Information Theory

Jayadev Naram

Contents

1 Entropy

Definition 1.1. The ***uncertainity or entropy*** of a discrete random variable U that takes values in the set $\mathcal{U}$ (also called alphabet $\mathcal{U}$) is defined as

$$H(U) = -\sum_{u\in\mathcal{U}} P_U(u) \log_b P_U(u),$$

where $P_U(\cdot)$ denotes the probability mass function of the random variable U.

Remark 1.2. It should be noted that when $P_U(u) = 0$, the corresponding term does not contribute to entropy because $\lim_{t\downarrow 0} t \log_b t = 0$. In view of this result, one can equivalently define entropy on the support of P_U which is defined as

$$\mathrm{supp}(P_U) = \{u \ : \ P_U(u) > 0\} \subseteq \mathcal{U}.$$

Remark 1.3. Entropy does not depend on different possible values that U can take on, but only on the probabilities of these values.

Definition 1.4. If U is binary with two possible values u_1 and u_2, such that $\mathbb{P}[U = u_1] = p$ and $\mathbb{P}[U = u_2] = 1 - p$, then

$$H(U) = H_b(p) = -p\log_2 p - (1-p)\log_2(1-p), p \in [0, 1],$$

where $H_b(\cdot)$ is called the ***binary entropy function***.

Lemma 1.5 (IT Inequality)**.** *For any base $b > 1$ and any $\xi > 0$,*

$$\left(1 - \frac{1}{\xi}\right) \log_b e \leq \log_b \xi \leq (\xi - 1)\log_b e,$$

with equalities on both sides hold iff $\xi = 1$.

Theorem 1.6. *If U has r possible values, then*

$$0 \le H(U) \le \log r,$$

where

$$\begin{aligned} H(U) = 0 \quad &\iff \exists\, u \in \mathcal{U}, P_U(u) = 1, \\ H(U) = \log r &\iff \forall\, u \in \mathcal{U}, P_U(u) = \frac{1}{r}. \end{aligned}$$

Definition 1.7. The ***conditional entropy*** of the random variable X given the event $Y = y$ is defined as

$$H(X|Y = y) = -\sum_{x \in \mathcal{X}} P_{X|Y}(x|y) \log P_{X|Y}(x|y) = -\mathbb{E}\Big[\log P_{X|Y}(X|Y)\Big| Y = y\Big],$$

where the conditional probability distribution is given by

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)}.$$

Corollary 1.8. *If X has r possible values, then*

$$0 \le H(X|Y = y) \le \log r,$$

where

$$\begin{aligned} H(X|Y = y) = 0 \quad &\iff \exists\, x \in \mathcal{X}, P_{X|Y}(x|y) = 1, \\ H(X|Y = y) = \log r &\iff \forall\, x \in \mathcal{X}, P_{X|Y}(x|y) = \frac{1}{r}. \end{aligned}$$

Definition 1.9. The ***conditional entropy*** of the random variable X given the random variable Y is defined as

$$\begin{aligned} H(X|Y) &= \sum_{y \in \mathcal{Y}} P_Y(y) H(X|Y = y) \\ &= \mathbb{E}_Y\big[H(X|Y = y)\big] \\ &= -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x,y) \log P_{X|Y}(x|y) \\ &= -\mathbb{E}\Big[\log P_{X|Y}(X|Y)\Big]. \end{aligned}$$

Corollary 1.10. *If X has r possible values, then*

$$0 \le H(X|Y) \le \log r,$$

where

$$\begin{aligned} H(X|Y) = 0 \quad &\iff \exists\, x \in \mathcal{X}, \forall\, y \in \mathcal{Y}, P_{X|Y}(x|y) = 1, \\ H(X|Y) = \log r &\iff \forall\, x \in \mathcal{X}, \forall\, y \in \mathcal{Y}, P_{X|Y}(x|y) = \frac{1}{r}. \end{aligned}$$

Remark 1.11. Generally, $H(X|Y) \neq H(Y|X)$.

