ECE 231A Discussion 1

TA: Hengjie Yang

Email: hengjie.yang@ucla.edu

04/03/2020

What is information theory all about?

A mathematical theory established by Claude E. Shannon in 1948 that answers two fundamental questions in communication theory:

1. What is the ultimate data compression? (answer: entropy $H(X)$)
2. What is the ultimate transmission rate of communication? (answer: channel capacity C)

Information theory now has a rich connection with

- communication theory,
- mathematics,
- statistics,
- computer science,
- artificial intelligence,
- economics,
- physics,
- linguistics,
- . . . .

Entropy

Setup: Let X, Y denote discrete random variables (r.v.'s); Let $\mathcal{X}, \mathcal{Y}$ denote their alphabets; $p(x) \triangleq \Pr\{X = x\}, x \in \mathcal{X}$.

Entropy $H(X)$:

$$
\begin{aligned}
H(X) &\triangleq -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\
&= -\mathbb{E}[\log p(X)]
\end{aligned}
$$

Properties of $H(X)$:

$$
0 \leq H(X) \leq \log |\mathcal{X}|
$$

Remarks:

(i) $H(X)$ measures the amount of uncertainty of r.v. X.

(ii) if base is 2, the unit is "bit"; if base is e, unit is "nat".

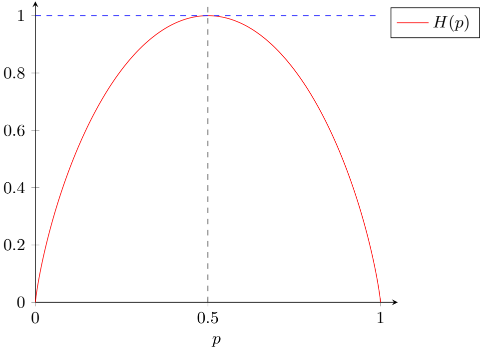
Binary entropy function

Scalar

Let $X \in \{0, 1\}$ and $p \triangleq \Pr\{X = 1\} \in [0, 1]$. The binary entropy function $H(p)$ is given by

$$H(p) \triangleq -p \log p - (1 - p) \log(1 - p).$$

The graph of $H(p)$:



Joint entropy, conditional entropy, chain rule

Joint entropy

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

Conditional entropy

$$H(Y|X) \triangleq \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \tag{1}$$
$$= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \tag{2}$$
$$= \underline{-\mathbb{E}[\log p(Y|X)]} \tag{3}$$

Chain rule

$\log p(X,Y) = \log p(X) + \log(Y|X)$

$$H(X,Y) = H(X) + H(Y|X)$$
$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$

Relative entropy and mutual information

Relative entropy (Kullback-Leibler divergence): for two probability mass functions $p(x)$ and $q(x)$,

$$D(\underbrace{p}\|\underbrace{q}) \triangleq \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$$

distributions

$$=\mathbb{E}_p\left[\log\frac{p(X)}{q(X)}\right]$$

Mutual information $I(X;Y)$

$$I(X;Y) \triangleq D(p(x,y)\|p(x)p(y)) \tag{4}$$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{5}$$

$$=\mathbb{E}_{p(x,y)}\left[\log\frac{p(X,Y)}{p(X)p(Y)}\right] \tag{6}$$

Remarks:

1. $D(p\|q) \neq D(q\|p)$ in general.
2. $I(X;Y)$: the amount of information X contains about Y.

Relation between mutual information and entropy

Theorem

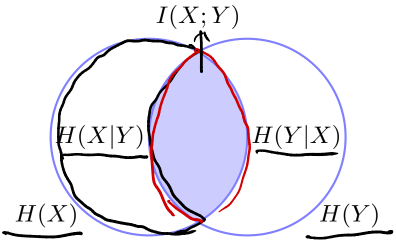
$$\log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}$$

$$\begin{aligned}
\underline{I(X;Y)} &= H(X) - H(X|Y) && (7)\\
&= H(Y) - H(Y|X) && (8)\\
&= I(Y;X) && (9)\\
I(X;X) &= H(X) && (10)\\
&= H(x) - \underbrace{H(x|x)}_{=0}
\end{aligned}$$

$I(X;Y)$: the reduction in the uncertainty of X due to the knowledge of Y.

