ITCT Lecture 2:
Entropy, Relative Entropy and Mutual Information

Prof. Ja-Ling Wu

Department of Computer Science
and Information Engineering
National Taiwan University

1

- Definition: The Entropy H(X) of a discrete random variable X is defined by

$$\underset{(H(P))}{H(x)} = -\sum_{x \in X} P(x) \log P(x)$$

$$\log \ : \ \text{base } 2 \ \rightarrow \ \text{H(P)} \ : \ \text{bits}$$

$$0 \log 0 = 0 \qquad (\text{xlogx as x} \rightarrow 0)$$

: adding terms of zero probability does not change the entropy

2

Note that *entropy* is a function of the *distribution* of X. It does not depend on the actual values taken by the *r.v.* X, but only on the *probabilities*.

If $(X, P(x))$, then the expected value of the *r.v.* g(x) is written as

Expectation value

$$\underset{(Eg(x))}{E_p g(x)} = \sum_{x \in X} g(x) P(x)$$

Remark : The entropy of X $\rightarrow$ the expected value of $\log \frac{1}{P(x)}$

$$H(x) = E\left[\log \tfrac{1}{P(x)}\right]$$

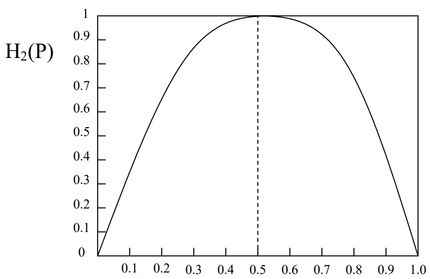
Self-information

- Lemma 1.1: H(x) $\geq$ 0
- Lemma 1.2: $H_b(x) = (\log_b a)\, H_a(x)$

Ex:

$$X = \begin{cases} 0 \quad , \quad P(0) = P \\ 1 \quad , \quad P(1) = 1 - P \end{cases}$$

$$H(X) = -P\log P - (1-P)\log(1-P) \overset{def}{=} H_2(P)$$



1) H(x)=1 bits when P=1/2
2) H(x) is a concave function of P
3) H(x)=0 if P=0 or 1
4) max H(x) occurs when P=1/2

Joint Entropy and Conditional Entropy

- Definition: The joint entropy H(X, Y) of a pair of discrete random variables (X, Y) with a joint distribution P(x, y) is defined as

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log P(x,y)$$

$$or$$

$$H(X,Y) = -E\log P(X,Y)$$

- Definition: The conditional entropy H(Y|X) is defined as

$$\begin{aligned} H(Y\mid X) &= \sum_{x\in X} P(x) H(Y\mid X=x) \text{ is defined as} \\ &= -\sum_{x\in X} P(x)\sum_{y\in Y} P(y\mid x)\log P(y\mid x) \\ &= -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log P(y\mid x) \\ &= -E_{P(x,y)}\log P(Y\mid X) \end{aligned}$$

- Theorem 1.1 (Chain Rule):

$$H(X,Y) = H(X) + H(Y\mid X)$$

$pf:$

$$\begin{aligned} H(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log P(x,y) \\ &= -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log P(x)P(y\mid x) \\ &= -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log P(x) - \sum_{x\in X}\sum_{y\in Y} P(x,y)\log P(y\mid x) \\ &= -\sum_{x\in X} P(x)\log P(x) - \sum_{x\in X}\sum_{y\in Y} P(x,y)\log P(y\mid x) \\ &= H(X) + H(Y\mid X) \end{aligned}$$

or equivalently, we can write

$$\log P(X,Y) = \log P(X) + \log P(Y\mid X)$$

Corollary:

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

Remark:

(i) $H(Y|X) \neq H(X|Y)$

(II) $H(X) - H(X|Y) = H(Y) - H(Y|X)$

Relative Entropy and Mutual Information

- The entropy of a random variable is a measure of the uncertainty of the random variable; it is a measure of the amount of information required on the average to describe the random variable.
- The relative entropy is a measure of the distance between two distributions. In statistics, it arises as an expected logarithm of the likelihood ratio. The relative entropy $D(p||q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p.

- Ex: If we knew the true distribution of the *r.v.*, then we could construct a code with average description length H(p). If instead, we used the code for a distribution q, we would need H(p)+D(p||q) bits on the average to describe the *r.v.*.

