Mutual Information and the KL Divergence

PUBLISHED
09 August 2020

This post is part of a series of notes on topics in information theory. The notes assume background knowledge equivalent to an introductory probability course and at some points will require knowledge of univariate calculus. I'd also recommend a refresher on logarithm rules for those who may have forgotten them. For those new to the series the first post is here.

Last time we ended our discussion with the *conditional entropy*. We saw how the conditional entropy measures the uncertainty of an event Y given knowledge of some other event X. Sometimes, instead of looking at the new uncertainty $H(Y|X)$, we instead want to understand the *change* in uncertainty caused by knowing X. This brings us to our first concept of the day, *mutual information*.

In our discussions so far we've assumed that we can measure the uncertainty of an event X accurately. In other words, we assume that we have knowledge of X's probability distribution $p(x)$. However, this assumption isn't always realistic, and we may be forced to guess a distribution q that may vary from the true distribution p. To measure these variations we introduce another quantity: the *relative entropy*.

Suppose we have two random variables $(X, Y) \sim p(x, y)$ with marginal distributions $p(x)$ and $p(y)$ respectively and we want to measure how much information the variables contain about each other. If X and Y are independent (i.e., $p(x, y) = p(x)p(y)$), then intuitively our variables are contain no

information about each other. On the other hand, for non-independent X and Y, knowing the value of one variable should provide some reduction in uncertainty of the other variable. One potential way to measure the amount of information X and Y provide about each other then would be to compare the ratio $p(x,y)/p(x)p(y)$ - in other words, how far the true joint distribution is from what independence would be.

We capture this intuition in the following definition

Definition: *The **mutual information** $I(X;Y)$ of two discrete random variables $(X,Y) \sim p(x,y)$ is defined as*

$$\begin{aligned} I(X;Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= E\left[\log \frac{p(x,y)}{p(x)p(y)} \right] \end{aligned}$$

We can read $I(X;Y)$ as the reduction in uncertainty of X given Y. Unfortunately, the notation for mutual information $I(X;Y)$ is very similar to that of the self-information $I(x)$ defined previously. When unclear from context we will make extra attention to be explicit as to which content we are referring to.

From the definition we can see that mutual information captures how different the joint distribution of X and Y is from what it would be if they were independent (and the joint was simply the product of marginals). The further away the joint is from the product of marginals, the more information X and Y convey about each other. On the other hand, if the variables are independent then we have $I(X;Y) = 0$ as we would expect.

As we hinted at in the end of the previous set of notes, the mutual information is intimately related to our previous definitions of entropy. In particular, starting from the definition above we have

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

and so the mutual information is *precisely* the reduction of uncertainty of X due to the knowledge of Y.

Let's consider an example before moving on.[1]

Example 1: *Suppose we roll a fair six-sided die. Let X represent the top face of the die, and let Y represent the side most facing us. Using our result from above we can calculate $I(X;Y)$ as*

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

Since our die is fair, we have

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) = -\sum_{i=1}^{6}\frac{1}{6}\log\frac{1}{6} = -\log\frac{1}{6} = \log 6$$

Now we calculate $H(X|Y)$. Starting from the definition we have

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

Given a particular value of Y (i.e., the side facing us), there are four potential values for X (the top side) that occur with equal, non-zero probability. This gives us

$$H(X|Y) = -\sum_{i=1}^{6} \frac{1}{6} \sum_{j=1}^{4} \frac{1}{4} \log \frac{1}{4}$$
$$= -\log \frac{1}{4} = \log 4 = 2$$

and going back to our formula for $I(X;Y)$ in terms of entropies we have

$$I(X;Y) = H(X) - H(X|Y) = \log 6 - 2 \approx 0.585 \text{ bits}$$

The reader is encouraged to verify that this result matches that from plugging in directly to the original definition of $I(X;Y)$.

The mutual information is related to entropy in more ways than the equality shown previously. By symmetry (you should verify this!) we also have

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

so X provides as much information about Y as vice versa. From the chain rule of entropy (proven previously) we have $H(X,Y) = H(X) + H(Y|X)$. Thus we can also write

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

It's also worth noting that

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

We collect these results into the following theorem

Theorem: Mutual Information and Entropy *For two discrete random variables $(X,Y) \sim p(x,y)$ we have*

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

We can also capture these relationships by filling in the center of our Venn diagram from last time. This diagram will prove *very* useful in keeping track of the relationships between our different measures of information.

