CS861: Theoretical Foundations of Machine Learning Lecture 11 - 09/27/2024
University of Wisconsin–Madison, Fall 2024

Lecture 11: Information Theory Review, Fano's Method

Lecturer: Kirthevasan Kandasamy *Scribed by: Thomas Zeng, Daisuke Yamada*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

Previously, we defined *Le Cam's method* and saw examples where we can reduce point estimation to binary hypothesis testing to establish the lower bound. However, we discussed how Le Cam's method is insufficient beyond binary hypothesis testing. In the previous lecture, we started a brief review of Information Theory and set the stage for Fano's method which will be more appropriate for the lower bounds of our interest. In this lecture, we wrap up the review, prove *Fano's inequality*, and apply it to define *Fano's method*.

1 Information Theory Review

1.1 Entropy

Definition 1 (Entropy, Joint Entropy, and Conditional Entropy)**.** *Let X be an RV with distribution P.*

$$H(X) = \mathbb{E}_{x\sim P}[-\log p(x)],$$

$$H(X,Y) := \mathbb{E}_{x,y\sim P_{XY}}[-\log p(x,y)]$$
$$H(X|Y) := \mathbb{E}_{x,y\sim P_{XY}}[-\log p(x|y)].$$

The conditional entropy captures how much information is left in X after knowing $Y = y$. That is,

$$H(X|Y=y) = -\sum_{x\in\mathcal{X}} p(x|y)\log(p(x|y)).$$

We also see

$$\begin{aligned}\mathbb{E}_{y\sim Y}[H(X|Y=y)] &= \sum_{y\in\mathcal{Y}} p(y)H(X|Y=y)\\ &= \sum_{y\in\mathcal{Y}} -p(x,y)\log p(x|y)\\ &= H(X|Y).\end{aligned}$$

Lemma 1 (Chain Rule for Entropy)**.**

$$H(X_1,...X_n) = \sum_{i=1}^{n} H(X_i|X_1...X_{i-1}) \tag{1}$$

$$H(X_1,...X_n|Y) = \sum_{i=1}^{n} H(X_i|X_1...X_{i-1},Y) \tag{2}$$

Proof *We will prove (1) by induction on $n \geq 2$. Note that,*

$$p(x_1, x_2) = p(x_1)p(x_2|x_1) \Rightarrow \log(p(x_1, x_2)) = \log(p(x_1)) + \log(p(x_2|x_1))$$

Taking expectation with respect to X_1, X_2, we get

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1).$$

Then, for the inductive case, we obtain

$$\begin{aligned} H(X_1, , X_n) &= H(X_1, \ldots, X_{n-1}) + H(X_n|X_1, , X_{n-1}) \\ &= H(X_1, \ldots, X_{n-2}) + H(X_{n-1}|X_1, \ldots, X_{n-2}) + H(X_n|X_1, \ldots, X_{n-1}) \\ &= \ldots\ldots \\ &= \sum_{i=1}^{n} H(X_i|X_1 \ldots X_{i-1}), \end{aligned}$$

which proves (1). The proof of (2) will follow a similar argument.

□

Definition 2 (KL divergence of distributions P and Q)**.**

$$KL\left(P, Q\right) = E_{X \sim P}\left[\log\left(\frac{p\left(X\right)}{q\left(X\right)}\right)\right]$$

1.2 Mutual Information

Definition 3 (Mutual Information (MI))**.** *Let X, Y be rv's with joint distribution P_{XY} and marginals P_X, P_Y. Then, MI is:*

$$I(X;Y) = KL(P_{XY}, P_X \times P_Y) = E_{P_{XY}}\left[\log\left(\frac{p(x,y)}{p(x)p(y)}\right)\right]$$

We will use the following properties:

1.) $KL(P, Q) \geq 0$ with equality iff $P = Q$

2.) $I(X;Y) \geq 0$ with equality iff $X \perp\!\!\!\perp Y$

3.) (Conditioning reduces entropy)$H(X|Y) \leq H(X)$

4.) (Symmetry) $I(X;Y) = I(Y;X)$

5.) $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

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

7.) $I(X;X) = H(X)$

Proof

1.) Apply Jensen's inequality

2.) Follows from Property (1)

