STAT 526

Topic X
Statistical Learning Thoery

Dr. Qifan Song

Learning Theory

Learning theory is a broad topic, which aims to explain general behavior of estimation and prediction accuary. I will go through some popular concepts and results

Distance/Divergence measure between two measures P and Q

- Total variation distance $TV = \sup_A |P(A) - Q(A)| = \int |p - q| dx/2 = \sup_{\|f\|_\infty \leq 1} E_P f - E_Q f$
- KL divergence $KL(P, Q) = \int \log(p/q) p dx$
- Hellinger distance $H = \sqrt{\int (\sqrt{p} - \sqrt{q})^2 dx}$
- f divergence $D_f(P, Q) = \int f(p/q) q dx$; KL is f div with $f = x \log x$
- $H^2/2 \leq TV \leq H \leq \sqrt{KL}$.

Testing

- $X \sim P_0$ (null distribution) or P_1 (alternative distribution), equivalently, $X \sim P_\theta$ with $\theta \in \{0, 1\}$
- A test/estimator $\phi(X) \to \{0, 1\}$
- We want to assess: Type I/II error probabilities $P_0(\phi(X) = 1)$ and $P_1(\phi(X) = 0)$, or estimation accuracy $P_\theta(\phi(X) \neq \theta)$. Especially, the lower bounds of these quantities
- Suppose that θ follow a uniform prior over $\{0, 1\}$, let $A = \phi^{-1}(1)$

  $P(\phi(X) \neq \theta) = \pi(\theta = 0)P_0(\phi(X) = 1) + \pi(\theta = 1)P_1(\phi(X) = 0)$
  $= [P_0(X \in A) + P_1(X \in A^c)]/2 = [1 + P_0(X \in A) - P_1(X \in A)]/2$
  $\geq [1 - TV(P_0, P_1)]/2$

Mini-Max

- More precisouly, $\inf_\phi P(\phi(X) \neq \theta) = [1 - TV(P_0, P_1)]/2$
- Back to frequentist setting, we can get

$$\inf_{\phi} \sup_{\theta \in \{0,1\}} P(\phi(X) \neq \theta) \geq [1 - TV(P_0, P_1)]/2,$$

  where the inf is taken over all possible test (or selector)

- The equality holds if using LRT
- Our Goal: Best accuracy under worst scenario (minimax risk)

$$R_n = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} E_P[L(\widehat{\theta}, \theta(P))],$$

  where $=\{P_\theta, \theta \in \Theta\}$, and $\theta(P_\theta) = \theta$, L is distance-based measure.

- Statistical optimality
- Gap: ϕ vs $\widehat{\theta}$, $\{\theta_0, \theta_1\}$ vs Θ, one sample vs a data set

Mini-Max Rate

- General strategy to obtain minimax rate
- Find an upper bound by defining a "good estimator" $\tilde{\theta}$, then $R_n \leq \sup_{P \in \mathcal{P}} E_P[L(\tilde{\theta}, \theta(P))]$
- Find a lower bound as in Le Cam's method (a challenging task)
- If upper and lower bounds match in order, mission accomplished.

Le Cam's method

- $X_1, \ldots, X_n \sim P_{\theta_1}$ or $\sim P_{\theta_2}$, for $\theta_1, \theta_2 \in \Theta$
- $P_{\theta_i}^n$ denote product distribution
- Then

$$\inf_{\phi} \sup_{\theta \in \{\theta_1, \theta_2\}} P(\phi(X_1, \ldots, X_n) \neq \theta) \geq \frac{1 - TV(P_{\theta_1}^n, P_{\theta_2}^n)}{2},$$

- $TV(P_{\theta_1}^n, P_{\theta_2}^n) \leq \sqrt{KL(P_{\theta_1}^n, P_{\theta_2}^n)} = \sqrt{nKL(P_{\theta_1}, P_{\theta_2})}$ for iid case
- $TV(P_{\theta_1}^n, P_{\theta_2}^n) \leq H(P_{\theta_1}^n, P_{\theta_2}^n)$

  with $H^2(P_{\theta_1}^n, P_{\theta_2}^n) = 2(1 - \int \prod_i^n p_{\theta_1}(x_i) p_{\theta_2}(x_i)) = 2(1 - (1 - H^2(P_{\theta_1}, P_{\theta_2})/2)^n) \leq nH^2(P_{\theta_1}, P_{\theta_2})$

Test function

Upper bound on testibility
Asymptotic Methods in Statistical Decision Theory (Le Cam, 1986, Lemma 4 on page 478)

