Lecture Notes 8: Information Measure Entropy

Professor: Zhihua Zhang *Scribe:*

8 Information Measure and Entropy

8.1 Discrete Cases

Definition 8.1 *Given discrete random variable X, the entropy $\mathbb{H}(X)$ of X is defined by* $\mathbb{H}(X) = -\sum\limits_{x\in\mathcal{X}} p(x)\log p(x)$

- $\log e = 1$,
- $0\log 0 = \lim\limits_{a\to 0^+} a\log a = 0$.

Lemma 8.1 *For any discrete random variable X, $\mathbb{H}(X) \geq 0$.*

Proof : Since $0 \leq p(x) \leq 1$, we have $p(x)\log p(x) \leq 0$. So $\mathbb{H}(x) \geq 0$ holds.

Example 8.1 *Given the random variable X with p.m.f that* $p(x) = \begin{cases} 1 & with \quad probability \quad p \\ 0 & with \quad probability \quad 1-p \end{cases}$ *Then,* $\mathbb{H}(x) = -p\log p - (1-p)\log(1-p)$.

8.2 Joint Entropy and Conditional Entropy

Definition 8.2 *The entropy $\mathbb{H}(X,Y)$ of (X,Y) is defined by*

$$\mathbb{H}(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) = -\mathbb{E}\left[\log p(x,y)\right].$$

Definition 8.3 *If $(X,Y) \sim p(x,y)$, then the conditional entropy*

$$\begin{aligned}
\mathbb{H}(Y|X) &= \sum_{x\in\mathcal{X}} p(x)\mathbb{H}(Y|X=x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\mathbb{E}\left[\log p(y|x)\right]
\end{aligned}$$

Theorem 8.1 (The Chain Rule) $\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y|X)$

Proof: Using $\log p(x,y) = \log p(x) + \log p(y|x)$, we compute the expectation in both sides about (X,Y).

Corollary 8.1 $\mathbb{H}(X,Y|Z) = \mathbb{H}(X|Z) + \mathbb{H}(Y|X,Z)$.

8.3 Relative Entropy and Mutual Information

Definition 8.4 *The relative entropy or KullbackLeibler Divergence(KLD) between p.m.f p(x) and q(x) is defined as follows:*

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right]$$

.

- $0 \log \frac{0}{q} = 0$,
- $a \log \frac{a}{0} = \infty$,
- $0 \log \frac{0}{0} = 0$.

Definition 8.5 *Given two variable X and Y with p.m.f $p(x, y)$ and the marginal p.m.f are $p(x)$ and $p(y)$. The mutual information $\mathbb{I}(X, Y)$ is*

$$\begin{aligned} \mathbb{I}(X, Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \log \frac{p(x, y)}{p(x)p(y)} p(x, y) \\ &= D(p(x, y)||p(x)p(y)) \end{aligned}$$

Generally the KullbackLeibler Divergence is not symmetric. But we can build $D'(p||q) = \frac{1}{2}D(p||q) + \frac{1}{2}D(q||p)$ to make the KLD symmetric.

Example 8.2 *Let $\mathcal{X} = \{0, 1\}$, $p(x)$ and $q(x)$ are p.m.f. let $p(X = 0) = 1-r$, $p(X = 1) = r$, $q(X = 0) = 1 - s$, $q(X = 1) = s$, Then*

$$\begin{aligned} D(p||q) &= (1 - r) \log \frac{1 - r}{1 - s} + r \log \frac{r}{s} \\ D(q||p) &= (1 - s) \log \frac{1 - s}{1 - r} + s \log \frac{s}{r} \end{aligned}$$

Theorem 8.2 (Mutual Information and Entropy)

$$\begin{aligned} \mathbb{I}(X, Y) &= \mathbb{I}(Y, X) \\ \mathbb{I}(X, X) &= \mathbb{H}(X) \\ \mathbb{I}(X, Y) &= \mathbb{H}(X) - \mathbb{H}(X|Y) \\ &= \mathbb{H}(Y) - \mathbb{H}(Y|X) \\ &= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X, Y) \end{aligned}$$

Definition 8.6 *The conditional mutual information of random variable X and Y given Z is*

$$\mathbb{I}(X, Y|Z) = \mathbb{H}(X|Z) - \mathbb{H}(X|Y, Z)$$

Definition 8.7 *The conditional relative entropy $D(p(y|x)||q(y|x))$ is*

$$D(p(y|x)||q(y|x)) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{p(y|x)}{q(y|x)}$$

.

