Lecture 13
Tuesday, November 04, 2014
10:30 AM

Conditional Entropy:

\[ H(X|Y) = E_{Y \sim p_Y} E_{X \sim p_{X|Y}} \log \frac{1}{p_{X|Y}(X|Y)} \]

\[ = E \log \frac{1}{p_{X|Y}(X|Y)} \]

Average conditional entropy, averaged over Y.

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

\[ H(X,Y) = E \log \frac{1}{p_{X,Y}(X,Y)} \]

Chain Rule:

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

\[ = E \log \frac{1}{p(X)p(Y|X)} \]

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

\[ = H(X) + H(Y|X) \]

\[ \Rightarrow H(X,Y) \geq H(X) \]

Property:

\[ H(X,Y) \leq H(X) + H(Y) \]

Proof momentarily.

Mutual Information:

\[ I(X;Y) = H(X) + H(Y) - H(X,Y) \]

\[ = H(X) - H(X|Y) \]

\[ = H(Y) - H(Y|X) \]

Property:

\[ D = I(X;Y) \leq \min \{ H(X), H(Y) \} \]
Property: $D = I(X;Y) = \min \{ H(X), H(Y) \}$

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

Relative entropy:
Kullback-Leibler divergence
KL 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. $\|p-g\|_1$.
If $p \neq q$, then $d_{KL}(p, q) = \infty$, even though $\|p-g\|_1$, may be arbitrarily small

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

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

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

\[ = - E_p \log \frac{q(x)}{p(x)} \geq - \log E_p \frac{q(x)}{p(x)} \]

Jensen's Ineq.

\[ = - \log \sum_x p(x) \frac{q(x)}{p(x)} \]

\[ = - \log \sum_x q(x) \]

\[ \geq - \log 1 = 0 \]

Combine

\[ \Rightarrow \text{Equality iff } \sqrt{p = q} \]
Mutual Information as Relative Entropy:

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

Measures how far the distribution is from being independent

\[ \text{Consequences: } I(X;Y) \geq 0 \text{ with equality iff } X \perp \perp Y \text{ independent} \]

Conditional Mutual Information:

\[ I(X_i;Y \mid Z) = \mathbb{E}_{p(x_i)} I(X_i;Y) \]

Properties (proven by simple algebra)

1. \[ I(X_i;Y \mid Z) = H(X_i \mid Z) + H(Y \mid Z) - H(X,Y \mid Z) \]

2. Chain rule: \[ I(X_i; Y \mid Z) = I(X_i; Z) + I(X_i; Y \mid Z) \]

Data Processing Inequality:

If \( X \rightarrow Y \rightarrow Z \) form a Markov chain.

(i.e. \( X \) and \( Z \) are conditionally independent given \( Y \leftrightarrow \)
\[ p(x,y,z) = p(x,y)p(z \mid y) \]
\[ = p(y)p(x \mid y)p(z \mid y) \]

then \[ I(X; Y) \geq I(X; Z) \]
Proof: Notice that \( I(X; Z|Y) = 0 \) ← Equivalent to Markov chain condition.

\[
\Rightarrow I(X; Y) = I(X; Y) + I(X; Z) = I(X; Y | Z) = I(X; Z) + I(X; Y | Z) = I(X; Z)
\]

\( I(X; Y) \geq I(X; f(Y)) \) because \( X-Y-f(Y) \)

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

Not true in general

Also notice that if \( X \perp Y \) 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 \( Y \).

\[
H(X | Y) \leq h(p_e) + p_e \log (1 \times 1 - 1) \leftarrow \text{Maximized by spreading probability uniformly on remaining support}
\]

\[
\leq 1 + p_e \log |X|
\]

**Application:** Lower bound on rate for lossless compression:

The 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 \( X^n \rightarrow M \rightarrow \hat{X}^n \) where \( M \in [2^R]^n \equiv \{1, 2, \ldots, 2^R\} \)

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.}, \text{ 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.}
\]
\[ I(X^n; Y^n) \quad \text{Data Processing Ineq.} \]
\[ = H(X^n) - H(X^n|Y^n) \quad \text{Def.} \]
\[ = nH(X) - H(X^n|Y^n) \quad \text{Prob. set - IID} \]
\[ \geq nH(X) - h(\epsilon) - \epsilon \log |X^n| \quad \text{Fano's Ineq.} \]
\[ = n \left( H(X) - \epsilon \log |X| - \frac{h(\epsilon)}{n} \right) \]

**Quantum Information and Entropy**: Ch 11

**von Neumann Entropy**: \[ H(A) = - \text{Tr}(\rho^A \log \rho^A) \]

Notice that if \( \rho \) is diagonal this is Shannon entropy.

**Canonical Decomposition**: \( \rho^A = \sum \rho(x) |x\rangle \langle x|^A \)

For any ensemble:
- If \( |\psi\rangle \) are orthogonal basis: \( H_S = H_{\text{VN}} \)
- If \( |\psi\rangle \) are not orthogonal: \( H_S > H_{\text{VN}} \)

Proof later: Think of the ensemble as a function \( H(\mathbf{x}) \geq H(f(\mathbf{x})) \)

**Example of non-orthogonal ensemble**:
\[ \{ (\frac{1}{\sqrt{4}}, 1|\rangle, (\frac{1}{\sqrt{4}}, 1\rangle, (\frac{1}{\sqrt{4}}, 1\rangle, (\frac{1}{\sqrt{4}}, 1\rangle) \} \]
\[ \rho = \Pi \quad H(\rho) = 1 \text{ bit} \]
\[ H(X) = 2 \text{ bits} \]

*(von Neumann) Entropy gives the fundamental limit for (Schumacher) compression.*
Properties:

1. \( H(\rho) \geq 0 \) with equality iff \( \rho \) 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: \( H(U^\dagger \rho U) = H(\rho) \) isometric

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

Von Neumann entropy and measurement:

\[
H(\rho) = \min_{\Lambda : \text{rank } 1} \quad H_\Sigma \left( \text{Tr}(\Lambda \rho) \right)
\]

\( \Lambda \) - Distribution induced by measurement
\( H_\Sigma \) - 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!\( \)

\[
H(B|A) = ? \quad \text{given } \rho^{AB}
\]