36-710: Advanced Statistical Theory **Fall 2018**

Lecture 2: August 29

Lecturer: Alessandro Rinaldo *Scribes: Natalia Lombardi de Oliveira*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

2.1 Examples regarding last lecture

Example 2.1 (Linear Discriminant Analysis) *Let $P_j \sim N(\mu_j, \Sigma), j = 1, 2$ and $X \in \mathbb{R}^d$ with distribution P_x. Goal: test $H_0 : P_x = P_1$ vs $H_a : P_x = P_2$. Let's use the Likelihood Ratio test statistic ($log \frac{dP_1}{dP_2}(X)$), in which we reject H_0 if $LR > 1$. This is equivalent to $\psi(X) =< \mu_1 - \mu_2, \Sigma^{-1}(X - \frac{\mu_1 - \mu_2}{2}) >< 0$. Assuming $P_x = .5P_1 + .5P_2$, the overall error $Err(\psi) = .5\mathbb{P}_1(\psi(X) \leq 0) + .5\mathbb{P}_2(\psi(X) \geq 0)$.*

If $\Sigma = \mathbb{I}_d$, we have $\phi(-\gamma/2)$, in which $\gamma = ||\mu_1 - \mu_2||$ and ϕ the cdf of a normal(0,1). Now, observe n samples from $P_1, X_1, \ldots, X_n$ and n samples form $P_2, Y_1, \ldots, Y_n$, and let $\hat{\mu}_1 and \hat{\mu}_2$ be the respective sample means. Using the plug-in rule, $\hat{\psi}(X) =< \hat{\mu}_1 - \hat{\mu}_2, (x - \frac{\hat{\mu}_1 - \hat{\mu}_2}{2}) >< 0$ and evaluate its error $Err(\hat{\psi})$. It's possible to show that $Err(\hat{\psi}) \xrightarrow{p} \Phi(\frac{-\gamma^2}{2\sqrt{\gamma^2 + 2\alpha}})$, where $\alpha = \lim_{n\to\infty} \frac{dn}{n}$. If $\alpha = 0$, then $\xrightarrow{p}$ higher error. If $\alpha = \infty$, then $\xrightarrow{p} 1/2$.

Example 2.2 (Many normal means problem) *Let $X \sim N_n(\mu, I_n), \mu \in \mathbb{R}, d = n$.*

Under square error loss, the mle X is not optimal estimator of μ ($n \geq 3$). (Option: use James-Stein estimator.) Now let's think about the problem of testing $H_0 : \mu = 0$ vs $H_a : \mu \neq 0$, which is equivalent to

$$H_0 : \cap_{i=1}^n H_{0i} \text{ vs } \cup_{i=1}^n H_{ai},$$

in which $H_{0i} : \mu_i = 0$ and $H_{ai} : \mu_i \neq 0$. Two cases:

1. **needle in haystack problem**: *there existis one coordinate such that $\mu_i \neq 0$ and $\mu_j = 0, j \neq i$. Optimal statistic is $\max_i |X_i| > t_\alpha$ and it is optimal if $\mu_i \geq (1 - \epsilon)\sqrt{2logn}$ for any fixed α and any $\epsilon > 0$ for $t_\alpha = \sqrt{2logn}$. (Optimal here means that power goes to 1 as $n \to \infty$.) If $|\mu_i| \leq (1-\epsilon)\sqrt{2logn}$, any $\epsilon > 0$, then the sum of type I and type II errors goes to 1 for any test as $n \to \infty$.*
2. **signal is weak but spread out**: *use χ^2 - test statistic and reject if $||X||^2$ is large.*

2.2 Basic concentration inequalities

Motivation: $X_1, \ldots, X_n \sim (\mu, \sigma^2)$, iid, $\bar{X}_n \xrightarrow{p} \mu$, $\bar{X}_n = \mu + o_p(1)$. We want to know how fast, $\forall t$, $P(|\bar{X}_n - \mu| \geq t) \to 0$.

By CLT, $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \xrightarrow{d} Z$, $Z \sim N(0, 1)$, so it's root-n consistent ($\bar{X}_n = \mu + O_p(1/\sqrt{n})$).

$\forall \epsilon > 0, \exists t = t(\epsilon)$ and $N = N(\epsilon, t)$ s.t. $P(|\bar{X}_n - \mu| \geq t\sigma/\sqrt{n}) \leq \epsilon$. Then, $\lim_{n\to\infty} P(\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \geq t) = P(Z \geq t) \leq .5e^{-t^2/2}$ (HW1).

