6.441 Transmission of Information Feb 14, 2006

Lecture 3

Lecturer: Madhu Sudan *Scribe: Daniel Kim (dskim116)*

1 Today's outline

- Property of information and entropy
- New notions: KL divergence, Markov chains
- results: non-negativity of mutual information, data processing inequality, Fano's inequality

2 Lecture 2's Review

Let us define marginal and joint distributions. $p(x)$ denotes a marginal probability that $X = x$, $p(y)$ denotes a marginal probability that $Y = y$ and $p(x, y)$ denotes a joint probability that $X = x$ and $Y = y$.

- **Entropy**:

$$H(X) = -\sum_x p(x) \log p(x)$$

- **Conditional entropy**:

$$H(X|Y) = \sum_{y \in \Omega_y} p_y(y) H(X|Y = y) = \sum_{x,y} p(x, y) \log \frac{p_y(y)}{p(x, y)}$$

- **Mutual information**:

$$I(x, y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x) \cdot p(y)} = I(y, x)$$

- **Chain rule**:

$$H(x, y) = H(x) + H(y|x)$$

Applying this iteratively, we derive:

$$\begin{aligned} H(x_1, x_2, \cdots, x_n) &= H(x_1) + H(x_2|x_1) + \cdots \\ &= \sum_{i=1}^{n} H(x_i|x_1, x_2, \cdots, x_{i-1}) \end{aligned}$$

3 Is $I(X,Y) \geq 0$?

Proving $I(X,Y) \geq 0$ is equivalent to proving that $H(X|Y) \leq H(X)$.

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

with equality when x and y are independent because:

$$p(x,y) = p(x) \cdot p(y) \Longrightarrow I(x,y) = 0$$

Before we prove Claim 3, let us define function convexity and state Jensen's Inequality.

Definition 1 *Function f is **convex** when either of following conditions holds:*

$$\begin{cases} f & : \mathbb{R} \to \mathbb{R} \text{ is convex if } f''(x) \geq 0 \ \forall x \\ f & : \mathbb{R} \to \mathbb{R} \text{ is strictly convex if } f''(x) > 0 \ \forall x \end{cases}$$

For example, x^2, e^x and $-\log x$ are convex functions.

Theorem 2 *Jensen's Inequality: $E[f(z)] \geq f[E[z]]$ provided f is convex.*

Now, here is the claim.

Claim 3 $E_{(x,y)\sim p}\left[\log \frac{p(x,y)}{q(x,y)}\right] \geq 0$ *with equality when $p(x,y) = q(x,y)$.*

Proof Let us define new variable $z = \frac{q(x,y)}{p(x,y)}$. Then,

$$\begin{aligned} E_{(x,y)\sim p}\left[\log \frac{p(x,y)}{q(x,y)}\right] &= E_z\left[\log \frac{1}{z}\right] \\ &= E\left[-\log z\right] \\ &\geq -\log E[z] (\because \text{ Jensen's Inequality}) \\ &= -\log\left[E_{(x,y)}\left[\frac{q(x,y)}{p(x,y)}\right]\right] \\ &= -\log\left[\sum_{x,y} p(x,y) \frac{q(x,y)}{p(x,y)}\right] = -\log\left[\sum_{x,y} q(x,y)\right] = -\log 1 = 0. \end{aligned}$$

■

Here, note that $E\left[\log \frac{p(x,y)}{q(x,y)}\right]$ shows how much similarity $q(x,y)$ and $p(x,y)$ share.

4 Relative Entropy

Definition 4 *The **relative entropy** or **Kullback-Liebler distance** between two probability mass functions $p(z)$ and $q(z)$ is defined as:*

$$D(p||q) = \sum_z p(z) \log \frac{p(z)}{q(z)}.$$

4.1 Example

Let us consider the case when $x \in \{0, 1\}$ with following distributions:

$$p : X = \begin{cases} 0 & \text{with probability } 1 \\ 1 & \text{with probability } 0 \end{cases}$$

$$q : X = \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases}$$

Based on the above scenario, we get $D(p||q) = \log 2$ and $D(q||p) = \infty$.

