STATS 300A: Theory of Statistics I **Autumn 2016/17**

Lecture 8: October 20

Lecturer: Joseph Romano *Scribes: Kenneth Tay*

8.1 Bayes Estimators

Recall the set-up for Bayes estimators:

Definition 8.1 *We are given a "weight" function Λ on Ω (usually a probability measure), $X \sim P_\theta$ with $\theta \in \Omega$. We wish to estimate $g(\theta)$, with loss function $L(\theta, d)$.*

*The **Bayes estimator** δ^* is the estimator which minimizes "average risk", i.e.*

$$\int_\theta \int_x L(\theta, \delta(x)) dP_\theta(x) d\Lambda(\theta),$$

*IfΛ is a probability distribution, it is called the **prior distribution** for θ.*

Assume that Λ is a probability distribution. We can rewrite the quantity to be minimized:

$$\int_\theta \int_x L(\theta, \delta(x)) dP_\theta(x) d\Lambda(\theta) = \mathbb{E}\left[L(\Theta, \delta(X))\right],$$

where the expectation is taken over the joint distribution of (x, θ), i.e. $\Theta \sim \Lambda$, and given $\Theta = \theta$, $X \sim P_\theta$.

We write the quantity in this way so that we can condition on X and use the Law of Iterated Expectation to find the Bayes estimator. The next theorem makes this precise:

Theorem 8.2 *In the set-up above, assume that*

1. *There exists an estimator δ_0 of $g(\theta)$ with finite average risk, and*
2. *For almost all x, $\delta_\Lambda(x)$ minimizes*
$$\mathbb{E}[L(\Theta, \delta(x)) \mid X = x].$$

Then δ_Λ is a Bayes estimator w.r.t. Λ, i.e. it minimizes average risk.

Proof: For any other δ,
$$\mathbb{E}[L(\Theta, \delta(x)) \mid X = x] \geq \mathbb{E}[L(\Theta, \delta_\Lambda(x)) \mid X = x]$$
for almost all x. Taking expectations on both sides and using the Law of Iterated Expectation, we get the desired result. ■

Some remarks on the theorem:

- The first condition in the Theorem ensures that the problem is meaningful. If it doesn't hold, then all estimators have infinite average risk, so any one of them would be a Bayes estimator.

- When we say "almost all" x in the second condition, "almost all" is with reference to the marginal (or unconditional) distribution of x, i.e. $Q(X \in E) = \int P_\theta(E) d\Lambda(\theta)$.

The theorem above gives the following examples:

- (Squared error loss) Suppose $L(\theta, d) = (d - \theta)^2$, $g(\theta) = \theta$. Then $\delta_\Lambda(X) = \mathbb{E}[\Theta \mid X]$, i.e. the mean of the **posterior distribution** (distribution of Θ given X).

  For general $g(\theta)$ we obtain the same result: $\delta_\Lambda(X) = \mathbb{E}[g(\Theta) \mid X]$.

- (Absolute error loss) Suppose $L(\theta, d) = |d - t|$. Then $\delta_\Lambda(X) =$ median of the conditional distribution of $g(\Theta) \mid X$.

The examples above show us that the posterior distribution is an important ingredient for determining the Bayes estimator. In general, we have

$$\text{posterior density of } \theta = \frac{\text{joint density of } (x, \theta)}{\text{marginal density of } x}.$$

Since the denominator does not depend on θ (it's just a constant that ensures that the joint density integrates to what it should), we can write

$$\begin{aligned}
\text{posterior density of } \theta &\propto \text{joint density of } (x, \theta), \\
&= \text{density of } (x \mid \theta) \times \text{density of } \theta \\
&=: \text{likelihood } \times \text{ prior density.}
\end{aligned}$$

We will be using this relation a lot in our calculations for Bayes estimators.

