A Primer on Information Theory

November 20, 2019

This note is written as I try to become familiar with terms in information theory such as entropy, mutual information, cross entropy, and the Kullback–Leibler divergence. It is based on two tutorial articles [4, 3] and Chapter 2 of Cover and Thomas [2].

1 Information

- **Definition 1.** *Let E be some event which occurs with probability $\Pr(E)$. If we are told that E has occurred, then we say that we have received*

  $$I(E) = \log \frac{1}{P(E)}$$

  *bits of **information** [1]. Here, the logarithm is in base* 2*.*

- Information can be thought of as the amount of "surprise" in the fact that E occcured.

- The expression $\log(1/P(E))$ can be motivated by searching for a function that satisfies a number of criteria [3]. Say, let $S(p)$ be a function that measures the amount of surprise associated with observing an event that occurs with probability p. The following are reasonable criteria to impose on S:

  - $S(1) = 0$. (Observing a certain event is no surprise.)
  - If $p < q$ then $S(p) > S(q)$. (Rarer events are more surprising.)
  - S varies continuously with p.
  - $S(pq) = S(p) + S(q)$. (Consider two independent events E and F that occurs with probability p and q, respectively. The surprise on seeing $E \cap F$ should be the surprise of seeing E plus the surprise of seeing F.)
  - $S(1/2) = 1$. (A normalizing condition.)

  Given the above criteria, the unique function S that statisfies it is $S(p) = \log(1/p)$.

2 Entropy

- **Definition 2.** *Let X be a discrete random variable that takes values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$, respectively. The **entropy** of X, denoted by $H(X)$ is given by:*

  $$H(X) = \sum_{i=1}^{n} \Pr(X = x_i) \log \frac{1}{\Pr(X = x_i)} = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log p_i.$$

- $H(X)$ can be interpreted as:

  - The average amount of surprise when we observe a realization of X.

  - The average amount of information of a value X can take.
  - The average number of bits needs to communicate the outcome of X.
  - The uncertainty an observer has before seeing the outcome of X.

- When Shannon first defined the concept, he was at a lost of what to call it. Von Neumann suggested the word entropy because (1) it is similar to entropy in statistical physics (but not exactly the same), and (2) nobody knows what entropy actually means, so Shannon would have an advantage when arguing with other people [3].

- Many properties of entropy can be proven by Jensen's inequality.

  Theorem 3 (Jensen's inequality). *Let $f : [a,b] \to \mathbb{R}$ be a continuous, concave function. Let $p_1, p_2, \dots, p_n$ be non-negative real numbers that sum to 1. For any $x_1, x_2, \dots, x_n \in [a,b]$, we have:*

$$\sum_{i=1}^{n} p_i f(x_i) \le f\left(\sum_{i=1}^{n} p_i x_i \right).$$

- **Proposition 4 (Maximality of the Uniform).** *For random variable X,*

$$H(X) \le \log |\mathrm{range}(X)|$$

  where $\mathrm{range}(X)$ *is the set of values that Xtakes on with positive probability.*

  Proof. Let X takes value $x_1, \dots, x_n$ with probability $p_1, \dots, p_n$. Let us also assume that all the p_i's are positive. We have that:

$$H(X) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} \le \log \left(\sum_{i=1}^{n} p_i \frac{1}{p_i} \right) = \log n = \log |\mathrm{range}(X)|.$$

  The inequality in the derivation is an application of Jensen's inequality. □

- **Theorem 5 (Gibb's Inequality).** *Let X be a random variable that takes value $x_1, \dots, x_n$ with probability $p_1, \dots, p_n$. Let $q_1, q_2, \dots, q_n$ be another probability mass function over the possible values of X. Then,*

$$H(X) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} \le \sum_{i=1}^{n} p_i \log \frac{1}{q_i}.$$

  Proof. WLOG, let us assume that all the p_i's are positive. We have that:

$$\sum_{i=1}^{n} p_i \log \frac{1}{p_i} - \sum_{i=1}^{n} p_i \log \frac{1}{q_i} = \sum_{i=1}^{n} p_i \log \frac{q_i}{p_i} \le \log \left(\sum_{i=1}^{n} p_i \frac{q_i}{p_i} \right) = \log \left(\sum_{i=1}^{n} q_i \right) = \log 1 = 0.$$

- We will use the above inequality when we define the Kullback–Leiber divergence.

3 Joint and Conditional Entropy

- In this section, let X and Y be two random variables with joint probability distribution $p(x,y)$.

- **Definition 6.** *The **joint entropy** $H(X,Y)$ is defined as:*

$$H(X,Y) = \sum_x \sum_y p(x,y) \log \frac{1}{p(x,y)} = -\sum_x \sum_y p(x,y) \log p(x,y).$$

- In other words, the cross entropy $H(X, Y)$ is the entropy of the tuple (X, Y), taken as a single random variable.