Theorem 1.12 (Conditioning Reduced Uncertainity)**.** *For any two discrete random variables X and Y,*

$$H(X|Y) \le H(X),$$

where equality holds iff X and Y are independent, i.e., $X \perp\!\!\!\perp Y$.

Remark 1.13. The conditioning reduces entropy-rule only applies to random variables, but not to events. In particular,

$$H(X|Y=y) \lesseqgtr H(X).$$

To understand why this is the case, consider the following example.

Theorem 1.14 (Chain Rule). *Let $X_1, \ldots, X_n$ be n discrete random variables. Then*

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1, \ldots, X_{n-1}) = \sum_{k=1}^{n} H(X_k|X^{(k-1)}),$$

where $X^{(k-1)} = X_{1_k-1}$.

Definition 1.15. The ***mutual information*** between the random variables X and Y is given by

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

Remark 1.16. Notice that

$$\begin{aligned} & H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \\ \Longrightarrow \quad & H(X) - H(X|Y) = H(Y) - H(Y|X) \\ \Longrightarrow \quad & I(X;Y) = I(Y;X). \end{aligned}$$

Remark 1.17. When $X \perp\!\!\!\perp Y$, we have $I(X;Y) = 0$. And also that $I(X;X) = H(X)$.

Remark 1.18. From the chain rule it follows that

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

and thus we obtain

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

Remark 1.19. The mutual information can be expressed as follows.

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \mathbb{E}\big[-\log P_X(X)\big] - \mathbb{E}\big[P_{X|Y}(X|Y)\big] \\ &= \mathbb{E}\left[\log \frac{P_{X|Y}(X|Y)}{P_X(X)}\right] \\ &= \mathbb{E}\left[\log \frac{P_{X,Y}(X,Y)}{P_X(X)P_Y(Y)}\right] \\ &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}. \end{aligned}$$

Theorem 1.20. *Let X and Y be two random variables. Then*

$$0 \leq I(X;Y) \leq \min\{H(X), H(Y)\}.$$

where equality holds on the left-hand side iff $P_{X,Y} = P_X P_Y$, i.e., iff $X \perp\!\!\!\perp Y$, and equality holds on the right-hand side iff X determines Y or vice versa.

Theorem 1.21 (Chain Rule). *Let $X, Y_1, \ldots, Y_n$ be $n+1$ discrete random variables. Then*

$$I(X;Y_1, \ldots, Y_n) = I(X;Y_1) + I(X;Y_2|Y_1) + \ldots + I(X;Y_n|Y_1, \ldots, Y_{n-1}) = \sum_{k=1}^{n} I(X;Y_k|Y^{(k-1)}).$$

Remark 1.22. Comments on Notation of Entropy and Mutual Information.

Theorem 1.23 (Uniqueness of the Definition of Entropy).

2 Relative Entropy

3 Typicality

Definition 3.1. We say that a sequence of random variables $\{X_n\}$ ***converges in probability*** to a random variable X if for all $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\big[|X_n - X| > \varepsilon\big] = 0.$$

Lemma 3.2 (Markov Inequality)**.** *Let X be a nonnegtaive random variable of finite mean $\mathbb{E}[X] < \infty$. Then for all $a > 0$, we have*

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.$$

Lemma 3.3 (Chebyshev Inequality)**.** *Let X be a random variable with finite mean μ and finite variance σ^2. Then for all $\varepsilon > 0$, we have*

$$\mathbb{P}[|X - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{\varepsilon}.$$

Lemma 3.4 (Weak Law of Large Numbers)**.** *Let $\{Z_n\}$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ an variance σ^2. Let*

$$S_n = \frac{1}{n}\sum_{k=1}^{n} Z_k$$

be the sample mean. Then $\{S_n\}$ converges in probability to μ. In particular,

$$\mathbb{P}[|S_n - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Definition 3.5 (Type)**.** Let $x^{(n)}$ be a sequence of n elements drawn from a finite-cardinality alphabet $\mathcal{X}$. The ***empirical probability mass function*** of $x^{(n)}$, also referred to as its ***type***, is defined for $x \in \mathcal{X}$ as

$$\pi(x|x^{(n)}) = \frac{|\{i \in [n] \ : \ x_i = x\}|}{n},$$

where $[n] = \{1, \ldots, n\}$.