4.2 Compression motivation example

Let us consider our satellite example with $x \sim p = (p_1, p_2, \cdots, p_N)$. Optimal compression should require $\left\lceil \log \frac{1}{p_i} \right\rceil$ bits long string. x with distribution q would require $\left\lceil \log \frac{1}{q} \right\rceil$ bits long string. By definition, average inefficiency of compressing by q when given distribution is p is $D(p||q)$.

4.3 Basic Property

- $D(p||q) \geq 0$ with equality only when $p = q$
- $I(X, Y) = D(p(x, y)||p(x) \cdot p(y)) \geq 0$
- $I(X, Y) = H(X) - H(X|Y) \geq 0$ ($\because$ conditioning reduces entropy)
- $H(X_1, X_2, \cdots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|(X1, X2)) + \cdots$
  Substituting the following:

$$\begin{aligned} &H(X_1) \leq H(X_1) \\ &H(X_2|X_1) \leq H(X_2) \\ &H(X_3|(X_1, X_2)) \leq H(X_3) \\ &\vdots \end{aligned}$$

  we can reduce it to:

$$\therefore H(X_1, X_2, \cdots X_n) \leq \sum_n H(X_n).$$

- $H(x) = \log(|\Omega_x|) - D(p||U)$ where U is uniform distribution on Ω_x. Because $D(p||q) \geq 0$, we derive that $H(x) \leq \log(|\Omega_x|)$.

4.4 Is entropy concave?

In order to prove whether entropy is concave or not, we need to show following:

$$H(\lambda p + (1 - \lambda)q) \geq \lambda H(p) + (1 - \lambda)H(q) \tag{1}$$

Proof Let us assume that $x \sim p$ and $y \sim q$ on set Ω. Also, let us define another variable b with following distribution.

$$b = \begin{cases} 0 & \text{with probability } \lambda \\ 1 & \text{with probability } 1-\lambda \end{cases}$$

Using these variables, let us define a new variable Z with following distribution :

$$Z : \text{if } b = 0 \text{ then } x; \text{ else } y.$$

Then, the left-hand side of Equation (1) is reduced to $H(Z)$ and the right-hand side of Equation (1) is reduced to $H(Z|b)$. Because conditioning reduces the uncertainty, $H(Z) \geq H(Z|b)$. This proves that the entropy is concave. ■

5 Data Processing Inequality (Markov Chain)

Let us consider three states, X, Y, and Z. $X \to Y \to Z$ forms a Markov chain if and only if X and Z are conditionally independent given Y. Let us put the definition into mathematical term. $X \to Y \to Z$ forms a Markov chain if and only if either of following conditions is true:

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

Also, $X \to Y \to Z \iff Z \to Y \to X$. Now let us consider the property of Markov chain.

Claim 5 *If $X \to Y \to Z$, then $I(X,Z) \leq I(X,Y)$.*

Proof

$$\begin{aligned} I(X,(Y,Z)) &= I(X,Z) + I((X,Y)|Z) \\ &= I(X,Y) + I((X,Z)|Y) \end{aligned}$$

Substituting the fact that $I((X,Z)|Y) = 0$ and $I((X,Y)|Z) \geq 0$, we get $I(x,z) \leq I(x,y)$. ■

6 Fano's Inequality

Let E be an event and let P_e denote the probability when $X \neq \widetilde{X}$.

Theorem 6 *When $H(X|Y)$ is large,*

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

Proof

$$
\begin{aligned}
H((E,X)|Y) &= H(X|Y) + H(E|(X,Y)) \\
&= H(E|Y) + H(X|(E,Y))
\end{aligned}
$$

Let us take a look at each term:

$$
\begin{aligned}
&H(E|(X,Y)) = 0 \\
&H(E|Y) \leq H(P_e)
\end{aligned}
$$

$$
\begin{aligned}
H(X|(E,Y)) &= P_e \cdot H(X|(E=1,Y)) + (1-P_e)H(X|(E=0,Y)) \\
&= P_e \cdot H(X|(E=1,Y)) \\
&\leq P_e \log(|\Omega_x| - 1)
\end{aligned}
$$

Substituting these into original equation, we prove the theorem. ■