SML Recitation Notes Week 6: Information Lower Bound (Minimax)

Min Xu

February 22, 2011

1 Problem Set-up

- Let $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$ be a set of distributions parametrized by $\theta \in \Theta \subset \mathcal{H}$, with support $\mathcal{X}$
- We define an estimator to be a function $\hat{\theta} : \mathcal{X}^n \to \Theta$

Definition 1. The *minimax risk* of estimating $\theta \in \Theta$ is

$$\inf_{\hat{\theta}} \sup_{\theta_0 \in \Theta} \mathbb{E}_{P_{\theta_0}}[d(\hat{\theta}(X), \theta_0)]$$

Where $X = (X_1, ..., X_n)$ are the data and are drawn iid from P_{θ_0}.

Remark:

- The risk $R(\hat{\theta}(X), \theta_0)$ is defined to be $\mathbb{E}_{P_{\theta_0}}[d(\hat{\theta}(X), \theta_0)]$.
- If we fix an estimator $\hat{\theta}$, then $\sup_{\theta_0 \in \Theta} \mathbb{E}_{P_{\theta_0}}[d(\hat{\theta}(X), \theta_0)]$ is the worst-case risk for that particular estimator
- This is a difficult problem; we have to minimize over ALL estimators!

To tame this beast of a problem, we used a strategy with two steps:

1. Choose a set of discrete estimators, a finite subset of $\mathcal{P}$; discretize the loss function $d(\hat{\theta}(X), \theta_0)$.
2. When everything is finite and discrete, apply information theory (Fano's Inequality)

Remark: Estimator $\hat{\theta}$ is a function that takes data X and outputs an estimate. If X is random, then $\hat{\theta}(X)$ could be random as well. We will sometimes abuse vocabulary and refer to both the data-independent function $\hat{\theta}$ and the data dependent estimate $\hat{\theta}(X)$ as **an estimator**

The main theorem we will prove is the following:

Theorem 2. *Let Θ be the whole space of parameters; let Θ_{finite} be a finite subset. Let $\alpha = \min_{j \neq k; \theta_j, \theta_k \in \Theta_{finite}} d(\theta_j, \theta_k)$, and let $\beta = \max_{j \neq k; \theta_j, \theta_k \in \Theta_{finite}} KL(P_{\theta_j}, P_{\theta_k})$. Let L be the size of Θ_{finite}. Then:*

$$\inf_{\hat{\theta}} \sup_{\theta_0 \in \Theta} \mathbb{E}_{P_{\theta_0}}[d(\hat{\theta}(X), \theta_0)] \geq \frac{\alpha}{2} \left(1 - \frac{\beta + \log 2}{\log L}\right)$$

2 Reduction to Finite Discrete Case

It is easy to justify that we can take a finite subset of $\mathcal{P}$: for a fixed estimator $\hat{\theta}$:

$$\sup_{\theta_0 \in \Theta} R(\hat{\theta}(X), \theta_0) \geq \max_{\theta_j \in \Theta_{\text{finite}}} R(\hat{\theta}(X), \theta_j)$$

where Θ_{finite} is a finite subset of Θ. Since for all estimators, finitizing Θ gives a lower bound on the worst-case risk, it also gives a lower bound for minimax risk. Let $\Theta_{\text{finite}} = \{\theta_1, ..., \theta_L\}$.

Let $Z(X) = \arg \min_{\theta_k \in \Theta_{\text{finite}}} d(\hat{\theta}(X), \theta_k)$ be a random variable whose distribution is **dependent on the data and on the estimator**. Note that Z is a **discrete function of the data**, it takes data and outputs a discrete estimate; if the data is random, $Z(X)$ is a multinomial.

Define $\alpha = \min_{j \neq k; \theta_j, \theta_k \in \Theta_{\text{finite}}} d(\theta_j, \theta_k)$.

Fix true distribution and let it be P_{θ_j}. If $Z(X) = j$, then $d(\hat{\theta}(X), \theta_j)$ could only be lower bounded by 0; if $Z(X) \neq j$, then $d(\hat{\theta}(X), \theta_j)$ could be lower bounded by $\frac{\alpha}{2}$ by triangle inequality.

Hence, for a fixed estimator and for a fixed true θ_j:

$$\mathbb{E}_{P_{\theta_j}}[d(\hat{\theta}(X), \theta_j)] \geq 0 \cdot P_{\theta_j}(Z(X) = j) + \frac{\alpha}{2} P_{\theta_j}(Z(X) \neq j)$$

Since this inequality hold for any fixed true θ_j:

$$\max_{\theta_j \in \Theta_{\text{finite}}} \mathbb{E}_{P_{\theta_j}}[d(\hat{\theta}(X), \theta_j)] \geq \max_{\theta_j \in \Theta_{\text{finite}}} \frac{\alpha}{2} P_{\theta_j}(Z(X) \neq j)$$