Given two convex sets $\mathcal{P}_0$ and $\mathcal{P}_1$ of probability measures, there exist tests ϕ_n, such that

$$\sup_{P\in\mathcal{P}_0} E_P\phi_n(X_1,\ldots,X_n) \le \exp(n\log\rho(\mathcal{P}_0,\mathcal{P}_1))$$
$$\sup_{P\in\mathcal{P}_1} E_P[1-\phi_n(X_1,\ldots,X_n)] \le \exp(n\log\rho(\mathcal{P}_0,\mathcal{P}_1))$$

where $\rho = 1 - h^2(\mathcal{P}_0,\mathcal{P}_1)/2$, and $H(\mathcal{P}_0,\mathcal{P}_1)$ is the Hellinger distance. Note that $\log\rho \le -H^2/2$

This theorem leads to

$$\inf_{\widehat{\theta}} \sup_{\theta\in\{\theta_1,\theta_2\}} P(\phi\neq\theta) \le \exp(-nH^2(P_0,P_1)/2),$$

Le Cam's method

- Let $\widehat{\theta}$ be any estimator

$$\begin{aligned}&\inf_{\widehat{\theta}} \sup_{\theta\in\{\theta_1,\theta_2\}} E\|\widehat{\theta}-\theta\|^2\\ \geq &\inf_{\widehat{\theta}} \sup_{\theta\in\{\theta_1,\theta_2\}} [\|\theta_1-\theta_2\|^2/4]P(\|\widehat{\theta}-\theta\|\geq\|\theta_1-\theta_2\|/2)\end{aligned}$$

- For any $\widehat{\theta}$, we can define a test $\phi = \arg\min_{\{\theta\in\theta_1,\theta_2\}} \|\widehat{\theta}-\theta\|$
- $\phi\neq\theta_j$ implies that $\|\widehat{\theta}-\theta_j\|\geq\|\theta_1-\theta_2\|/2$, that is $P(\|\widehat{\theta}-\theta_j\|\geq\|\theta_1-\theta_2\|/2)\geq P(\phi\neq\theta_j)$, thus

$$\begin{aligned}&\inf_{\widehat{\theta}} \sup_{\theta\in\{\theta_1,\theta_2\}} E\|\widehat{\theta}-\theta\|^2\\ \geq &\inf_{\widehat{\theta}} \sup_{\theta\in\{\theta_1,\theta_2\}} [\|\theta_1-\theta_2\|^2/4]P(\phi\neq\theta)\\ \geq &[\|\theta_1-\theta_2\|^2/4]\frac{1-TV(P^n_{\theta_1},P^n_{\theta_2})}{2},\end{aligned}$$

Apply Le Cam's method

- For any two θ_1 and $\theta_2 \in \Theta$, then

  $\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} E\|\widehat{\theta}-\theta\|^2 \geq \inf_{\widehat{\theta}} \sup_{\theta\in\{\theta_1,\theta_2\}} E\|\widehat{\theta}-\theta\|^2 \geq \|\theta_1-\theta_2\|^2 F(n, P_{\theta_1}, P_{\theta_2}),$

  where $F = \frac{1}{8}(1 - \sqrt{nKL(P_{\theta_1}, P_{\theta_2})})$ or $\frac{1}{8}(1 - \sqrt{n}H(P_{\theta_1}, P_{\theta_2}))$

- $$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} E\|\widehat{\theta} - \theta\|^2 \geq \max_{\theta_1,\theta_2} \|\theta_1 - \theta_2\|^2 F(n, P_{\theta_1}, P_{\theta_2}),$$

- Quadratic approximation (2-order Taylor) $KL(P_{\theta_1}, P_{\theta_2}) \approx c\|\theta_1 - \theta_2\|^2$ and $H^2(P_{\theta_1}, P_{\theta_2}) \approx c\|\theta_1 - \theta_2\|^2$ holds for most parametric families

- $$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} E\|\widehat{\theta} - \theta\|^2 \geq \max_{\Delta\theta} (\Delta\theta)^2(1 - c\sqrt{n}\Delta\theta)/8 \asymp \frac{1}{n},$$

Try some examples

1. Estimating the mean of normal distribution

2. Estimating parameter of uniform distribution $U(\theta, \theta + 1)$

Le Cam's results are usually independent to dimension of the parameter space. Assouad introduced a hyper-rectangle trick to resolve it.

