Basic Concepts from Information Theory

11-761: Language and Statistics

1.1.1 Properties of entropy

For a random variable X that can take the values $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, we denote

$$\text{Prob}(X = x_i) = p(x_i) = p_i \,. \tag{1}$$

The *entropy* of X, denoted $H(X)$, or equivalently the entropy of the set of probabilities $\{p_1, p_2, \ldots, p_n\}$, denoted $H(p)$, is defined as follows:

$$H(p) = -\sum_i p_i \log p_i \tag{2}$$

$$H(X) = -\sum_{x_i \in \mathcal{X}} \text{Prob}(X = x_i) \log \text{Prob}(X = x_i) \,. \tag{3}$$

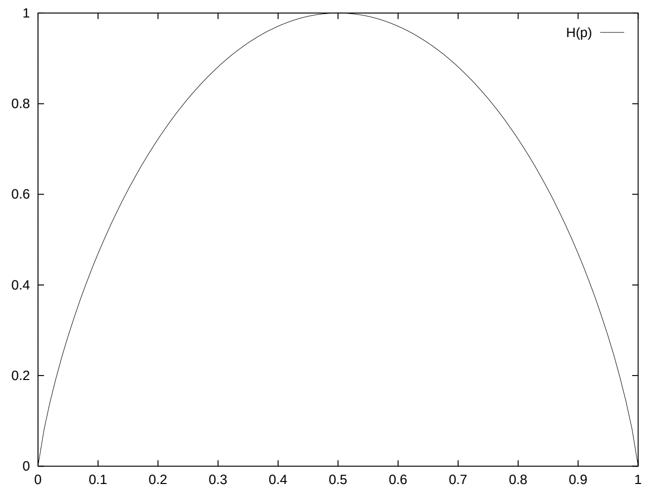

Figure 1: The Shannon entropy

The *cross-entropy* $H(p, q)$ of a distribution q with respect to another distribution p defined over the same event space is given by

$$H(p, q) = E_p[-\log q] = -\sum_{x \in \mathcal{X}} p(x) \log q(x) \,. \tag{4}$$

The *Kullback-Leibler divergence*, or *relative entropy* is defined by

$$D(p \,\|\, q) = E_p[\log(p/q)] = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{5}$$

$$= E_p[-\log q] - E_p[-\log p] \tag{6}$$

One measure of the quality of a model is how well it predicts a new sample of data. Suppose we have a sample $Y = (y_1, \ldots, y_N)$ such that each y_i is sampled independently. Then

$$p(Y) = \prod_{i=1}^{N} p(y_i)$$

Since the log function is monotonic and increasing, we can just as well consider the log of this probability:

$$\begin{aligned} \mathit{logprob}(Y) &= \sum_{i=1}^{N} \log p(y_i) & (7)\\ &= N \sum_{y \in \mathcal{S}} \tilde{p}(y) \log p(y) & (8) \end{aligned}$$

where $\tilde{p}$ is the *empirical distribution*

$$\tilde{p}(y) = \frac{1}{N} \sum_{i=1}^{N} \delta(y_i, y) \tag{9}$$

Equivalently, we might also consider the average logprob

$$\frac{1}{N} \sum_{i=1}^{N} \log p(y_i) = \sum_{y \in \mathcal{S}} \tilde{p}(y) \log p(y) \tag{10}$$

Note that the average logprob is equal to minus the cross-entropy between $\tilde{p}$ and p.

The *perplexity* of a model p, with respect to a sample with empirical distribution $\tilde{p}$ is defined as

$$\mathit{perplexity}(p) = 2^{-\sum_y \tilde{p}(y) \log p(y)} \tag{11}$$

Note that logprob, cross-entropy, and perplexity are all monotonically related to one another. That is, as a measure of "model quality" they are equivalent.

Imagine that you have some events distributed according to some distribution p, and you want to send a sequence of these events to someone else and he knows the distribution p. The entropy $H(p)$ is the number of bits it will take per event assuming an optimal code, and the number of events to be communicated is large. This is essentially the channel coding theorem from information theory.

The entropy of a fair coin is $H(p) = 1$. It takes one bit per toss to communicate the outcomes of a fair coin: 0 for heads, 1 for tails. For a "fair" four-sided die,

$$p(A) = p(B) = p(C) = p(D) = \frac{1}{4} \tag{12}$$

and

$$H(p) = -\log \frac{1}{4} = 2\,. \tag{13}$$

An efficient code to communicate the outcomes is

$$A = 00,\ B = 01,\ C = 10,\ D = 11 \tag{14}$$

Now suppose that we instead have a lopsided die:

$$p(A) = \frac{1}{2},\ p(B) = \frac{1}{4},\ p(C) = p(D) = \frac{1}{8} \tag{15}$$

Then the entropy of this die is

$$H(p) = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = 1.75 \tag{16}$$

To code the outcomes of this die, we can use

$$A = 0,\ B = 10,\ C = 110,\ D = 111 \tag{17}$$

This is a *prefix code*. So, we send 1 bit for A half the time, 2 bits for B a quarter of the time, etc., for an average of 1.75 bits.