- Definition:

  The relative entropy or Kullback Liebler distance between two probability mass functions p(x) and q(x) is defines as

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

$$= E_p \log \frac{p(x)}{q(x)} = E_p\left[\log \frac{1}{q(x)} - \log \frac{1}{p(x)}\right]$$

$$= E_p\left[\log \frac{1}{q(x)}\right] - E_p\left[\log \frac{1}{p(x)}\right]$$

- Definition:

  Consider two *r.v.*'s X and Y with a joint probability mass function p(x,y) and marginal probability mass functions p(x) and p(y). The mutual information I(X;Y) is the relative entropy between the joint distribution p(x,y) and the product distribution p(x) p(y) i.e.,

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$
$$=D\big(p(x,y)\,\|\,p(x)p(y)\big)$$
$$=E_{p(x,y)}\left[\log\frac{P(X,Y)}{P(X)P(Y)}\right]$$

Information Theory

11

- Ex: Let X = {0, 1} and consider two distributions p and q on X. Let p(0)=1-r, p(1)=r, and let q(0)=1-s, q(1)=s. Then

$$D(p\,\|\,q)=p(0)\log\frac{p(0)}{q(0)}+p(1)\log\frac{p(1)}{q(1)}$$
$$=(1-r)\log\frac{1-r}{1-s}+r\log\frac{r}{s}$$
$$and\;\;D(q\,\|\,p)=q(0)\log\frac{q(0)}{p(0)}+q(1)\log\frac{q(1)}{p(1)}$$
$$=(1-s)\log\frac{1-s}{1-r}+s\log\frac{s}{r}$$

$\Rightarrow$ If r=s, then D(p||q)=D(q||p)=0

While, in general,

D(p||q) $\neq$ D(q||p)

Information Theory

12

Relationship between Entropy and Mutual Information

Rewrite I(X;Y) as

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y)\log\frac{p(x\mid y)}{p(x)} \\
&= -\sum_{x,y} p(x,y)\log p(x) + \sum_{x,y} p(x,y)\log p(x\mid y) \\
&= -\sum_{x} p(x)\log p(x) - \left(-\sum_{x,y} p(x,y)\log p(x\mid y)\right) \\
&= H(X) - H(X\mid Y)
\end{aligned}
$$

Thus the mutual information I(X;Y) is the reduction in the uncertainty of X due to the knowledge of Y.

By symmetry, it follows that

I(X;Y) = H(Y) – H(Y|X)

→ X says much about Y as Y says about X

Since H(X;Y) = H(X) + H(Y|X)

→ I(X;Y) = H(X) + H(Y) – H(X,Y)

I(X;X) = H(X) + H(X|X) = H(X)

The mutual information of a r.v. with itself is the entropy of the r.v. ---> entropy : self-information

- Theorem: (Mutual information and entropy):

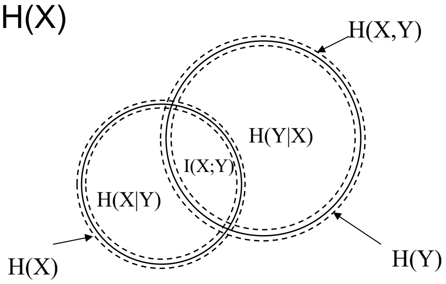
  i. $I(X;Y) = H(X) - H(X|Y)$

  $= H(Y) - H(Y|X)$

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

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

  iii. $I(X;X) = H(X)$



Chain Rules for Entropy, Relative Entropy and Mutual Information

- Theorem: (Chain rule for entropy)

  Let $X_1, X_2, \ldots, X_n$, be drawn according to $P(x_1, x_2, \ldots, x_n)$.

  Then

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

- Proof

(1)

$$
\begin{aligned}
H(X_1, X_2) &= H(X_1) + H(X_2 \mid X_1) \\
H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3 \mid X_1) \\
&= H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_2, X_1) \\
&\vdots \\
H(X_1, X_2, \cdots, X_n) &= H(X_1) + H(X_2 \mid X_1) + \cdots H(X_n \mid X_{n-1}, \cdots, X_1) \\
&= \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \cdots, X_1)
\end{aligned}
$$

(2) We write

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} P(x_i \mid x_{i-1}, \cdots, x_1)$$

then

$$
\begin{aligned}
&H(X_1, X_2, \cdots, X_n) \\
&= -\sum_{X_1, X_2, \cdots, X_n} P(x_1, x_2, \cdots, x_n) \log P(x_1, x_2, \cdots, x_n) \\
&= -\sum_{X_1, X_2, \cdots, X_n} P(x_1, x_2, \cdots, x_n) \log \prod_{i=1}^{n} P(x_i \mid x_{i-1}, \cdots, x_1) \\
&= -\sum_{X_1, X_2, \cdots, X_n} \sum_{i=1}^{n} P(x_1, x_2, \cdots, x_n) \log P(x_i \mid x_{i-1}, \cdots, x_1) \\
&= -\sum_{i=1}^{n} \sum_{X_1, X_2, \cdots, X_n} P(x_1, x_2, \cdots, x_n) \log P(x_i \mid x_{i-1}, \cdots, x_1) \\
&= -\sum_{i=1}^{n} \sum_{X_1, X_2, \cdots, X_i} P(x_1, x_2, \cdots, x_i) \log P(x_i \mid x_{i-1}, \cdots, x_1) \\
&= \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \cdots, X_1)
\end{aligned}
$$