Theorem 8.3 *Let $p(x)$ and $q(x)$ with $x \in \mathcal{X}$ be two p.m.f. Then $D(p||q) \geq 0$ with the equality if and only if $p(x) = q(x)$, for all $x \in \mathcal{X}$*

Lemma 8.2 *Let $\sum a_i$ and $\sum b_i$ be convergent sequence of non-negative numbers. Then the following hold:*

- $\sum a_i \log\frac{b_i}{a_i} + \sum(a_i - b_i) \leq 0$ *or* $\sum a_i \log\frac{a_i}{b_i} + \sum(b_i - a_i) \geq 0$.
- *If* $\sum a_i \geq \sum b_i$, *then* $\sum a_i \log\frac{b_i}{a_i} \leq 0$ *with equality iff* $a_i = b_i$.
- *Further more, if* $a_i \leq 1$ *and* $b_i \leq 1$ *for all* i, *then* $2\sum a_i \log\frac{a_i}{b_i} \geq \sum a_i(a_i - b_i)^2$

Proof : Considering the taylor expansion of $\log x$ at $x = 1$, we have $\log x = (x-1) - \frac{(x-1)^2}{2}\frac{1}{\theta^2}$, where θ is between 1 and x. Hence, $\log\frac{b_i}{a_i} = (\frac{b_i}{a_i} - 1) - \frac{1}{2\theta_i^2}(\frac{b_i}{a_i} - 1)^2$, then $a_i \log\frac{b_i}{a_i} = (b_i - a_i) - \frac{a_i^3}{2a_i^2\theta_i^2}(\frac{b_i}{a_i} - 1)^2$. So $\sum a_i \log\frac{b_i}{a_i} = \sum(b_i - a_i) - \sum\frac{a_i^3}{2a_i^2\theta_i^2}(\frac{b_i}{a_i} - 1)^2$. Notice that $\theta_i \in \left[1, \frac{b_i}{a_i}\right]$, we have $a_i\theta_i \in [a_i, b_i]$, hence $\sum\frac{a_i^3}{2a_i^2\theta_i^2}(\frac{b_i}{a_i} - 1)^2 \geq 0$. So $\sum a_i \log\frac{b_i}{a_i} + \sum(a_i - b_i) \leq 0$. And the equality holds when $\frac{a_i}{b_i} = 1$. Further more, $\sum\frac{a_i^3}{2a_i^2\theta_i^2}(\frac{b_i}{a_i} - 1)^2 \leq \sum\frac{a_i}{2}(a_i - b_i)^2$, accordingly we obtain $2\sum a_i \log\frac{a_i}{b_i} \geq \sum a_i(a_i - b_i)^2$.

Lemma 8.3 *Let $\sum a_i$ and $\sum b_i$ be convergent sequences. Then*

$$\sum a_i \log\frac{a_i}{b_i} \geq (\sum a_i)\log\frac{\sum a_i}{\sum b_i}$$

Proof : $\frac{\sum a_i \log\frac{b_i}{a_i}}{\sum a_i} \leq \log\sum\frac{a_i}{\sum a_i}\frac{b_i}{a_i} = \log\frac{\sum b_i}{\sum a_i}$. Both sides multiplies -1, we have $\sum a_i \log\frac{a_i}{b_i} \geq (\sum a_i)\log\frac{\sum a_i}{\sum b_i}$. The equality holds when $\frac{a_i}{b_i}$ are the same for all i, that is $\frac{a_i}{b_i}$ are constant.

