CS258: Information Theory

Fan Cheng

Spring, 2018. chengfan@sjtu.edu.cn

- Entropy
- Mutual Information

Random Variable and Entropy

Support Set

For a random variable X, denote its alphabet by $\mathscr{X}$. The probability distribution of X is $p(x)$. The support set of X is defined as

$$supp(X) := \{x : p(x) > 0, x \in \mathscr{X}\}$$

- $supp(X) \subseteq \mathscr{X}$
- $x \to 0$, $x \log x \to 0$

Entropy

For a random variable X with probability density function $p(x)$, its entropy is defined as

$$H(X) := -\sum_{x \in supp(X)} p(x) \log p(x) = -\mathscr{E} \log p(x)$$

- $H(X) \geq 0$
- $H(X) \leq \log |\mathscr{X}|$

Joint Entropy and Conditional Entropy

Joint Entropy

The joint entropy $H(X,Y)$ of a pair of random variables X and Y is defined by

$$H(X,Y) := -\sum_{x,y} p(x,y) \log p(x,y) = -\mathscr{E} \log p(X,Y)$$

Conditional Entropy

For random variables of X and Y, the conditional entropy of Y given X is defined by

$$H(Y|X) := -\sum_{x,y} p(x,y) \log p(y|x) = -\mathscr{E} \log p(Y|X)$$

$H(X,Y)$, $H(X)$, $H(Y)$

Proposition

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

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

Proof from definitions.

Mutual Information

For random variables X and Y, the mutual information between X and Y is defined by

$$I(X;Y) := \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \mathscr{E} \log \frac{p(X,Y)}{p(X)p(Y)}$$

$I(X;Y) = I(Y;X)$, $I(X;X) = H(X)$

$I(X;Y)$, $H(X)$, $H(Y)$

Proposition

$$I(X;Y) = H(X) - H(X|Y)$$
$$I(X;Y) = H(Y) - H(Y|X)$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

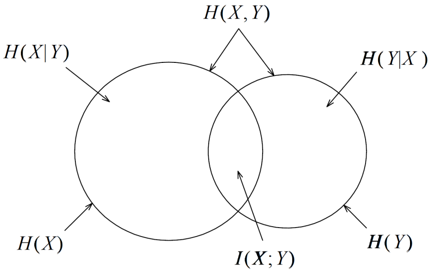

Figure: Relationship between entropies and mutual information for two random variables

Conditional Mutual Information

For random variables X, Y and Z, the mutual information between X and Y conditioning on Z is defined by

$$I(X;Y|Z) := \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = \mathscr{E} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}$$

- $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$
- $I(X;Y|Z) = H(Y|Z) - H(Y|X,Z)$
- $I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$

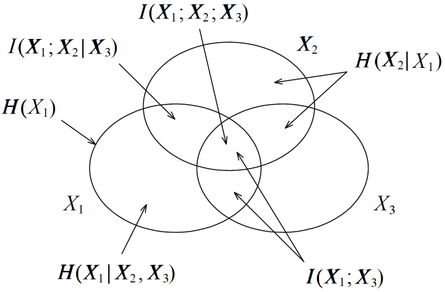

Figure: Information diagram for three random variables

Let X_1 and X_2 be independent binary random variables with

$$P(X_i = 0) = P(X_i = 1) = 0.5,$$

$i = 1,2$. Let

$$X_3 = (X_1 + X_2) \bmod 2.$$

Calculate $I(X_1;X_2;X_3)$ ($I(X_1;X_2;X_3)$ is not an information measure)

Chain rule for entropy

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

Proof by induction.

Chain rule for conditional entropy

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

$p(x_1, x_2, \ldots, x_n) = \prod p(x_i | x_1, \ldots, x_{i-1})$

Chain Rule: Mutual Information

Chain rule for mutual information

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

Proof by induction.

Chain rule for conditional mutual information

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

Information Divergence, Relative Entropy, Kullback-Leibler Distance

$D(p||q)$

The informational divergence between two probability distributions p and q on a common alphabet $\mathscr{X}$ is defined as

$$D(p||q) := \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathscr{E}_p \log \frac{p(X)}{q(X)},$$

where $\mathscr{E}_p$ denotes expectation with respect to p.

- In convention, $p(x) \log \frac{p(x)}{q(x)} = \infty$ if $q(x) = 0$.
- $D(p||q)$ is not symmetric.
- $D(p||q)$ is not a metric. It does not satisfy the triangular inequality.
- $D(p||q) \geq 0$ (Proof via $\ln a \geq 1 - \frac{1}{a}$)

Two Inequalities on $D(p||q)$

Log-sum Inequality

For positive numbers $a_1, a_2, \ldots$ and nonnegative numbers $b_1, b_2, \ldots$ such that $\sum_i a_i < \infty$ and $0 < \sum_i b_i < \infty$,

$$\sum_i a_i \log \frac{a_i}{b_i} \geq (\sum_i a_i) \log \frac{\sum_i a_i}{\sum_i b_i}.$$

Moreover, equality holds if and only if $\frac{a_i}{b_i} = \text{constant}$ for all i.