And finally, since the inequality immediately above hold for any fixed estimator $\hat{\theta}$:

$$\inf_{\hat{\theta}} \max_{\theta_j \in \Theta_{\text{finite}}} \mathbb{E}_{P_{\theta_j}}[d(\hat{\theta}(X), \theta_j)] \geq \inf_{\hat{\theta}} \max_{\theta_j \in \Theta_{\text{finite}}} \frac{\alpha}{2} P_{\theta_j}(Z(X) \neq j)$$

Notice that with these manuevers, we have effectively replaced the loss $d(\hat{\theta}(X), \theta_j)$ with $\frac{\alpha}{2} P_{\theta_j}(Z(X) \neq j)$

Let's consider RHS $\inf_{\hat{\theta}} \max_{\theta_j \in \Theta_{\text{finite}}} \frac{\alpha}{2} P_{\theta_j}(Z(X) \neq j)$. We can lower bound this by taking infimum over all discrete valued functions of the data: $\mathcal{Z} = \{Z : Z \text{ is discrete function of data}\}$.

$$\inf_{\hat{\theta}} \max_{\theta_j \in \Theta_{\text{finite}}} \frac{\alpha}{2} P_{\theta_j}(Z(X) \neq j) \geq \inf_{Z \in \mathcal{Z}} \max_{\theta_j \in \Theta_{\text{finite}}} \frac{\alpha}{2} P_{\theta_j}(Z(X) \neq j)$$

In summary, we have the bound:

$$\inf_{\hat{\theta}} \sup_{\theta_0 \in \Theta} \mathbb{E}_{P_{\theta_0}}[d(\hat{\theta}(X), \theta_0)] \geq \frac{\alpha}{2} \inf_{Z \in \mathcal{Z}} \max_{\theta_j \in \Theta_{\text{finite}}} P_{\theta_j}(Z(X) \neq j)$$

On the RHS, we have a lower bound in which replaced original Θ with Θ_{finite}, the original $\inf_{\hat{\theta}}$ with inf over a space of discrete estimators, and replaced the original loss function.

3 Information Inequalities

Now, our goal is to lower bound:

$$\inf_{Z \in \mathcal{Z}} \max_{\theta_j \in \Theta_{\text{finite}}} P_{\theta_j}(Z(X) \neq j)$$

Fix estimator Z. We are going to get a lower bound by expanding our probability space a little and define a prior over the possible true parameters. Thus, we will analyze a probability space where

$$\begin{aligned} Y &\sim \text{Uniform Multinomial} \\ X|Y &\sim P_{\theta_Y} \end{aligned}$$

and Z is a function of the data X.

More explicitly, let Y be a multinomial random variable uniform among $\{1, ..., L\}$, then

$$\begin{aligned} \max_{\theta_j \in \Theta_{\text{finite}}} P_{\theta_j}(Z(X) \neq j) &\geq \frac{1}{L} \sum_{j=1}^{L} P_{\theta_j}(Z(X) \neq j) \\ &= \sum_{j=1}^{L} P(Y = j) P_{\theta_j}(Z(X) \neq j) \\ &\triangleq \sum_{j=1}^{L} P(Y = j) P(Z(X) \neq j | Y = j) \\ &= P(Z(X) \neq Y) \end{aligned}$$

Where we have also represented distribution $P_{\theta_j}(\cdot)$ as a conditional distribution $P(\cdot|Y = j)$.

We must now lower bound $P(Z(X) \neq Y)$ where P is probability induced by $\{P_{\theta_1}, ..., P_{\theta_L}\}$ and the uniform prior Y, and $Z(X)$ is a discrete estimate of Y based on the data generated from P_{θ_Y}.

The distribution $\{P_{\theta_1}, ..., P_{\theta_L}\}$ can be arbitrary so is an lower bound even posssible? We will give two intuition that show the lower bound, though difficult, is possible.

1. On one hand, suppose P_{θ_j}'s are all the same, then the data X says nothing about Y and any estimate $Z(X)$ of Y cannot do better than random guess. Thus $P(Z(X) \neq Y) \geq \frac{L-1}{L}$.
2. On the other hand, suppose P_{θ_j} have disjoint support, then data X reveals everything about Y and $P(Z(X) \neq Y) \geq 0$ is the tightest lower bound we can get.
3. If P_{θ_j}'s are between the two extremes, then the lower bound for $P(Z(X) \neq Y)$ should also vary from $\frac{L-1}{L}$ to 0

Intuitively, the data X carries **information** about Y and the more distinct the P_{θ_j}'s are, the more information X carries. We will formalize this notion of information and lower bound the probability of error by a function of the information.

3.1 Entropy

Definition 3. We need at least $\log N$ bits to encode N distinct objects. We will say that a collection of N distinct objects has $\log N$ bits of information.

Definition 4. Let p be the density of a distribution, we define **entropy** of p to be $H(p) = -\int p(x) \log p(x) dx$. If p is finite and discrete taking on values among $\{1, ..., K\}$, then $H(p) = -\sum_{k=1}^{K} p_k \log p_k$ where p_k is probability of value k.