Theorem 8.4 $\mathbb{H}(X) \leq \log|\mathcal{X}|$, *where $|\mathcal{X}|$ denotes the number of elements in the range of X with equality iff X has a uniform distribution over $\mathcal{X}$.*

Proof Suppose $p(x)$ and $q(x)$ are p.m.f of random variable X, the KullbackLeibler Divergence(KLD) between p and g are

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \geq 0$$

Alternatively, $-\sum_{x\in\mathcal{X}} p(x)\log p(x) + \sum_{x\in\mathcal{X}} p(x)\log q(x) \leq 0$, that is $\mathbb{H}(X) \leq -\sum_{x\in\mathcal{X}} p(x)\log q(x)$. Let $q(x) = \frac{1}{|\mathcal{X}|}$, we have $\mathbb{H}(X) \leq \log|\mathcal{X}|$, which complete the proof.

Theorem 8.5 (Condition Reduces Entropy)

$$\mathbb{H}(X|Y) \leq \mathbb{H}(X)$$

with the equality iff X and Y are independent.

Definition 8.8 (Differential Entropy)

$$\mathbb{H}(X) = -\int_{\mathcal{S}} f(x) \log f(x) dx$$

, where $\mathcal{S}$ is the support set of random variable X(if f(x) exists) and $f(x)$ is p.d.f of X.

Example 8.3 *Suppose random variable X is uniformly distributed on $(0, a)$. Then $\mathbb{H}(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} dx = \log a$. $\mathbb{H}(X) \leq 0$, when $0 \leq a \leq 1$.*

Similarly, suppose the p.d.f of $X_1, X_2, \cdots, X_n$ is $f(x_1, x_2, \cdots, x_n)$, then

$$\mathbb{H}(x_1, x_2, \cdots, x_n) = -\int f(x_1, x_2, \cdots, x_n) \log f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n$$

.

$$\mathbb{H}(X|Y) = -\int f(x, y) \log f(x|y) dx dy$$

Example 8.4 *Let $X = (X_1, \cdots, X_n)$ is gaussian distribution, that is, $X \sim N(\mu, \Sigma)$.*

$$\begin{aligned}
\mathbb{H}(X) &= -\int f_X(x)(-\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(X-\mu)^{\mathbf{T}}\Sigma^{-1}(X-\mu))dX \\
&= \frac{1}{2}(n \log 2\pi + \log|\Sigma| + \int f_X(x)\mathrm{tr}((X-\mu)^{\mathbf{T}}\Sigma^{-1}(X-\mu))dX) \\
&= \frac{1}{2}(n \log 2\pi + \log|\Sigma| + \int f_X(x)\mathrm{tr}(\Sigma^{-1}(X-\mu)(X-\mu)^{\mathbf{T}})dX) \\
&= \frac{1}{2}(n \log 2\pi + \log|\Sigma| + \mathrm{tr}(\Sigma^{-1}\int f_X(x)(X-\mu)(X-\mu)^{\mathbf{T}}dX)) \\
&= \frac{1}{2}(n \log 2\pi + \log|\Sigma| + \mathrm{tr}(\Sigma^{-1}\Sigma)) \\
&= \frac{1}{2}(n \log 2\pi + \log|\Sigma| + n)
\end{aligned}$$

Definition 8.9 *Suppose $X \sim f(X)$ and $Y \sim g(X)$, then*

$$D(f||g) = \int f(X) \log \frac{f(X)}{g(X)} dX$$

Note that $D(f||g)$ is finite only if the support of f is contained in the support of g.

Definition 8.10 (Mutual Information)

$$\mathbb{I}(X, Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy = D(f(x, y)||f(x)f(y))$$

$$\mathbb{I}(X, Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X)$$

Theorem 8.6

$$D(f||g) \geq 0$$

with the equality iff $f = g$ at almost everywhere.

