Entropy and Mutual Information

Notes on Information Theory

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

Introduction

- Given two random variables, what can we say about one when we know the other? This is the central problem in information theory.
- Information theory answers two fundamental questions in communication theory:
  1. What is the ultimate lossless data compression?
  2. What is the ultimate transmission rate of reliable communication?
- Information theory is more: it gives insight into the problems of statistical inference, computer science, investments and many other fields.

Entropy

- The most fundamental concept of information theory is the entropy. The entropy of a random variable X is defined by

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)}$$

- The entropy is non-negative. It is zero when the random variable is “certain” to be predicted.

Joint and Conditional Entropy

- For two random variables X and Y, the joint entropy is defined by

$$H(X,Y) = \sum_{x,y} p(x,y) \log \frac{1}{p(x,y)}.$$

- The conditional entropy is defined by

$$H(X|Y) = \sum_{y} p(y) H(X|Y=y)$$

$$= \sum_{x,y} p(x,y) \log \frac{1}{p(x|y)}.$$

Mutual Information

- The mutual information of X and Y is defined by

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

- Note that the mutual information is symmetric in the arguments. That is,

$$I(X;Y) = I(Y;X).$$

- Mutual information is also non-negative, as we will show in a minute.

Mutual Information and Entropy

- It follows from definition of entropy and mutual information that

$$I(X;Y) = H(X) - H(X|Y).$$

- The mutual information is the reduction of entropy of X when Y is known.

Conditional Mutual Information

- The conditional mutual information of X and Y given Z is defined by

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

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

Chain Rules of Entropy

- From the definition of entropy, it can be shown that for two random variables X and Y, the joint entropy is the sum of the entropy of X and the conditional entropy of Y given X,

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

- More generally, for n random variables,

$$H(X_{1:n}) = \sum_{i=1}^{n} H(X_i|X_{1:i-1}).$$

Chain Rules of Mutual Information

- It can be shown from the definitions that the mutual information of (X, Y) and Z is the sum of the mutual information of X and Z and the conditional mutual information of Y and Z given X. That is,

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

- More generally, for n random variables,

$$I(X_{1:n}; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{1:i-1}).$$

Relative Entropy

- The relative entropy of two distributions is defined by

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

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

- The conditional relative entropy of two conditional distributions is defined by

$$D(p(y|x)||q(y|x)) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)}.$$

Relative Entropy

- We have the chain rule for relative entropy,

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

- There are several synnonyms for relative entropy.

Examples

- entropy: Example 2.1.1, 2.1.2
- joint and conditional entropy: Example 2.2.1
- relative entropy: Example 2.3.1

Relationships between H and I

- For random variables X and Y, the mutual information and the relative entropy are related by

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

- The entropy and the mutual information are related by

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

Convexity

- We will derive some inequalities in information theory. We begin by the concept of convexity.
- A set is convex if every line segment between two points in the set is a subset of the set.
- A function $f(x)$ is convex over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$, it is true that

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

  f is strictly convex if the equality holds only when $\lambda = 0, 1$.
- A function $f(x)$ is concave if $-f(x)$ is convex.

Sufficient Condition for Convexity

- If a function f has non-negative (positive) second derivatives everywhere, then f is (strictly) convex.
- This can be shown by Taylor's expansion of $f(x)$

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2$$

around the point $x_0 = \lambda x_1 + (1 - \lambda)x_2$ and evaluate at the points $x = x_1, x_2$.

Jensen's Inequality

- If f is a convex function and X is a random variable, then

$$f(EX) \leq Ef(X).$$

  Moreover, if f is strictly convex, then equality implies that $X = EX$.

- The Jensen's inequality is used in many proofs.

Proof of Jensen's Inequality

- We prove by mathematical induction on $|\mathcal{X}|$, the number of elements in the set $\mathcal{X}$.
  - When $|\mathcal{X}| = 2$, the inequality follows from the convexity of f.
  - Suppose it is true for $|\mathcal{X}| = k$. For $|\mathcal{X}| = k+1$,

$$\sum_{i=1}^{k+1} p_i f(x_i) = p_{k+1} f(x_{k+1}) + \sum_{i=1}^{k} p_i f(x_i) = p_{k+1} f(x_{k+1}) + (1-p_{k+1}) \sum_{i=1}^{k} q_i f(x_i)$$

$$\geq p_{k+1} f(x_{k+1}) + (1-p_{k+1}) f(\sum_{i=1}^{k} q_i x_i)$$

$$\geq f(p_{k+1} x_{k+1} + (1-p_{k+1}) \sum_{i=1}^{k} q_i x_i)$$

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

Information Inequality

- Let $p(x), q(x)$ be two probability functions defined for random variable X, then

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

- To prove, let A be the support set of $p(x)$. Then

$$\begin{aligned} -D(p||q) &= -\sum_{x\in A} p(x) \log \frac{p(x)}{q(x)} = \sum_{x\in A} p(x) \log \frac{q(x)}{p(x)} \\ &\leq \log(\sum_{x\in A} p(x) \frac{q(x)}{p(x)}) = \log(\sum_{x\in A} q(x)) \\ &\leq \log(\sum_{x\in \mathcal{X}} q(x)) = \log 1 = 0. \end{aligned}$$

Bound on Entropy

- Let $\mathcal{X}$ be the range of random variable X, then

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

- To prove, use the uniform distribution $u(x) = \frac{1}{|\mathcal{X}|}$ and apply the information inequality $D(p||u) \geq 0$.
- $H(X|Y)$ is bounded by $H(X)$. To prove, note

$$H(X|Y) + I(X;Y) = H(X) \Rightarrow H(X|Y) \leq H(X),$$

since $I(X;Y)$ is non-negative.

