Statistics 210A Lecture 8 Notes

Daniel Raban

September 21, 2021

1 Bayes Estimation

1.1 Recap: Lower bound for unbiased estimation

Last time, we talked about the **score function**

$$\nabla \ell(\theta; x),$$

where $\ell(\theta; x) = \log p_\theta(x)$ is a log-likelihood. We saw some properties of the score function, like

$$\mathbb{E}_\theta[\nabla \ell(\theta; x)] = 0.$$

The Fisher information was

$$J(\theta) = \mathrm{Var}_\theta(\nabla \ell(\theta; x)) = -\mathbb{E}[\nabla^2 \ell(\theta; x)].$$

If $g(\theta) = \mathbb{E}_\theta[\delta(X)]$ with $g : \mathbb{R}^d \to \mathbb{R}$, then

$$\nabla g(\theta) = \mathrm{Cov}_\theta(\delta(X), \nabla \ell(\theta; X)).$$

Combining this with Cauchy-Schwarz gives the **Cramér-Rao lower bound**

$$\mathrm{Var}_\theta(\delta(X)) \geq \frac{\dot{g}(\theta)^2}{J(\theta)}, \qquad d = 1$$

with multivariate form

$$\mathrm{Var}_\theta(\delta(X)) \geq \nabla g(\theta)^\top J(\theta)^{-1} \nabla g(\theta), \qquad d \geq 1.$$

This gives us a lower bound on how small we can make our risk with unbiased estimation.

Example 1.1. Let $X \sim \mathrm{Binom}(n, \theta)$. Consider two estimators $\delta_0(x) = x/n$ and $\delta_1(X) = \frac{x+3}{n+6}$. The second estimator weights the estimation more towards $1/2$. How can we say that one is better than the other?

To compare these estimators, we previously ruled out all unbiased estimators. However, we can alternatively try to reduce the *average risk*.

1.2 Some problems with unbiased estimation

Unbiased estimation is not always desirable.

Example 1.2. Suppose $X \sim \text{Binom}(50, \theta)$ and $g(\theta) = \mathbb{P}_\theta(X \geq 25)$. The UMVU estimator is

$$\delta(X) = \mathbb{1}_{\{X \geq 25\}},$$

which is somewhat ridiculous because if we saw $X = 25$, we would assume this probability is 1.

Example 1.3. Suppose $X \sim N_d(\theta, I_d)$, where we want to estimate $\|\theta\|_2^2$. The UMVU estimator is $\|X\|_2^2 - d$ because

$$\mathbb{E}[\|X\|_2^2] = \|\theta\|_2^2 + d.$$

This estimator can be < 0, while $\|\theta\|_2^2$ cannot be. So we can always improve on the estimator by instead considering $(\|X\|^2 - d)^+$ instead.

1.3 Bayes estimation from a frequentist viewpoint

We have the model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ for the data X, a loss function $L(\theta; d)$, and the risk $R(\theta; \delta) = \mathbb{E}_\theta[L(\theta; \delta(X))]$.

Definition 1.1. The **Bayes risk** is

$$\begin{aligned} R_{\text{Bayes}}(\Lambda; \delta) &= \int_\Omega R(\theta; \delta)\, d\Lambda(\theta) \\ &= \mathbb{E}[R(\Theta; \delta(X))] \\ &= \mathbb{E}[L(\Theta; \delta(X))], \end{aligned}$$

where $\Theta \sim \Lambda$ and $X \mid \Theta = \theta \sim P_\theta$. This is the average-case risk, integrated with respect to a measure Λ on Ω, called the **prior**.

For now, we assume $\Lambda(\Omega) = 1$. Later, we will allow for $\Lambda(\Omega) = \infty$, which is called an **improper prior**.

Definition 1.2. $\delta(X)$ is a **Bayes estimator** if it minimizes $R_{\text{Bayes}}(\Lambda, \delta)$.

This definition depends on $\mathcal{P}$, Λ, and L. How do we find a Bayes estimator? Fortunately, they are easy to find.

Theorem 1.1. *Suppose $\Theta \sim \Lambda$ and $X \mid \Theta = \theta \sim P_\theta$. Assume that $L(\theta; d) \geq 0$ for all θ, d and that $R_{\text{Bayes}}(\Lambda; \delta_0) < \infty$ for some $\delta_0(X)$. Then*

$$\delta_\Lambda(x) \in \arg\min_d \mathbb{E}[L(\Theta; d) \mid X = x] \text{ for a.e. } x \iff \delta_\Lambda(X) \text{ is Bayes.}$$

So we split up the problem by solving it for any fixed x.

Proof. ($\Longrightarrow$): Let δ be any other estimator. Then

$$\begin{aligned} R_{\text{Bayes}}(\Lambda; \delta) &= \mathbb{E}[L(\Theta; \delta(X))] \\ &= \mathbb{E}[\mathbb{E}[L(\Theta; \delta(X)) \mid X]] \\ &\geq \mathbb{E}[\mathbb{E}[L(\Theta; \delta_\Lambda(X)) \mid X]] \\ &= R_{\text{Bayes}}(\Lambda; \delta_\Lambda). \end{aligned}$$

In particular, δ_Λ has finite Bayes risk because we could plug in δ_0 for δ.