Proof : $\int f(x)\log\frac{g(x)}{f(x)}dx \leq \log\int f(x)\frac{g(x)}{f(x)}dx = \log\int g(x)dx = 0$, So $\int f(x)\log\frac{f(x)}{g(x)}dx \geq 0$, which complete the proof.

Corollary 8.2

- $\mathbb{I}(X,Y) \geq 0$, *with the equality iff X and Y are independent.*
- $\mathbb{H}(X|Y) \leq \mathbb{H}(X)$, *with the equality iff X and Y are independent.*

Theorem 8.7 *The chain rule for differential entropy*

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

Corollary 8.3

$$\mathbb{H}(X_1,\cdots,X_n) \leq \sum_{i=1}^{n}\mathbb{H}(X_i)$$

Example 8.5 *Suppose $\Sigma \in \mathbf{S}^n_{++}$, where $\Sigma = [\sigma_{ij}]$ then*

$$\det\Sigma \leq \prod_{i=1}^{n}\sigma_{ii}$$

Proof : Suppose $X = (X_1,\cdots,X_n) \sim N(0,\Sigma)$, so $X_i \sim N(0,\sigma_{ii})$. $\mathbb{H}(X_1,\cdots,X_n) = \frac{1}{2}(n\log 2\pi + \log\det\Sigma + n)$, $\mathbb{H}(X_i) = \frac{1}{2}(\log 2\pi + \log\sigma_{ii} + 1)$. Since $\mathbb{H}(X_1,\cdots,X_n) \leq \sum\limits_{i=1}^{n}\mathbb{H}(X_i)$, we have $\frac{1}{2}(n\log 2\pi + \log\det\Sigma + n) \leq \frac{n}{2}\log 2\pi + \frac{1}{2}\sum\limits_{i=1}^{n}\log\sigma_{ii} + \frac{n}{2}$, thus $\log\det\Sigma \leq \sum\limits_{i=1}^{n}\log\sigma_{ii}$. So $\det\Sigma \leq \prod\limits_{i=1}^{n}\sigma_{ii}$ holds.

Theorem 8.8

$$\mathbb{H}(\alpha X + c) = \mathbb{H}(X) + \log|\alpha|$$

, where $\alpha \geq 0$

Proof : Let $Y = \alpha X + c$, then $f_Y(y) = \frac{1}{|\alpha|}f_X(\frac{Y-c}{\alpha})$

$$\begin{aligned}
\mathbb{H}(\alpha X + c) &= -\int f_Y(y)\log f_Y(y)dy \\
&= -\int \frac{1}{|\alpha|}f_X(\frac{Y-c}{\alpha})(\log\frac{1}{|\alpha|} + \log f_X(\frac{Y-c}{\alpha}))dy \\
&= -\int f_X(X)(\log\frac{1}{|\alpha|} + \log f_X(X))dx \\
&= \mathbb{H}(X) + \log|\alpha|
\end{aligned}$$

Corollary 8.4 *Suppose* $\mathbf{A}$ *is nonsingular, then* $\mathbb{H}(\mathbf{A}X) = \mathbb{H}(X) + \log|\mathbf{A}|$.

Theorem 8.9 *Let* $X \in \mathbb{R}^m$ *have zero mean and covariance* $\Sigma = \mathbb{E}[XX^{\mathbf{T}}]$*, then*

$$\mathbb{H}(X) \leq \frac{1}{2}\log((2\pi)^n|\Sigma|) + \frac{n}{2}$$

Proof : Suppose $g(X)$ is p.d.f of X, we also let $f(X) = N \sim (0, \Sigma)$.

$$\begin{aligned} 0 &\leq D(g||f) \\ &= \int g \log \frac{g}{f} \\ &= \int g \log g - \int g \log f \\ &= -\mathbb{H}(X) - \int g \log f \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{H}(X) &\leq -\int g \log f \\ &= -\int g(x)(-\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(X-\mu)^{\mathbf{T}}\Sigma^{-1}(X-\mu))dX \\ &= \frac{1}{2}(n\log 2\pi + \log|\Sigma| + \mathrm{tr}(\Sigma^{-1}\int g(x)(X-\mu)(X-\mu)^{\mathbf{T}}dX)) \\ &= \frac{1}{2}\log((2\pi)^n|\Sigma|) + \frac{n}{2} \end{aligned}$$