Venn diagram representation

area = the amount of uncertainty of the corresponding r.v.



More chain rules

Chain rule for entropy

$$\log p(X_1 \cdots X_n) = \sum_{i=1}^{n} \log p(X_i | X_{i-1} \cdots X_1)$$

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) = \log \prod_{i=1}^{n} P(X_i | X_{i-1} \cdots X_1)$$

Chain rule for information $= H(X_1 \cdots X_n) - H(X_1 \cdots X_n | Y)$

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1)$$

Chain rule for relative entropy

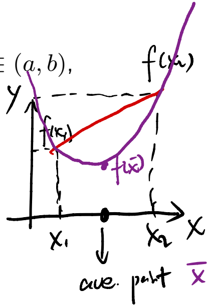
$$D(p(x,y) \| q(x,y)) = D(p(x) \| q(x)) + D(p(y|x) \| q(y|x))$$

Convexity and Jensen's inequality

Convex function: f is said to be convex over (a,b) if $\forall x_1, x_2 \in (a,b)$, $\forall \lambda \in [0,1]$,

"average point"

$$f(\underline{\lambda x_1 + (1-\lambda)x_2}) \leq \underline{\lambda f(x_1) + (1-\lambda) f(x_2)}.$$



f is strictly convex if equality holds only for $\lambda = 0, 1$.

Concave function: f is concave if $-f$ is convex.

Theorem: If $f'' \geq 0$ over (a,b), then f is convex over (a,b).

Jensen's inequality: if f is convex and X is a r.v.,

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Proof: mathematical induction on $k-1$ mass points.

Proof of Jensen's inequality

Base: 2 mass points $\Rightarrow$ def of convex functions ✓.

Assume theorem holds for $k-1$ mass points

want to show that theorem holds for k mass points.

Suppose $X \in \{x_1 \cdots x_k\}$ $\quad P_X = \{p_1 \cdots p_k\}$ $\quad \sum_{i=1}^{k} p_i = 1$

$$\mathbb{E}[f(X)] = \sum_{i=1}^{k} p_i f(x_i) \qquad \Rightarrow \sum_{i=1}^{k-1} p_i = 1 - p_k$$

$$= \sum_{i=1}^{k-1} p_i f(x_i) + p_k f(x_k)$$

$$= (1-p_k) \sum_{i=1}^{k-1} \frac{p_i}{1-p_k} f(x_i) + p_k f(x_k)$$

hypothesis

$$\geq (1-p_k)\, f\left(\sum_{i=1}^{k-1} \frac{p_i x_i}{1-p_k} \right) + p_k f(x_k)$$

Jensen's

"single point"

$$\geq f\left(\cancel{(1-p_k)} \sum_{i=1}^{k-1} \frac{p_i x_i}{\cancel{1-p_k}} + p_k x_k \right) = f\left(\sum_{i=1}^{k} p_i x_i \right) \checkmark$$

Information inequality

Information inequality: for two probability mass functions $p(x)$ and $q(x)$,

$$D(p\|q) \geq 0.$$

Proof: apply Jensen's inequality to $-D(p\|q)$.

Corollary

$$\begin{aligned} I(X;Y) =& D(p(x,y)\|p(x)p(y)) \geq 0 \\ I(X;Y|Z) =& D(p(x,y|z)\|p(x|z)p(y|z)) \geq 0 \\ H(X) \geq& H(X|Y) \quad \text{(conditioning reduces entropy)} \end{aligned}$$

$I(X;Y) = H(X) - H(X|Y) \geq 0$

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

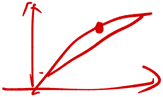
equality holds iff $p = u$.

p: arbitrary distribution

u: uniform distribution $\left(\frac{1}{|\mathcal{X}|}, \dots, \frac{1}{|\mathcal{X}|}\right)$