- Definition:
  The conditional mutual information of rv's. X and Y given Z is defined by

$$I(X;Y\,|\,Z) = H(X\,|\,Z) - H(X\,|\,Y,Z)$$
$$= E_{p(x,y,z)} \log \frac{P(X,Y\,|\,Z)}{P(X\,|\,Z)\cdot P(Y\,|\,Z)}$$

19

- Theorem: (chain rule for mutual-information)

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

proof:

$$\begin{aligned}
& I(X_1, X_2, \cdots, X_n; Y) \\
&= H(X_1, X_2, \cdots, X_n) - H(X_1, X_2, \cdots, X_n\,|\,Y) \\
&= \sum_{i=1}^{n} H(X_i\,|\,X_{i-1}, \cdots, X_1) - \sum_{i=1}^{n} H(X_i\,|\,X_{i-1}, \cdots, X_1, Y) \\
&= \sum_{i=1}^{n} I(X_i; Y\,|\,X_1, X_2, \cdots, X_{i-1})
\end{aligned}$$

20

- **Definition:**
  The conditional relative entropy D(p(y|x) || q (y|x)) is the average of the relative entropies between the conditional probability mass functions p(y|x) and q(y|x) averaged over the probability mass function p(x).

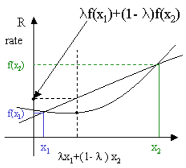
$$D(p(y\mid x)\,\|\,q(y\mid x)) = \sum_x p(x)\sum_y p(y\mid x)\log\frac{p(y\mid x)}{q(y\mid x)}$$

$$= E_{p(x,y)}\log\frac{p(Y\mid X)}{q(Y\mid X)}$$

- **Theorem:** (Chain rule for relative entropy)
  D(p(x,y)||q(x,y)) = D(p(x)||q(x))+ D(p(y|x)||q(y|x))

Jensen's Inequality and Its Consequences

- Definition: A function is said to be convex over an interval (a,b) if for every x_1, $x_2 \in$ (a,b) and $0 \le \lambda \le 1$, $f(\lambda x_1+(1-\lambda)x_2) \le \lambda f(x_1)+(1-\lambda)f(x_2)$
  A function f is said to be strictly convex if equality holds only if $0< \lambda <1$.



- Definition: A function is concave if –f is convex.
  Ex: convex functions: X^2, |X|, e^X, XlogX (for $X \ge 0$)
  concave functions: logX, $X^{1/2}$ for $X \ge 0$
  both convex and concave: ax+b; linear functions

- Theorem:
  If the function f has a second derivative which is non-negative (positive) everywhere, then the function is convex (strictly convex).

$$\begin{cases} EX = \sum_{x \in X} p(x)x & : \quad \text{discrete case} \\ EX = \int p(x)xdx & : \quad \text{continuous case} \end{cases}$$

Information Theory

23

- Theorem : (Jensen's inequality):
  If f(x) is convex function and X is a random variable, then Ef(X) $\geq$ f(EX).
  Proof: For a two mass point distribution, the inequality becomes
  $p_1 f(x_1)+p_2 f(x_2) \geq f(p_1x_1+p_2x_2)$, $p_1+p_2=1$
  which follows directly from the definition of convex functions.
  Suppose the theorem is true for distributions with K-1 mass points.
  Then writing $P'_i=P_i/(1-P_K)$ for i = 1, 2, …, K-1, we have

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

$$\geq p_k f(x_k) + (1-p_k) f(\sum_{i=1}^{k-1} p'_i x_i)$$

$$\geq f(p_k x_k + (1-p_k)\sum_{i=1}^{k-1} p'_i x_i)$$

$$= f((1-p_k)p'_k x_k + (1-p_k)\sum_{i=1}^{k-1} p'_i x_i)$$

$$= f((1-p_k)\sum_{i=1}^{k} p'_i x_i)$$

$$= f(\sum_{i=1}^{k} (1-p_k)p'_i x_i)$$

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

The proof can be extended to continuous distributions by continuity arguments.
(Mathematical Induction)

Information Theory

24

- Theorem: (Information inequality):

  Let p(x), q(x) ($x \in X$), be two probability mass functions. Then

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

  with equality iff p(x)=q(x) for all x.

  Proof: Let A={x:p(x)>0} be the support set of p(x). Then

$$
\begin{aligned}
-D(p\|q) &= -\sum_{x\in A} p(x)\log\frac{p(x)}{q(x)} \\
&= \sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} = E\left\{\log\frac{q(x)}{p(x)}\right\} \leq \left\{\log E\left(\frac{q(x)}{p(x)}\right)\right\} \\
&= \log\sum_{x\in A} p(x)\frac{q(x)}{p(x)} \qquad\qquad (\log t \text{ is concave}) \\
&= \log\sum_{x\in A} q(x) \\
&\leq \log\sum_{x\in X} q(x) \\
&= \log 1 = 0
\end{aligned}
$$

- Corollary: (Non-negativity of mutual information):

  For any two rv's., X, Y,

  $I(X;Y) \geq 0$

  with equality iff X and Y are independent.

