EE376A/STATS376A Information Theory Lecture 3 - 01/16/2018

Lecture 3: Entropy, Relative Entropy, and Mutual Information

Lecturer: Tsachy Weissman *Scribe: Yicheng An, Melody Guan, Jacob Rebec, John Sholar*

In this lecture, 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. **Discrete Random Variable:** U
2. **Alphabet:** $\mathcal{U} = \{u_1, u_2, ..., u_M\}$ (An alphabet of size M)
3. **Specific Value:** u, u_1, etc.

For discrete random variables, we will write (interchangeably) $P(U = u)$, $P_U(u)$ or most often just, $p(u)$ Similarly, for a pair of random variables X, Y we write $P(X = x \mid Y = y)$, $P_{X|Y}(x \mid y)$ or $p(x \mid y)$

2 Entropy

Definition 1. *"Surprise" Function:*

$$S(u) \triangleq \log \frac{1}{p(u)} \tag{1}$$

A lower probability of u translates to a greater "surprise" that it occurs.
Note here that we use log to mean $\log_2$ by default, rather than the natural log ln, as is typical in some other contexts. This is true throughout these notes: log is assumed to be $\log_2$ unless otherwise indicated.

Definition 2. ***Entropy****: Let U a discrete random variable taking values in alphabet $\mathcal{U}$. The* ***entropy*** *of U is given by:*

$$H(U) \triangleq \mathbb{E}[S(U)] = \mathbb{E}\left[\log\left(\frac{1}{p(U)}\right)\right] = \mathbb{E}\big[-\log\left(p(U)\right)\big] = -\sum_u p(u)\log p(u) \tag{2}$$

Where U represents all u values possible to the variable.

The entropy is a property of the underlying distribution $P_U(u), u \in \mathcal{U}$ that measures the amount of randomness or surprise in the random variable.

Lemma 3. ***Jensen's Inequality:*** *Let Q denote a function on a random variable X. Jensen's inequality states that $\forall Q$ that are convex:*

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

Further, if Q is strictly convex, equality holds iff X is deterministic.
Conversely, if Q is a concave *function, then*

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

Proof:

If Q is a convex function then its graph $\{(X, Q(X)) : X \in \mathbb{R}\}$ can be seen as an upper bound on the set of affine functions that lie below it. Written more concretely,

$$Q(X) = \sup_{L \in \mathcal{L}} L(X)$$

where

$$\mathcal{L} = \{L : L(u) = au + b \leq Q(u) \text{ for all } -\infty < u < \infty\}$$

Thus, by linearity:

$$\begin{aligned}
\mathbb{E}[Q(X)] &= \mathbb{E}[\sup_{L \in \mathcal{L}} L(x)] && (5)\\
&\geq \sup_{L \in \mathcal{L}} \mathbb{E}[L(X)] && (6)\\
&= \sup_{L \in \mathcal{L}} L(\mathbb{E}[X]) && (7)\\
&= Q(\mathbb{E}[X]) && (8)
\end{aligned}$$

(6) holds because of the monotonicity of $\mathbb{E}$.
The same argument for concave functions can be done using the infimum instead of the supremum.

2.1 Properties of Entropy

Suppose $\mathcal{U} = \{u_1, u_2, ..., u_M\}$

1. $H(U) \leq \log M$, with equality iff U is uniformly distributed i.e. $p(u) = \frac{1}{M}\ \forall u$

   Proof:

$$\begin{aligned}
H(U) &= \mathbb{E}\left[\log \frac{1}{p(U)}\right] && (9)\\
&\leq \log \mathbb{E}\left[\frac{1}{p(U)}\right] \text{ (By Jensen's inequality because log is concave)} && (10)\\
&= \log \sum_u p(u) \cdot \frac{1}{p(u)} && (11)\\
&= \log M. && (12)
\end{aligned}$$

   Equality by Jensen's inequality iff $\frac{1}{p(U)}$ is deterministic, iff $p(u) = \frac{1}{M}\ \forall u \in U$

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

   Proof:

$$H(U) = \mathbb{E}\left[\log \frac{1}{p(U)}\right] \geq 0 \text{ because } \log \frac{1}{p(U)} \geq 0 \quad (13)$$

   The equality occurs iff $\log \frac{1}{p(u)} = 0$ with probability 1 so U must be deterministic.

3. For a PMF q define

$$H_q(U) \triangleq \mathbb{E}\left[\log \frac{1}{q(U)}\right] = \sum_{u \in \mathcal{U}} p(u) \log \frac{1}{q(u)}. \tag{14}$$

Then:

$$H(U) \leq H_q(U), \tag{15}$$

with equality iff $q = p$.

Proof:

$$H(U) - H_q(U) = \mathbb{E}\left[\log \frac{1}{p(u)}\right] - \mathbb{E}\left[\log \frac{1}{q(u)}\right] \tag{16}$$

$$H(U) - H_q(U) = \mathbb{E}\left[\log \frac{q(u)}{p(u)}\right] \tag{17}$$

$$\leq \log \mathrm{E}\left[\frac{q(u)}{p(u)}\right] \tag{18}$$

$$= \log \sum_{u \in \mathcal{U}} p(u) \frac{q(u)}{p(u)} \tag{19}$$

$$= \log \sum_{u \in \mathcal{U}} q(u) \tag{20}$$

$$= \log 1 \tag{21}$$

$$= 0 \tag{22}$$

Thus,

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

Equality only holds when $\frac{q(u)}{p(u)}$ is deterministic, which occurs when $q = p$ (distributions are identical).

