COS597D: Information Theory in Computer Science	September 26, 2011

Lecture 3

Lecturer: Mark Braverman *Scribe: Pawel Przytycki**

Theorem 1 (Fano's Inequality)**.** *Let $\hat{X}$ be an estimator for X such that $P_e = Pr(X = \hat{X})$ then $H(P_e) + P_e log|\chi| \geq H(X|\hat{X}) \geq H(X|Y)$.*

Proof [of the first part of the inequality]

Define $\mathcal{E} = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{if } \hat{X} = X \end{cases}$

$H(\mathcal{E}X|\hat{X}) = H(X|\hat{X}) + H(\mathcal{E}|X\hat{X}) = H(X|\hat{X})$, since $\mathcal{E}$ is completely determined by $X\hat{X}$,

$H(\mathcal{E}X|\hat{X}) = H(\mathcal{E}|\hat{X}) + H(X|\mathcal{E}\hat{X}) \leq H(\mathcal{E}) + (1-P_e)H(X|\hat{X}, \mathcal{E} = 0) + P_e H(X|\hat{X}, \mathcal{E} = 1) \leq H(P_e) + P_e log|\mathcal{X}|$.

■

1 Relative Entropy

The *relative entropy*, also known as the *Kullback-Leibler divergence*, between two probability distributions on a random variable is a measure of the distance between them. Formally, given two probability distributions $p(x)$ and $q(x)$ over a discrete random variable X, the relative entropy given by $D(p||q)$ is defined as follows:

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

In the definition above $0 \log \frac{0}{0} = 0$, $\log \frac{0}{q} = 0$, and $p \log \frac{1}{0} = \infty$.

Example 2. $D(p||p) = 0$.

Example 3. *Consider a random variable X with the law $q(x)$. We assume nothing about $q(x)$. Now consider a set $E \subseteq \mathcal{X}$ and define $p(x)$ to be the law of $X|_{X \in E}$. The divergence between p and q:*

Solution

$$\begin{aligned} D(p||q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x \in E} p(x) \log \frac{q(x|x \in E)}{q(x|x \in E) Pr_q[X \in E]} \\ &= \sum_{x \in E} p(x) \log \frac{1}{Pr_q[X \in E]} \\ &= \log \frac{1}{Pr[E]}. \end{aligned}$$

In the extreme case with $E = \mathcal{X}$, the two laws p and q are identical with a divergence of 0.

■

We will henceforth refer to relative entropy or Kullback-Leibler divergence as divergence.

*Based on lecture notes by Anup Rao and Prasang Upadhyaya

1.1 Properties of Divergence

1. Divergence is not symmetric. That is, $D(p||q) = D(q||p)$ is not necessarily true. For example, unlike $D(p||q)$, $D(q||p) = \infty$ in the example mentioned in the previous section, if $\exists x \in \mathcal{X} \setminus E : q(x) > 0$.

2. Divergence is always non-negative. This is because of the following:

$$\begin{aligned} D(p||q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= -\sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \\ &= -\mathbb{E}\left[\log \frac{q}{p}\right] \\ &\geq -\log\left(\mathbb{E}\left[\frac{q}{p}\right]\right) \\ &= -\log\left(\sum_{x \in \mathcal{X}} p(x) \frac{q(x)}{p(x)}\right) \\ &= 0, \end{aligned}$$

   where the inequality follows by the convexity of $-\log x$.

3. Divergence is a convex function on the domain of probability distributions.

   Theorem 4 (Log-sum Inequality)**.** *If $a_1, \ldots, a_n, b_1, \ldots, b_n$ are non-negative numbers, then* $\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq (\sum_{i=1}^n a_i) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$

   Lemma 5 (Convexity of divergence)**.** *Let p_1, q_1 and p_2, q_2 be probability distributions over a random variable X and $\forall \lambda \in (0, 1)$ define*

$$\begin{aligned} p &= \lambda p_1 + (1 - \lambda) p_2 \\ q &= \lambda q_1 + (1 - \lambda) q_2 \end{aligned}$$

   Then, $D(p||q) \leq \lambda D(p_1||q_1) + (1 - \lambda) D(p_2||q_2)$.

