Lecture 3

Note Title 2/1/2016

Today

- Basic Quantities in Information Theory
  - Entropy, Conditional Entropy (already known)
  - Information, KL Divergence (new)
- Basic Relationships
  - Positivity of Information
  - Upper bound on Entropy
  - Convexity of Entropy
  - Data Processing Inequalities
  - Positivity of Divergence
  - Fano's Inequality ...

ERRATA: Proof of Kraft's Inequality ($\Leftarrow$)

$l_1 \ldots l_n$ pattern of prefix-free C

$\Leftarrow \sum_{i=1}^{n} 2^{-l_i} \leq 1$

Proof Claimed: Can just pick C greedily ...

But clearly wrong

Example: $l_1 = l_2 = n$; $l_3 = 1$

$C(1) = 000000$; $C(1) = 1111111$

$C(3) = ?$ Can't start with 0, Can't start with 1.

Fix: Assume $l_1 \leq l_2 \leq l_3 \ldots \leq l_n$

Now greedy works.

After assigning $C(i)$, claim # alive nodes at level l_i or deeper $= 1 - \sum_{j=}^{i} 2^{-l_j} > 0$

$\boxtimes$

Review

- $X \sim P_X = (p_1, \dots p_k)$ $\quad \Omega = \{1, \dots, k\}$
- $H(X) \triangleq \sum_{i=1}^{k} p_i \log \frac{1}{p_i} \triangleq \mathop{\mathbb{E}}_{x \sim P_X} \left[\log \frac{1}{P_x} \right]$
- $H(X|Y) \triangleq \mathop{\mathbb{E}}_{y \sim P_Y} \left[H(X|Y=y) \right]$
- Axioms

(?) • $H(X) \leq \log k$ (not yet proved!)

• $H(U_\Omega) = \log k$ ($U_\Omega \triangleq$ uniform dist. on Ω)

(?) • $H(X|Y) \leq H(X)$ (not yet proved!)

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

New Concepts

(Mutual) Information:

Suppose X, Y jointly distributed

$I(X;Y)$ = Amount of information that Y contains about X

Intuitive Examples:

$X \perp Y$: $I(X,Y) = 0$

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

$X = g(Y)$: $I(X,Y) = H(X)$

$Y = f(X)$: $I(X,Y) = H(Y)$

$X = (A,B)$
$Y = (A,C)$
$A \perp B \perp C$
} : $I(X,Y) = H(A)$

Formal Definition $I(X;Y) \triangleq H(X) - H(X|Y)$

$$I(X;Y|Z) \triangleq \mathop{\mathbb{E}}_{z \sim P_z}\left[I(X;Y \mid Z=z) \right]$$

Questions: • Is Information (as defined) non-negative?

(follows from unproven axiom (3))

• Is there some monotonicity to Information wrt conditioning?

Example: $X \perp Y \in \cup\{0,1\}$

$Z = X \oplus Y$

$$I(X;Y) = 0 \quad ; \quad I(X;Y|Z) = 1$$

$$I(X;Y|Z) > I(X;Y)$$

Example: $X = Y = Z \in \cup\{0,1\}$

$$I(X;Y) = 1$$

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

$$I(X;Y|Z) < I(X;Y)$$

Conclusion: No monotonicity.

Chain Rule For Information

$$I(X_1 \dots X_n ; Y)$$

$$= I(X_1 ; Y) + I(X_2 ; Y \mid X_1)$$

$$\vdots$$

$$+ \dots \quad I(X_n ; Y \mid X_1 \dots X_{n-1})$$

Now: for some proofs.

Mother of all inequalities

- Let P, Q be distributions of Ω

then $\underset{x \sim P}{E}\left[-\log \frac{Q(x)}{P(x)}\right] \geq 0$

What? Why? How?

Divergence Inequality

Intuition

$$\mathop{\mathbb{E}}_{x\sim p}\left[-\log \frac{1}{p(x)}\right] \approx \text{(optimal) length of compressing } X\sim p$$

$$\mathop{\mathbb{E}}_{x\sim p}\left[-\log \frac{1}{Q(x)}\right] \approx \text{length of some compression scheme for } X\sim p, \text{ one that pretends } X\sim Q$$

Inequality asserts

"Some" Compression $\geq$ "Minimum" Compression

Food for thought

Is $\mathop{\mathbb{E}}_{x\sim p}\left[-\log \frac{1}{p(x)}\right] \leq \mathop{\mathbb{E}}_{x\sim Q}\left[-\log \frac{1}{p(x)}\right]$?

