ST06 - LECTURE 3

Note Title 2/14/2006

TODAY

- Properties of Information & Entropy.
- Tool : Convexity & Jensen's Inequality ; Informational Divergence
- Results : - Positivity of Mutual Information
  - Data Processing theorem
  - Fano's Theorem (Prob. Error high if Conditional Entropy high)

Review of last time

Entropy $X \in \{1 \ldots N\}$

$$\Pr[X=i] = P_i$$

$$H(X) = H(P_1 \ldots P_N) = \sum_{i=1}^{N} -P_i \log P_i$$

Conditional Entropy

$$H(X|Y) = \sum_{y} P_Y(y) \sum_{x} \Pr(X=x \mid Y=y) \cdot \log(\quad)$$

$$= \sum_{y} \sum_{x} P(x,y) \log \frac{P_Y(y)}{P(x,y)}$$

Mutual Information

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

$$= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P_X(x) \cdot P_Y(y)}$$

Chain Rule of Entropy

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

Today: Some more complex properties

Two objectives:

- $I(X;Y) \geq 0$?
- Conditional Entropy $\Rightarrow$ Unpredictability ?

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P_X(x) \cdot P_Y(y)}$$

Question: $I(X;Y) \geq 0$?

Why is $\displaystyle\sum_{x,y} p(x,y) \log \frac{p(x,y)}{P_X(x) \cdot P_Y(y)} \geq 0$

$\Updownarrow$

$$E\left[\log \frac{p(x,y)}{P_X(x) \cdot P_Y(y)}\right] \geq 0$$

Expression is of the form

$$\mathbb{E}_{n}\left[f(x) \right] \geq 0 \ ?$$

"Pattern matching" points to Jensen's Inequality

<u>Jensen's Inequality</u>

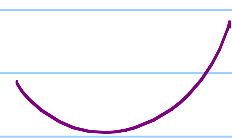
for every "convex function" $f : \mathbb{R}^n \to \mathbb{R}$

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

<u>What is a convex function?</u>

- x^2 good example; Also e^x;

  $-\log x$;



- Calculus defn: $f : \mathbb{R} \to \mathbb{R}$ is convex if $f''(x) \geq 0$

- More generally

$f : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$f(\lambda \cdot x + (1-\lambda)\cdot y) \leq \lambda f(x) + (1-\lambda) f(y)$$

Jensen's Inequality: Trivial for distribution on finite random-variables from ~~first~~ defn.

But Calculus definition $\Rightarrow$ general defn. ...

takes a little work won't do here

Implication

$$E[-\log Z] \geq -\log E[Z]$$

Apply to $Z = \frac{q(X)}{p(X)}$ for any distr p, q & r.v. X with distr p

$$E\left[\log \frac{p(x)}{q(x)}\right] = E\left[-\log Z\right] \geq -\log E[Z]$$

$$= -\log \sum_x p(x) \cdot \frac{q(x)}{p(x)}$$

$$= \log 1$$

$$= 0$$

Conclude: $\forall$ pair of distributions p, q

$$E\left[\log \frac{q(x)}{p(x)}\right] \geq 0$$

When is $E\left[\log \frac{q(x)}{p(x)}\right] = 0$?

iff $\forall x$, $\frac{q(x)}{p(x)} = \quad = 1$

i.e. if $q(x) = p(x)$

Motivates "Relative Entropy"

"(Informational / KL) Divergence"

$\uparrow$

Kullback-Liebler

$$D(p \| q) = E\left[\log \frac{q(x)}{p(x)}\right]$$

- $D(p\|q) \geq 0$
- $D(p\|q) = 0 \Rightarrow p = q.$
- $I(X;Y) = D\left(p(x,y) \,\|\, p_X(x) \cdot p_Y(y)\right)$
- $H(X) = \log|X| - D(p \| U)$

(Introduction a little premature ...)

Later we will see that the "best" compression of x should take $\approx \left| \log \frac{1}{P_x} \right|$ bits.

So if we compress X thinking it comes from dist q will use

$$- \sum p(x) \log \frac{1}{q(x)} \text{ bits}$$

– Should have taken

$$\sum p(x) \log \frac{1}{p(x)} \text{ bits}$$

Inefficiency $= \sum p(x) \log \frac{p(x)}{q(x)} = D(p \| q)$.