  Proof:

  $I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0$ with equality iff p(x,y)=p(x)p(y), i.e., X and Y are independent

- Corollary:

  $D(p(y|x)||q(y|x)) \geq 0$

  with equality iff p(y|x)=q(y|x) for all x and y with p(x)>0.

- Corollary:

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

  with equality iff X and Y are conditionally independent given Z.

- Theorem:
  H(x)≤ log|***X***|, where |***X***| denotes the number of elements in the range of X, with equality iff X has a uniform distribution over ***X***.

Proof:

Let u(x)=1/|***X***| be the uniform probability mass function over ***X***, and let p(x) be the probability mass function for X. Then

$$D(p\,\|\,u) = \sum p(x)\log\frac{p(x)}{u(x)} = \log|\mathbf{X}| - H(x)$$

Hence by the non-negativity of relative entropy

$$0 \le D(p\,\|\,u) = \log|\mathbf{X}| - H(x)$$

- Theorem: (conditioning reduces entropy):

  H(X|Y) ≤ H(X)

  with equality iff X and Y are independent.

  Proof: 0 ≤ I(X;Y)=H(X) – H(X|Y)

Note that this is true only on the average; specifically, H(X|Y=y) may be greater than or less than or equal to H(X), but on the average H(X|Y)=Σ p(y)H(X|Y=y) ≤ H(X).

- Ex: Let (X,Y) have the following joint distribution

Y \ X	1	2
1	0	3/4
2	1/8	1/8

Then, H(X)=H(1/8, 7/8)=0.544 bits

H(X|Y=1)=0 bits

H(X|Y=2)=1 bits > H(X)

However, H(X|Y) = 3/4 H(X|Y=1)+1/4 H(X|Y=2)

= 0.25 bits < H(X)

- Theorem: (Independence bound on entropy):

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

$$H(X_1, X_2, \cdots, X_n) \le \sum_{i=1}^{n} H(X_i)$$

with equality iff the Xi are independent.

Proof: By the chain rule for entropies,

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

$$\le \sum_{i=1}^{n} H(X_i)$$

with equality iff the X_i's are independent.

The LOG SUM INEQUALITY AND ITS APPLICATIONS

- Theorem: (Log sum inequality)

  For non-negative numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2 .. b_n$

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

  with equality iff a_i/b_i = constant.

$$\left(\begin{array}{ll} \text{some conventions:} & 0\log 0 = 0,\ a\log\tfrac{a}{0} = \infty \text{ if } a > 0 \\ & 0\log\tfrac{0}{0} = 0 \end{array} \right)$$

Proof:

Assume w.l.o.g that $a_i>0$ and $b_i>0$. The function f(t)=tlogt is strictly convex, since $f''(t) = \frac{1}{t}\log e > 0$ for all positive t. Hence by Jensen's inequality, we have

$$\sum \alpha_i f(t_i) \geq f\left(\sum \alpha_i t_i\right)$$

for $\alpha_i \geq 0, \sum_i \alpha_i = 1$. Setting $\alpha_i = \frac{b_i}{\sum_{i=1}^{n} b_i}$ and $t_i = \frac{a_i}{b_i}$,

we obtain $\sum \frac{b_i}{\sum_i b_i} \cdot \frac{a_i}{b_i} \log \frac{a_i}{b_i} \geq \sum \frac{b_i}{\sum_i b_i} \cdot \frac{a_i}{b_i} \log \left(\sum \frac{b_i}{\sum_i b_i} \cdot \frac{a_i}{b_i} \right)$

$$\sum \frac{b_i}{\sum_i b_i} \cdot \frac{a_i}{b_i} \log \frac{a_i}{b_i} \geq \sum \frac{a_i}{\sum b_i} \log \sum \frac{a_i}{\sum b_i} \qquad (\text{note that } \sum_i b_i = 1)$$

$$\Rightarrow \sum a_i \log \frac{a_i}{b_i} \geq \sum a_i \log \frac{\sum a_i}{\sum b_i}$$

which is the log sum inequality. (Sum bi greater than (

- Reproving the theorem that D(p||q) ≥ 0, with equality iff p(x)=q(x)

$$
\begin{aligned}
D(p\,\|\,q) &= \sum p(x)\log\frac{p(x)}{q(x)} \\
&\geq \left(\sum p(x)\right)\log\frac{\sum p(x)}{\sum q(x)} \qquad (\text{from log-sum inequality}) \\
&= 1\log\frac{1}{1} = 0
\end{aligned}
$$

with equality iff p(x)/q(x)=c. Since both p and q are probability mass functions, c=1 ⇒ p(x)=q(x), ∀ x.