Proof of info. inequality

Let $A = \{x : p(x) > 0\}$. (support set)



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

$$E[f(x)] = \sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} \qquad (\log t \text{ is concave})$$

$$\overset{\text{Jensen's}}{\le} \log\left(\sum_{x\in A} \cancel{p(x)}\cdot\frac{q(x)}{\cancel{p(x)}}\right)$$

$$= \log\left(\sum_{x\in A} q(x)\right) \qquad A \subset \mathcal{X}$$

$$\le \log\left(\underbrace{\sum_{x\in \mathcal{X}} q(x)}_{=1}\right)$$

$$= 0$$

Log-sum inequality and convexity of relative entropy

Log-sum inequality: For nonnegative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

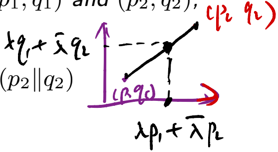
with equality iff $a_i/b_i = c$, where c is some constant. $f(t) = t\log t$

Theorem (Convexity of relative entropy)

$D(p\|q)$ is convex in pair (p, q). Specifically, for two pairs (p_1, q_1) and (p_2, q_2), and $0 \leq \lambda \leq 1$

distribution pair

$$D(\lambda p_1 + \bar{\lambda} p_2 \| \lambda q_1 + \bar{\lambda} q_2) \leq \lambda D(p_1\|q_1) + \bar{\lambda} D(p_2\|q_2)$$



where $\bar{\lambda} = 1 - \lambda$.

Corollary distribution $H(p) = H(X)$

1. $H(p) = \log|\mathcal{X}| - D(p\|u)$ is concave in distribution p.
2. $I(X;Y) = D(p(x)p(y|x)\|p(x)\sum_{x'} p(x')p(y|x'))$ is concave in $p(x)$ for fixed $p(y|x)$, and is convex in $p(y|x)$ for fixed $p(x)$.

$I(X;Y) = D(p(X,Y) \| p(x)\, p(Y))$

$= p(x)p(y|x) \qquad = \sum_{x'} p(x')\,p(y|x')$

Data processing inequality and Markov chain

Markov chain: R.v.'s X, Y, Z are said to form a Markov chain (denoted $X \to Y \to Z$) if

$$p(z|y,x) = p(z|y)$$

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{\underline{p(z|x,y)}\,\underline{p(x,y)}}{p(y)} = \frac{p(z|y)\,\cancel{p(y)}\,p(x|y)}{\cancel{p(y)}} = p(z|y)\,p(x|y)$$

Namely, we have $p(x,z|y) = \frac{p(z|x,y)p(x|y)p(y)}{p(y)} = p(z|y)p(x|y)$.

Data processing inequality: If $X \to Y \to Z$, then

$$I(X;Y) \geq I(X;Z)$$

Proof: by chain rule,

$$\begin{aligned} \underline{I(X;Y,Z)} &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + \underline{I(X;Z|Y)} \end{aligned}$$

and notice that $I(X;Z|Y) = \underbrace{\mathbb{E}\left[\log \frac{p(X,Z|Y)}{p(X|Y)p(Z|Y)}\right]}_{=0} = 0$ since $X \to Y \to Z$.

Exercise

Let $X \sim p(x)$, $x = 1, 2, \ldots, m$. We are given a set $S \subseteq \{1, 2, \ldots, m\}$. We ask whether $X \in S$ and receive the answer

$$Y = \begin{cases} 1, & \text{if } X \in S \\ 0, & \text{otherwise.} \end{cases}$$

Suppose $\Pr\{X \in S\} = \alpha$. Find the decrease in uncertainty $H(X) - H(X|Y)$.

$$\begin{aligned} H(X) - H(X|Y) &= I(X;Y) \\ &= H(Y) - \cancel{H(Y|X)} \\ &= H(\Pr\{X \in S\}) - {}^{=0} \\ &= H(\alpha) \end{aligned}$$