Information Theory: Lecture Notes 1

zqy1018

June 10, 2020

Contents

1 Nomenclature

$X, Y, Z, \cdots$: random variables.

$(X_1, X_2, \cdots, X_n)$: a n-dimensional random vector.

$\mathcal{X}, \mathcal{Y}, \cdots$: alphabets (a.k.a. sample space).

$x, y, z, \cdots$: elements in the sample space.

$p(x)$: probability mass function. A shorthand for $P(X = x)$.

$\mathrm{E}_p g(X)$: shorthand for $\mathrm{E}(g(X))$, where the probability mass function is p. That is, $\mathrm{E}_p g(X) = \sum_{x \in \mathcal{X}} p(x) g(x)$.

2 Entropy

2.1 Basic Definitions

Definition. The **entropy** $H(X)$ of X is defined by

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

Also can be written as $H(p)$.

Note.

(1) Usually the log is to the base 2 and entropy is expressed in **bits**. If the base of the logarithm is b, we denote the entropy as $H_b(X)$. If the base of the logarithm is e, the entropy is measured in **nats**.

(2) The function is well-defined since $\lim_{x \to 0} x \log x = 0$.

(3) $H(X)$ has nothing to do with a specific $\mathcal{X}$, but is related to p.

(4) $H(p)$ (or $H(p, 1-p)$) is a short-hand for $-p \log p - (1-p) \log(1-p)$ (i.e. the entropy of a two-point distribution) if $p \in [0, 1]$.

Intuitively, entropy can be seen as *a measure of uncertainty a random variable.*

Definition. An equivalent definition is that the **entropy** $H(X)$ of X is the expected value of $\log \frac{1}{p(X)}$. So $H(X) = -\mathrm{E}_p \log p(X)$.

This definition of entropy is related to the definition of entropy in physics.

2.2 Basic Properties of Entropy

Theorem 1. (Uniform distribution maximizes entropy) $\forall X, \log |\mathcal{X}| \geq H(X) \geq 0$.

Proof. $\forall x, 1 \geq p(x) \geq 0 \implies H(X) \geq 0$.

By Jensen's inequality: $\sum p_i f(x_i) \geq f(\sum p_i x_i)$, if f is convex. Then since $x \log x$ is convex, we have:

$$-\sum_{x\in\mathcal{X}} \frac{1}{|\mathcal{X}|} \cdot p(x) \log p(x) \leq -\left(\sum_{x\in\mathcal{X}} \frac{1}{|\mathcal{X}|} \cdot p(x)\right) \log \left(\sum_{x\in\mathcal{X}} \frac{1}{|\mathcal{X}|} \cdot p(x)\right) = \frac{1}{|\mathcal{X}|} \log |\mathcal{X}|$$

The equality holds iff $p(x) = \frac{1}{|\mathcal{X}|}$. □

3 Joint Entropy and Conditional Entropy

3.1 Basic Definitions

Definition. The **joint entropy** $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

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

Also can be written as $H(X, Y) = -\mathrm{E}_{p(x,y)} \log p(X, Y)$.

Generally, for a n-dimensional random vector $(X_1, X_2, \cdots, X_n)$ with joint distribution $p(x_1, \cdots, x_n)$, its joint entropy is defined as

$$H(X_1, X_2, \cdots, X_n) = -\sum p(x_1, \cdots, x_n) \log p(x_1, \cdots, x_n) = -\mathrm{E}_p \log p(x_1, \cdots, x_n)$$

We know that $P(Y|X = x)$ is also a probability distribution, so we can write its entropy as $H(Y|X = x)$. Now we define the conditional entropy for a joint distribution.

Definition. The **conditional entropy** $H(Y|X)$ is defined as

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

Note. Generally $H(X|Y) \neq H(Y|X)$. Easy to check.

The definition above shows two ways of calculating $H(Y|X)$.