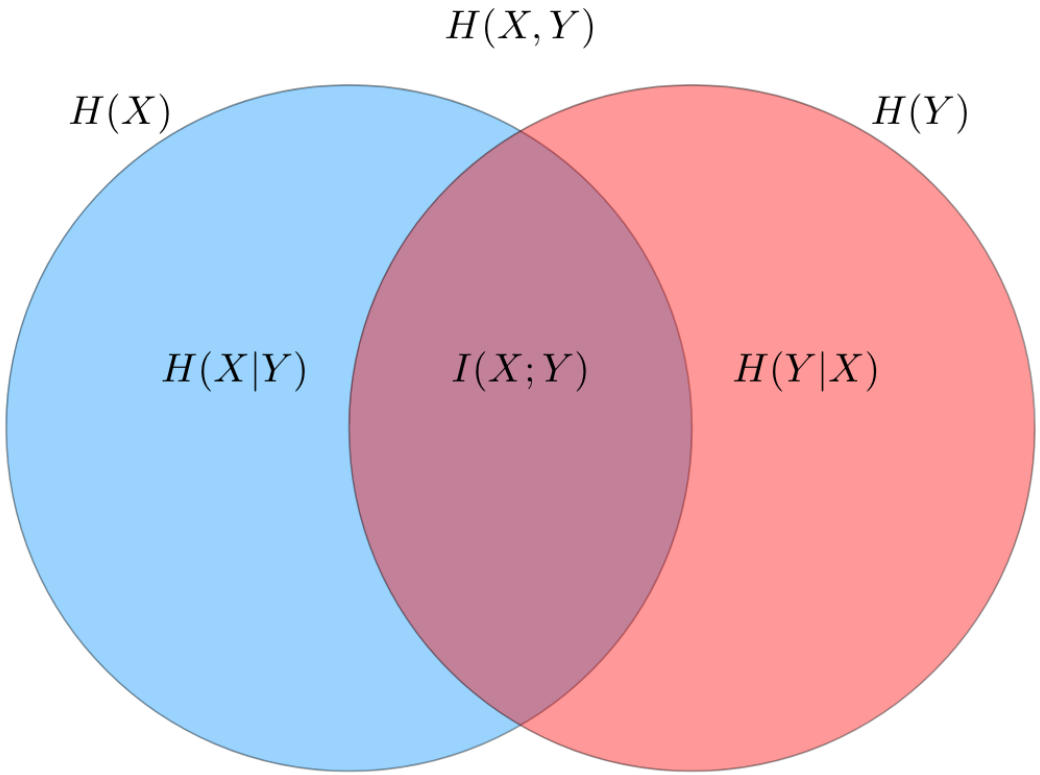


Figure 1. A visual representation of the relationships between mutual information and our measures of entropy. Note that this is the same diagram as in the previous notes with the middle now filled in.

Now, suppose we're given the value of a third random variable Z. How might this affect the reduction in uncertainty of X provided by Y? We can define a new quantity to capture this

Definition: *The **conditional mutual information** $I(X;Y|Z)$ of two discrete random variables X and Y given Z is defined as*

$$I(X;Y|Z) = \sum_z p(z) I(X;Y|Z=z)$$

With some manipulation we can rewrite this to obtain

$$\begin{aligned}
I(X;Y|Z) &= \sum_z p(z) I(X;Y|Z=z) \\
&= \sum_z p(z) \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \\
&= \sum_{x,y,z} p(x,y,z) \left[\log \frac{p(x,y|z)}{p(y|z)} - \log p(x|z) \right] \\
&= -\left[\sum_{x,z} p(x,z) \log p(x|z) - \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(y|z)} \right] \\
&= -\left[\sum_{x,z} p(x,z) \log p(x|z) - \sum_{x,y,z} p(x,y,z) \log p(x|y,z) \right] \\
&= H(X|Z) - H(X|Y,Z)
\end{aligned}$$

Using this result we can show that the mutual information, like entropy, also satisfies a chain rule.

