EE290 Mathematics in Data Science Lecture 3 - 01/13/2015

Lecture 3: Entropy, Relative Entropy, and Mutual Information

Lecturer: Jiantao Jiao *Scribe: Alon Devorah, David Hallac, Kevin Shutzberg*

In this lecture[1], we will introduce certain key measures of information, that play crucial roles in theoretical and operational characterizations throughout the course. These include the entropy, the mutual information, and the relative entropy. We will also exhibit some key properties exhibited by these information measures.

1 Notation

A quick summary of the notation

1. **Random Variables (objects):** used more "loosely", i.e. X, Y, U, V
2. **Alphabets:** $\mathcal{X}, \mathcal{Y}, \mathcal{U}, \mathcal{V}$
3. **Specific Values:** x, y, u, v

For discrete random variable (object), U has p.m.f: $P_U^{(u)} \triangleq P(U = u)$. Often, we'll just write $p(u)$. Similarly: $p(x, y)$ for $P_{X,Y}^{(x,y)}$ and $p(y|x)$ for $P_{Y|X}^{(y|x)}$, etc.

2 Entropy

Definition 1. *"Surprise Function":*

$$s(u) \triangleq \log \frac{1}{P_U^{(u)}} \tag{1}$$

Definition 2. ***Entropy****: Let U a discrete R.V. taking values in $\mathcal{U}$. The* ***entropy*** *of U is defined by:*

(2)

Note: The entropy $H(U)$ is not a random variable. In fact it is not a function of the object U, but rather a functional (or property) of the underlying distribution $P_U^{(u)}, u \in \mathcal{U}$. An analogy is $E[U]$, which is also a number (the mean) corresponding to the distribution.

Jensen's Inequality: Let Q denote a *convex* function, and X denote any random variable. Jensen's inequality states that

$$E[Q(X)] \geq Q(E[X]). \tag{3}$$

Further, if Q is strictly convex, equality holds iff X is deterministic.

Example: $Q(x) = e^x$ is a convex function. Therefore, for a random variable X, we have by Jensen's inquality:

$$\mathbb{E}[e^X] \geq e^{\mathbb{E}[X]}$$

Conversely, if Q is a *concave* function, then

$$\mathbb{E}[Q(X)] \leq Q(\mathbb{E}[X]). \tag{4}$$

Example: $Q(x) = \log x$ is a concave function. Therefore, for a random variable $X \geq 0$,

$$\mathbb{E}[\log X] \leq \log \mathbb{E}[X] \tag{5}$$

[1] *Reading:* Chapter 2 of Cover and Thomas.

2.1 Properties of Entropy

W.L.O.G suppose $\mathcal{U} = \{1,2,...,m\}$

1. $H(U) \leq \log m$, with equality iff $P(u) = \frac{1}{m} \forall u$ (i.e. uniform).

   Proof:

$$H(U) = \mathbb{E}[\log \frac{1}{P(U}] \tag{6}$$

$$\leq \log \mathbb{E}[\frac{1}{P(U)}] \text{ (Jensen's inequality, since log is concave)} \tag{7}$$

$$= \log \sum_u P(U) \cdot \frac{1}{P(U)} \tag{8}$$

$$= \log m. \tag{9}$$

   Equality in Jensen, iff $\frac{1}{P(U)}$ is deterministic, iff $p(u) = \frac{1}{m}$

2. $H(U) \geq 0$, with equality iff U is deterministic.

   Proof:

$$H(U) = \mathbb{E}[\log \frac{1}{P(U)}] \geq 0 \text{ since } \log \frac{1}{P(U)} \geq 0 \tag{10}$$

   The equality occurs iff $\log \frac{1}{P(U)} = 0$ with probability 1, iff P(U) = 1 w.p. 1 iff U is deterministic.

3. For a PMF q, defined on the same alphabet as p, define

$$H_q(U) \triangleq \sum_{u \in \mathcal{U}} p(u) \log \frac{1}{q(u)}. \tag{11}$$

   Note that this is the expected surprise function, but instead of the surprise associated with p, it is the surprise associated U, which is distributed according to PMF p, but incorrectly assumed to be having the PMF of q. The following result stipulates, that we will (on average) be more surprised if we had the wrong distribution in mind. This makes intuitive sense! Mathematically,

$$H(U) \leq H_q(U), \tag{12}$$

   with equality iff $q = p$.

   Proof:

$$H(U) - H_q(U) = \mathbb{E}\left[\log \frac{1}{p(u)}\right] - \mathbb{E}\left[\log \frac{1}{q(u)}\right] \tag{13}$$

$$H(U) - H_q(U) = \mathbb{E}\left[\log \frac{q(u)}{p(u)}\right] \tag{14}$$

By Jensen's, we know that $\mathbb{E}\left[\log \frac{q(u)}{p(u)}\right] \leq \log \mathbb{E}\left[\frac{q(u)}{p(u)}\right]$, so

$$\begin{aligned} H(U) - H_q(U) &\leq \log \mathbb{E}\left[\frac{q(u)}{p(u)}\right] && (15)\\ &= \log \sum_{u \in \mathcal{U}} p(u) \frac{q(u)}{p(u)} && (16)\\ &= \log \sum_{u \in \mathcal{U}} q(u) && (17)\\ &= \log 1 && (18)\\ &= 0 && (19) \end{aligned}$$

Therefore, we see that

$$H(U) - H_q(U) \leq 0.$$

Equality only holds when Jensen's yields equality. That only happens when $\frac{q(u)}{p(u)}$ is deterministic, which only occurs when $q = p$, i.e. the distributions are identical.