More on $D(p \| q)$

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

$D(p \| q)$ could be infinite

Example

$$\underbrace{\begin{aligned} X &= 0 \quad \text{w.p. } 1 \\ &= 1 \quad \text{w.p. } 0 \end{aligned}}_{p} \qquad \underbrace{\begin{aligned} X &= 0 \quad \text{w.p. } \tfrac{1}{2} \\ &= 1 \quad \text{w.p. } \tfrac{1}{2} \end{aligned}}_{q}$$

one of $D(p \| q)$ / $D(q \| p)$ is infinite

Back to consequences:

1. $H(X) \leq \log |\Omega_X|$

since $H(X) = \log |\Omega_X| - D(p \| u)$

(i) $H(X|Y) \leq H(X)$ [Conditioning reduces uncertainity]

(ii) $H(X_1 \ldots X_n) \leq \sum_{i=1}^{n} H(X_i)$

s.t. $H(X_1 \ldots X_n) = \sum H(X_i | X_1 \ldots X_{i-1})$

$\leq \sum H(X_i)$

Concavity of Entropy

$H(\lambda \cdot p + (1-\lambda) \cdot q)$

$\geq \lambda \cdot H(p) + (1-\lambda) H(q)$

$X \sim p$; $Y \sim q$; $b = 0$ w.p. λ

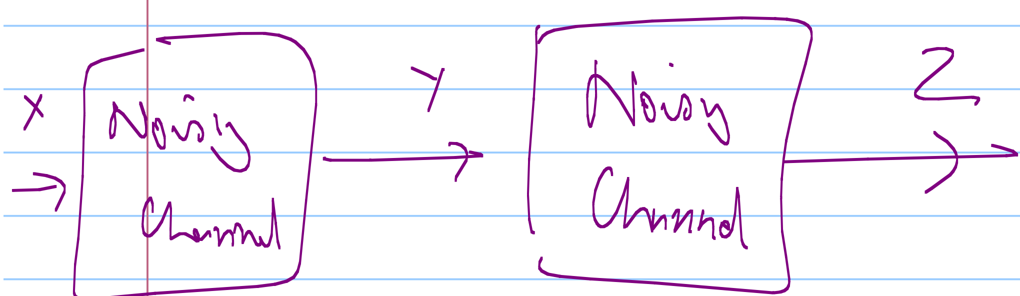
1 o.w.

$Z =$ X if $b=0$ & Y if $b=1$

$$H(Z) = H(\lambda p + (1-\lambda) q) \geq H(Z|b)$$

$$= \lambda H(X) + (1-\lambda) H(Y)$$

Data Processing Theorem



would expect

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

To formalize: "Markov Chains

$X, Y; Z$ form M.C $(X \to Y \to Z)$

if

$$P_{Z|(X,Y)}(z|(x,y)) = P_{Z|Y}(z|y)$$

Equivalently

$$P_{(X,Z)|Y}((x,z)|y) = P_{X|Y} \cdot P_{Z|Y}$$

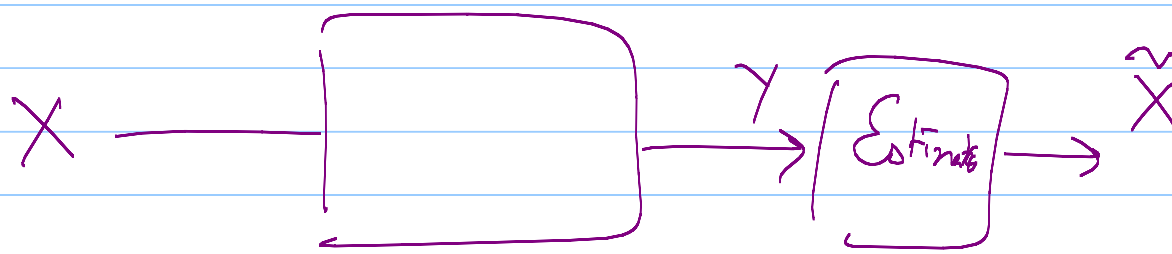
$$X \to Y \to Z \iff Z \to Y \to X$$

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

$$= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0}$$

$$\Rightarrow \quad I(X;Y) = I(X;Z) + I(X;Y|Z)$$
$$\geq I(X;Z)$$

Finally .... Fano's Lemma



$P_e = \Pr[X \neq \tilde{X}]$; I : indicator of event $\tilde{X} \neq X$

Fano's Lemma:

$$H(I) + P_e \cdot \log(|\Omega_X| - 1) \geq H(X|Y)$$

Corollary $\quad P_e \geq \dfrac{H(X|Y) - 1}{\log |\Omega_X|}$

Proof:

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

$$H(X \mid (I,Y))$$
$$= P_e \cdot H(X \mid (I=1, Y))$$
$$+ (1-P_e) \cdot \underbrace{H(X \mid (I=0, Y))}_{=0}$$
$$= P_e \cdot H(X \mid (I=1, Y))$$

$$H(X \mid (I=1, Y)) \le \log(|\Omega_x| - 1)$$

$$\le P_e \cdot \log(|\Omega_x| - 1)$$

Application :

$X = n$ random bits $= X_1 \dots X_n$

$$Y_i = X_i \quad \text{w.p.} \quad 1-p$$
$$= \bar{X_i} \quad \text{w.p.} \quad p$$

$$H(X \mid Y) = n \cdot H(p)$$

Fano's Lemma : $P_e \ge \dfrac{n \cdot H(p) - 1}{n}$

$$\approx H(p)$$