# STA 385 Lecture 9

Recap: decision theory.

- Bayes criteria.

$$r_\pi(\hat\theta) = \int_\Theta R(\theta; \hat\theta)\,\pi(\theta)\,d\theta.$$

- Minimax criteria.

$$\bar R(\hat\theta) = \sup_{\theta\in\Theta} R(\theta; \hat\theta).$$

"minimax optimal":

$$\boxed{\inf_{\hat\theta}}\ \boxed{\sup_{\theta\in\Theta}}\ R(\theta, \hat\theta)$$

statistician chooses $\hat\theta$ first

nature (adversary) choose $\theta$ after observing $\hat\theta$.

Optimal $\hat\theta$ is from equilibrium of this game.

Analogy to rock-scissor-paper game:

If we randomize, adversary cannot make advantage out of it.

Basic results from game theory:

Alice choose $a \in A$

Bob choose $b \in B$.

$$\underbrace{\sup_{b \in B} \inf_{a \in A} R(a,b)}_{\text{Alice chooses after Bob}} \leq \underbrace{\inf_{a \in A} \sup_{b \in B} R(a,b)}_{\text{Alice choose before Bob.}}$$

. von Neumann's minimax theorem

$$\sup_{\pi_b} \inf_{\pi_a} \mathbb{E}_{\substack{a \sim \pi_a \\ b \sim \pi_b}} \left[ R(a,b) \right] = \inf_{\pi_a} \sup_{\pi_b} \mathbb{E}_{\substack{a \sim \pi_a \\ b \sim \pi_b}} \left[ R(a,b) \right]$$

(under some regularity conditions).

---

For minimax estimation.

$$\inf_{\hat{\theta}} \sup_{\pi} r_\pi(\hat{\theta}) = \sup_{\pi} \underbrace{\boxed{\inf_{\hat{\theta}} r_\pi(\hat{\theta})}}_{\text{optimal Bayes risk under prior } \pi.}$$

Implication: under certain regularity conditions, minimax estimator is the Bayes estimator under "least favorable prior".

---

Simple special cases.

- Constant-risk Bayes estimator $\hat{\theta}$ is minimax optimal.

Proof: Suppose $\tilde{\theta}$ is another estimator

$$\sup_{\theta \in \Theta} R(\theta; \tilde{\theta}) \geq \int_{\Theta} R(\theta; \tilde{\theta}) \pi(\theta) d\theta$$

$$\geq \int_{\Theta} R(\theta; \hat{\theta}) \pi(\theta) d\theta \quad \text{(since } \hat{\theta} \text{ is Bayes optimal under } \pi\text{).}$$

$$= \sup_{\theta \in \Theta} R(\theta; \hat{\theta}).$$

eg. $X \sim Binom(n, p)$, $R(p; \hat{p}) = \mathbb{E}_p[|p - \hat{p}|^2]$.

Natural choice: $\hat{p} = \frac{X}{n}$. $X = \sum_{i=1}^{n} Z_i$

$$R(p; \hat{p}) = \frac{var_p(Z_1)}{n} = \frac{p(1-p)}{n}$$

$$\bar{R}(\hat{p}) = \frac{1}{4n}, \text{ achieved at } p = \frac{1}{2}.$$

To construct prior distribution, we use conjugate families.

Conjugate family:

Suppose we work with $(P_\theta : \theta \in \Theta)$. $X_1, X_2 \dots X_n \overset{iid}{\sim} P_\theta$

Let $\mathcal{F}$ to be a class of prior distributions

We call it conjugate if $\forall \pi \in \mathcal{F}$

$$\Pi(\cdot \mid x_1 \dots x_n) \in \mathcal{F}.$$

(useful for computational convenience).

Beta - Binomial conjugacy.

If $\pi = \text{Beta}(\alpha, \beta)$

then $\Pi(\cdot \mid X) = \text{Beta}(\alpha + X, \beta + n - X)$.

( interpreted as "pseudo-counts").

