Lecture 13

Tuesday, November 04, 2014
10:30 AM

Conditional Entropy:

$$H(X|Y) = E_Y \, H_{P_{X|Y}}(X) = E_Y \, E_{X|Y} \log \frac{1}{P_{X|Y}(X|Y)}$$

$$= \boxed{E \log \frac{1}{P_{X|Y}(X|Y)}} \leftarrow \text{Towering Property}$$

Average conditional entropy, averaged over Y.

Joint Entropy: (same as entropy, but random variable is a vector)

$$H(X,Y) = E \log \frac{1}{P_{XY}(X,Y)}$$

Chain Rule:

$$H(X,Y) = E \log \frac{1}{p(X,Y)}$$

$$= E \log \frac{1}{p(X)\,p(Y|X)}$$

$$= E \log \frac{1}{p(X)} + E \log \frac{1}{p(Y|X)}$$

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

$$\Rightarrow \boxed{H(X,Y) \geq H(X)}$$

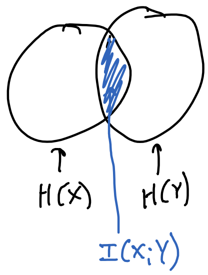
Property:

$$H(X,Y) \leq H(X) + H(Y)$$

Proof momentarily.

$$\Rightarrow \begin{array}{l} H(Y) \geq H(Y|X) \\ H(X) \geq H(X|Y) \end{array} \text{ by the chain.}$$

Only true because this an average quantity



Mutual Information:

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

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

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

Property: $0 \leq I(X;Y) \leq \min\{H(X), H(Y)\}$

Property: $0 \le I(X;Y) \le \min\{H(X), H(Y)\}$

Equality only if X is a function of Y or Y is a function of X (i.e. $H(X|Y)=0$ or $H(Y|X)=0$)

Relative entropy:
Kullback-Leibler divergence
K-L divergence
Divergence

$$d_{KL}(p,q) = E_p \log \frac{p(X)}{q(X)} = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

This is another metric of distance, but not a true distance.
- Not symmetric
- No triangle inequality.

Important to note that $d_{KL}(p,q)$ is discontinuous w.r.t. $\|p-q\|_1$

If $p \not\ll q$ then $d_{KL}(p,q)=\infty$, even though $\|p-q\|_1$ may be arbitrarily small.

K-L divergence plays many important roles.
- Fundamental limits of detection theory
- Penalty for encoding $\sim q$ when true distribution is $\sim p$.
- Pinsker inequality: $\|p-q\|_1 \le \sqrt{2 d_{KL}(p,q)}$

Property: $d_{KL}(p,q) \ge 0$ with equality iff $p=q$

Proof:

$$\begin{aligned}
d_{KL}(p,q) &= E_p \log \frac{p(X)}{q(X)} \\
&= -E_p \log \frac{q(X)}{p(X)} \\
&\ge -\log E_p \frac{q(X)}{p(X)} \qquad \text{Jensen's Ineq.} \\
&= -\log \sum_x \cancel{p(x)} \frac{q(x)}{\cancel{p(x)}} \\
&= -\log \sum_{x \in \mathcal{X}_p} q(x) \\
&\ge -\log 1 \\
&= 0
\end{aligned}$$

$-\log$ is strictly convex

Equality iff $\frac{q(X)}{p(X)}$ is deterministic.

$\Rightarrow p(x) = \alpha q(x) \quad \forall x$

Combine

$\alpha = 1$

$\Rightarrow$ Equality iff $\boxed{p=q}$

Mutual Information as Relative Entropy:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$
$$= E\left(\log\frac{1}{p(X)} + \log\frac{1}{p(Y)} - \log\frac{1}{p(X,Y)}\right)$$
$$= \underset{p(x,y)}{E} \log\frac{p(X,Y)}{p(X)p(Y)}$$
$$= d_{KL}\left(p(x,y),\ p(x)p(y)\right)$$

Measures how far the distribution is from being independent

Consequence: $I(X;Y) \geq 0$ with equality iff $X \perp\!\!\!\perp Y$

independent

Conditional Mutual Information:

$$I(X;Y|Z) = \underset{Z}{E} \underset{p(x,y|z)}{I}(X;Y)$$

Properties (proven by simple algebra)

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

2.) Chain rule: $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$

Same definition for M.I.
Interpret as a vector

Data Processing Inequality:

If $X - Y - Z$ form a Markov chain.

(i.e. X and Z are conditionally independent given $Y \Leftrightarrow p(x,y,z) = p(x,y)p(z|y)$
$= p(y)p(x|y)p(z|y)$)

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

Proof: Notice that $I(X;Z|Y)=0$ ← Equivalent to Markov chain condition.

$$\Rightarrow I(X;Y) = I(X;Y) + I(X;Z|Y) \quad \text{Markovity}$$
$$= I(X;YZ) \quad \text{Chain rule}$$
$$= I(X;Z) + I(X;Y|Z) \quad \text{chain rule}$$
$$\geq I(X;Z) \quad \text{non-negativity.}$$

$$\boxed{I(X;Y) \geq I(X;f(Y)) \quad \text{because} \quad X - Y - f(Y)}$$

Notice: $I(X;Y) \geq I(X;Y|Z)$ from same proof

Not true in general.