1. To calculate the expected value of a new random vector $\log p(Y|X)$.
2. To calculate the weighted average of $H(Y|X = x)$ (the weight is $p(x)$).

Intuitively, when X is known, the uncertainty of Y will not increase; that is, it should hold that $H(Y|X) \leq H(X)$. We will prove it later.

3.2 Basic Properties

Theorem 2. $\forall X, Y, H(X, X) = H(X), H(X, Y) = H(Y, X)$.

Note that the first equation shows that duplication of information can not reduce uncertainty.

Theorem 3. (Chain rule) $\forall X, Y, H(X|Y) + H(Y) = H(Y|X) + H(X) = H(X, Y)$.

Proof.

$$
\begin{aligned}
& p(x, y) = p(x|y)p(y) = p(y|x)p(x) \\
\Longrightarrow\ & \log p(x, y) = \log p(x|y) + \log p(y) = \log p(y|x) + \log p(x) \\
\Longrightarrow\ & -\mathrm{E}_{p(x,y)} \log p(x, y) = -\mathrm{E}_{p(x,y)} \log p(x|y) - E_{p(x,y)} \log p(y) = -\mathrm{E}_{p(x,y)} \log p(y|x) - \mathrm{E}_{p(x,y)} \log p(x) \\
\Longrightarrow\ & H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)
\end{aligned}
$$

Note that here we use the logarithm function to *linearize* the equation, i.e. change the multiplication into addition, in order to use the linearity of expected values. □

Corollary.

(1) If X and Y are independent, $H(X, Y) = H(X) + H(Y)$.

(2) (Bayesian rule) $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$.

(3) (Conditioning reduces entropy) $H(X|Y) \leq H(X)$, since $H(Y) \geq 0$.

Warning. For some particular X and Y, there may $\exists y \in \mathcal{Y}$, such that $H(X|Y = y) \geq H(X)$.

What if $H(Y|X) = 0$?

Theorem 4. (Problem 2.5 in [Cover]) $H(Y|X) = 0$ iff Y is a function of X, i.e. for all x with $p(x) > 0$, there is only one possible value of y with $p(x, y) > 0$.

Proof. We only show the " $\Longrightarrow$ " part here.

$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$. Since the condition entropy is always non-negative, we have: $\forall x \in \mathcal{X}, p(x) > 0 \implies H(Y|X = x) \geq 0$.

Then we prove the following proposition:

$$\forall x \in \mathcal{X}, p(x) > 0, \exists! y \in \mathcal{Y}, P(Y = y|X = x) = 1$$

We can prove it easily by contradiction. Suppose $\exists y_1, y_2 \in \mathcal{Y}, 1 > P(Y = y_1|X = x), P(Y = y_2|X = x) > 0$. Then $H(Y|X = x) \geq P(Y = y_1|X = x) \log \frac{1}{P(Y=y_1|X=x)} > 0$, which yields a contradiction.

Then we can lead to the conclusion that $\forall x \in \mathcal{X}, p(x) > 0 \implies \exists! y \in \mathcal{Y}, p(x, y) = p(x)P(Y = y|X = x) = p(x) > 0$. In other words, Y is a function of X. □

4 Relative Entropy

4.1 Basic Definition

Till now, we should notice that although we use random variables, we do not care about their values. We only need their probability mass functions, or, distributions.

So it may help if we just consider a distribution as *a point in a high dimensional space.*

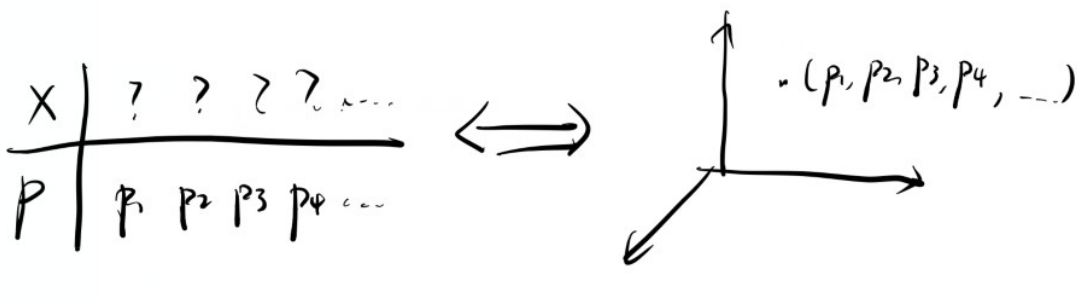

Figure 1: The accordance of a distribution and a point.

