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

Lecture 9: October 25

Lecturer: Joseph Romano *Scribes: Kenneth Tay*

9.1 Bayes Estimators for Different Loss Functions

9.1.1 0-1 Loss

Assume that there are only 2 possible distributions having densities f_0 and f_1, i.e. either $X \sim f_0$ or $X \sim f_1$. Loss is 0 if you guess correctly, otherwise loss is 1. (For concreteness, let $\theta(f_0) = 0$ and $\theta(f_1) = 1$. Then this is the same as trying to estimate θ.)

Suppose that the prior puts mass π on f_0 and $1 - \pi$ on f_1.

Without any data, the best estimate is

$$\theta = \begin{cases} 0 & \text{if } \pi > 1 - \pi, \text{ i.e. } \pi > \frac{1}{2}, \\ 1 & \text{if } \pi < \frac{1}{2}, \\ \text{anything} & \text{if } \pi = \frac{1}{2}. \end{cases}$$

The Bayes risk in this case would be $\min(\pi, 1 - \pi)$.

Now, if data X is observed, then the best estimate would be

$$\theta = \begin{cases} 0 & \text{if posterior probability for } (\theta = 1) < \frac{1}{2}, \\ 1 & \text{if posterior probability for } (\theta = 1) > \frac{1}{2}. \end{cases}$$

By Bayes rule, we would choose $\theta = 0$ if

$$\begin{aligned} \text{posterior probability for } \theta = 0 &> \text{posterior probability for } \theta = 1, \\ \frac{\pi f_0(x)}{\pi f_0(x) + (1-\pi) f_1(x)} &> \frac{(1-\pi) f_1(x)}{\pi f_0(x) + (1-\pi) f_1(x)}, \\ \pi f_0(x) &> (1-\pi) f_1(x), \\ \frac{f_0(x)}{f_1(x)} &> \frac{\pi}{1-\pi}. \end{aligned}$$

The fraction on the LHS is called the **likelihood ratio**.

9.1.2 Weighted Squared Loss

In this case, the loss function has the form $L(\theta, d) = w(\theta)(d - \theta)^2$ for some weight function w.

First, consider the case where no data has been observed. What would be the Bayes estimator be? If the prior has density λ, then we would choose the constant d which minimizes

$$\int w(\theta)(d - \theta)^2 \lambda(\theta) d\theta,$$

or equivalently, d which minimizes

$$\frac{\int w(\theta)(d-\theta)^2\lambda(\theta)d\theta}{\int w(\theta)\lambda(\theta)d\theta}.$$

With this formulation, we can view $w\lambda$ as a density: letting $p(\theta) = \dfrac{w(\theta)\lambda(\theta)}{\int w(\theta)\lambda(\theta)d\theta}$, we are looking for d which minimizes $\int (d-\theta)^2 p(\theta)d\theta$. We are back to the case of squared error loss, and we know that the value of d which minimizes this quantity is

$$d^* = \int \theta p(\theta)d\theta = \frac{\int \theta w(\theta)\lambda(\theta)d\theta}{\int w(\theta)\lambda(\theta)d\theta} = \frac{\mathbb{E}[\Theta w(\Theta)]}{\mathbb{E}[w(\Theta)]}, \tag{9.1}$$

where $\Theta \sim \lambda$. If data is observed, we use $\Theta \sim$ posterior distribution instead.

9.1.2.1 Binomial setting

Suppose we are in a binomial setting where Θ has prior Beta(a, b), and $L(\theta, d) = \dfrac{(d-\theta)^2}{\theta(1-\theta)}$.

Since $\theta \sim$ Beta(a, b), we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1-\Theta}\right] &= \int_0^1 \frac{1}{1-t} t^{a-1}(1-t)^{b-1}dt \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\
&= \int_0^1 t^{a-1}(1-t)^{b-2}dt \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\
&= \frac{\Gamma(a)\Gamma(b-1)}{\Gamma(a+b-1)} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \\
&= \frac{a+b-1}{b-1}.
\end{aligned}$$

Similarly, we can compute

$$\mathbb{E}\left[\frac{1}{\Theta(1-\Theta)}\right] = \frac{(a+b-1)(a+b-2)}{(a-1)(b-1)}.$$

Thus, using Equation 9.1, in the absence of data the Bayes estimator is

$$\begin{aligned}
\frac{\mathbb{E}[\Theta w(\Theta)]}{\mathbb{E}[W(\Theta)]} &= \frac{\mathbb{E}\left[\frac{1}{1-\Theta}\right]}{\mathbb{E}\left[\frac{1}{\Theta(1-\Theta)}\right]} \\
&= \frac{a+b-1}{b-1} \cdot \frac{(a-1)(b-1)}{(a+b-1)(a+b-2)} \\
&= \frac{a-1}{a+b-2}.
\end{aligned}$$

Now, if we observe data X, the posterior distribution of Θ is Beta(a', b') with $a' = x + a$ and $b' = n - x + b$. Thus, the Bayes estimator with data is

$$\frac{a'-1}{a'+b'-2} = \frac{x+a-1}{n+a+b-2}.$$

(Note: If $a = b = 1$, then $\frac{x}{n}$ is Bayes.)