3.) Follows from Property (5)

5.) We have

$$\begin{aligned} I(X;Y) &= \mathbb{E}_{X,Y}\left[\log \frac{P(X,Y)}{P(X)P(Y)}\right] \\ &= \mathbb{E}_{X,Y}\left[\log \frac{P(X)P(Y|X)}{P(X)P(Y)}\right] \\ &= -\mathbb{E}_{Y}[\log P(Y)] + \mathbb{E}_{X,Y}[\log P(Y|X)] \\ &= H(Y) - H(Y|X) \end{aligned}$$

6.) By Chain Rule (Lemma 1), we have

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

7.) We have

$$\begin{aligned} I(X;X) &= H(X) - H(X|X) \\ &= H(X) - 0 \end{aligned}$$

□

Definition 4 (Conditional Mutual Information)**.**

$$\begin{aligned} I(X;Y|Z) &:= H(X|Z) - H(X|Y,Z) \\ &= \mathbb{E}_{X,Y,Z}\left[\log \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}\right] \end{aligned}$$

Lemma 2 (Chain Rule for Mutual Information)**.**

$$I(X_1,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1,\ldots,X_{i-1})$$

Proof We prove for $n = 2$. The same argument will apply for all $n > 2$.

$$\begin{aligned} I(X_1,X_2;Y) &= H(X_1,X_2) - H(X_1,X_2|Y) \\ &= H(X_1) + H(X_2|X_1) - (H(X_1|Y) + H(X_2|X_1,Y)) \\ &= I(X_1;Y) + I(X_2;Y|X_1), \end{aligned}$$

where the first equality follows from Property 5 and the last equality follows by applying Lemma 1 for entropy and conditional entropy. □

Lemma 3 (Data Processing Inequality)**.** *Let X, Y, Z form a Markov chain $X \to Y \to Z$ such that $X \perp\!\!\!\perp Z|Y$. Then,*

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

Hence,

$$H(X|Y) \leq H(X|Z).$$

Proof We apply the chain rule for MI Lemma 2 in two ways:

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

Since $X \perp\!\!\!\perp Z|Y$, we have $I(X;Z|Y) = 0$. Also, $I(X;Y|Z) \geq 0$. Thus, we have $I(X;Y) \leq I(X;Y)$. To prove the second statement, we observe

$$H(X) - H(X|Z) = I(X;Z) \leq I(X;Y) = H(X) - H(X|Y).$$

□

Remark 1.1. *Data Processing Inequality make the following connections to Hypothesis Testing:*

- *In hypothesis testing, we will assume a prior on the alternatives $\{P_1, \dots, P_N\} \subseteq \mathcal{P}$. Let $X \in [N]$ forms a selection from $\{P_1, \dots, P_N\}$. Then, the data Y is generated from the chosen P_X.*
- *Now we have a test Z to estimate X from data Y.*
- *$I(X;Z) \leq I(X;Y)$ means that the test Z contains no more information about X than the data Y already does.*

We will now present Fano's inequality, which we will then use to develop Fano's method.

Theorem 1.1. *(Fano's inequality) Let X be a discrete random variable with a finite support $\mathcal{X}$. Let $X, Y, \hat{X}$ form a Markov chain $X \to Y \to \hat{X}$. Denote $P_e = \mathbb{P}(\hat{X} \neq X)$ and*

$$h(P_e) = -P_e \log(P_e) - (1 - P_e)\log(1 - P_e).$$

Then,

$$H(X \mid Y) \leq H(X \mid \hat{X}) \leq P_e \log(|\mathcal{X}|) + h(P_e). \tag{3}$$

Hence,

$$\mathbb{P}(X \neq \hat{X}) \geq \frac{H(X \mid Y) - \log(2)}{\log(|\mathcal{X}|)}.$$

Remark 1.2 (Connection to Testing)**.** *We can view X as defining a prior on $\{P_1, \dots, P_N\}$, and $\hat{X}$ as trying to guess X after observing data Y. If Y uniquely determines X, then we expect to perfectly predict X from Y i.e. $H(X|Y) = 0$ and $P_e = 0$. Fano's inequality quantifies how well we can estimate X from Y based on the conditional entropy (where $H(X|Y) > 0$).*