What if we send the events with the wrong distribution? Suppose we design the code with the distribution

$$p(A) = p(B) = \frac{1}{8},\ p(C) = \frac{1}{4},\ p(D) = \frac{1}{2} \tag{18}$$

but we use it according to

$$p(A) = \frac{1}{2},\ p(B) = \frac{1}{4},\ p(C) = p(D) = \frac{1}{8} \tag{19}$$

Then we will send

$$-\sum_y \tilde{p}(y) \log p(y) = \frac{3}{2} + \frac{3}{4} + \frac{2}{8} + \frac{1}{8} = 2.625 \tag{20}$$

bits on average. This is a lousy code! Mathematically, this is explained by the inequality $\sum \tilde{p}(y) \log \tilde{p}(y) \geq \sum \tilde{p}(y) \log q(y)$, as we'll now discuss.

1.1.2 A useful inequality

Suppose that f is a convex function. By definition this means that if $p_1, \ldots p_n$ is a probability distribution then

$$f\left(\sum_i x_i\, p_i\right) \leq \sum_i f(x_i)\, p_i \tag{21}$$

In other words, if X is a random variable, then

$$f(E[X]) \leq E[f(X)]\,. \tag{22}$$

This fact is known as *Jensen's inequality*.

One of the most useful facts in this business is the concavity of the logarithm, which can be expressed in various forms. Using Jensen's inequality, we can write

$$\begin{aligned}
-D(p\,\|\,q) &= \sum_{x\in\mathcal{X}} p(x) \log \frac{q(x)}{p(x)} && (23)\\
&\leq \log \sum_{x\in\mathcal{X}} p(x) \frac{q(x)}{p(x)} && (24)\\
&= \log \sum_{x\in\mathcal{X}} q(x) && (25)\\
&= 0 && (26)
\end{aligned}$$

Since log is strictly concave, this shows that $D(p \| q) \geq 0$ with equality if and only $p = q$.

Equivalently, we can employ the very useful inequality

$$x - 1 \geq \log(x) \tag{27}$$

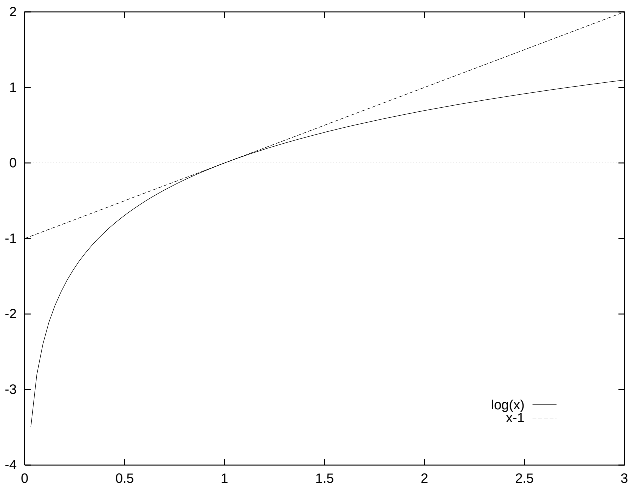

Figure 2: A useful inequality

Using this, it follows that

$$\sum_y p(y) \log q(y) \geq \sum_y p(y) \log p(y) \tag{28}$$

with equality if and only if $p = q$. This is because

$$0 = \sum_y q(y) - \sum_y p(y) \quad = \quad \sum_y p(y) \left(\frac{q(y)}{p(y)} - 1 \right) \tag{29}$$

$$\geq \quad \sum_y p(y) \log \frac{q(y)}{p(y)} \tag{30}$$

This means that the empirical distribution always gives the lowest cross-entropy, lowest perplexity, and the greatest logprob.

1.1.3 Interpretation of the KL divergence

Jensen's inequality shows that the Kullback-Leibler divergence is always non-negative. It is interpreted as the average number of bits that are wasted by encoding events from a distribution p with a code based on a "wrong" distribution q.