- **Definition 7.** *If E is any event, we define the entropy of X given E to be*

$$H(X|E) = \sum_x p(x|E) \log \frac{1}{p(x|E)}.$$

- **Definition 8.** *For a pair of random variables X and Y, the conditional entropy of X given Y is given by:*

$$H(X|Y) = E_Y[H(X|\{Y = y\})] = \sum_y p(y) \left(\sum_x p(x|y) \log \frac{1}{p(x|y)} \right).$$

- **Proposition 9 (Chain Rule).**

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

Proof. We have that:

$$\begin{aligned}
H(X,Y) - H(X) &= \sum_x \sum_y p(x,y) \log \frac{1}{p(x,y)} - \sum_x p(x) \log \frac{1}{p(x)} \\
&= \sum_x \sum_y p(x)p(y|x) \log \frac{1}{p(x)p(y|x)} - \sum_x p(x) \left(\sum_y p(y|x) \right) \log \frac{1}{p(x)} \\
&= \sum_x p(x) \left(\sum_y p(y|x) \log \frac{1}{p(x)p(y|x)} \right) - \sum_x p(x) \left(\sum_y p(y|x) \log \frac{1}{p(x)} \right) \\
&= \sum_x p(x) \left[\sum_y p(y|x) \left(\log \frac{1}{p(x)p(y|x)} - \frac{1}{p(x)} \right) \right] \\
&= \sum_x p(x) \left(\sum_y p(y|x) \log \frac{1}{p(y|x)} \right) \\
&= H(Y|X).
\end{aligned}$$

- It can be shown by induction that:

$$H(X_1, \ldots, H_k) = H(X_1) + H(X_2|X_1) + H(H_3|H_1, H_2) + \cdots + H(X_n|X_1, \ldots, X_{n-1}).$$

- The chain rule relates the entropy of a random vector to the entropy of revealing the components one by one.

- **Proposition 10 (Dropping Conditioning).** *For random variables X and Y,*

$$H(X|Y) \leq H(X).$$

Also, for random variable Z,

$$H(X|Y, Z) \leq H(X|Y).$$

Proof. We will prove only the first inequality. The second is very similar. We have that:

$$\begin{aligned}
H(X|Y) &= \sum_y p(y)\left(\sum_x p(x|y)\log\frac{1}{p(x|y)}\right)\\
&= \sum_y\sum_x p(y)p(x|y)\log\frac{1}{p(x|y)}\\
&= \sum_x\sum_y p(x)p(y|x)\log\frac{1}{p(x|y)}\\
&= \sum_x p(x)\left(\sum_y p(y|x)\log\frac{1}{p(x|y)}\right)\\
&\le \sum_x p(x)\log\left(\sum_y \frac{p(y|x)}{p(x|y)}\right)\\
&= \sum_x p(x)\log\left(\sum_y \frac{p(x,y)}{p(x)}\frac{p(y)}{p(x,y)}\right)\\
&= \sum_x p(x)\log\left(\sum_y \frac{p(y)}{p(x)}\right)\\
&= \sum_x p(x)\log\frac{1}{p(x)}\\
&= H(X).
\end{aligned}$$

- The above proposition should be intuitive. The surprise of knowing X after knowing Y should be less than the surprise of knowing X without having any other information because more information can only reduce surprise.

- **Proposition 11 (Subadditivity).** *For random variables $X_1, X_2, \ldots, X_n$, we have that:*

$$H(X_1, X_2, \ldots, X_n) \le \sum_{i=1}^{n} H(X_i).$$

Proof.

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

The first step is an application of the chain rule, and the second step is just dropping conditioning. □

4 Mutual Information

- Let p and q be two probability mass functions on the set $x_1, x_2, \ldots, x_n$.

- **Definition 12.** *The **relative entropy** or **Kullback–Leiblier divergence** from q to p, denoted by $D(p\|q)$, is given by:*

$$D(p\|q) = \sum_x p(x)\log\frac{p(x)}{q(x)}.$$

- The KL divergence is a non-negative number. This is because

  $$\sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_x p(x) \log \frac{1}{q(x)} - \sum_x p(x) \log \frac{1}{p(x)}.$$

  We know from Gibb's inequality that the above expression is greater than or equal to 0. We can also show that it is zero if and only if $p = q$. As such, it can be a measure of how q is different from p.

- Let X and Y be two random variables with joint probability distribution $p(x, y)$ and marginal distribution $p(x)$ and $p(y)$.

  Definition 13. *The* ***mutual information*** *between X and Y, denoted by $I(X;Y)$, is the KL divergence between the joint distribution and the product distribution $p(x)p(y)$. That is,*

  $$I(X;Y) = D(p(x,y) \| p(x)p(y)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$

- Clearly, $I(X;Y) = I(Y;X)$.