Parallel 2-hypothesis Test

- A family of 2^m hypothesis P_v, indexed by $v \in \{-1, +1\}^m$
- We need to connect the index space and the parameter space
- Given θ_1, θ_2, assume that $d(\theta_1, \theta_2) = \sum_j d_j(\theta_1, \theta_2)$ (e.g., L^1 loss)
- For any u, v differ only at jth entry, assume $d_j(\theta(P_u), \theta(P_v)) \geq \delta$ for all j
- Follow Le Cam's idea (assign a uniform prior over v)
- $$\sup_{\theta \in \Theta} Ed(\hat{\theta}, \theta) = \sup_{\theta \in \Theta} \sum_j Ed_j(\hat{\theta}, \theta(P_v)) \geq \sup_{P_v} \sum_j Ed_j(\hat{\theta}, \theta)$$
$$\geq E_{P_v, v \sim \text{Unif}} \sum_j Ed_j(\hat{\theta}, \theta(P_v)) \geq \sum_j E_{v_{-j} \sim \text{Unif}} [E_{v_j \sim \text{Unif}} d_j(\hat{\theta}, \theta(P_v))]$$

Parallel 2-hypothesis Test

- Now we can directly apply Le Cam's result

$$E_{v_j \sim \text{Unif}} d_j(\hat{\theta}, \theta(P_v)) \geq \frac{\delta}{2} \frac{1 - TV(P_{v_j=1, v_{-j}} || P_{v_j=-1, v-j})}{2}$$
$$\geq \frac{\delta}{2} \frac{1 - \Delta_j}{2}$$

  where $\Delta_j = \max_{\{u,v \text{ diffs only at jth entry}\}} TV(P_u || P_v)$

- Then

$$\sup_{\theta \in \Theta} E d(\hat{\theta}, \theta) \geq \sum_j E_{v_{-j} \sim \text{Unif}} \frac{\delta}{2} \frac{1 - \Delta_j}{2} \geq \frac{m\delta}{2} \frac{1 - \Delta}{2}$$

  where $\Delta = \max \Delta_j = \max_{\{u,v \text{ diffs only by one entry}\}} TV(P_u || P_v)$

- If n samples, then

$$\Delta = \max_{\{u,v \text{ diffs only by one entry}\}} TV(P_u^n || P_v^n)$$

  which is further bounded by $\sqrt{nKL}$ or $\sqrt{n}H$

Assouad's Method for non-metric or non-decomposable $\tilde{d}(\cdot, \cdot)$

- related $\tilde{d}$ with another decomposable metric d: (L_2 norm $\geq$ $(1/\sqrt{d})$ L_1 norm)
- $\tilde{d} \geq C \sum_j d_j$ where d_j is a metric
- $\tilde{d} = \sum_j \tilde{d}_j$ where $\tilde{d}_j$ is not a metric, but satisfies if $d_j(\theta_0, \theta_1) < d_j(theta_0, \theta_2)$, then $d_j(theta_0, \theta_2) \geq C d_j(\theta_1, \theta_2)$

  Under weak triangle inequality, $d_j(\theta_0, \theta_2) \geq [d_j(\theta_0, \theta_2) + d_j(\theta_0, \theta_1)]/2 \geq C d_j(\theta_1, \theta_2)/2$

Apply Assouad's method

- $\Theta = \mathbb{R}^d$
- Routine choice of P_v: P_θ with θ_i's $= (\pm\delta, \pm\delta, \ldots, \pm\delta) = \delta * v$
- By continuity of KL, $KL(P_u||P_v), H^2(P_u||P_v) \approx c\delta^2$ for u and v differing by one entry.
- $$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} E\|\widehat{\theta} - \theta\|_1 \geq m\delta \frac{1 - \sqrt{cn}\delta}{2}$$
- Maximizing RHS w.r.t. δ leads to an order of $d/\sqrt{n}$
- We can also derive lower bounds for L_2 and L_2^2 ($\sqrt{d/n}$ and d/n) respectively

Multiple testing problem

- X is a discrete r.v. whose sample space is of size $M \geq 2$, Y is another r.v.
- $\tilde{X} = f(Y)$ for some function f, $E = 1(\tilde{X} \neq X)$, i.e., using Y to test X
- $H(X|Y) = -\sum_{ij} P(x_i, y_i) \log P(x_i|y_j)$ denotes the conditional entropy, $H(X) = -\sum_i P(x_i) \log P(x_i)$ denotes the entropy.
- By definition $H(E, X|\tilde{X}) = H(X|\tilde{X}) + H(E|X, \tilde{X}) = H(E|\tilde{X}) + H(X|E, \tilde{X})$, which implies $H(X|\tilde{X}) = H(E|\tilde{X}) + H(X|E, \tilde{X})$
- By definition $H(X|E, \tilde{X}) = H(X|E = 0, \tilde{X})P(E = 0) + H(X|E = 1, \tilde{X})P(E = 1) = H(X|E = 1, \tilde{X})P(E = 1)$