- Theorem:
D(p||q) is convex in the pair (p,q), i.e., if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then

$$D(\lambda p_1 + (1-\lambda)p_2 \,\|\, \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1\,\|\,q_1) + (1-\lambda)D(p_2\,\|\,q_2)$$
$$\textit{for all } 0 \leq \lambda \leq 1$$

- Proof:

$$
\begin{aligned}
& D(\lambda p_1 + (1-\lambda)p_2 \,\|\, \lambda q_1 + (1-\lambda)q_2) \\
&= \sum (\lambda p_1 + (1-\lambda)p_2)\log\frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2} \cdots (1) \\
& \textit{Let} \quad a_1 = \lambda p_1, \quad a_2 = (1-\lambda)p_2 \\
& \qquad\;\; b_1 = \lambda q_1, \quad b_2 = (1-\lambda)q_2 \\
& \textit{then } (1) \Rightarrow \sum\left(\sum_{i=1}^{2} a_i\right)\log\frac{\left(\sum_{i=1}^{2} a_i\right)}{\left(\sum_{i=1}^{2} b_i\right)} \\
& \overset{\textit{log-sum}}{\leq} \sum\left[\sum_{i=1}^{2} a_i \log\frac{a_i}{b_i}\right] = \sum\left(\lambda p_1 \log\frac{\lambda p_1}{\lambda q_1} + (1-\lambda)p_2 \log\frac{(1-\lambda)p_2}{(1-\lambda)q_2}\right) \\
&= \lambda \sum p_1 \log\frac{p_1}{q_1} + (1-\lambda)\sum p_2 \log\frac{p_2}{q_2} \\
&= \lambda D(p_1\,\|\,q_1) + (1-\lambda)D(p_2\,\|\,q_2)
\end{aligned}
$$

- Theorem: (concavity of entropy):
  H(p) is a concave function of P.
  That is: $H(\lambda_1 p_1+(1-\lambda)p_2) \geqq \lambda H(p_1)+(1-\lambda)H(p_2)$

  Proof:

  $H(p)=\log|\boldsymbol{\mathcal{X}}| - D(p||u)$

  where u is the uniform distribution on $|\boldsymbol{\mathcal{X}}|$ outcomes. The concavity of H then follows directly from the convexity of D.

- Theorem: Let $(X,Y)\sim p(x,y) = p(x)p(y|x)$.
  The mutual information I(X;Y) is
  (i) a concave function of p(x) for fixed p(y|x)
  (ii) a convex function of p(y|x) for fixed p(x).

Proof:

(1) $I(X;Y)=H(Y)-H(Y|X)=H(Y) - \Sigma_x p(x)H(Y|X=x) \ldots (\Delta)$

if p(y|x) is fixed, then p(y) is a linear function of p(x). ($p(y) = \Sigma_x p(x,y) = \Sigma_x p(x)p(y|x)$)
Hence H(Y), which is a concave function of p(y), is a concave function of p(x). The second term of (Δ) is a linear function of p(x). Hence the difference is a concave function of p(x).

(2) We fix p(x) and consider two different conditional distributions $p_1(y|x)$ and $p_2(y|x)$. The corresponding joint distributions are $p_1(x,y)=p(x)\,p_1(y|x)$ and $p_2(x,y)=p(x)\,p_2(y|x)$, and their respective marginals are $p(x)$, $p_1(y)$ and $p(x)$, $p_2(y)$.

Consider a conditional distribution

$$p_\lambda(y|x)= \lambda p_1(y|x)+(1-\lambda)p_2(y|x)$$

that is a mixture of $p_1(y|x)$ and $p_2(y|x)$. The corresponding joint distribution is also a mixture of the corresponding joint distributions,

$$p_\lambda(x,y) = \lambda p_1 (x,y)+(1-\lambda)p_2(x,y)$$

when p(x) is fixed, $p_\lambda(x,y)$ is linear with $p_i(y|x)$

and the distribution of Y is also a mixture $p_\lambda(y)= \lambda p_1 (y)+(1-\lambda)p_2(y)$. Hence if we let $q_\lambda(x,y)=p(x)p_\lambda(y) \Rightarrow q_\lambda(x,y)= \lambda q_1 (x,y)+(1-\lambda)q_2(x,y)$.

The product of the marginal distributions

$q_\lambda(x,y)$ is also linear with $p_i(y|x)$ when p(x) is fixed.

$I(X;Y) = D(p_\lambda||q_\lambda) \rightarrow$ convex of (p,q)

$\Rightarrow$ the mutual information is a convex function of the conditional distribution. Therefore, the convexity of I(X;Y) is the same as that of the $D(p_\lambda||q_\lambda)$ w.r.t. $p_i(y|x)$ when p(x) is fixed.