- **Proposition 14.**

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

  Proof.

  $$\begin{aligned} I(X;Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x|y)}{p(x)} \\ &= \sum_x \sum_y p(x,y) \log \frac{1}{p(x)} - \sum_x \sum_y p(x,y) \log \frac{1}{p(x|y)} \\ &= \sum_x p(x) \log \frac{1}{p(x)} - \sum_y p(y) \bigg(\sum_x p(x|y) \log \frac{1}{p(x|y)} \bigg) \\ &= H(X) - H(X|Y). \end{aligned}$$

  The equation $I(X;Y) = H(Y) - H(Y|X)$ can be proved similarly. Now, we have that:

  $$I(X,Y) = H(X) - H(X|Y) = H(X) - [H(X,Y) - H(Y)] = H(X) + H(Y) - H(X,Y).$$

- Let's interpret what the mutual information means. Consider the equation

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

  $H(X)$ is the information about X, and $H(X|Y)$ is the extra information about X given that we already know about Y. Subtracting $H(X|Y)$ from $H(X)$ gives us the information about X that we already know when we know Y. So, $I(X;Y)$ is the information the random variables have about each other.

- **Definition 15.** *The* ***conditional mutual information*** *of random variable X and Y given Z is defined by:*

  $$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = \sum_{x,y,z} p(x,y,z) \frac{p(x,y|z)}{p(x|z)p(y|z)}.$$

- **Proposition 16 (Chain Rule for Mutual Information).**

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

Proof.

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

- Now, consider two probability distribions $p(\cdot, \cdot)$ and $q(\cdot, \cdot)$ over the set of tuples $\{x_1, \ldots, x_n\} \times \{y_1, \ldots, y_m\}$. Note that the KL divergence from q to p is given by:

$$D(p\|q) = D(p(x, y)\|q(x, y)) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{q(x, y)}.$$

- **Definition 17.** *The* ***conditional Kullback–Leibler divergence*** $D(p(y|x)\|q(y|x))$ *is the expected value with respect to* x *of the KL divergence from* $q(y|x)$ *to* $p(y|x)$*. That is,*

$$D(p(y|x)\|q(y|x)) = \sum_{x} p(x) \left(\sum_{y} p(y|x) \log \frac{p(y|x)}{q(y|x)} \right) = \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)}.$$

- **Proposition 18 (Chain Rule for KL Divergence).**

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

Proof.

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

5 Cross Entropy

- So far, we have only defined entropy of a random variable. However, we can also think of entropy as a function of the probability distribution.

  Definition 19. *Let $p(\cdot)$ be a probability distribution over the set $x_1, \ldots, x_n$. The* ***entropy*** *$H(p)$ of p is given by:*

  $$H(p) = \sum_x p(x) \log \frac{1}{p(x)}.$$

  That is, it is the entropy of the random variable X where $\Pr(X = x_i) = p(x_i)$.

- **Definition 20.** *The* ***cross entropy*** *$H(p,q)$ of probability distribution q with respect to p is given by:*

  $$H(p,q) = \sum_x p(x) \log \frac{1}{q(x)}.$$

- **Proposition 21.**

  $$H(p,q) = H(p) + D(p\|q)$$

  Proof.

  $$\begin{aligned} H(p) + D(p\|q) &= \sum_x p(x) \log \frac{1}{p(x)} + \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_x p(x) \left(\log \frac{1}{p(x)} + \log \frac{p(x)}{q(x)} \right) \\ &= \sum_x p(x) \log \frac{1}{q(x)} \\ &= H(p,q). \end{aligned}$$

- The typical interpretation of the cross entropy involves trying to estimate p with q. For each element x_i, the number $\log(1/q(x_i))$ is the amount of bits needed to encode the outcome x_i when the generating distribution is q. The cross entropy measures the average number of bits when the outcomes are sampled according to p but encoded according to q.

  By Gibb's inequality, the most succinct encoding is when $p = q$, resulting in the average number of bits of $H(p)$. By trying to minimize the cross entropy, the process would make q as close to p as possible. Equivalently, it would try to force the KL divergence down to 0.

References

[1] Abramson, N. *Information Theory and Coding.* McGraw–Hill, 1963.

[2] Cover, T. M., and Thomas, J. A. *Elements of Information Theory.* Wiley–Interscience, 1991.

[3] Galvin, D. Three tutorial lectures on entropy and counting. `https://arxiv.org/abs/1406.7872`.

[4] Rosenfeld, R. A gentle tutorial on information theory and learning. `http://www.cs.cmu.edu/~roni/10601-slides/info-theory.pdf`.