8.1.1 Uniqueness of Bayes Estimators

For squared error loss, we have the following theorem:

Theorem 8.3 *δ_Λ is unique (a.e. $\mathcal{P}$) if:*

1. *Its average risk w.r.t. Λ is finite, and*
2. *Almost everywhere w.r.t. $Q \Rightarrow$ almost everywhere w.r.t. $\mathcal{P}$, i.e. $Q(N) = 0 \Rightarrow P_\theta(N) = 0$ for all θ. (Q is the marginal distribution of X.)*

(**Remark**: Condition 2 is satisfied if the parameter space Ω is an open set and equal to the support of Λ, and $P_\theta(E)$ is continuous in θ for every E.)

Example: Binomial setting where $X_i \sim \text{Binom}(n, \theta)$. Suppose Λ puts all its mass on $\{0, 1\}$, i.e. we will only observe $X = 0$ or $X = n$. Then $\delta(X)$ will be Bayes as long as $\delta(0) = 0$ and $\delta(n) = 1$.

The following theorem tells us that when determining the "best" estimator, you can't automatically rule out unique Bayes estimators:

Theorem 8.4 *If $\delta_\Lambda(X)$ is a unique Bayes estimator, then $\delta_\Lambda(X)$ is admissible.*

Proof: Suppose δ' is any other estimator which dominates δ_Λ, i.e.

$$\begin{aligned} R(\theta,\delta') &\le R(\theta,\delta_\Lambda) && \text{for all } \theta, \\ R(\theta,\delta') &< R(\theta,\delta_\Lambda) && \text{for some } \theta. \end{aligned}$$

Averaging both sides w.r.t Λ, we get avg risk of $\delta' \le$ avg risk of δ_Λ. Since δ_Λ uniquely minimizes average risk, we must have $\delta' = \delta_\Lambda$. ■

8.1.2 Bayes Estimators are Biased

Under squared error loss, Bayes estimators cannot be unbiased (except in exceptional settings):

Theorem 8.5 *Under squared error loss, no unbiased estimator $\delta(X)$ is Bayes unless its average risk is 0, i.e.*

$$\mathbb{E}[(\delta(X) - g(\Theta))^2] = 0.$$

(Here, the expectation is taken over the joint distribution of X and Θ.)

Proof: Because δ is unbiased, $\mathbb{E}[\delta(X) \mid \Theta] = g(\Theta)$. Because δ is Bayes, $\delta(X) = \mathbb{E}[g(\Theta) \mid X]$.

We can compute $\mathbb{E}[\delta(X)g(\Theta)]$ in 2 ways:

$$\begin{aligned} \mathbb{E}[\delta(X)g(\Theta)] &= \mathbb{E}[\mathbb{E}[\delta(X)g(\Theta) \mid \Theta]] \\ &= \mathbb{E}[g(\Theta)\mathbb{E}[\delta(X) \mid \Theta]] \\ &= \mathbb{E}[g(\Theta)g(\Theta)], \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[\delta(X)g(\Theta)] &= \mathbb{E}[\mathbb{E}[\delta(X)g(\Theta) \mid X]] \\ &= \mathbb{E}[\delta(X)\mathbb{E}[g(\Theta) \mid X]] \\ &= \mathbb{E}[\delta(X)\delta(X)]. \end{aligned}$$

Thus,

$$\mathbb{E}[(\delta(X) - g(\Theta))^2] = \mathbb{E}[\delta(X)^2 - 2\delta(X)g(\Theta) + g(\Theta)^2] = 0.$$

■

8.1.3 Examples of Bayes Estimators for Squared Error Loss

8.1.3.1 Binomial setting

Assume $X \mid \Theta = \theta \sim \text{Binom}(n,\theta)$, prior $\Lambda = \text{Beta}(a,b)$, i.e. the prior has density

$$\lambda(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

Calculate the posterior density of θ:

$$\begin{aligned} \text{posterior density of } \theta &\propto \text{likelihood } \times \text{ prior density}, \\ &\propto \binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot \theta^{a-1}(1-t)^{b-1} \\ &\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1}, \end{aligned}$$

i.e. Θ has posterior distribution Beta(a', b') with $a' = x + a$, $b' = n - x + b$.