Definition. The **relative entropy**, **information divergence** or **Kullback–Leibler distance** between two probability mass functions $p(x)$ and $q(x)$ *over the same alphabet* is defined as

$$\begin{aligned} D(p\|q) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \\ &= \mathrm{E}_p \log\frac{p(X)}{q(X)} \end{aligned}$$

Note that in the second line, the random variable $\log\frac{p(X)}{q(X)}$ is weighted by p, not q.

In such a view, we can treat the relative entropy defined here as a measure of the distance between two distributions.

Note.

(1) $0\log\frac{0}{0} = 0, 0\log\frac{0}{q} = 0, p\log\frac{p}{0} = \infty(p, q > 0)$. So if there is any $x \in X$ such that $p(x) > 0$ and $q(x) = 0$, then $D(p\|q) = \infty$.

(2) $D(p\|q) = -\mathrm{E}_p \log q(x) + \mathrm{E}_p \log p(x) = -\mathrm{E}_p \log q(x) - H(p)$.

Remark. More formally, the relative entropy $D(p\|q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p. For example, if we knew the true distribution p of the random variable, we could construct a code with average description length $H(p)$. If, instead, we used the code for a distribution q, we would need $H(p) + D(p\|q)$ bits on the average to describe the random variable.

We can also add conditions to the relative entropy.

Definition. For joint probability mass functions $p(x, y)$ and $q(x, y)$, the **conditional relative entropy** $D(p(y|x)\|q(y|x))$ is the average of the relative entropy between the conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over the probability mass function $p(x)$. More precisely,

$$\begin{aligned} D(p(y|x)\|q(y|x)) &= \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum_x \sum_y p(x)p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= \mathrm{E}_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)} \end{aligned}$$

Note. Both the relative entropy and the conditional relative entropy are averaged on *the first term without conditions*. For example, $D(p\|q)$ is averaged on $p(x)$, and $D(p(y|x)\|q(y|x))$ is averaged on $p(x, y)$ (no conditions here).

4.2 Basic Properties

Definition. A **metric** $d : X \times Y \to \mathbb{R}^+$ between two elements satisfies:

1. $d(x, y) \geq 0$.
2. $d(x, y) = d(y, x)$.
3. $d(x, y) = 0 \iff x = y$.
4. $d(x, y) + d(y, z) \geq d(x, z)$.

For example, the Euclidean distance is a metric.

Note. $\exists p, q, D(p\|q) \neq D(q\|p)$. Thus KL-distance is not a metric. Also, the KL-distance does not necessarily satisfy the triangular inequality.

Definition. The **variation distance** between p and q is denoted as

$$V(p, q) = \sum_{x \in \mathcal{X}} |p(x) - q(x)|$$

Easy to check it is a metric.

Theorem 5. (Pinsker's inequality)

$$\forall p, q, D(p\|q) \geq \frac{1}{2 \ln 2} V^2(p, q)$$

So although KL-distance is not a real metric, it is useful to be treated as a "metric". And it can be bounded below by a real metric.

Theorem 6. $D(p\|q) \geq 0$. The equality holds iff $p(x) = q(x), \forall x$.