Definition 3. ***Relative Entropy.*** *An important measure of distance between probability measures is relative entropy, or the Kullback–Leibler divergence:*

$$D(p||q) \triangleq \sum_{u \in \mathcal{U}} p(u) \log \frac{p(u)}{q(u)} = \mathbb{E}\left[\log \frac{p(u)}{q(u)}\right] \tag{20}$$

Note that property 3 is equivalent to saying that the relative entropy is always greater than or equal to 0, with equality iff $q = p$ (convince yourself).

4. If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) \tag{21}$$

Proof:

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= \mathbb{E}\left[\log \frac{1}{p(x_1, x_2, \ldots, x_n)}\right] && (22)\\ &= \mathbb{E}\left[-\log p(x_1, x_2, \ldots, x_n)\right] && (23)\\ &= \mathbb{E}\left[-\log p(x_1)p(x_2)\ldots p(x_n)\right] && (24)\\ &= \mathbb{E}\left[-\sum_{i=1}^{n} \log p(x_i)\right] && (25)\\ &= \sum_{i=1}^{n} \mathbb{E}\left[-\log p(x_i)\right] && (26)\\ &= \sum_{i=1}^{n} H(X_i). && (27) \end{aligned}$$

Therefore, the entropy of independent random variables is the sum of the individual entropies. This is also intuitive, since the uncertainty (or surprise) associated with each random variable is independent.

Definition 4. ***Conditional Entropy of*** X ***given*** Y

$$H(X|Y) \triangleq \mathbb{E}\big[\log \frac{1}{P(X|Y)}\big] \tag{28}$$

$$= \sum_{x,y} P(x,y) \frac{1}{\log P(x|y)} \tag{29}$$

$$= \sum_{y} P(y) \big[\sum_{x} P(x|y) \frac{1}{\log P(x|y)}\big] \tag{30}$$

$$= \sum_{y} P(y) H(X|y). \tag{31}$$

Note: The conditional entropy is a functional of the joint distribution of (X, Y). Note that this is also a number, and denotes the "average" surprise in X when we observe Y. Here, by definition, we also average over the realizations of Y. Note that the conditional entropy is NOT a function of the random variable Y. In this sense, it is very different from a familar object in probability, the conditional expectation $E[X|Y]$ which is a random variable (and a function of Y).

5. $H(X|Y) \leq H(X)$, equal iff $X \perp Y$

   Proof:

$$H(X) - H(X|Y) = \mathbb{E}\big[\log \frac{1}{P(X)}\big] - \mathbb{E}\big[\log \frac{1}{P(X|Y)}\big] \tag{32}$$

$$= \mathbb{E}\big[\log \frac{P(X|Y)}{P(X)} \frac{P(Y)}{P(Y)}\big] = \mathbb{E}\big[\log \frac{P(X,Y)}{P(X)P(Y)}\big] \tag{33}$$

$$= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{34}$$

$$= D(P_{x,y} || P_x \times P_y) \tag{35}$$

$$\geq 0 \qquad \text{equal iff } X \perp Y. \tag{36}$$

The last step follows from the non-negativity of relative entropy. Equality holds iff $P_{x,y} \equiv P_x \times P_y$, i.e. X and Y are independent.

Definition 5. ***Joint Entropy of*** X ***and*** Y

$$H(X,Y) \triangleq \mathbb{E}\big[\log \frac{1}{P(X,Y)}\big] \tag{37}$$

$$= \mathbb{E}\big[\log \frac{1}{P(X)P(Y|X)}\big] \tag{38}$$

6. Chain rule for entropy:

$$H(X,Y) = H(X) + H(Y|X) \tag{39}$$

$$= H(Y) + H(X|Y) \tag{40}$$

7. Sub-additivity of entropy

$$H(X,Y) \leq H(X) + H(Y), \tag{41}$$

with equality iff $X \perp Y$ (follows from the property that conditioning does not increase entropy)

Definition 6. ***Mutual information between*** X ***and*** Y

We now define the mutual information between random variables X and Y distributed according to the joint PMF $P(x, y)$:

$$\begin{aligned} I(X,Y) &\triangleq H(X) + H(Y) - H(X,Y) && (42)\\ &= H(Y) - H(Y|X) && (43)\\ &= H(X) - H(X|Y) && (44)\\ &= D(P_{x,y}||P_x \times P_y) && (45) \end{aligned}$$

The mutual information is a canonical measure of the information conveyed by one random variable about another. The definition tells us that it is the reduction in average surprise, upon observing a correlated random variable. The mutual information is again a functional of the joint distribution of the pair (X, Y). It can also be viewed as the relative entropy between the joint distribution, and the product of the marginals.