, 2026

Series: [A Statistical View of Data](#)

3.3 min read

Recap

In the previous post ([linked here](#)), we discussed entropy and information.

To quickly recap,

Recap

1. Information is defined as $I(x) = -\log(p(x))$
2. Entropy is defined as $H(X) = \sum_{i=1}^n p_i \times -\log(p_i)$ where p_i is the probability for x_i and x_i is a possible outcome of the random variable X
3. Entropy is defined on mutually exclusive outcomes.

Now, we move on to discussing multivariate entropy, conditional entropy and mutual information.

Multivariate Entropy

Recall

$$P(x, y) = P(x)P(y|x) = P(y)P(x|y)$$

Suppose we have two random variables X, Y . Calculating the entropy follows quite naturally from the univariate case. Entropy is defined as follows:

$$\begin{aligned}
H(X, Y) &= - \sum_x \sum_y P(x, y) \log(P(x, y)) \\
&= - \sum_x \sum_y P(x, y) [\log P(x) + \log(P(y|x))] \\
&= - \sum_x \sum_y P(x, y) \log P(x) + \sum_x \sum_y P(x, y) \log(P(y|x)) \\
&= - \sum_x \sum_y P(x, y) \log P(x) - \sum_x \sum_y P(x, y) \log(P(y|x)) \\
&= - \sum_x P(x) \log P(x) - \sum_x \sum_y P(x, y) \log(P(y|x)) \\
&= H(X) + - \sum_x \sum_y P(x, y) \log(P(y|x))
\end{aligned}$$

Independent Random Variables

Now, if the two variables are independent, then the second term evaluates to

$$\begin{aligned}
&- \sum_x \sum_y P(x, y) \log(P(y|x)) \\
&= - \sum_x \sum_y P(x) P(y) \log(P(y)) \\
&= - \sum_x P(x) \sum_y P(y) \log(P(y)) \\
&= - \sum_y P(y) \log(P(y)) \\
&= H(Y)
\end{aligned}$$

and the full equation evaluates to $H(X) + H(Y)$.

The interpretation for this is fairly straightforward, since my random variables are independent, knowing one does not tell me anything about the other, so the amount of surprise I get is just the sum of the surprise I would get individually.

Dependent Random Variables

In the case where they are not independent we get:

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

To see how we arrive at this result; consider this:

$$\begin{aligned}
& - \sum_x \sum_y P(x, y) \log(P(y|x)) \\
& = - \sum_x \sum_y P(x) P(y|x) \log(P(y|x)) \\
& = - \sum_x P(x) \sum_y P(y|x) \log(P(y|x)) \\
& = - \sum_x P(x) H(Y|X = x) \\
& = H(Y|X)
\end{aligned}$$

(Note the use of lower case letters to denote an outcome, and upper case letters to denote a random variable).

Of course, these joint distributions could be extended to more variables but the algebra gets tedious.

The interpretation in this case, is how much surprise do I get knowing X , and given that I know X and I find out about Y , how surprised am I? The total surprise I get, is the sum of the two.

Conditional Entropy

The term that we derived above is what we call conditional entropy. So, conditional entropy is defined as

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

Note the term outside the log and inside the log are different (see above for how they reconcile into the same thing).

The interpretation in this case, is how surprised am I when I find out about Y given that I know about X . An important note is that conditioning cannot increase entropy.

Mutual Information

Here, suppose we have two random variables X, Y . I know nothing about their relationship. Mutual information aims to answer, the question - how much information can I get about X , by only observing Y .

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

The interpretation here is as follows:

The first term $H(X)$ tells me all there is to know about X . The second term, $H(X|Y)$ tells how surprised I'd be to find out about X if I knew about Y . The difference between the two, is how much information I get about X from observing Y alone.

Of course the same thing can be said about observing X and gaining information about Y . After some algebra, we can work out that $H(X) - H(X|Y) = H(Y) - H(Y|X)$.

