Lecture 2

Class website up now

Handout: HW #1 due next Thur

Entropy

Let \( X \) be a discrete random variable defined on an alphabet \( \mathcal{X} \) drawn according to the pmf \( p(x) \).

The entropy of \( X \) is defined as

\[
H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} = -\mathbb{E} \left[ \log \frac{1}{p(x)} \right]
\]

Notation

\( X \sim p(x) \)

\( p \) is prob measure / law distribution

Convention: \( 0 \log 0 = 0 \)

\[ x \log x \]
Sometimes we write $H(x) = H(p(x))$.

Example: $X \sim \text{Bern}(p)$

$$X = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$$

Then:

$$H(x) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$$

$$= H(p), \quad p \in (0, 1)$$

$H(p)$, $p \in (0, 1)$, is a binary entropy function.

The binary entropy function is concave in $p$.

Convex
Properties of $H(x)$

1. $H(x) > 0$
2. $H(p(x))$ is concave in $p(x)$
3. $H(x) \leq \log |x|

Proof for 3. Jensen's inequality

If $f(x)$ is convex, then

$E[f(x)] \geq f(E[x])$

If $f(x)$ is concave, then

$E[f(x)] \leq f(E[x])$

Convex:

$\frac{f(a) + f(b)}{2} \geq f(\frac{a + b}{2})$

$H(x)$

$E[\log \frac{1}{p(x)}] \leq \log (E[\frac{1}{p(x)}])$

$E(\frac{1}{p(x)}) = \sum_{x} p(x) \cdot \frac{1}{p(x)} = 1$\times 1

$\Rightarrow E[\log (\frac{1}{p(x)})] \leq 1$\times 1
Let \((X, Y)\) be a pair of discrete random variables. Then, the conditional entropy is given:

\[
H(Y|X) = \sum_{x \in \mathcal{X}} H(p(y|x)) \cdot p(x)
\]

\[
= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{1}{p(y|x)}
\]

\[
= E \left[ \log \frac{1}{p(Y|X)} \right]
\]

Can also write \(H(Y|X)\) as \(H(Y|X=x)\):

\[
H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(p(y|x))
\]

\[
\leq H \left( \sum_{x \in \mathcal{X}} p(x) p(y|x) \right)
\]

\[
\Rightarrow H(Y|X) \leq H(Y)
\]

Equality holds true if \(X\) and \(Y\) are independent.
\((x, y) \sim p(x, y)\)

Then their joint entropy is:

\[
H(x, y) = E \left[ \log \frac{1}{p(x, y)} \right] = E \left[ \log \frac{1}{p(x)} \right] + E \left[ \log \frac{1}{p(y|x)} \right] = H(x) + H(y|x) \]

\[\implies H(x, y) = H(x) + H(y|x) = H(y) + H(x|y)\]

Let \(x^n = (x_1, \ldots, x_n)\) (always in ECE 287)

\[
H(x^n) = H(x_1) + H(x_2|x_1) + \ldots + H(x_n|x_1, \ldots, x_{n-1}) = \sum_{i=1}^{n} H(x_i; 1 \ x^{i-1})
\]

\[
H(x^n) \leq \sum_{i=1}^{n} I(x_i; x_i)\]

Equality holds if \((x_1, \ldots, x_n)\) are mutually independent.
Relative entropy

Let $p(x)$ and $q(x)$ be pmfs on $\mathbb{X}$.

Then:

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

$$= \mathbb{E}_p \log \frac{p(x)}{q(x)}$$

$D(p \| q)$ is called relative entropy (or KL divergence).

We say $p(x)$ is absolutely continuous w.r.t. $q(x)$.

If $q(x) = 0 \Rightarrow p(x) = 0$

If $p(x)$ is not absolutely continuous w.r.t. $q(x)$

$$\Rightarrow D(p \| q) = \infty$$

Can think of $D(p \| q)$ as distance between $p(x)$ & $q(x)$ (divergence).
Properties of $D(p||q)$

1. $D(p||q) = 0$ if and only if $p(x) = q(x)$ for all $x \neq x^*$.

2. $D(p||q) \neq D(q||p)$ in general.