Fano's Lemma

- $H(X|E=1,\tilde{X}) \leq \log(M-1)$ (a r.v. with a possible values has max entropy $\log a$)
- $H(E|\tilde{X}) \leq H(E)$ (HW: conditional entropy always is smaller than entropy)
- $H(X|Y) \leq H(X|\tilde{X})$ ($H(X|Y) = H(X,Y) - H(X)$, $H(X,\tilde{X}) = H(X,\tilde{X}) - H(X)$)
- $$H(X|Y) \leq H(E) + P(E=1)\log(M-1)$$

Fano's Method

- θ has a uniform prior over $\{\theta_1, \ldots, \theta_M\}$
- By Fano's Lemma, for any estimator $\phi(X)$, let $E = 1(\phi(X) \neq \theta)$

$$H(\theta|X) \leq H(E) + P(E = 1)\log(M-1); H(E) \leq \log 2$$

$$P(E = 1) \geq \frac{H(\theta|X) - \log 2}{\log(M-1)}$$

- Using similar argument in Le Cam's method

$$\inf_{\widehat{\theta}} \sup_{\theta \in \{\theta_1, \ldots, \theta_M\}} E\|\widehat{\theta} - \theta\|^2 \geq \frac{\min_{\theta_i, \theta_j} \|\theta_i - \theta_j\|^2}{4} P(E = 1)$$

Handle $H(\theta|X)$

- $$H(\theta|X) = E - \log \frac{P(\theta, X)}{P(X)P(\theta)} - E \log P(\theta)$$
$$= - KL(P(\theta, X)||P(\theta)P(Y)) + \log M$$
$$= - \frac{1}{M} \sum_i KL(P_{\theta_i}||\bar{P}) + \log M$$

  where $\bar{P} = \sum P_{\theta_i}/M$

- By convexity of KL

$$H(\theta|X) = -\frac{1}{M} \sum_i KL(P_{\theta_i}||\bar{P}) + \log M$$
$$\geq - \frac{1}{M^2} \sum_{i,j} KL(P_{\theta_i}||P_{\theta_j}) + \log M$$

- Finally

$$\begin{aligned}
&\inf_{\widehat{\theta}} \sup_{\theta \in \{\theta_1, \ldots, \theta_M\}} E\|\widehat{\theta} - \theta\|^2 \\
\geq &\frac{\min_{\theta_i, \theta_j} \|\theta_i - \theta_j\|^2}{4} \frac{\log M - \log 2 - \frac{1}{M^2} \sum_{i,j} KL(P_{\theta_i} || P_{\theta_j})}{\log(M-1)} \\
\geq &\frac{\min_{\theta_i, \theta_j} \|\theta_i - \theta_j\|^2}{4} \left(1 - \frac{\log 2 + \frac{1}{M^2} \sum_{i,j} KL(P_{\theta_i} || P_{\theta_j})}{\log(M-1)}\right)
\end{aligned}$$

- If we observe n data point $X_1, \ldots, X_n$

$$\begin{aligned}
&\inf_{\widehat{\theta}} \sup_{\theta \in \{\theta_1, \ldots, \theta_M\}} E\|\widehat{\theta} - \theta\|^2 \\
\geq &\frac{\min_{\theta_i, \theta_j} \|\theta_i - \theta_j\|^2}{4} \left(1 - \frac{\log 2 + \frac{1}{M^2} \sum_{i,j} nKL(P_{\theta_i} || P_{\theta_j})}{\log(M-1)}\right)
\end{aligned}$$

Apply Fano's method

- Find multiple θ_i's $\in \Theta$
- $\|\theta_i - \theta_j\| \geq 2\delta$, while $KL(P_{\theta_i}||P_{\theta_j}) \leq \epsilon$ for all i, j
- Then $\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} E\|\widehat{\theta} - \theta\|^2 \geq \delta^2 \left(1 - \frac{\log 2 + n\epsilon}{\log(M-1)}\right)$
- To obtain a sharper bound (i.e., larger bound), we need a big M (the concept of packing number)
- Fano vs Le Cam vs Assouad: Le Cam's method reduces Θ to one-dim case, thus the derived rate is usually dimension-indep. Assouad's method usually applies to that case where $\Theta \subset \mathbb{R}^d$ with $dim(\Theta) = d$. Fano's method may improve, and applies to nonrequlars shaped Θ.