Data processing inequality:

No clever manipulation of the data can improve the inferences that can be made from the data

- Definition:

  Rv's. X,Y,Z are said to form a Markov chain in that order (denoted by $X\rightarrow Y\rightarrow Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. That is $X\rightarrow Y\rightarrow Z$ form a Markov chain, then

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

  (ii) $p(x,z|y)=p(x|y)p(z|y)$: X and Z are conditionally independent given Y

- $X\rightarrow Y\rightarrow Z$ implies that $Z\rightarrow Y\rightarrow X$

  If Z=f(Y), then $X\rightarrow Y\rightarrow Z$

- **Theorem: (Data processing inequality)**

  if $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$

  No processing of Y, deterministic or random, can increase the information that Y contains about X.

Proof:

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) \quad \text{: chain rule}$$
$$= I(X;Y) + I(X;Z|Y) \quad \text{: chain rule}$$

Since X and Z are independent given Y, we have $I(X;Z|Y)=0$. Since $I(X;Y|Z) \geq 0$, we have $I(X;Y) \geq I(X;Z)$ with equality iff $I(X;Y|Z)=0$, i.e., $X \to Z \to Y$ forms a Markov chain. Similarly, one can prove $I(Y;Z) \geq I(X;Z)$.

- **Corollary:**

  If $X \to Y \to Z$ forms a Markov chain and if $Z=g(Y)$, we have $I(X;Y) \geq I(X;g(Y))$

  : functions of the data Y cannot increase the information about X.

- **Corollary:** If $X \to Y \to Z$, then $I(X;Y|Z) \leq I(X;Y)$

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

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

By Markovity, $I(X;Z|Y)=0$
and $I(X;Z) \geq 0 \Rightarrow I(X;Y|Z) \leq I(X;Y)$

$\Rightarrow$ The dependence of X and Y is decreased (or remains unchanged) by the observation of a “downstream” r.v. Z.

- Note that it is possible that I(X;Y|Z)>I(X;Y) when X,Y and Z do not form a Markov chain.

Ex: Let X and Y be independent fair binary rv's, and let Z=X+Y. Then I(X;Y)=0, but

$$I(X;Y|Z)=H(X|Z) - H(X|Y,Z)$$
$$=H(X|Z)$$
$$=P(Z=1)H(X|Z=1)=1/2 \text{ bit.}$$

Fano's inequality:

- Fano's inequality relates the probability of error in guessing the r.v. X to its conditional entropy H(X|Y).

Note that:

The conditional entropy of a r.v. X given another random variable Y is zero iff X is a function of Y.

proof: HW H(X|Y)=0 implies there is no uncertainty about X if we know Y
⇒ for all x with p(x)>0, there is only one possible value of y with p(x,y)>0

⇒ we can estimate X from Y with zero probability of error iff H(X|Y)=0.

⇒ we expect to be able to estimate X with a low probability of error only if the conditional entropy H(X|Y) is small.

Fano's inequality quantifies this idea.

- Suppose we wish to estimate a r.v. X with a distribution p(x). We observe a r.v. Y which is related to X by the conditional distribution p(y|x). From Y, we calculate a function $g(Y) = \hat{X}$ which is an estimate of X. We wish to bound the probability that $\hat{X} \neq X$. We observe that $X \to Y \to \hat{X}$ forms a Markov chain.

  Define the probability of error

$$P_e = P_r\left\{\hat{X} \neq X\right\} = P_r\{g(Y) \neq X\}$$

Information Theory

43

- Theorem: (Fano's inequality)

  For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$ with $P_e = P_r(X \neq \hat{X})$, we have

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

  $H(P_e) \leq 1$, E: binary r.v.
  $\log(|\boldsymbol{X}|-1) \leq \log|\boldsymbol{X}|$

  This inequality can be weakened to

  $1 + P_e \log(|\boldsymbol{X}|) \geq H(X|Y)$

  or

$$P_e \geq \frac{H(X|Y) - 1}{\log|\mathbf{X}|}$$

Remark: $P_e = 0 \Rightarrow H(X|Y) = 0$

Information Theory

44

Proof: Define an error rv.

$$E = \begin{cases} 1 & ,\text{if } \hat{X} \neq X \\ 0 & ,\text{if } \hat{X} = X \end{cases}$$

By the chain rule for entropies, we have

$H(E,X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X,\hat{X})}_{=0}$

$= \underbrace{H(E|\hat{X})}_{\leq H(P_e)} + \underbrace{H(X|E,\hat{X})}_{\leq P_e \log(|\mathcal{X}|-1)}$

Since conditioning reduces entropy, $H(E|\hat{X}) \leq H(E) = H(P_e)$. Now since E is a function of X and $\hat{X} \Rightarrow H(E|X,\hat{X})=0$. Since E is a binary-valued r.v., $H(E)= H(P_e)$.