Definition 4. ***Relative Entropy***[1] *An measure of distance between probability distributions is relative entropy:*

$$D(p \parallel q) \triangleq \sum_{u \in \mathcal{U}} p(u) \log \frac{p(u)}{q(u)} = \mathbb{E}\left[\log \frac{p(u)}{q(u)}\right] \tag{23}$$

Note that by property 3, the relative entropy is always greater than or equal to 0, with equality iff $q = p$. For now, relative entropy can be thought of as a measure of discrepancy between two probability distributions. We will soon see that it is central to information theory.

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

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) \tag{24}$$

Note: $H(X_1, X_2, \ldots, X_n)$ is called the **joint entropy** of $X_1, X_2, \ldots, X_n$.

[1]Some students may be familiar with relative entropy as Kullback-Leibler (KL) divergence

Proof:

$$H(X_1, X_2, \ldots, X_n) = \mathbb{E}\left[-\log p(x_1, x_2, \ldots, x_n)\right] \tag{25}$$
$$= \mathbb{E}\left[-\log \prod_{i=1}^{n} p(x_i)\right] \tag{26}$$
$$= \mathbb{E}\left[-\sum_{i=1}^{n} \log p(x_i)\right] \tag{27}$$
$$= \sum_{i=1}^{n} \mathbb{E}\left[-\log p(x_i)\right] \tag{28}$$
$$= \sum_{i=1}^{n} H(X_i). \tag{29}$$

5.

Definition 5. ***Conditional Entropy of*** X ***given*** Y

$$H(X \mid Y) \triangleq \mathbb{E}\left[\log \frac{1}{p(X \mid Y)}\right] \tag{30}$$
$$= \sum_{x,y} p(x,y) \log \frac{1}{p(x \mid y)} \tag{31}$$
$$= \sum_{y} p(y) \left[\sum_{x} p(x \mid y) \log \frac{1}{p(x \mid y)}\right] \tag{32}$$
$$= \sum_{y} p(y) H(X \mid Y = y). \tag{33}$$

$H(X \mid Y) \leq H(X)$ with equality iff X and Y are independent.

Proof:

$$H(X) - H(X \mid Y) = \mathbb{E}\left[\log \frac{1}{p(X)}\right] - \mathbb{E}\left[\log \frac{1}{p(X|Y)}\right] \tag{34}$$
$$= \mathbb{E}\left[\log \frac{p(X \mid Y)}{p(X)} \frac{p(Y)}{p(Y)}\right] \tag{35}$$
$$= \mathbb{E}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] \tag{36}$$
$$= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{37}$$
$$= D(P_{x,y} \parallel P_x \times P_y) \tag{38}$$
$$\geq 0 \tag{39}$$

$D(P_{x,y} \parallel P_x \times P_y) \geq 0$ because relative entropy can never be negative. Equality holds iff $P_{x,y} \equiv P_x \times P_y$, ($X$ and Y are independent).

6. Chain Rule:

$$H(X,Y) \triangleq \mathbb{E}\big[\log \frac{1}{P(X,Y)}\big] \tag{40}$$
$$= \mathbb{E}\big[\log \frac{1}{P(Y)P(X \mid Y)}\big] \tag{41}$$
$$= H(Y) + H(X \mid Y) \tag{42}$$

We can take this one step further with (5):

$$H(X,Y) = H(Y) + H(X \mid Y) \leq H(X) + H(Y), \tag{43}$$

with equality holding iff X and Y are independent.

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 $P(x,y)$:

$$I(X;Y) \triangleq D(P_{x,y} \parallel P_x \times P_y) \tag{44}$$
$$= H(X) - H(X|Y) \tag{45}$$
$$= H(X) + H(Y) - H(X,Y) \tag{46}$$

(May find any of these in the literature) The mutual information tells how helpful one variable is at reducing uncertainty in the other.
Note: while relative entropy **is not** symmetric, mutual information **is**.

3 Exercises

1. "Data processing decreases entropy" (note that this statement only applies to deterministic functions)
$Y = f(X) \Rightarrow H(Y) \leq H(X)$ with equality when f is one-to-one.
Note: Proof is part of homework 1.

2. "Data processing on side information increases entropy"
$Y = f(X) \Rightarrow H(Z|X) \leq H(Z|Y)$
True more generally:
whenever $Y - X - Z$ (Markov Relation), i.e., $p(Z|X,Y) = p(Z|X)$, then $H(Z|X) \leq H(Z|Y)$
Note: Proof is part of homework 1.

3.

Definition 7. ***Conditional mutual information***

$$I(X;Y|Z) \triangleq H(X|Z) - H(X|Y,Z) \tag{47}$$

Show that: $I(X;Y_1,Y_2) = I(X;Y_1) + I(X;Y_2|Y_1)$
Proof:

$$I(X;Y_1,Y_2) = H(X) - H(X|Y_1,Y_2) \tag{48}$$
$$= H(X) - H(X|Y_1,Y_2) - H(X|Y_1) + H(X|Y_1) \tag{49}$$
$$= [H(X) - H(X|Y_1)] + [H(X|Y_1) - H(X|Y_1,Y_2)] \tag{50}$$
$$= I(X;Y_1) + I(X;Y_2|Y_1) \tag{51}$$