Proof. With Jensen's inequality:

$$\begin{aligned} D(p\|q) &= \sum_{x\in\mathcal{X}} p(x)\left(-\log\frac{q(x)}{p(x)}\right) \\ &\geq -\log\left(\sum_{x\in\mathcal{X}} p(x)\frac{q(x)}{p(x)}\right) \\ &\geq -\log\left(\sum_{x\in\mathcal{X}} q(x)\right) \geq \log 1 = 0 \end{aligned}$$

We write $-\log\sum_{x\in\mathcal{X}} p(x)\frac{q(x)}{p(x)} \geq -\log\sum_{x\in\mathcal{X}} q(x)$ since some $p(x)$ may be 0.
With $\log x \leq x - 1 (1 \geq x > 0)$:

$$\begin{aligned} D(p\|q) &= \sum_{x\in\mathcal{X}} p(x)\left(-\log\frac{q(x)}{p(x)}\right) \\ &\geq \sum_{x\in\mathcal{X}} p(x)\left(1-\frac{q(x)}{p(x)}\right) \\ &\geq \left(\sum_{x\in\mathcal{X}} p(x)\right) - \left(\sum_{x\in\mathcal{X}} q(x)\right) = 0 \end{aligned}$$

We write $\sum_{x\in\mathcal{X}} p(x)\left(1-\frac{q(x)}{p(x)}\right) \geq \left(\sum_{x\in\mathcal{X}} p(x)\right) - \left(\sum_{x\in\mathcal{X}} q(x)\right)$ since some $p(x)$ may be 0. □

5 Mutual Information

5.1 Basic Definition

Usually one random variable contains some information about another random variable. We use mutual entropy to measure it.

Definition. Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution

$p(x)p(y)$:

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

Note. Do not write $I(X,Y)$ or $H(X;Y)$.

Remark. We can interpret the mutual information $I(X;Y)$ as the reduction of the uncertain of X after Y is observed.

We now define the conditional mutual information as the reduction in the uncertainty of X due to knowledge of Y when Z is given.

Definition. The **conditional mutual information** of random variables X and Y given Z is defined by

$$\begin{aligned} I(X;Y|Z) &= \sum_{z\in\mathcal{Z}} p(z)\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)} \\ &= \mathrm{E}_{p(x,y,z)}\log\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)} \\ &= H(X|Z) - H(X|Y,Z) \end{aligned}$$

Note. The priority between operators: , (joint distribution) is higher than ; (separation of two distributions). | (conditions) is the lowest.

We can also define the multivariate mutual information of three random variables. It can be generalized to n random variables.

Definition. Define $I(X;Y;Z) = I(X;Y) - I(X;Y|Z)$.

Warning. The conditional mutual information is not necessarily less than or equal to the mutual information, and therefore the multivariate mutual information is *not necessarily non-negative*. For example, assume X, Y are independent and uniformly distributed on $\{0,1\}$. Let $Z = X \operatorname{xor} Y$. Then $I(X;Y|Z) > I(X;Y)$.

5.2 Basic Properties

Theorem 7. $\forall X, Y, I(X;Y) = I(Y;X), I(X;X) = H(X)$.

So sometimes entropy is also called **self-information**.

Theorem 8. If X and Y are independent, $I(X;Y) = 0$.

Theorem 9. $\forall X, Y, I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.

Proof. We only prove $I(X;Y) = H(X) - H(X|Y)$. The other is similar.

Using the chain rule:

$$\begin{aligned} & p(X,Y) = p(Y)p(X|Y) \\ \implies & \frac{p(X,Y)}{p(X)p(Y)} = \frac{p(X|Y)}{p(X)} \\ \implies & \log \frac{p(X,Y)}{p(X)p(Y)} = \log p(X|Y) - \log p(X) \end{aligned}$$

Take the expected value on both sides. Then it is finished. □

Corollary. $\forall X, Y, H(X,Y) = H(X) + H(Y) - I(X;Y)$.

Proof. Use $H(X,Y) = H(X) + H(Y|X)$. □

We can use the information diagram to help remember those equations.