Theorem: The Chain Rule for Mutual Information *For a collection of discrete random variables* $(X_1, X_2, \ldots, X_n, Y) \sim p(x_1, x_2, \ldots, x_n, y)$ *we have*

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

Proof:

$$I(X_1, X_2, \ldots X_n; Y) = H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n | Y)$$

Using the chain rule for entropy we then have

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

In the previous sections we've talked a lot about different measures of information for a random variable X (or variables X, Y, Z, etc.) with known distribution p. However, what if (as is often the case) we don't know the true distribution p, and instead can only approximate it with some other distribution q? In such situations we'll want a way to measure how wrong our assumption is from the truth. This brings us to our final measure of information for the time being, the *relative entropy*.

Definition: *The* ***relative entropy*** $D(p||q)$ *of two distributions* $p(x)$ *and* $q(x)$ *is defined as*

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

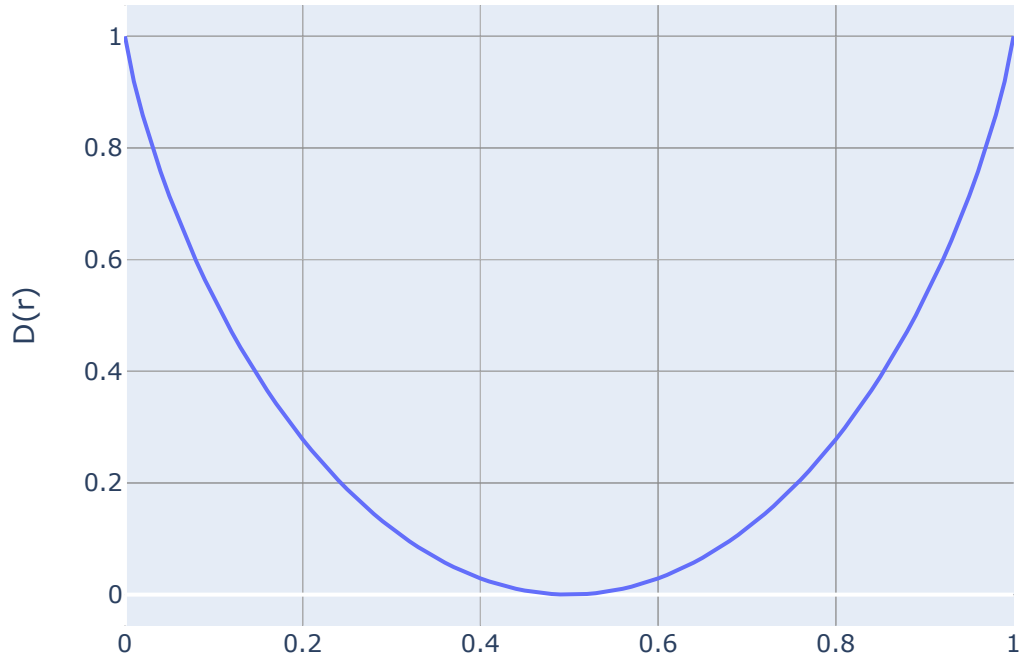
The relative entropy is also known as the *Kullback-Leibler divergence (KL divergence)* after the mathematicians who originally studied the quantity. We'll use the terms interchangeably in these notes. By convention we let (justified using continuity arguments) $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$, and $p \log \frac{p}{0} = \infty$.

Let's look at a quick example to illustrate the concept.

Example 2(a): *Suppose we have some (not necessarily fair) coin, the flips of which we can represent as a Bernoulli distribution p with unknown parameter $r \in [0, 1]$. Since we don't know r, we naively assume that our coin is fair, which we can represent by another Bernoulli distribution q with parameter $1/2$. Using the KL divergence we can quantify how wrong our assumption would be for different values of r:*

$$D(p||q) = (1 - r) \log \frac{1 - r}{1/2} + r \log \frac{r}{1/2}$$

And graphing our KL divergence as a function of r yields:



r

Figure 2. $D(p||q)$ as a function of r. Note that $D(r) = 0$ when $r = 1/2$. This makes sense intuitively, as $p = q$ in this case and so there's no "distance" between the distributions. On the other hand, $D(r)$ is maximized when our coin is as biased as possible at $r = 0$ or $r = 1$.