Independence Bound on Entropy

- Let $X_{1:n}$ be n random variables, then

$$H(X_{1:n}) \leq \sum_{i=1}^{n} H(X_i).$$

- To prove, expand $H(X_{1:n})$ as the sum of conditional entropies and then bound the conditional entropy by unconditional entropy.

Log Sum Inequality

- Let $a_{1:n}, b_{1:n}$ be non-negative numbers, then

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

  with equality if and only if $\frac{a_i}{b_i}$ is constant.

- To prove, first note that the function $f(x) = x \log x$ is strictly convex for positive x. Define $\alpha_i = \frac{b_i}{\sum_j b_j}$ and $x_i = \frac{a_i}{b_i}$. From $\sum \alpha_i f(x_i) \geq f(\sum \alpha_i x_i)$, the log sum inequality follows.

Convexity of Relative Entropy

- $D(p||q)$ is convex in (p, q): For $0 \leq \lambda \leq 1$ and probability mass functions p_1, p_2, q_1, q_2,

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

- To prove, apply the log sum inequality to each $x \in \mathcal{X}$ with $a_i = \lambda_i p_i(x)$ and $b_i = \lambda_i q_i(x)$, where $\lambda_1 = \lambda$ and $\lambda_2 = 1 - \lambda$. That is,

  $$(\sum_{i=1,2} \lambda_i p_i(x)) \log \frac{\sum_{i=1,2} \lambda_i p_i(x)}{\sum_{i=1,2} \lambda_i q_i(x)} \leq \sum_{i=1,2} \lambda_i p_i(x) \log \frac{\lambda_i p_i(x)}{\lambda_i q_i(x)}.$$

  Summing over all x leads to the desired property.

Concavity of Entropy

- $H(X)$ is a concave function of $p(x)$.
- This follows from $H(X) = \log |\mathcal{X}| - D(p(x)||u(x))$ and the convexity of the relative entropy. Here $u(x)$ is the uniform distribution.

Convexity and Concavity of MI

- $I(X;Y)$ is a concave function of $p(x)$ given $p(y|x)$ and a convex function of $p(y|x)$ given $p(x)$.
- To prove $I(X;Y)$ is concave in $p(x)$ given $p(y|x)$, we use $I(X;Y) = H(Y) - H(Y|X)$.

$$
\begin{aligned}
H(Y)\{\lambda p_1(x) + (1-\lambda)p_2(x)\} &= H(Y)\{\lambda p_1(y) + (1-\lambda)p_2(y)\} \\
&\geq \lambda H(Y)\{p_1(y)\} + (1-\lambda)H(Y)\{p_2(y)\} \\
&= \lambda H(Y)\{p_1(x)\} + (1-\lambda)H(Y)\{p_2(x)\},
\end{aligned}
$$

So the first term $H(Y)$ is concave in $p(x)$. The second term $H(Y|X)$ is linear in $p(x)$, which is both concave and convex. Therefore the rhs is a concave functions of $p(x)$.

Convexity and Concavity of MI

- To prove the other part, that $I(X;Y)$ is convex in $p(y|x)$ given $p(x)$,

$$
\begin{aligned}
& I(X;Y)\{p(x), \lambda p_1(y|x) + (1-\lambda)p_2(y|x)\} \\
&= D(\lambda p_1(x,y) + (1-\lambda)p_2(x,y)||p(x)(\lambda p_1(y) + (1-\lambda)p_2(y))) \\
&\leq \lambda D(p_1(x,y)||p(x)p_1(y)) + (1-\lambda)D(p_2(x,y)||p(x)p_2(y)) \\
&= \lambda I(X;Y)\{p(x), p_1(y|x)\} + (1-\lambda)I(X;Y)\{p(x), p_2(y|x)\}.
\end{aligned}
$$

  So it is indeed convex.

Data Processing Inequality

- Three random variables X, Y, Z are said to form a Markov chain, denoted by $X \to Y \to Z$, if their joint probability can be factorized as

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

- If X, Y, Z form a Markov chain, then

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

- To prove, note $I(X; Z|Y) = 0$, and

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

Fano's Inequality

- Let $\hat{X} = g(Y)$ be an estimator of X based on Y.
  Define the probability of error $P_e \triangleq Pr\{\hat{X} \neq X\}$.
  According to Fano, P_e and $H(X|Y)$ are related by

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y),$$

  where $H(p) = -p \log p - (1-p) \log(1-p)$ is the entropy of a Bernoulli random variable.
- Note that if $P_e = 0$ then $H(X|Y) = 0$.

Proof of Fano's Inequality

- Define the error event indicator $E = I_{\hat{X} \neq X}$. We have

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

- The above terms can be bounded by

$$H(E|X, Y) = 0; \quad H(E|Y) \leq H(E) = H(P_e)$$

$$\begin{aligned} H(X|E, Y) &= Pr(E = 0)H(X|0, Y) + Pr(E = 1)H(X|1, Y) \\ &\leq 0 + P_e \log(|\mathcal{X}| - 1). \end{aligned}$$

- Putting things together, we have

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(E|Y) + H(X|E, Y) = H(X|Y).$$