The remaining term, $H(X|E,\hat{X})$, can be bounded as follows:

$H(X|E,\hat{X}) = P_r(E=0)H(X|\hat{X},E=0)+P_r(E=1)H(X|\hat{X},E=1)$

$\leq (1- P_e)0 + P_e \log(|\mathcal{X}|-1)$,

Since given E=0, $X=\hat{X}$, and given E=1, we can upper bound the conditional entropy by the log of the number of remaining outcomes ($|\mathcal{X}|-1$).

$H(P_e)+P_e\log|X| \geq H(X|\hat{X})$. By the data processing inequality, we have $I(X;\hat{X}) \leq I(X;Y)$ since $X \rightarrow Y \rightarrow \hat{X}$, and therefore $H(X|\hat{X}) \geq H(X|Y)$. Thus we have $H(P_e)+P_e\log|X| \geq H(X|\hat{X}) \geq H(X|Y)$.

Remark:

Suppose there is no knowledge of Y. Thus X must be guessed without any information. Let $X \in \{1,2,\ldots,m\}$ and $P_1 \geq P_2 \geq \ldots \geq P_m$. Then the best guess of X is X=1 and the resulting probability of error is $P_e = 1 - P_1$.

Fano's inequality becomes

$$\mathbf{H(P_e) + P_e\log(m-1) \geq H(X)}$$

The probability mass function

$$\mathbf{(P_1, P_2,\ldots, P_m) = (1-P_e, P_e/(m-1), \ldots, P_e/(m-1))}$$

achieves this bound with equality.

Some Properties of the Relative Entropy

1. Let μ_n and μ'_n be two probability distributions on the state space of a Markov chain at time n, and let μ_{n+1} and μ'_{n+1} be the corresponding distributions at time n+1. Let the corresponding joint mass function be denoted by p and q.

   That is,

   $p(x_n, x_{n+1}) = p(x_n)\, r(x_{n+1} | x_n)$

   $q(x_n, x_{n+1}) = q(x_n)\, r(x_{n+1} | x_n)$

   where

   $r(\cdot \mid \cdot)$ is the probability transition function for the Markov chain.

Then by the chain rule for relative entropy, we have the following two expansions:

$$D(p(x_n, x_{n+1}) || q(x_n, x_{n+1}))$$
$$= D(p(x_n) || q(x_n)) + D(p(x_{n+1}|x_n) || q(x_{n+1}|x_n))$$
$$= D(p(x_{n+1}) || q(x_{n+1})) + D(p(x_n|x_{n+1}) || q(x_n|x_{n+1}))$$

Since both p and q are derived from the same Markov chain, so

$$p(x_{n+1}|x_n) = q(x_{n+1}|x_n) = r(x_{n+1}|x_n),$$

and hence

$$D(p(x_{n+1}|x_n)) \,||\, q(x_{n+1}|x_n)) = 0$$

That is,

$$D(p(x_n) \,||\, q(x_n))$$
$$= D(p(x_{n+1}) \,||\, q(x_{n+1})) + D(p(x_n|x_{n+1}) \,||\, q(x_n|x_{n+1}))$$

Since $D(p(x_n|x_{n+1}) \,||\, q(x_n|x_{n+1})) \geq 0$

$$\Rightarrow \boxed{D(p(x_n) \,||\, q(x_n)) \geq D(p(x_{n+1}) \,||\, q(x_{n+1}))}$$

or $D(\mu_n || \mu'_n) \geq D(\mu_{n+1} || \mu'_{n+1})$

Conclusion:

The distance between the probability mass functions is decreasing with time n for any Markov chain.

2. Relative entropy $D(\mu_n || \mu)$ between a distribution μ_n on the states at time n and a stationary distribution μ decreases with n.
   In the last equation, if we let μ'_n be any stationary distribution μ, then μ'_{n+1} is the same stationary distribution. Hence

$$D(\mu_n || \mu) \geq D(\mu_{n+1} || \mu)$$

$\Rightarrow$ Any state distribution gets closer and closer to each stationary distribution as time passes. $\lim_{n\to\infty} D(\mu_n \,||\, \mu) = 0$

3. Def:A probability transition matrix $[P_{ij}]$,
$P_{ij} = P_r\{x_{n+1}=j|x_n=i\}$ is called doubly stochastic if

$$\Sigma_i P_{ij}=1,\ i=1,2,\ldots,\ j=1,2,\ldots$$

and

$$\Sigma_j P_{ij}=1,\ i=1,2,\ldots,\ j=1,2,\ldots$$

The uniform distribution is a stationary distribution of P iff the probability transition matrix is doubly stochastic.

4. The conditional entropy $H(X_n|X_1)$ increase with n for a stationary Markov process.
If the Markov process is stationary, then $H(X_n)$ is constant. So the entropy is non-increasing. However, it can be proved that $H(X_n|X_1)$ increases with n. This implies that:

   the conditional uncertainty of the future increases.