Also notice that if $X \perp\!\!\!\perp Z$ then $I(X;Y) \leq I(X;Y|Z)$

Fano's inequality:

Entropy continuously goes to zero as uncertainty becomes less and less.
Exact bound depends on cardinality, which must be accounted for in proof.

Let p_e be the probability of error in estimating X from Y.

$$H(X|Y) \leq h(p_e) + p_e \log(|\mathcal{X}|-1)$$ ← Maximized by spreading probability uniformly on remaining support

$$\leq 1 + p_e \log |\mathcal{X}|$$

Application: Lower bound on rate for lossless compression:

$Th^{\underline{m}}$ Minimum data needed for compressing an iid signal $\sim P_X$ is $H(X)$.

Lower bound: Denote $X_1, X_2, \ldots X_n$ as X^n

Assume encoding $X^n \to M \to \hat{X}^n$ where $M \in [2^{nR}] \triangleq \{1, 2, \ldots, 2^{nR}\}$

and $P(X^n \neq \hat{X}^n) < \varepsilon$ for arbitrarily small $\varepsilon > 0$.

$$nR \geq H(M) \quad \text{Entropy upper bound}$$
$$\geq I(X^n;M) \quad \text{Def., non-negativity}$$
$$\geq I(X^n;\hat{X}^n) \quad \text{Data Processing Ineq.}$$
$$= H(X^n) - H(X^n|\hat{X}^n) \quad \text{Def.}$$

$\to 0$

$$\geq I(X^n ; \hat{X}^n) \qquad \text{Data Processing Ineq.}$$
$$= H(X^n) - H(X^n | \hat{X}^n) \qquad \text{Def.}$$
$$= nH(X) - H(X^n|\hat{X}^n) \qquad \text{Prob. set} - \text{IID}$$
$$\geq nH(X) - h(\varepsilon) - \varepsilon \log |X^n| \qquad \text{Fano's Ineq.}$$
$$= n\left(H(X) - \varepsilon \log |X| - \frac{h(\varepsilon)}{n}\right)$$

$$\Rightarrow R \geq H(X) - \varepsilon \log |X| - \frac{h(\varepsilon)}{n}$$

$\to 0$

Quantum Information and Entropy: Ch. 11

von Neumann Entropy : $H(A) = -\mathrm{Tr}\,(\rho^A \log \rho^A)$

Notice that if ρ is diagonal this is Shannon entropy.

Canonical Decomposition: $\rho^A = \sum_x p(x)\, |\varphi_x\rangle\langle\varphi_x|^A$

orthonormal basis

Shannon entropy of $p(x)$ = von Neumann entropy.

This implies an ensemble preparation $\{(p(x), |\varphi_x\rangle)\}$

For any ensemble:

If $|\varphi_x\rangle$ are orthogonal basis : $H_S = H_{VN}$

If $|\varphi_x\rangle$ are not orthogonal : $H_S > H_{VN}$

deterministic

Proof later: Think of the ensemble as a function

$$H(X) \geq H(f(X))$$

Example of non-orthogonal ensemble:

$$\left\{ \left(\tfrac{1}{4}, |0\rangle\right), \left(\tfrac{1}{4}, |1\rangle\right), \left(\tfrac{1}{4}, |+\rangle\right), \left(\tfrac{1}{4}, |-\rangle\right) \right\}$$

$\rho = \pi \qquad H(\rho) = 1$ bit

$H(X) = 2$ bits

$\boxed{\text{(von Neumann) Entropy gives the fundamental limit for (Schumacher) compression.}}$

Properties: 1) $H(\rho) \geq 0$ with equality iff ρ is a pure state

Properties :

1.) $H(\rho) \geq 0$ with equality iff ρ is a pure state

2.) $H(\rho) \leq \log D$

3.) Concavity : $H(\rho) \geq \sum_x p(x) H(\rho_x)$ Proof later

4.) Unitary invariant : $H(U\rho U^\dagger) = H(\rho)$

isometric

- Think of this like a classical one-to-one mapping.

von Neumann entropy and measurement

$$H(\rho) = \min_{\substack{\Lambda_y \\ \text{rank } 1}} H_s\left(\operatorname{Tr}(\Lambda_y \rho)\right)$$

Distribution induced by measurement

Shannon entropy

How do we define conditional entropy ?

No natural notion of conditional density.

Unless conditioning on classical information

This is the first major anomaly of quantum information theory !

$H(B|A) = ?$ given ρ^{AB}