9.2 Admissibility

Problem: Suppose $X_1, \dots, X_n$ iid, $X_i \sim \mathcal{N}(\theta, \sigma^2)$ for all i, where σ^2 is known. (Without loss of generality, we let $\sigma^2 = 1$.) Find all real number pairs (a, b) such that $a\bar{X} + b$ is admissible under squared error loss.

Case 1: $0 < a < 1$**.**

We know that if $\Theta \sim \mathcal{N}(\mu, c^2)$, then the posterior distribution is normal with

$$\text{mean } \frac{nc^2}{\sigma^2 + nc^2}\bar{X} + \frac{\mu c^2}{\sigma^2 + nc^2}, \quad \text{variance } \left(\frac{n}{\sigma^2} + \frac{1}{c^2}\right)^{-1}.$$

In this case, the Bayes estimator is the mean, and is admissible. For any $0 < a < 1$ and any b, we can find corresponding μ and c such that

$$a = \frac{nc^2}{\sigma^2 + nc^2}, \quad b = \frac{\mu c^2}{\sigma^2 + nc^2}.$$

Hence, $a\bar{X} + b$ is admissible in this case.

Case 2: $a = 0$**.**

In this case, $a\bar{X} + b$ is a constant (just b), and hence is admissible (by our argument in Lecture 1).

Case 3: $a = 1$**,** $b \neq 0$**.**

In this case, $\bar{X} + b$ is dominated by $\bar{X}$ ($\bar{X}$ has same variance but no bias), and so is inadmissible.

Case 4: $a > 1$**.**

Let $\rho_\theta(a, b)$ denote the risk of $a\bar{X} + b$, i.e.

$$\rho_\theta(a, b) = \mathbb{E}_\theta[(a\bar{X} + b - \theta)^2] = \frac{a^2\sigma^2}{n} + (a\theta + b - \theta)^2.$$

Then $\rho_\theta(a, b) \geq \dfrac{a^2\sigma^2}{n} \geq \dfrac{\sigma^2}{n} = \rho_\theta(1, 0)$, i.e. $a\bar{X} + b$ is dominated by $\bar{X}$, and so is inadmissible.

Case 5: $a < 0$**.**

$$\begin{aligned}
\rho_\theta(a, b) &> [(a-1)\theta + b]^2 \\
&= (a-1)^2\left[\theta + \frac{b}{a-1}\right]^2 \\
&> \left[\theta + \frac{b}{a-1}\right]^2 \\
&= \rho_\theta\left(0, -\frac{b}{a-1}\right),
\end{aligned}$$

i.e. $a\bar{X} + b$ is dominated by the constant estimator $-\dfrac{b}{a-1}$, and so is inadmissible.

Case 6: $a = 1$**,** $b = 0$**.** We will prove that $\bar{X}$ is admissible.

Proof: Assume otherwise, i.e. there exists another estimator δ^* such that

$$\begin{aligned}
R(\theta, \delta^*) &\leq \frac{1}{n} \qquad \text{for all } \theta, \\
R(\theta, \delta^*) &< \frac{1}{n} \qquad \text{for some } \theta.
\end{aligned}$$

Since exponential families have continuous risk functions, there is a small enough $\varepsilon > 0$ and an interval (θ_0, θ_1) such that

$$R(\theta, \delta^*) < \frac{1}{n} - \varepsilon$$

for all $\theta \in (\theta_0, \theta_1)$.