1.2 Relationship of Divergence with Entropy

Intuitively, the entropy of a random variable X with a probability distribution $p(x)$ is related to how much $p(x)$ diverges from the uniform distribution on the support of X. The more $p(x)$ diverges the lesser its entropy and vice versa. Formally,

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= \log |\mathcal{X}| - \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{\frac{1}{|\mathcal{X}|}} \\ &= \log |\mathcal{X}| - D(p||uniform) \end{aligned}$$

1.3 Conditional Divergence

Given the joint probability distributions $p(x,y)$ and $q(x,y)$of two discrete random variables X and Y, the conditional divergence between two conditional probability distributions $p(y|x)$ and $q(y|x)$ is obtained by computing the divergence between p and q for all possible values of $x \in \mathcal{X}$ and then averaging over these values of x. Formally,

$$D(p(y|x)||q(y|x)) = \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

Given the above definition we can prove the following chain rule about divergence of joint probability distribution functions.

Lemma 6 (Chain Rule)**.**

$$D\left(p(x,y)||q(x,y)\right) = D\left(p(x)||q(x)\right) + D\left(p(y|x)||q(y|x)\right)$$

Proof

$$\begin{aligned}
D\left(p(x,y)||q(x,y)\right) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)} \\
&= D(p(x)||q(x)) + \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\
&= D\left(p(x)||q(x)\right) + D\left(p(y|x)||q(y|x)\right)
\end{aligned}$$

■

2 Mutual Information

Mutual information is a measure of how correlated two random variables X and Y are such that the more independent the variables are the lesser is their mutual information. Formally,

$$\begin{aligned}
I(X;Y) &= D(p(x,y)||p(x)p(y)) \\
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log p(x,y) - \sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y) \\
&= -H(X,Y) + H(X) + H(Y) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X)
\end{aligned}$$

2.1 Conditional Mutual Information

We define the conditional mutual information when conditioned upon a third random variable Z to be

$$\begin{aligned}
I(X;Y|Z) &= \mathbb{E}_z[I(X;Y|Z=z)] \\
&= H(X|Z) - H(X|YZ)
\end{aligned}$$

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \sum_y p(y) \sum_x p(x|y) \log \frac{p(x,y)/p(y)}{p(x)} = E_y D(p(x|y)||p(x))$$

Example 7. *X,Y,Z uniform, conditioned on X+Y+Z = 0 mod 2*
$I(X;Y) = H(X) - H(X|Y) = 0;$
$I(X;YZ) = H(X) - H(X|YZ) = 1;$
$I(X;Y|Z) = H(X|Z) - H(X|YZ) = 1.$

Conditioning can decrease (or eliminate) or increase mutual information:

Example 8. *$X = x_1x_2$, $Y = y_1y_2$, random bits s.t. $x_1 \oplus x_2 = y_1 \oplus y_2$. Let $Z := x_1 \oplus x_2 = y_1 \oplus y_2$, then*
$I(X;Y) = H(X) - H(X|Y) = 2 - 1 = 1;$
$I(X;Y|Z) = H(X|Z) - H(X|YZ) = 1 - 1 = 0.$

Lemma 9 (Chain Rule)**.** $I(XY;Z) = I(X;Z) + I(Y;Z|X)$

Proof

$$\begin{aligned} I(XY;Z) &= H(XY) - H(XY|Z) \\ &= H(X) + H(Y|X) - H(X|Z) - H(Y|XZ) \\ &= I(X;Z) + I(Y;Z|X) \end{aligned}$$

■

2.2 Convexity/Concavity of Mutual Information

Let (X,Y) have a joint probability distribution $p(x,y) = p(x)p(y|x)$. Write $\alpha = \alpha(x) = p(x)$ and $\pi = \pi(x,y) = p(y|x)$. Then the pair (α, π) specifies the distribution $p(x,y)$.

Lemma 10 (Mutual information is concave in p)**.**
Let I_1 be $I(X;Y)$ where $(X,Y) \sim (\alpha_1, \pi)$,
let I_2 be $I(X;Y)$ where $(X,Y) \sim (\alpha_2, \pi)$,
let I be $I(X;Y)$ where $(X,Y) \sim (\lambda\alpha_1 + (1-\lambda)\alpha_2, \pi)$, for some $0 \le \lambda \le 1$.
then $I \ge \lambda I_1 + (1-\lambda)I_2$.

Proof Let S be a B_λ random variable such that S is 1 with probability λ and and 0 with probability $1-\lambda$. If $S = 1$ we select X using α_1, and otherwise we select X using α_2. In both cases, we select Y conditioned on X using π. Note that $I(X;Y) = I$, and that conditioned on X, Y and S are independent.
$I(SX;Y) = I(X;Y) + I(S;Y|X) = I;$
$I(SX;Y) = I(S;Y) + I(X;Y|S) \ge I(X;Y|S) = \lambda I(X;Y|S=1) + (1-\lambda)I(X;Y|S=0) = \lambda I_1 + (1-\lambda)I_2.$
■