Apply Fano's method

- Routine choice of θ_i's $= (\pm\delta, \pm\delta, \ldots, \pm\delta)$, totally $M = 2^d$
- Unfortunately, it fails to obtain a good dim-dependent bound: $KL(P_{\theta_i}||P_{\theta_j}) \leq cd\delta^2$

- Instead, find M θ_i's within $B(0, \epsilon)$, such that pairwise distance is greater than $\epsilon/3$.
- Then $KL(P_{\theta_i}||P_{\theta_j}) \leq c\epsilon^2$, $\|\theta_j - \theta_i\| \geq \epsilon/3$

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} E\|\widehat{\theta} - \theta\|^2 \geq \frac{\epsilon^2}{36}\left(1 - \frac{\log 2 + nc\epsilon^2}{\log(M-1)}\right)$$

- Maximizing RHS w.r.t. δ leads to an order of $\log M/n$
- How large can M be?

Packing Number and Covering Number

- ϵ-covering: a set of θ_i's such that for any $\theta \in \Theta$, there exists j, such that $\|\theta - \theta_j\| \leq \epsilon$
- ϵ-packing: a set of θ_i's such that for $\|\theta_i - \theta_j\| > \epsilon$

- Maximal ϵ-packing is a ϵ-covering, hence its size is greater than or equal to the minimal ϵ-covering size.
- Size of maximal 2ϵ-packing is smaller than or equal to the minimal ϵ-covering size. (if not, centers of two packing balls must be inside one covering ball)
- The minimal ϵ-covering size $\geq$ volume of Θ divided by volume of ϵ-radius ball
- Back to our question: the largest possible $M \geq$ volume of $B(0, \epsilon)$ divided by volume of $\epsilon/3$-radius ball. Thus $\log M$ at least of $O(d)$

Apply Fano's method

- Let us try one example $X \sim \mathrm{N}(\theta, \sigma^2\mathrm{I})$, where $\theta \in \mathbb{R}^n$ and $\|\theta\|_0 \leq s$, i.e., sparse mean regression
- We need to find as many θ_i's as possible inside $\Theta = \{\|\theta\|_2 \leq \epsilon, \|\theta\|_2 \leq s\}$ with $\|\theta_i - \theta_j\| \geq \delta$ for some ϵ and δ
- Let all these θ_i's be of sparsity $2s/3$, all of its nonzero entry be $\pm a$.For every pair of θ_i and θ_j, they share at most $s/3$ nonzero entry index
- Then for all $\|\theta_i - \theta_j\| \leq 2\sqrt{2s/3}a$, $KL(P_{\theta_i}||P_{\theta_j}) \leq 4sa^2/3$
- Then for all $\|\theta_i - \theta_j\| \geq \sqrt{2s/3}a$
- L_2^2 rate low bound reduce to $\log M$

Apply Fano's method

- The problem reduces to a combinatorial problem for n-dim binary vectors.
- There are totally $\binom{n}{k}$ k-sparse n-dim binary vector (here $k = 2s/3$)
- For every k-sparse vector, there are $\binom{k}{k/2}\binom{n-k}{k/2}+\cdots+\binom{k}{k}\binom{n-k}{0}$ k-sparse vectors that shares more than $d/2$ nonzero entry index
- So at least $M \geq \binom{n}{k}/(\binom{k}{k/2}\binom{n-k}{k/2} + \cdots + \binom{k}{k}\binom{n-k}{0})$
- $\binom{k}{k/2}\binom{n-k}{k/2} + \cdots + \binom{k}{k}\binom{n-k}{0} \leq 2^k\binom{n}{k/2}$
- $\log M \geq \log \binom{n}{k} - k \log 2 - \log \binom{n}{k/2}$
- When n/k is sufficient large, by Sterling approximation ($n! \sim n^{n+0.5}e^{-n}$), we have $\log M$ at least of order $k \log(n/k) \asymp s \log(n/s)$

Minimax Rate of Density estimation

- $\Theta = \{\text{pdf } f \text{ on } [0,1]\}$.
- $x_j = (j-(1/2))/m$, $j = 1, \ldots, m$, $g_j(x) = c\sin(m(x-x_j))/m^2$ for a fixed small c
- Define $f_v = 1 + \sum_{j=1}^m v_j g_j$ with $v \in \{-1,+1\}^m$
- d_j satisfies weak triangle inequality
- Assouad's method and Fano method both apply; then we take maximum over m

$$\inf \sup H^2(\widehat{f}, f) \geq n^{-4/5}$$