Goal: $X_1, \ldots, X_n$ independent. Let $Z = f(X_1, \ldots X_n)$. We want to establish upper bounds on $P(|Z - m| \geq t)$ for some m $(\mathbb{E}(Z), median(Z), \ldots)$ at each finite n. Typically, $f(X) = \frac{\sum_{i=1}^n X_i}{n}$. We want to be agnostic with respect to the distribution of $X_1, \ldots, X_n$.

2.2.1 Basic bounds

- **Markov**: Let $X \geq 0$ and $t > 0$. $P(X \geq t) \leq \mathbb{E}(X)/t$.
- **Chebyshev**: Let X be a RV with finite variance and $t > 0$. $P(|X - \mathbb{E}(X)| \geq t) \leq V(X)/t^2$. Also, $P(|X - \mathbb{E}(X)| \geq t) \leq \min_{k\in\mathbb{N}} (X - \mathbb{E}(X))^k/t^k$.

2.2.2 Chernoff bounds

Let $t > 0$ and assume that the function $\lambda \in \mathbb{R} \mapsto \psi_{X-\mu}(\lambda) = log(\mathbb{E}(e^{\lambda(X-\mu)})$ exists $\forall \lambda \in (-b, b), b < \infty$. Then, for all $\lambda \in [0, b)$,

$$\begin{aligned} P(X - \mu \geq t) &= P(e^{X-\mu} \geq e^t) \\ &\leq P(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \\ &\leq \mathbb{E}[e^{\lambda(X-\mu)}]e^{-\lambda t} \\ &= e^{\psi_{X-\mu}(\lambda) - \lambda t} \\ \implies P(X - \mu \geq t) &\leq \exp -\psi^*_{X-\mu}(t), \end{aligned}$$

where $\psi^*_{X-\mu} = \sup_{\lambda\in[0,b)} e^{\lambda t - \psi_{X-\mu}(\lambda)}$.

In fact, we can extend the supremum over all $\lambda \in (-b, b)$ because $\psi_{X-\mu}(0) = 0, \psi^*_{X-\mu}(t) \geq 0, \forall t$, and $\psi_{X-\mu}(\lambda) \geq \lambda\mathbb{E}[X - \mu]$ by Jensen's inequality. Now we can get a concentration inequality:

$$P(|X - \mu| \geq t) \leq 2 \exp -\psi^*_{X-\mu}(t).$$

Example 2.3 (Normal) $X \sim N(\mu, \sigma^2)$.

$\mathbb{E}[e^{\lambda X}] = e^{\mu\lambda + \sigma^2\lambda^2/2}, \lambda \in \mathbb{R}$.

$\sup_{\lambda\geq 0} \lambda t - \psi_{X-\mu}(\lambda) = \frac{t^2}{2\sigma^2} \implies P(|X - \mu| \geq t) \leq 2 \exp \frac{-t^2}{2\sigma^2}$.

$\sup_{t\geq 0} P(Z \geq t) \exp \frac{t^2}{2\sigma^2}$

Remark: bound is of the form $c_1 \exp -t^2 c_2$, so we call it a Gaussian tail bound.

Definition 2.4 (Sub-gaussian RV) *A R.V. X s. t. $\mu = \mathbb{E}[X]$ exists and is sub-gaussian with parameter σ^2, $X \in SG(\sigma^2)$, when $\mathbb{E}[e^{\lambda(X-\mu)}] \leq \exp \frac{\lambda^2\sigma^2}{2}, \forall \lambda \in \mathbb{R}$*

Remarks:

1. $X \in SG(\sigma^2)$, then $-X \in SG(\sigma^2)$
2. Repeating the calculation for gaussian case: $\forall t > 0$, $P(|X - \mu| \geq t) \leq 2 \exp \frac{-t^2}{2\sigma^2}$.

3. For R.V. X, let $\sigma = \sigma(X) = \inf_{\sigma>0} : \mathbb{E}[\exp \lambda(X - \mu)] \leq \exp \frac{\lambda^2\sigma^2}{2}, \forall\lambda \in \mathbb{R}$. It turns out $V(X) \leq \sigma^2$ and, in some cases, $V(X) < \sigma^2$.

Properties of $SG(\sigma^2)$:

1. $V(X) \leq \sigma^2$;
2. if $a \leq X - \mu \leq b$, a, b finite, then $X \in SG((\frac{b-a}{2})^2)$;
3. if $X \in SG(\sigma^2)$, then $\alpha X \in SG(\alpha^2\sigma^2), \alpha \in \mathbb{R}$;
4. if $X \in SG(\sigma^2)$ and $Y \in SG(\tau^2)$, then $X + Y \in SG((\sigma + \tau)^2)$. If X and Y are independent, $X + Y \in SG(\sigma^2 + \tau^2)$.