Recall that the mean of Beta(a, b) is given by $\frac{a}{a+b}$. Hence, the Bayes estimator of θ is

$$\begin{aligned}\delta_\Lambda(x) &= \frac{a'}{a'+b'} \\ &= \frac{x+a}{x+a+n-x+b} \\ &= \frac{x+a}{n+a+b} \\ &= \frac{x}{n} \cdot \frac{n}{n+a+b} + \frac{a}{a+b} \cdot \frac{a+b}{n+a+b}.\end{aligned}$$

The last line gives us the following interpretation: Recall that $\frac{X}{n}$ is the UMVU estimator for θ, while $\frac{a}{a+b}$ is the prior estimator for θ (i.e. best estimate without any data). Viewed in this way, the Bayes estimator is a convex combination of the UMVU estimator and prior estimator.

Also note that $\frac{n}{n+a+b} \to 1$ as $n \to \infty$, i.e. as the sample size grows, the Bayes estimator tends to the frequentist UMVU estimator.

8.1.3.2 Binomial setting vs. Geometric setting

Consider the binomial setting where we observe the data "TTTH". We have

$$\text{posterior} \propto \binom{4}{1}(1-\theta)^3\theta \cdot \text{prior}.$$

If we were in the geometric setting instead and observed the same data "TTTH", we would have

$$\text{posterior} \propto (1-\theta)^3\theta \cdot \text{prior},$$

which gives the same posterior as the binomial setting! Hence, in some sense, the Bayes estimator is the same regardless of the sampling rule.

8.1.3.3 Normal setting

$X_1, \ldots, X_n$ iid, $X_i \sim \mathcal{N}(\theta, \sigma^2)$, σ^2 fixed and known. Let $\Lambda = \mathcal{N}(\mu, b^2)$.

$$\begin{aligned}\text{posterior} &\propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i-\theta)^2\right] \cdot \frac{1}{\sqrt{2\pi}b} \exp\left[-\frac{(\theta-\mu)^2}{2b^2}\right] \\ &\propto \exp\left\{\frac{\theta\sum x_i}{\sigma^2} - \frac{n\theta^2}{2\sigma^2} - \frac{\theta^2}{2b^2} + \frac{\mu\theta}{b^2}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\theta^2\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right) - 2\theta\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{b^2}\right)\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)\left[\theta^2 - 2\theta\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{b^2}}{\frac{n}{\sigma^2}+\frac{1}{b^2}}\right]\right\}.\end{aligned}$$

Hence, the posterior distribution of θ is

$$\mathcal{N}\left(\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{b^2}}{\frac{n}{\sigma^2}+\frac{1}{b^2}}, \left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)^{-1}\right).$$

The Bayes estimator under squared error loss is

$$\frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu}{b^2}}{\frac{n}{\sigma^2} + \frac{1}{b^2}} = \bar{x}\frac{n/\sigma^2}{\frac{n}{\sigma^2} + \frac{1}{b^2}} + \mu\frac{1/b^2}{\frac{n}{\sigma^2} + \frac{1}{b^2}}.$$

Note that as in the binomial setting earlier, the Bayes estimator is a convex combination of the UMVU estimator and the prior estimator.

Note also that as $b \to \infty$, the Bayes estimator tends to $\bar{X}$. So, while $\bar{X}$ is never Bayes w.r.t. any probability distribution, it is Bayes w.r.t. to the improper prior Lebesgue measure on $\mathbb{R}$.

8.1.3.4 Improper priors

A proper prior is a probability distribution, while an improper prior is one that is not a probability distribution.

For improper priors, we can calculate the Bayes estimator in the same way, since we still have

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Typically, the posterior ends up being a probability distribution, in which case we can take the mean to be the Bayes estimator.

As an example, consider the improper prior Lebesgue measure on $\mathbb{R}$ for the Normal setting.

$$\begin{aligned}\text{posterior} &\propto \exp\left[-\frac{1}{2\sigma^2}\sum(x_i - \theta)^2\right] \cdot 1 \\ &\propto \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{x})^2\right],\end{aligned}$$

i.e. the posterior is $\mathbb{N}(\bar{x}, \sigma^2/n)$. Thus, the Bayes estimator is $\bar{X}$.