Proof of Divergence Inequality

Key Ingredient: Jensen's Inequality

if $f: \mathbb{R} \to \mathbb{R}$ is nice & convex

$$\mathop{E}_{x}[f(x)] \geq f\left(\mathop{E}_{x}[x]\right)$$

(measure on x s.t. $\mathop{E}_{x}[x]$ exists;

$\mathop{E}_{x}[f(x)]$ exists etc.)

for us all true: x will have

finite support

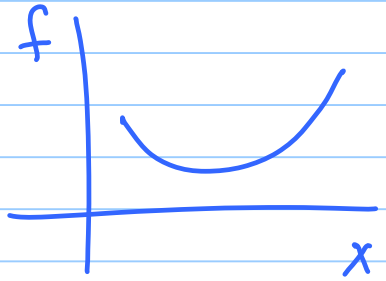
$f(x)$ bounded

Convex: $\forall\ x, y,\ \lambda \in [0,1]$

$$f(\lambda x + (1-\lambda) y) \geq \lambda f(x) + (1-\lambda) f(y)$$

Pictorially

Cup Shaped



For us:

$-\log x$ is convex.

Divergence Inequality.

$$\mathop{E}_{x\sim P}\left[-\log\frac{Q(x)}{P(x)}\right] \geq -\log\left[\mathop{E}_{x\sim P}\left[\frac{Q(x)}{P(x)}\right]\right]$$

$$= -\log\left[\sum_x \frac{P(x)\cdot Q(x)}{P(x)}\right]$$

$$= -\log\left[\sum_x Q(x)\right]$$

$$= -\log 1 = 0$$

⊠

Using Divergence Inequality

① $H(X) \leq \log \Omega$

$$\Leftrightarrow \mathop{\mathbb{E}}_{x \sim P}\left[-\log \frac{1}{P(x)}\right] \leq \mathop{\mathbb{E}}_{x \sim P}\left[-\log \frac{1}{\Omega}\right]$$

Let $U(x) = \frac{1}{\Omega} \quad \forall x \in \Omega$

$$\Leftrightarrow \mathop{\mathbb{E}}_{x \sim P}\left[-\log \frac{U(x)}{P(x)}\right] \geq 0 \qquad \boxtimes$$

(Divergence between P & Uniform)

② $H(X|Y) \leq H(X)$

$\Leftrightarrow H(X,Y) \leq H(X) + H(Y)$ [Chain Rule]

$$\Leftrightarrow \mathop{\mathbb{E}}_{(x,y) \sim P_{XY}}\left[-\log \frac{1}{P_{XY}(x,y)}\right] \leq \mathop{\mathbb{E}}_{(x,y)}\left[-\log \frac{1}{P_X(x)}\right] + \mathop{\mathbb{E}}_{(x,y)}\left[-\log \frac{1}{P_Y(y)}\right]$$

$$\Leftrightarrow \mathop{\mathbb{E}}_{(x,y) \sim P_{XY}}\left[-\log \frac{P_X(x) \cdot P_Y(y)}{P_{XY}(x,y)}\right]$$

Apply Divergence Inequality to

$$P = P_{XY} \quad \& \quad Q = P_X \times P_Y$$

$$\iff \mathop{\mathbb{E}}_{x,y \sim P} \left[-\log \frac{Q(x,y)}{P(x,y)} \right] \geq 0 \qquad \boxtimes$$

Kullback-Leibler Divergence aka Relative Entropy

$$D(P \| Q) \triangleq \mathop{\mathbb{E}}_{x \sim P} \left[-\log \frac{Q(x)}{P(x)} \right]$$

Key Measure of "distance between distributions"
?

$$D(P \| Q) \neq D(Q \| P)$$

$$D(P \| R) \leq D(P \| Q) + D(Q \| R) \; ?$$

(I think not ... exercise)

But key, key quantity! Why?

① Divergence Inequality : Miracle ?

② Chain Rule :

$$D(P_{XY} \| Q_{XY}) = D(P_X \| Q_X) + D(P_{Y|X} \| Q_{Y|X})$$

$$\geq \mathop{\mathbb{E}}_{x \sim X} \left[D(P_{Y|x} \| Q_{Y|x}) \right]$$

in general cumbersome.

But for i.i.d variable

$X_1 \dots X_n$ $\qquad X_i \sim P$?

$\qquad\qquad$ or $X_i \sim Q$?

$$D(P^n \| Q^n) = n \cdot D(P \| Q).$$

$$\mathbb{P}_r \left[x \sim P \text{ typical for } Q \right] \leq \exp(-D(P \| Q))$$

[will see later]

Markov Chains & Data Processing Inequality:

Notation: $X \to Y \to Z$

implies $(X \perp Z) \mid Y$

conditioned on Y, X & Z independent

Data Processing Inequality

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

Fano's Inequality (Entropy ⇒ Unpredictability)

Let $X \to Y \to \hat{X}$ & $P_e = Pr[X \neq \hat{X}]$

then $H(P_e) + P_e \log|\Omega| \geq H(X|Y)$

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

Proof: $E = 1$ if $\hat{X} \neq X$

$= 0$ o.w.

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

$\geq H(X|Y)$

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

$\blacksquare$