Proof Let $\mathbb{E} = \mathbf{1}\left\{X \neq \hat{X}\right\}$ (thus, if $E = 1$, there is an error). Using the chain rule in two different ways,

$$\begin{aligned} H(E, X \mid \hat{X}) &= H(X \mid \hat{X}) + H(E \mid X, \hat{X}) \\ &= H(E \mid \hat{X}) + H(X \mid E, \hat{X}). \end{aligned} \tag{4}$$

Here, since E is a function of $X, \hat{X}$,

$$H(E \mid X, \hat{X}) = 0. \tag{5}$$

Also, since conditioning reduces entropy,

$$H(E \mid \hat{X}) \leq H(E) = h(P_e), \tag{6}$$

where the equality follows by the definition of $h(\cdot)$. By the definition of the conditional entropy,

$$\begin{aligned} H(X \mid E, \hat{X}) &= \underbrace{\mathbb{P}(E=0)}_{=1-P_e}\underbrace{H(X \mid \hat{X}, E=0)}_{=0} + \underbrace{\mathbb{P}(E=1)}_{=P_e}\underbrace{H(X \mid \hat{X}, E=1)}_{\leq H(X)} \\ &\leq P_e H(X) \leq P_e \log(|\mathcal{X}|), \end{aligned} \tag{7}$$

where $H(X \mid \hat{X}, E=0)$ follows since $E=0$ implies $X=\hat{X}$ and the last inequality follows from the property of discrete X. Thus, (4)-(7) together imply that $H(X \mid \hat{X}) \leq h(P_e) + P_e \log(|\mathcal{X}|)$. We have proved the second inequality in the theorem.

Next, we will show the first inequality in the theorem. By the data processing inequality, $I(X, \hat{X}) \leq I(X, Y)$. Since we also have $I(X, \hat{X}) = H(X) - H(X \mid \hat{X})$ and $I(X,Y) = H(X) - H(X \mid Y)$, the next inequality is implied:

$$H(X \mid Y) \leq H(X \mid \hat{X}).$$

Since $h(P_e)$ is a entropy for a binary random variable, $h(P_e) \leq \log(2)$. This and (3) together imply the third inequality in the theorem:

$$P_e \geq \frac{H(X \mid Y) - \log(2)}{\log(|\mathcal{X}|)}.$$

□

2 Fano's method

Given Fano's inequality, we will derive two different types of lower bounds of the minimax risk. The first bound is called the global Fano's method since it contains a Kullback-Leibler divergence between one distribution and the mixture of all other distributions in the alternatives. The second bound is called the local Fano's method since it contains only the KL divergences between the alternatives.

Theorem 2.1. *Let S be drawn (not necessary i.i.d.) from some joint distribution $P \in \mathcal{P}$. Let $\{P_1, \cdots, P_N\} \subseteq \mathcal{P}$, and denote $\bar{P} = \frac{1}{N}\sum_{j=1}^{N} P_j$ (an equally weighted mixture distribution). Denote the minimax risk*

$$R^* = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi \circ \rho\left(\theta(P), \hat{\theta}(S)\right)\right].$$

Let $\delta = \min_{j \neq k} \rho\left(\theta\left(P_j\right), \theta\left(P_k\right)\right)$. Then, the following statements are true:

(I) (Global Fano's method)

$$R^* \geq \Phi\left(\frac{\delta}{2}\right)\left(1 - \frac{\frac{1}{N}\sum_{j=1}^{N} \mathrm{KL}\left(P_j, \bar{P}\right) + \log(2)}{\log(N)}\right).$$

(II) (Local Fano's method)

$$R^* \geq \Phi\left(\frac{\delta}{2}\right)\left(1 - \frac{\frac{1}{N^2}\sum_{1 \leq j,k \leq N} \mathrm{KL}\left(P_j, P_k\right) + \log(2)}{\log(N)}\right).$$

Remark 2.1. *The global Fano's method is stronger (tighter), but the local Fano's method is easier to apply since $\mathrm{KL}\left(P_j, P_k\right)$ is easier to compute than $\mathrm{KL}\left(P_j, \bar{P}\right)$.*