Larger entropy means the distribution contains more information. We can interpret entropy as the average number of bits to encode a sample from distribution p. Why?

Consider the case where p is multinomial with probabilities $p_1, ..., p_K$. Suppose we drawn $X_1, ..., X_n \sim p$. As n gets very very large, we know that with overwhelming probability, we will expect to see np_k samples with value k.

Hence, although the support of $X_1, ..., X_n$ is the set of all length n K-ary string (with size K^n), its *effective* support has size $\binom{n}{np_1; np_2; ...; np_K} = \frac{n!}{(np_1)!(np_2)!...(np_K)!}$; everything outside of effective support has negligibly small probability.

How many bits at minimum does it take to encode $\binom{n}{np_1; np_2; ...; np_K}$ objects?

$$\begin{aligned}
\log \binom{n}{np_1; np_2; ...; np_K} &= \log n! - \log(np_1)! - ... - \log(np_K)! \\
&= n \log n - (np_1) \log(np_1) - ... - (np_K) \log(np_K) \\
&= n((p_1 + ... + p_K) \log n - p_1 \log(np_1) - ... - p_K \log(np_K)) \\
&= n(p_1 \log \frac{n}{np_1} + ... + p_K \log \frac{n}{np_K} \\
&= nH(p)
\end{aligned}$$

Where we used Stirling's approximation $\log n! = n \log n - n + \frac{1}{2} \log(2\pi n)$ for large n on the second equality.

Since it takes at least $nH(p)$ bits to encode n samples for large n, we see that $H(p)$ is the asymptotic number of bits required to encode one sample from distribution p.

Abusing notation again, if X is a random variable, we will use $H(X)$ to denote the entropy of the distribution of X.

Definition 5. Let X, Y be random variables, the conditional entropy $H(X|Y = y)$ is the entropy of the distribution of X on condition that $Y = y$. The overall **conditional entropy** $H(X|Y) = \sum_y p(Y = y) H(X|Y = y)$

We can interpret $H(X|Y)$ as, given a pair of samples (X, Y), if you can see Y and use any information in Y, how many bits to do you need to encode X.

To get some more intuition about entropy, the following are true:

- $H(X)$ also measures amount of randomness in X
- For bernoulli X, $H(X)$ is maximized at 1 when $P(X = 1) = P(X = 0)$
- If X is non-random, $H(X) = 0$

- If X, Y are independent, $H(X,Y) = H(X) + H(Y)$
- More generally, $H(X,Y) = H(X|Y) + H(Y)$

3.2 Fano's Inequality

Theorem 6. *Let Y be a multinomial with values $\{1, ..., L\}$. Let $\{P_1, ..., P_L\}$ be a set of distributions and let $X|Y$ be drawn from P_Y. Let Z be a discrete function of X so that $Z(X)$ is a multinomial with values $\{1, ..., L\}$.*

Let E be an error indicator, $E = 1$ if $Z(X) \neq Y$, 0 else

$$\begin{aligned} H(Y|X) &\leq P(E=1)\log(L-1) + H(E) \\ &\leq P(E=1)\log(L-1) + \log 2 \\ &\leq P(E=1)\log L + 1 \end{aligned}$$

and hence

$$P(Z(X) \neq Y) \geq \frac{H(Y|X) - 1}{\log L}$$

We give a proof intuition here. Suppose we have data X and we would like to use X to encode the effective support of Y. One encoding scheme is to use $Z(X)$ and first encode error indicator E. If $E = 0$, we can just read Y off of $Z(X)$ and require no additional bits. If $E = 1$, then we still need to encode Y. More precisely,

$$H(Y|X) = P(E=1)H(Y|X, E=1) + H(E)$$

$H(Y|X, E=1)$ is upper bounded by $\log(L-1)$ and $H(E)$ is upper bounded by 1 and we get the result as desired.

To get a better intuition, we consider two cases again:

- Given data X, if we know that there exist a very accurate estimator $Z(X)$, then $Y|X$ cannot be too random and we have an upper bound constraint on $H(Y|X)$.
- If we know $H(Y|X)$ is very high, then $Y|X$ is highly random, and there cannot exist a very accurate estimator. Hence we have a lower bound constraint on $P(Z(X) \neq Y)$

To complete the proof of our overall theorem, we just need to upper bound $H(Y|X)$

$$\begin{aligned} H(Y|X) &= H(Y) - I(Y;X) \\ &= \log L - \frac{1}{L}\sum_{j=1}^{L} KL(P_{\theta_j}, \bar{P}) \\ &= \log L - \frac{1}{N^2}\sum_{j,k}^{N} KL(P_{\theta_j}, P_{\theta_k}) \\ &\geq \log L - \beta \end{aligned}$$

Where $I(Y;X)$ is the mutual informationed defined as $I(X,Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$. For us, it is also convenient to note that $p(x,y) = p(x|y)p(y)$ and use an equivalent form $I(Y;X) = \sum_y p(y)\sum_x p(x|y)\log\frac{p(x|y)}{p(x)}$.