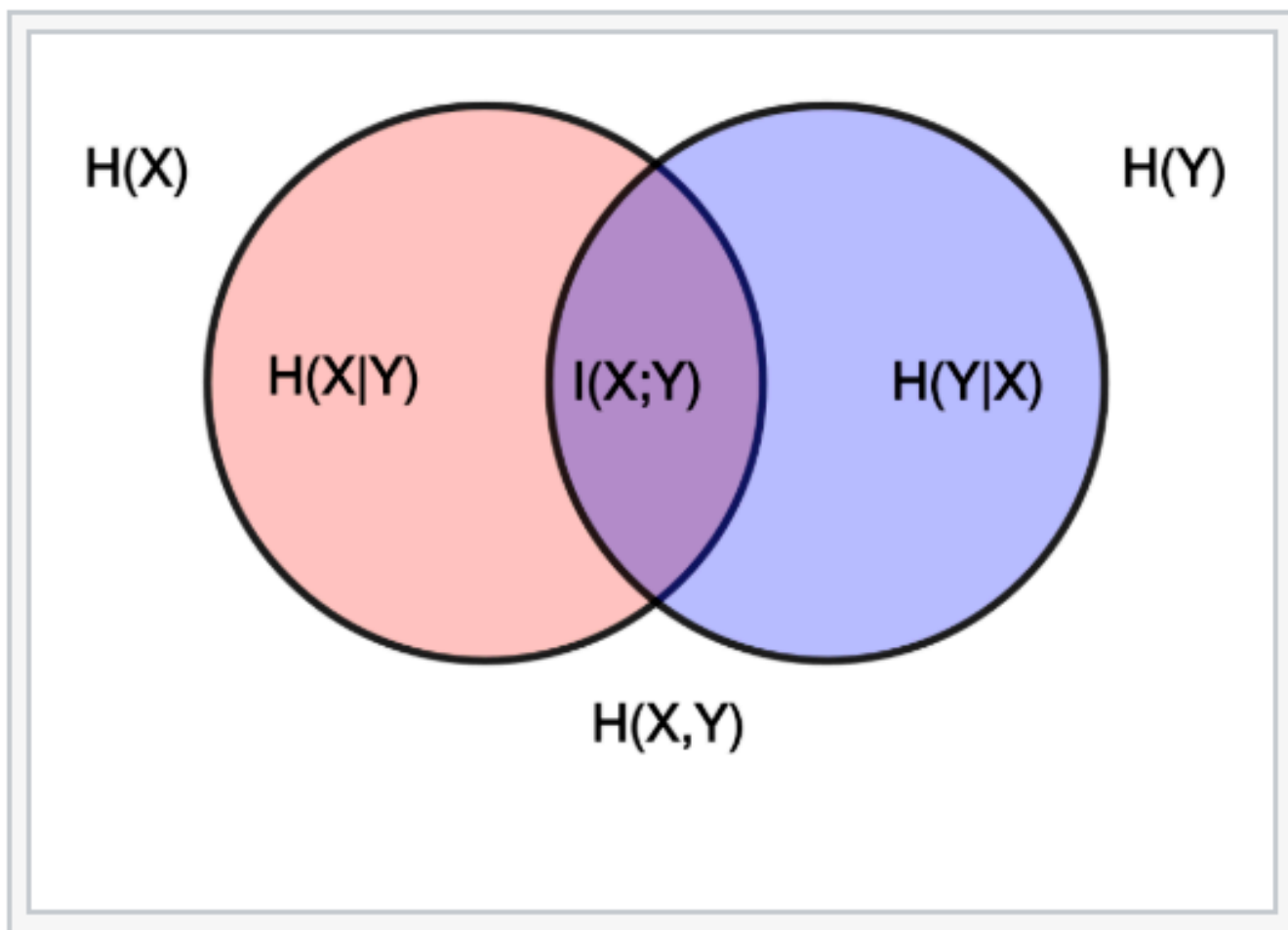
This tells us that mutual information (as the name suggests) isn't one variable informing about the other, it's a shared quantity between the variables.

This notion can also be extended to more variables. $I(X, Y_1, Y_2) = I(X, Y_1) + I(X, Y_2|Y_1)$

Naturally from this definition it follows that $I(X, X) = H(X) - H(X|X) = H(X)$ which is consistent with our reasoning. Observing X tells us all that we need to know about X . This is what we call self information.

Summary

The diagram below shows how joint entropy decomposes into the information shared by X and Y , together with what remains uncertain about each variable after observing the other.



Venn diagram showing the relationship between joint entropy, conditional entropy, and mutual information.

 Image source

This image is from Wikipedia. Clicking the image opens the Wikipedia article on conditional entropy.

Sources

1. Information Theory Lecture Notes
2. Wikipedia — Conditional entropy