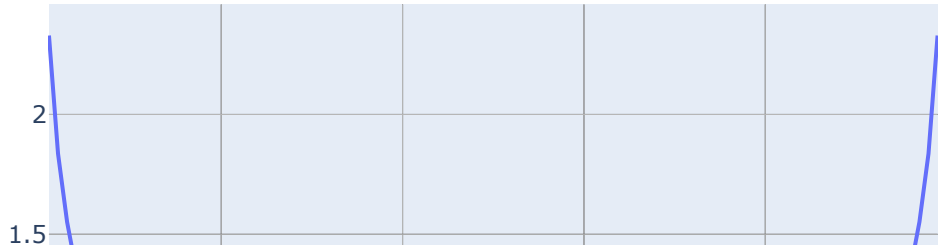
As we might expect $D(p||q)$ is minimized with a value of 0 when $r = 1/2$ - i.e., when our assumption of a fair coin is correct. On the other hand, our KL divergence is maximized when p is deterministic (a completely unfair coin) at $r = 0$ and $r = 1$.

The KL divergence satisfies some nice properties that we would expect from a measure of distance between distributions. Namely, $D(p||q) >= 0$ with equality if and only if $p = q$ at all points. To prove this properties we'll need some additional mathematical tools that we'll introduce in the next set of notes. However, it's important to note now that the KL distance is *not* symmetric. We illustrate this property using the previous example

Example 2(b): *Consider the same p and q from Example 1(a). Calculating $D(q||p)$ instead yields*

$$D(p||q) = \frac{1}{2}\log\frac{1/2}{1-r} + \frac{1}{2}\log\frac{1/2}{r}$$

which we can visualize as



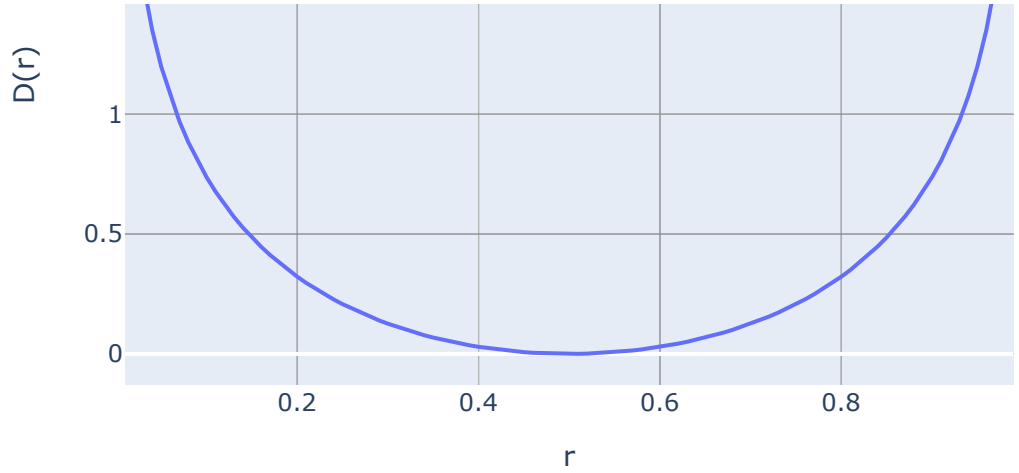


Figure 3. $D(q||p)$ as a function of r. Note the differences in shape from $D(p||q)$.

When $r = 1/2$ (and our distributions are equal) we have $D(p||q) = D(q||p) = 0$. However, from our graphs we can see that $D(p||q)$ is not equal to $D(q||p)$ in general.

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

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y))$$

Thus we can see that $I(X;Y)$ measures the difference (in the KL sense) between the true joint distribution of X and Y and the product of marginal distributions.

As with entropy and mutual information, we can define a *conditional* relative entropy to understand how knowledge of an additional variable changes the KL divergence between two distributions.

Definition: *For two joint probability distributions $p(x,y)$ and $q(x,y)$, the **conditional relative entropy** is defined as*

$$\begin{aligned} D(p(y|x)||q(y|x)) &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \\ &= E\left[\log \frac{p(y|x)}{q(y|x)} \right] \end{aligned}$$

Moreover, as with our previous measures of information the KL divergence also satisfies a chain rule.

Theorem: The Chain Rule for KL Divergence *For two joint probability distributions $p(x,y)$ and $q(x,y)$ we have*

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(x|y)||q(x||y))$$

Proof:

$$\begin{aligned} D(p(x,y)||q(x,y)) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum_{x} p(x) \log \frac{p(x)}{q(x)} + \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \\ &= D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \square \end{aligned}$$

and this proof concludes these notes.

Footnotes:

1. Adapted from https://www.maths.tcd.ie/~houghton/MA3466/PS-09-10/soln4.xq2.c2.pdf ↩