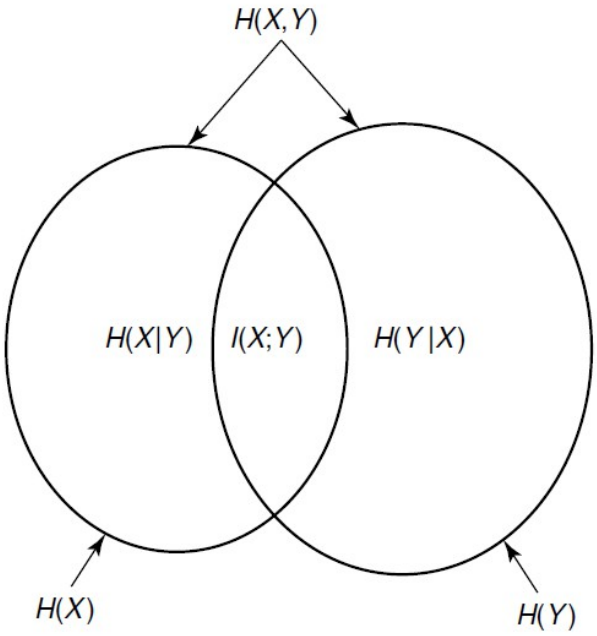

Figure 2: The information diagram of X, Y.

6 Chain Rule

6.1 For Joint Entropy

We know $H(X,Y) = H(Y) + H(X|Y)$. Can we generalize it?

Theorem 10. Let $(X_1, X_2, \cdots, X_n)$ be a random vector and $p(x_1, \cdots, x_n)$ be the probability mass function. Then

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

Proof. It is easy to prove with the take-log-and-then-take-expected-value approach we have used many times above.

We can also prove it with Bayesian rule. For example, $H(X, Y, Z) = H(X) + H(Y, Z|X) = H(X) + H(Y|X) + H(Z|X, Y)$. □

Corollary. (Independence bound on entropy) $H(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$. The equality holds iff all the X_i are independent.

6.2 For Conditional Mutual Information

Theorem 11. Let $(X_1, X_2, \cdots, X_n)$ be a random vector and $p(x_1, \cdots, x_n)$ be the probability mass function. Then

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

Proof.

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

□

6.3 For Conditional Relative Entropy

Theorem 12. $D(p(x, y) \| q(x, y)) = D(p(x) \| q(x)) + D(p(y|x) \| q(y|x)) = D(p(y) \| q(y)) + D(p(x|y) \| q(x|y))$.

Proof. By definition,

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

Similarly for y being the condition. □

7 Review

We can see that the information theory is based on the probability theory. So when doing proofs, we can either use the facts in the probability theory (at a lower level) or use the facts in the information theory (at a higher level).

Also, it may be useful to *compress* a list of random variables as a single random vector. It will transform the n-variable case into the easier 2 or 3-variable case, which will be easier to think about.

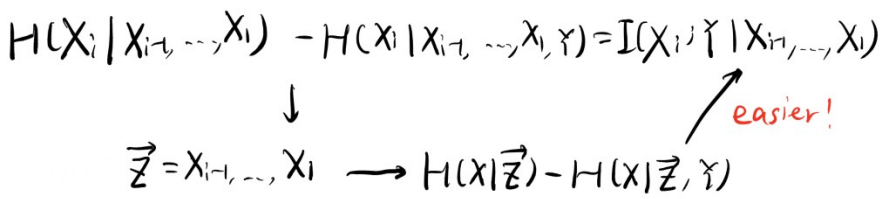

Figure 3: An example of compressing a list of random variables into a random vector.

In this part, maybe the most difficult thing to remember is the chain rule. In fact, we can relate it with the conditional probability.

$$p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i|x_1, \cdots, x_{i-1}) \iff H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1)$$

Acknowledgment

The contents are mainly based on the course materials of CS258, 2020 Spring, Shanghai Jiao Tong University and *Elements of Information Theory* by Thomas M. Cover and Joy A. Thomas.