Proof:

$$\begin{aligned} H(X_n|X_1) &\geq H(X_n|X_1, X_2) && \text{(conditioning reduces entropy)} \\ &= H(X_n|X_2) && \text{(by Markovity)} \\ &= H(X_{n-1}|X_1) && \text{(by stationarity)} \end{aligned}$$

Similarly: $H(X_0|X_n)$ is increasing in n for any Markov chain.

Sufficient Statistics

Suppose we have a family of probability mass function $\{f_\theta(x)\}$ indexed by θ, and let X be a sample from a distribution in this family. Let T(X) be any statistic (function of the sample) like the sample mean or sample variance. Then

$$\theta \to X \to T(X),$$

And by the data processing inequality, we have

$$I(\theta;T(X)) \le I(\theta;X)$$

for any distribution on θ. However, if equality holds, no information is lost.

A statistic T(X) is called sufficient for θ if it contains all the information in X about θ.

- Def:
  A function T(X) is said to be a sufficient statistic relative to the family $\{f_\theta(x)\}$ if X is independent of θ give T(X), i.e., $\theta \to T(X) \to X$ forms a Markov chain.

  or:

  $$I(\theta;X) = I(\theta; T(X))$$

  for all distributions on θ

  Sufficient statistics preserve mutual information.

Some examples of Sufficient Statistics

1. Let $X_1, X_2, \ldots, X_n, X_i \in \{0,1\}$ be an i.i.d. sequence of coin tosses of a coin with unknown parameter $\theta = Pr(X_i = 1)$.

   Given n, the number of 1's is a sufficient statistics for θ.

   Here $T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} X_i$. $\Rightarrow$

   Given T, all sequences having that many 1's are equally likely and independent of the parameter θ.

Information Theory

55

$$\Pr\left\{ (X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n) \middle| \sum_{i=1}^{n} x_i = k \right\}$$

$$= \begin{cases} 1 \Big/ \binom{n}{k} & , if \sum x_i = k \\ 0 & , otherwise \end{cases}$$

$$Thus,\ \theta \to \sum X_i \to (X_1, X_2, \ldots, X_n)$$

and T is a sufficient statistics for θ.

Information Theory

56

2. If X is normally distributed with mean θ and variance 1; that is,

if $f_\theta = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\theta)^2}{2}} = N(\theta,1)$

and $X_1, X_2, \ldots, X_n$ are drawn independently according to f_θ, a sufficient statistic for θ is the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

This can be verified that $P(X_1, X_2, \ldots, X_n \mid \overline{X}_n, n)$ is independent of θ.

The minimal sufficient statistics is a sufficient statistics that is a function of all other sufficient statistics.

<u>Def:</u>

A static T(X) is a minimal sufficient statistic related to $\{f_\theta(X)\}$ if it is a function of every other sufficient statistic $U: \theta \to T(X) \to U(X) \to X$

Hence, a minimal sufficient statistic maximally compresses the information about θ in the sample. Other sufficient statistics may contain additional irrelevant information.

The sufficient statistics of the above examples are minimal.

Shuffles increase Entropy:

If T is a shuffle (permutation) of a deck of cards and X is the initial (random) position of the cards in the deck and if the choice of the shuffle T is independent of X, then

$$H(TX) \geq H(X)$$

where TX is the permutation of the deck induced by the shuffle T on the initial permutation X.

Proof: $H(TX) \geq H(TX|T)$

$= H(T^{-1}TX|T)$ (why?)

$= H(X|T)$

$= H(X)$

if X and T are independent!

If X and X' are i.i.d. with entropy H(X), then $P_r(X=X') \geq 2^{-H(X)}$ with equality iff X has a uniform distribution.

pf: suppose X~p(x). By Jensen's inequality, we have $2^{E\log p(x)} \leq E2^{\log p(x)}$

which implies that $2^{-H(X)}=2^{\sum p(x)\log p(x)} \leq \sum p(x)2^{\log p(x)}= \sum p^2(x)=P_r(X=X')$

(Let X and X' be two i.i.d. rv's with entropy H(X). The prob. at X=X' is given by $P_r(X=X')= \sum_x p^2(x)$)

Let X, X' be independent with X~p(x), X'~r(x), x, x' $\in \chi$

Then $P_r(X=X') \geq 2^{-H(p)-D(p||r)}$

$P_r(X=X') \geq 2^{-H(r)-D(r||p)}$

pf: $2^{-H(p)-D(p||r)}= 2^{\sum p(x)\log p(x)+\sum p(x)\log r(x)/p(x)}=2^{\sum p(x)\log r(x)} \leq \sum p(x)2^{\log r(x)} = \sum p(x)r(x) =P_r(X=X')$