Posterior mean (i.e. Bayes optimal estimator)

$$\hat{p}_{n,\pi} = \frac{\alpha + X}{\alpha + \beta + n}$$

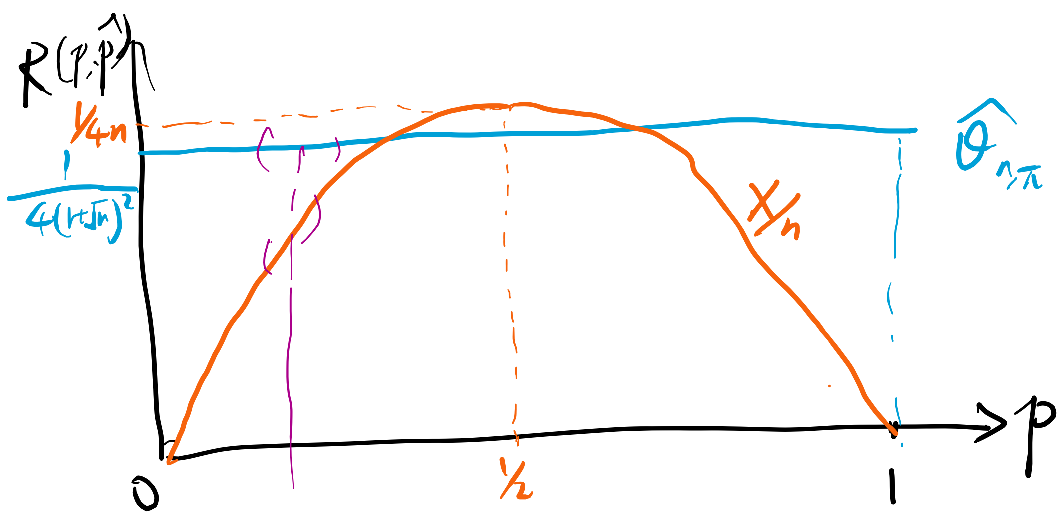
$$R(p;\hat{p}_{n,\pi}) = \frac{np(1-p)+(\alpha(1-p)-\beta p)^2}{(\alpha+\beta+n)^2}$$

Want it to be indep of $p$.

Solve for $(\alpha,\beta) \Rightarrow \begin{cases} \alpha = \sqrt{n}/2 \\ \beta = \sqrt{n}/2 \end{cases}$

$$\bar{R}(\hat{\theta}_n) = r_\pi(\hat{\theta}_n) = \frac{1}{4(1+\sqrt{n})^2}$$

Does it mean that we should use $\frac{\sqrt{n}}{2}$ pseudo-count?




This shows limitation of minimax criteria.

(In literature, people considered more refined optimality notion, e.g. local minimax)

---

Extension of this "simple case"

Fact: Suppose that we have a sequence of priors $(\pi_j)_{j=1}^{+\infty}$

$r_j$ = optimal Bayes risk under $\pi_j$

If $\exists$ an estimator $\hat{\theta}$

s.t. $\lim_{j\to+\infty} r_j \geq \sup_{\theta} R(\theta;\hat{\theta})$

then $\hat{\theta}$ is minimax optimal.

Proof: consider any other estimator $\tilde{\theta}$

$$\sup_{\theta\in\Theta} R(\theta;\tilde{\theta}) \geq \lim_{j\to+\infty} \int_{\Theta} R(\theta;\tilde{\theta})\,\pi_j(\theta)\,d\theta$$

$$\geq \lim_{j\to+\infty} r_j \geq \sup_{\theta} R(\theta,\hat{\theta}).$$

e.g. $X_1, X_2, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$

Construct prior $\pi_j = N(0, j^2)$

$$\pi(\cdot \mid X_1 \dots X_n) = N\left(\frac{j^2 n \cdot \bar{X}_n}{j^2 n+1}, \frac{j^2}{j^2 n+1}\right).$$

$$r_j = \frac{j^2}{j^2 n+1} \xrightarrow{j\to+\infty} \frac{1}{n}.$$

We also know $\sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta\left[|\bar{X}_n-\theta|^2\right] = \frac{1}{n}$

So $\bar{X}_n$ is minimax optimal.

---

Admissibility.

Let $\hat{\theta}$ be an estimator, if $\exists\, \tilde{\theta}$

s.t. $R(\theta; \tilde{\theta}) \le R(\theta; \hat{\theta})$ $(\forall \theta \in \Theta)$

and furthermore, $\exists\, \theta_0 \in \Theta$, s.t.

$R(\theta_0; \tilde{\theta}) < R(\theta_0; \hat{\theta})$

then we call $\hat{\theta}$ inadmissible.

We call $\hat{\theta}$ admissible when not inadmissible.

e.g. $X_1, \cdots X_n \overset{iid}{\sim} N(\theta, 1)$, $\hat{\theta}_n = 0$ admissible.

e.g. $X_1, \cdots, X_n \overset{iid}{\sim} N(\theta, I_d)$, $\theta \in \mathbb{R}^d$

$\bar{X}_n$ is the natural estimator

(unbiased, minimax, MLE, ...).

Stein's phenomenon: $\bar{X}_n$ inadmissible for $d \ge 3$.