8.4 The Exponential Family

Consider the p.d.f $p(x)$ which satisfies the k (independent) constraints,

$$\int_{\mathcal{X}} h_i(x)p(x)dx = m_i < \infty, \quad i = 1, \cdots, k$$

, where $m_1, \cdots, m_k$ are specified constants. We want to find certain p.d.f $p(x)$ that is closest to $f(x)$. That is,

$$\min_p \quad \int p(x)\log\frac{p(x)}{f(x)}dx \qquad \text{s.t.} \quad \int_{\mathcal{X}} h_i(x)p(x)dx = m_i < \infty, \quad i = 1, \cdots, k, and \int p(x)dx = 1.$$

This is an optimization problem with the object function

$$F(p) = \int p(x)\log\frac{p(x)}{f(x)}dx + \sum_{i=1}^{k}\theta_i(\int_{\mathcal{X}} h_i(x)p(x)dx - m_i) + c(\int_{\mathcal{X}} p(x)dx - 1)$$

, where $\theta_i, i = 1, \cdots, k$ *and* c are lagrange multipliers. Besides, $f(x)$ is known.

Theorem 8.10 *The function defined above is minimized by*

$$
\begin{aligned}
p(x) &= E_{f_k}(X|f, g, \vec{h}, \vec{\phi}, \vec{\theta}, \vec{c}) \\
&= \frac{1}{g(\theta)} f(x) \exp(\sum_{i=1}^{k} \theta_i h_i(x)),
\end{aligned}
$$

where $c_i = 1$, *and* $\vec{\phi} = \vec{\theta} = (\theta_1, \cdots, \theta_k)$.

Proof :

$$
\begin{aligned}
dF(p) = \lim_{\alpha \to 0} \int p(x + \alpha\tau(x)) \log \frac{p(x + \alpha\tau(x))}{f(x)} dx - \lim_{\alpha \to 0} \int p(x) \log \frac{p(x)}{f(x)} \\
+ \sum_{i=1}^{k} \theta_i (\int h_i(x)(p(x) + \alpha\tau(x) - p(x))dx) + c(\int p(x) + \alpha\tau(x) - p(x)dx) \\
= \lim_{\alpha \to 0} (\int p(x) \log(1 + \alpha \frac{\tau(x)}{p(x)})dx + \alpha \sum_{i=1}^{k} \int \theta_i h_i(x)\tau(x)dx + \alpha c \int \tau(x)dx)
\end{aligned}
$$

So

$$
\begin{aligned}
\frac{dF(p)}{dp} &= \int p(x) \lim_{\alpha \to 0} \frac{\log(1 + \alpha \frac{\tau(x)}{p(x)})}{\alpha} dx + \int \tau(x) \log \frac{p(x)}{f(x)} dx + \sum_{i=1}^{k} \int \theta_i h_i(x)\tau(x)dx + c \int \tau(x)dx) \\
&= (c + 1)(\int \tau(x)dx) + \int \tau(x) \log \frac{p(x)}{f(x)} dx + \sum_{i=1}^{k} \int \theta_i h_i(x)\tau(x)dx
\end{aligned}
$$

For any small $\tau(x)$, $\frac{dF(p)}{dp} = 0$. Thus

$$
c + 1 + \log \frac{p(x)}{f(x)} + \sum_{i=1}^{k} \theta_i h_i(x) = 0
$$

, which means $p(x) = \frac{1}{g(\theta)} f(x) \exp(\sum_{i=1}^{k} \theta_i h_i(x))$, where $g(\theta) = \int_{x \in \mathcal{X}} f(x) \exp(\sum_{i=1}^{k} \theta_i h_i(x))dx$.