In terms of equations, we can calculate that

$$D(p \| q) \quad = \quad \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{31}$$

$$= \quad \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) \tag{32}$$

Therefore,

$$-\sum_x p(x)\log q(x) = H(p,q) = H(p) + D(p\,\|\,q) \tag{33}$$

We observe the following conventions:

$$\begin{aligned} 0\log 0 &= 0 &(34)\\ p\log\frac{p}{0} &= \infty &(35)\end{aligned}$$

The KL divergence is not symmetric and does not satisfy the triangle inequality. However, it has many of the properties of a metric. For example,

$$\begin{aligned} D(p\,\|\,q) &\geq 0 &(36)\\ D(p\,\|\,q) = 0 &\Leftrightarrow \quad p = q &(37)\end{aligned}$$

It also has nice convexity properties:

$$D(\alpha_1 p_1 + \alpha_2 p_2 || \alpha_1 q_1 + \alpha_2 q_2) < \alpha_1 D(p_1\,\|\,q_1) + \alpha_2 D(p_2\,\|\,q_2) \tag{38}$$

for $\alpha_1 + \alpha_2 = 1$. Finally, it satisfies *Pinsker's inequality*:

$$\max_x |p(x) - q(x)| \leq \sum_x |p(x) - q(x)| \leq C\sqrt{D(p\,\|\,q)} \tag{39}$$

1.1.4 Conditional entropy

The *conditional entropy* of a random variable Y conditioned on another, X, is denoted $H(Y\,|\,X)$ and defined as follows:

$$H(Y\,|\,X) = -\sum_x p(x)\sum_y p(y\,|\,x)\log p(y\,|\,x) \tag{40}$$

The *joint entropy* $H(X,Y)$ is defined as (beware the notational confusion with cross-entropy)

$$H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y) \tag{41}$$

Using the definition of conditional probability, $p(Y\,|\,X) = \frac{p(X,Y)}{p(X)}$, we can express a relationship between joint and conditional entropy:

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)}p(x) &(42)\\ &= -\sum_{x,y} p(x,y)(\log p(y\,|\,x) + \log p(x)) &(43)\\ &= -\sum_x p(x)\log p(x) - \sum_{x,y} p(x)p(y\,|\,x)\log p(y\,|\,x) &(44)\\ &= H(X) + H(Y\,|\,X) = H(Y) + H(X\,|\,Y) &(45)\end{aligned}$$

More generally,

$$H(X_1,\ldots,X_n) = \sum_{i=1}^{n} H(X_i\,|\,X_{i-1},\ldots,X_1) \tag{46}$$

This is sometimes called the *chain rule* for conditional entropy.

1.1.5 Mutual information

Mutual information $I(X;Y)$ gives us a measure of how much information is shared between two random variables X and Y. It is the relative entropy between the joint distribution and product distribution of the two random variables, *i.e.*,

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{47}$$

It is symmetric:

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \tag{48}$$

The interpretation of mutual information is the average reduction in the length of a codeword for Y given that X is known.

We define the *conditional mutual information* of two random variables X and Y with respect to a third, Z, as

$$\begin{aligned} I(X;Y \mid Z) &= H(X \mid Z) - H(X \mid Y, Z) && (49) \\ &= H(Y \mid Z) - H(Y \mid X, Z) && (50) \end{aligned}$$

The notation for conditional mutual information is unfortunate because $I(X;Y \mid Z)$ looks like it takes two arguments, the second a "conditional random variable" $Y \mid Z$; but there's no such thing.

The definition generalizes to an arbitrary collection of random variables:

$$\begin{aligned} I(X_1, X_2, \ldots, X_n; Y) &= H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_{n-1}, \ldots, X_1) \\ &\quad - H(X_1 \mid Y) - H(X_2 \mid X_1, Y) - \cdots - H(X_n \mid X_{n-1}, \ldots X_1 \mid Y) && (51) \\ &= \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, \ldots, X_1) && (52) \end{aligned}$$

1.1.6 The data processing inequality

Let X, Y, and Z be random variables such that X and Z are conditionally independent given Y, *i.e.*,

$$\begin{aligned} p(z, x \mid y) &= p(z \mid y) p(x \mid y) && (53) \\ p(x, y, z) &= p(x) p(y \mid x) p(z \mid y) && (54) \end{aligned}$$

In this case X, Y, and Z are said to form a *Markov chain*, and this is written

$$X \longrightarrow Y \longrightarrow Z\,. \tag{55}$$

It implies that $Z \longrightarrow Y \longrightarrow X$ (the Markov property is time-reversible). Also, if $Z = g(Y)$, then we have a Markov chain $X \longrightarrow Y \longrightarrow g(Y)$.

If $X \longrightarrow Y \longrightarrow Z$ then we can show that $I(X;Y) \geq I(X;Z)$:

$$\begin{aligned} I(X;Y,Z) &= I(X;Y) + I(X;Z \mid Y) && (56) \\ &= I(Z;X) + I(X;Y \mid Z) && (57) \end{aligned}$$

But since X and Z are independent given Y, we have that $I(X;Z|Y) = 0$. As a consequence, since mutual information cannot be negative,

$$I(X;Y) \geq I(Z;X) \tag{58}$$

By applying this *data processing inequality* to the case where $Z = f(Y)$, we can interpret this as saying that *no amount of fiddling with the data will improve the inferences that we can make from the data.*

For a lot more information [sic] on all of this, see the book *Elements of Information Theory*, by Thomas Cover and Joy Thomas (Wiley, 1991).