James-Stein estimator

$$\hat{\theta}_{JS} = \left(1 - \frac{(d-2)}{n\|\bar{X}_n\|_2^2}\right) \bar{X}_n$$

"adaptive shrinkage towards 0"!

$$\mathbb{E}_\theta\left[\|\hat{\theta}_{JS}-\theta\|_2^2\right] = \frac{d}{n} - \mathbb{E}_\theta\left[\left(\frac{d-2}{n\|\bar{X}_n\|_2}\right)^2\right]$$

$$< \frac{d}{n} = \mathbb{E}_\theta\left[\|\bar{X}_n-\theta\|_2^2\right].$$

---

Fact. Bayes estimators are admissible. (for discrete case).

Fact. Under some regularity conditions admissible estimators are Bayes.

---

Bayesian stats (ct'd).

terval estimation in Bayesian setting:

Given $\pi$, $\theta|X_1 \cdots X_n \sim \pi(\cdot|X_1 \cdots X_n)$.

"Posterior interval / credible interval".

Find an interval $C_n$ that depends on data

s.t. $\pi(\theta \in C_n | X_1 \cdots X_n) = 1-\alpha.$

Warning: this is difference from confidence interval!

Conditionally on data, $C_n$ is deterministic, $\theta$ is random.

Asymptotic properties for Bayes estimators.

Given prior $\pi$, suppose $X_1 \cdots X_n \overset{iid}{\sim} P_{\theta^*}$

(We assume a frequentist model, Bayes estimator is just a choice of method).

$\theta^*$ is deterministic but unknown.

Under two conditions:

- $\pi(\theta$ at a local neighborhood of $\theta^*) > 0$.

  (e.g. when $\pi$'s density is positive around $\theta^*$)

- For the testing problem:

  $H_0: \theta = \theta^*$ v.s. $H_1: \|\theta - \theta^*\| > \varepsilon$

  $\exists$ a test $\phi_n$ s.t. $n \to +\infty$

  type-I, type-II err $\longrightarrow 0$

Thm (Schwartz). Under above two conditions,

posterior distr $\xrightarrow{P}$ point mass at $\theta^*$.

i.e. $\forall \varepsilon > 0$, $\Pi(\|\theta - \theta^*\| > \varepsilon \mid X_1 \cdots X_n) \xrightarrow{P} 0$.

Asymptotic shape of posterior:

Thm (Bernstein–von-Mises).

$X_1, \dots, X_n \overset{iid}{\sim} P_{\theta^*}$. under regularity conditions

(same conditions as required by MLE asymptotic normality)

(In particular, $I(\theta^*)$ invertible)

$$\text{Posterior} \approx N\left(\hat{\theta}_n, \left(nI(\theta^*)\right)^{-1}\right).$$

i.e. $$d_{TV}\left(\Pi(\cdot \mid X_1, \dots, X_n), N\left(\hat{\theta}_n, \left(nI(\theta^*)\right)^{-1}\right)\right) \xrightarrow{P} 0.$$

---

Implication. (in 1-D).

$$\Pi(\cdot \mid X_1, \dots, X_n) \approx N\left(\hat{\theta}_n, \left(nI(\theta^*)\right)^{-1}\right).$$

Credible interval $C_n \approx \left[\hat{\theta}_n \pm \frac{z_{1-\alpha/2}}{\sqrt{nI(\theta^*)}}\right]$. $\approx$ CI for $\theta$ using frequentist MLE

---

Sometimes we use "improper prior":

Given a density function $\pi$ (may not be a pdf)

(we may have $\int \pi(x)dx = +\infty$)

$$\Pi(\theta \mid X_1 \cdots X_n) = \frac{\pi(\theta) \cdot p_\theta(X_1) \cdots p_\theta(X_n)}{\int \pi(\theta') p_{\theta'}(X_1) \cdots p_{\theta'}(X_n)\, d\theta'}$$

Can be still well-defined.

---

Computational methods.

Usually $\Pi(\cdot \mid X_1 \cdots X_n)$ is difficult to compute.

Mainstream solutions:

- MCMC. Construct a Markov chain $(Z_t)_{t \geq 0}$

  s.t. $Z_t \xrightarrow{d} \Pi(\cdot \mid X_1 \cdots X_n) \quad (t \to +\infty)$

  We can simulate posterior by simulating $(Z_t)_{t \geq 0}$.

- VI. Given a class $\mathcal{Q}$ of tractable distributions

  $$\underset{Q \in \mathcal{Q}}{\text{minimize}}\ D_{KL}(Q \,\|\, \Pi(\cdot \mid X_1 \cdots X_n)).$$