3. $D(p||q)$ is convex in $p$ and $q$, i.e., for any $(p_1, q_1)$, $(p_2, q_2)$, $\lambda, \overline{\lambda} = 1 - \lambda$,

   $$\lambda D(p_1||q_1) + \overline{\lambda} D(p_2||q_2) \geq D(\lambda p_1 + \overline{\lambda} p_2 || \lambda q_1 + \overline{\lambda} q_2).$$

4. Chain rule: For any $p(x,y)$, $q(x,y)$,

   $$D(p(x,y) || q(x,y)) = D(p(x)||q(x)) + \sum_{y} p(x) D(p(y|x) || q(y|x)).$$
For any prob. measures $P$ & $Q$

$$D(P \| Q) = \begin{cases} \int \log \frac{dP}{dQ} \, dP & \text{Random-Nikodym derivative} \\ \infty & \text{w.r.t \&} \ Q \text{ w.r.t} \ P \text{ is a.c w.r.t} \ Q \end{cases}$$

Special case, if both $P$ & $Q$ are a.c w.r.t $\sigma$-finite measure $\mu$ (Lebesgue measure) then

$$D(P \| Q) = D(Q \| P) = \int p(x) \log \frac{q(x)}{P(x)} \, d\mu(x)$$

derivatives w.r.t $\mu(x)$

$\uparrow$

$\uparrow$ can define relative entropy for anything.

Let $\mathcal{X}$ be finite & $p(x)$ be the uniform pmf on $\mathcal{X}$. Then:

$$D(p \| q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

$$= \log |\mathcal{X}| - H(x)$$

$$= H(u(x)) - H(p(x))$$
$f$-divergence

Let $f: [0, \infty) \to \mathbb{R}$ & $f(1) = 0$

Then for $p(x)$ & $q(x)$

$$D_f(p \parallel q) = \mathbb{E}_q \left[ f\left( \frac{p(x)}{q(x)} \right) \right]$$

$$= \int f\left( \frac{p(x)}{q(x)} \right) q(x) \, dq(x)$$

If $f(p) = p \log p$

$$D_f(p \parallel q) = \int q(x) \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} \, dq(x)$$

$$= \int p(x) \log \frac{p(x)}{q(x)} \, dq(x) = D(p \parallel q)$$
Mutual information

Let \((x, y) \sim p(x, y)\) be a pair of discrete random variables.

Mutual information between \(x\) & \(y\) is

\[
I(x; y) = D(p(x, y) \mid \mid p(x)p(y))
\]

\[
= \sum p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = E \left[ \log \frac{p(x, y)}{p(x)p(y)} \right]
\]

\[
= \sum_x p(x) D(p(y | x) \mid \mid p(y))
\]

Properties

1. \(I(x; y) \geq 0\)

2. \(I(x; y) = 0\) iff \(x\) & \(y\) are independent

3. As a function of \((p(x), p(y | x))\)

\(I(x; y)\) is concave in \(p(x)\) for a fixed \(p(y | x)\) & convex in \(p(y | x)\) for a fixed \(p(x)\)

4. \(I(x; x) = H(x)\)

5. \(I(x; y) = H(x) + H(y) - H(x, y)\)

\[
= H(x) - x(x | y)
\]

\[
= H(y) - H(y | x)
\]
(5) Variational characterization

\[ I(X; Y) = \min_{p(x)} D\left( p(y|x) \parallel p(y) \right) \leq \min_{p(x)} D\left( p(y|x) \parallel p(y) \right) \]

Proof:

\[ I(X; Y) = \sum_{x} p(x) D\left( p(y|x) \parallel p(y) \right) \]

\[ = \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{p(y|x)}{p(y)} \leq \sum_{x} p(x) D\left( p(y|x) \parallel p(y) \right) \]

\[ \leq \sum_{x} p(x) D\left( p(y|x) \parallel p(y) \right) \]

\[ \leq \sum_{x} p(x) D\left( p(y|x) \parallel p(y) \right) \]

\[ \Rightarrow \text{iff } p(y) = p(y|x) \]

Suppose:

\[ p(y|x) \text{ is fixed and we are interested in: } \max_{p(x)} I(X; Y) \]

Information capacity (Capacity in short)
\[
\max \pm (x' y) = \max_{p(x)} \min_{q(y)} \left( \frac{1}{p(x)} D(p(y|x) \parallel q(y|x)) \right)
\]

\[
(\max_{\alpha} \min_{\beta} f(\alpha, \beta) \leq \min_{\beta} \max_{\alpha} f(\alpha, \beta))
\]

by minimax theorem:

\[
\max_{p(x)} I(x; y) = \min_{q(y)} \max_{p(x)} \left( \frac{1}{p(x)} D(p(y|x) \parallel q(y|x)) \right)
\]

\[
= \min_{q(y)} \max_{p(x)} D(p(y|x) \parallel q(y|x))
\]