Theorem 3.6. *Let $\{X_n\}$ be an i.i.d. sequence of random variables with $X_i \sim P_X(x_i)$. Then $\forall\, x \in \mathcal{X}$ and for all $\varepsilon > 0$, we have*

$$\lim_{n\to 0} \mathbb{P}[|\pi(x|X^{(n)}) - P_X(x)| > \varepsilon] = 0,$$

or in other words, $\{\pi(x|X^{(n)})\}$ converges in probability to $P_X(x)$ for all $x \in \mathcal{X}$.

Definition 3.7 (Typical Set)**.** The ***set of ε-typical n-sequences*** for a random variable $X \sim P_X$ and $\varepsilon \in (0,1)$ (simply typical set) is defined as

$$\mathcal{T}_\varepsilon^{(n)}(X) = \{x^{(n)} \ : \ |\pi(x|x^{(n)}) - P_X(x)| \leq \varepsilon P_X(x), \forall\, x \in \mathcal{X}\}.$$

Remark 3.8. For an element $x \in \mathcal{X}$ which has $P_X(x)$ cannot be a part of typical sequence. Suppose such an x belonged to a sequence $x^{(n)}$, then $\pi(x|x^{(n)}) > 0$. Consequently, we have $|\pi(x|x^{(n)}) - P_X(x)| = \pi(x|x^{(n)}) > 0 = \varepsilon P_X(x)$ for all $\varepsilon > 0$, which shows that $x^{(n)}$ is not a typical sequence.

Lemma 3.9 (Typical Average Lemma)**.** *Consider a typical sequence $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(X)$. Then for any **nonnegtaive** function $g(\cdot)$ on $\mathcal{X}$, we have*

$$(1-\varepsilon)\mathbb{E}[g(X)] \leq \frac{1}{n}\sum_{k=1}^{n} g(x_k) \leq (1+\varepsilon)\mathbb{E}[g(X)].$$

4 Source Coding

5 Joint Typical

Definition 5.1 (Joint Type)**.** Let $(x^{(n)}, y^{(n)})$ be a sequence of a pair of n length sequences from a finite-cardinality alphabet $(\mathcal{X}, \mathcal{Y})$. The ***joint empirical probability mass function*** of $(x^{(n)}, y^{(n)})$, also referred to as its ***joint type***, is defined for $x \in \mathcal{X}$ as

$$\pi(x, y|x^{(n)}, y^{(n)}) = \frac{|\{i \in [n] \ : \ (x_i, y_i) = (x, y)\}|}{n}.$$

Remark 5.2. The X-marginal of X, Y-joint empirical probability mass function is the X-empirical probability mass function.

Definition 5.3 (Jointly Typical Set)**.** The ***set of*** ε***-jointly typical*** n***-sequences*** for a random variable $(X, Y) \sim (P_X, P_Y)$ and $\varepsilon \in (0, 1)$ (simply jointly typical set) is defined as

$$\mathcal{T}_\varepsilon^{(n)}(X, Y) = \{(x^{(n)}, y^{(n)}) \ : \ |\pi(x, y|x^{(n)}, y^{(n)}) - P_{X,Y}(x, y)| \leq \varepsilon P_{X,Y}(x, y), \forall \ x \in \mathcal{X}, y \in \mathcal{Y}\}.$$

Remark 5.4. If $(x^{(n)}, y^{(n)}) \in \mathcal{T}_\varepsilon^{(n)}(X, Y)$, then $x^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(X)$ and $y^{(n)} \in \mathcal{T}_\varepsilon^{(n)}(Y)$.

6 Channel Coding

References