Let p and q be two probability distributions on a common alphabet $\mathscr{X}$. The variational distance between p and q is defined by

$$d(p,q) := \sum_{x \in \mathscr{X}} |p(x) - q(x)|$$

Pinsker's inequality

$$D(p||q) \geq \frac{1}{2 \ln 2} d^2(p,q).$$

Chain Rule for Relative Entropy

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

Proof.

$$\begin{aligned}
&D(p(x,y)||q(x,y)) \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) log \frac{p(y|x)}{q(y|x)} \\
&= D(p(x)||q(x)) + D(p(y|x)||q(y|x))
\end{aligned}$$

More on $D(p||q)$

Convexity of relative entropy

$D(p||q)$ is convex in the pair (p,q); that is, if (p_1,q_1) and (p_2,q_2) are two pairs of probability mass functions, then

$$D(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$$

for all $0 \leq \lambda \leq 1$.

Checked by log-sum inequality.

Concavity of entropy

$H(p)$ is a concave function of p.

$H(p) = \log|\mathscr{X}| - D(p||u)$, where u is the uniform distribution on $|\mathscr{X}|$ outcomes.

Data Processing Inequality

DPI

If X, Y, Z form a Markov chain $X \to Y \to Z$ (i.e, $p(x,y,z) = p(x)p(y|x)p(z|y)$), then

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

$I(X;Z|Y) = 0$

Corollary

- For any function g, $I(X;Y) \geq I(X;g(Y))$
- If $X \to Y \to Z$, then $I(X;Y|Z) \leq I(X;Y)$.

Basic Inequality

$$I(X;Y|Z) \geq 0$$

$$I(X;Y|Z) = \sum p(z) D(P_{XY|z} || P_{X|z} P_{Y|z})$$

In the information diagram, except $I(X;Y;Z)$, every region is non-negative

$H(X) = 0$ if and only if X is deterministic

$I(X;Y) = 0$ if and only if X and Y are independent

$H(Y|X) = 0$ if and only if Y is a function of X

Equivalent condition

More Information Inequalities

(Conditioning reduces entropy)(Information cant hurt)

$$H(X|Y) \leq H(X)$$

with equality if and only if X and Y are independent.

(Independence bound on entropy)

Let $X_1, X_2, ..., X_n$ be drawn according to $p(x_1, x_2, ..., x_n)$. Then

$$H(X_1, X_2, ..., X_n) \leq \sum_i H(X_i)$$

with equality if and only if the X_i are independent

Chain rule + conditioning

Question

Conditioning reduce mutual information?

$$I(X;Y|Z) \leq I(X;Y)$$

Axiomatic definition of entropy

If a sequence of symmetric function $H_m(p_1, ..., p_m)$ satisfies the following properties:

- Normalization: $H_2(\frac{1}{2}, \frac{1}{2}) = 1$.
- Continuity: $H_2(p, 1-p)$ is continuous function of p.
- Grouping: $H_m(p_1, p_2, ..., p_m) = H_{m-1}(p_1 + p_2, p_3, ..., p_m) + (p_1 + p_2) H_2(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2})$, H_m must be of the form of entropy function.

Rényi Entropy and Shannon Entropy

Renyi Entropy

For a discrete random variable X with probability density function $p(x)$, its Renyi entropy with index α is defined as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_x p^\alpha(x)$$

When $\alpha \to 1$, $H_\alpha(X) \to H(X)$

Differential Entropy

For a continuous random variable $X \sim f(x)$, its differential entropy is defined as

$$h(x) := -\int_x f(x) \log f(x) dx$$

$h(x)$ may be negative

Uniform Distribution

If X is uniformly distributed from 0 to a, then

$$h(X) = \log a$$

Gaussian Distribution

If $X \sim \mathcal{N}(0, \sigma^2)$, then

$$h(X) = \frac{1}{2} \log 2\pi e \sigma^2$$

More on $h(X)$

$$h(X+c) = h(X).$$

Translation does not change the differential entropy.

$$h(aX) = h(X) + \log|a|$$

Checked by definition

For vector-valued random variable X,

$$h(AX) = h(X) + \log|det(A)|$$

Proof is not required

Reading

- Ch. 2 (Yeung), Ch. 2 (Cover)
- Facets of entropy:
  `http://www.inc.cuhk.edu.hk/EII2013/entropy.pdf`