Let r_b^* be the average risk of δ^* with respect to the $\mathcal{N}(0, b^2)$ prior, and let r_b be the average risk of the Bayes estimator δ_b with respect to the same $\mathcal{N}(0, b^2)$ prior. Then

$$\begin{aligned}
r_b &= \mathbb{E}[(\delta_b(X) - \Theta)^2] \\
&= \mathbb{E}\left[\mathbb{E}[(\delta_b(X) - \Theta)^2 \mid X]\right] \\
&= \mathbb{E}\left[\mathbb{E}[(\mathbb{E}\Theta - \Theta)^2]\right] \\
&= \mathbb{E}[\text{posterior variance}] \\
&= \mathbb{E}\left[\left(n + \frac{1}{b^2}\right)^{-1}\right] \\
&= \left(n + \frac{1}{b^2}\right)^{-1}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{\frac{1}{n} - r_b^*}{\frac{1}{n} - r_b} &= \frac{\frac{1}{b\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[\frac{1}{n} - R(\theta, \delta^*)\right] \exp\left[-\frac{1}{2b^2}\theta^2\right] d\theta}{\frac{1}{n} - \left(\frac{1}{n} + \frac{1}{b^2}\right)^{-1}} \\
&= \frac{n(1 + nb^2)}{b\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[\frac{1}{n} - R(\theta, \delta^*)\right] \exp\left[-\frac{1}{2b^2}\theta^2\right] d\theta \\
&\geq \frac{n(1 + nb^2)}{b\sqrt{2\pi}} \int_{\theta_0}^{\theta_1} \varepsilon e^{-\frac{1}{2b^2}\theta^2} d\theta.
\end{aligned}$$

As $b \to \infty$, $\displaystyle\int_{\theta_0}^{\theta_1} \varepsilon e^{-\frac{1}{2b^2}\theta^2} d\theta \to \varepsilon(\theta_1 - \theta_0)$ and $\frac{n(1+nb^2)}{b\sqrt{2\pi}} \to \infty$. Hence, for large enough b, we have

$$\begin{aligned}
\frac{1}{n} - r_b^* &> \frac{1}{n} - r_b, \\
r_b &> r_b^*,
\end{aligned}$$

which contradicts the fact that δ_b is the Bayes estimator. Hence, $\bar{X}$ is admissible. ■

9.2.1 Multivariate normal setting

Let $X_1, \dots, X_n$ iid, $X_1 \sim \mathcal{N}(\theta, \Sigma)$, where

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix}, \quad \Sigma = I_d.$$

The goal is to estimate θ with loss function $L(\theta, d) = \displaystyle\sum_{i=1}^{d} (\theta_i - d_i)^2$. We can ask the same question as in the univariate case: is $\bar{X}$ admissible? It turns out that:

- If $d < 3$, $\bar{X}$ is admissible.
- If $d \geq 3$, $\bar{X}$ is not admissible.

9.3 Minimax Estimation

In the minimax estimation setting, we are looking for the estimator δ^* which minimizes $\sup_\theta R(\theta, \delta)$.

Proposition 9.1 *If a Bayes estimator has constant risk, it is minimax.*

Proof: Let δ_Λ be the Bayes estimator with constant risk. If δ is an estimator with better worst case risk, then

$$\begin{aligned} & R(\theta, \delta) < R(\theta, \delta_\Lambda) && \text{for all } \theta, \\ \Rightarrow \quad & \int R(\theta, \delta) d\Lambda(\theta) < \int R(\theta, \delta_\Lambda) d\Lambda(\theta), \end{aligned}$$

which contradicts the definition of a Bayes estimator. ∎

The proposition above can be generalized to the following theorem:

Theorem 9.2 *For a prior distribution Λ, let $r_\Lambda := \int R(\theta, \delta_\Lambda) d\Lambda(\theta)$, i.e. the Bayes risk of the Bayes estimator w.r.t. Λ.*

If $r_\Lambda = \sup_\theta R(\theta, \delta_\Lambda)$, then:

1. *δ_Λ is minimax.*
2. *If δ_Λ is uniquely Bayes, then it is uniquely minimax as well.*
3. *Λ is "least favorable", i.e. $r_\Lambda \geq r_{\Lambda'}$ for any Λ'.*

Proof:

1. If δ is any other estimator, then

$$\begin{aligned} \sup_\theta R(\theta, \delta) &\geq \int R(\theta, \delta) d\Lambda(\theta) \\ &\geq \int R(\theta, \delta_\Lambda) d\Lambda(\theta) \\ &= \sup_\theta R(\theta, \delta_\Lambda). \end{aligned}$$

2. If δ_Λ is uniquely Bayes, then the second inequality above is strict, and so it is uniquely minimax as well.

3. For any other prior Λ',

$$\begin{aligned} r_{\Lambda'} &= \int R(\theta, \delta_{\Lambda'}) d\Lambda'(\theta) \\ &\leq \int R(\theta, \delta_{\Lambda}) d\Lambda'(\theta) \\ &\leq \sup_\theta R(\theta, \delta_\Lambda) \\ &= r_\Lambda. \end{aligned}$$

■

Note: The assumption in the theorem above holds if the risk of the estimator is constant. It can also hold if Λ puts all of its mass on θ values where the risk function is worst.