($\Longleftarrow$): By contradiction. Let $E_x(d) := \mathbb{E}[L(\Theta; d) \mid X = x]$. Define

$$\delta^*(x) = \begin{cases} \delta_\Lambda(x) & \text{if } \delta_\Lambda(x) \in \arg\min E_x(d) \\ \delta_0(x) & \text{if } E_x(\delta_0(x)) < E_x(\delta_\Lambda(x)) \\ d^*(x) & \text{otherwise,} \end{cases}$$

where $E_x(d^*(x)) < E_x(\delta_\Lambda(x))$. By construction, we have

$$E_x(\delta^*(X)) \leq E_x(\delta_0(X))$$

a.s., so $R_{\text{Bayes}}(\Lambda, \delta^*) < \infty$. We also have

$$E_x(\delta^*(X)) \leq E_x(\delta_\Lambda(X))$$

a.s., with $<$ on a positive measure set. So

$$R_{\text{Bayes}}(\Lambda, \delta^*) \leq R_{\text{Bayes}}(\delta_\Lambda(X)),$$

which is a contradiction. $\square$

1.4 Posterior distributions

Definition 1.3. The conditional distribution of Θ given X is called the **posterior distribution**.

Definition 1.4. When we have densities $\lambda(\theta)$ for a prior and the likelihood $p_\theta(x)$, then the **marginal density** for X is

$$q(x) = \int_\Lambda \lambda(\theta) p_\theta(x) \, d\mu(\theta).$$

The **posterior density** is

$$\lambda(\theta \mid x) = \frac{\lambda(\theta) p_\theta(x)}{q(x)}.$$

In this case, the Bayes estimator is given by

$$\delta_\Lambda = \arg\min_d \int_\Omega L(\theta; d)\lambda(\theta \mid x)\, d\theta.$$

Proposition 1.1. *If $L(\theta; d) = (g(\theta) - d)^2$ is the squared error, then the Bayes estimator is the posterior mean $\mathbb{E}[g(\Theta) \mid X]$ of $g(\Theta)$.*

Proof.

$$\begin{aligned}\mathbb{E}[(g(\Theta) - \delta(X))^2 \mid X] &= \mathbb{E}[(g(\Theta) - \mathbb{E}[g(\Theta) \mid X] + \mathbb{E}[g(\theta) \mid X] - \delta(X))^2 \mid X] \\ &= \mathrm{Var}(g(\Theta) \mid X) + (\mathbb{E}[g(\Theta) \mid X] - \delta(X))^2,\end{aligned}$$

where the cross term is 0 because $\mathbb{E}[g(\Theta) - \mathbb{E}[g(\Theta) \mid X] \mid X] = 0$. This equals $\mathrm{Var}(g(\Theta) \mid X)$ if $\delta(X) \overset{\text{a.s.}}{=} \mathbb{E}[g(\Theta) \mid X]$. □

Let's now consider the **weighted square error** $L(\theta; d) = w(\theta)(g(\theta) - d)^2$. For example, we might take the relative error $L(\theta; d) = (\frac{\theta - d}{\theta})^2$.

Proposition 1.2. *For the weighted square error $L(\theta; d) = w(\theta)(g(\theta) - d)^2$, the Bayes estimator is*

$$\delta_\Lambda(X) = \frac{\mathbb{E}[w(\Theta)g(\Theta) \mid X]}{\mathbb{E}[w(\Theta)]}.$$

Example 1.4 (Beta-Binomial)**.** Suppose $X \mid \Theta = \theta \sim \mathrm{Binom}(n, \theta) = \theta^x (1-\theta)^{n-x}\binom{n}{x}$ with prior $\Theta \sim \mathrm{Beta}(\alpha, \beta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Note that in $X \mid \Theta = \theta$, θ is a parameter, whereas in the prior, we are giving a distribution over values of θ. The posterior distribution is

$$\lambda(\theta \mid x) = \frac{\lambda(\theta)p_\theta(x)}{q(x)}$$

Since this will integrate to 1 in θ, we will ignore the quantities not related to θ.

$$\begin{aligned}&\propto_\theta \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^x(1-\theta)^{n-x} \\ &= \theta^{x+\alpha-1}(1-\theta)^{n-x+\alpha-1} \\ &\propto_\theta \mathrm{Beta}(x+\alpha, n-x+\beta).\end{aligned}$$

So the posterior distribution is a different Beta distribution. Using what we know about the Beta distribution, we have

$$\mathbb{E}[\Theta \mid X] = \frac{X + \alpha}{n + \alpha + \beta}$$

The interpretation is that we have $k = \alpha + \beta$ "pseudo-trials" with α successes. We can write

$$\delta_\Lambda(x) = \frac{x}{n} \cdot \frac{n}{n+\alpha+\beta} + \frac{\alpha}{\alpha+\beta} \cdot \frac{\alpha+\beta}{n+\alpha+\beta}$$

If $n \gg \alpha+\beta$, we can say "the data swamps the prior," whereas for $n \ll \alpha+\beta$, we can say "the prior swamps the data."

Example 1.5 (Normal mean)**.** Suppose $X \mid \Theta = \theta \sim N(\theta, \sigma^2) \propto_\theta e^{-(x-\theta)^2/(2\sigma^2)}$, where σ^2 is known. Take the prior $\Theta \sim N(\mu, \tau^2) \propto_\theta e^{-(\theta-\mu)^2/(2\sigma^2)}$. The posterior is

$$\lambda(\theta \mid x) \propto_\theta \exp\left(\theta\left(\frac{x}{\sigma^2}+\frac{\mu}{\tau^2}\right)-\frac{\theta^2}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\right).$$

After some algebra,

$$\propto_\theta N\left(\frac{x/\sigma^2+\mu/\tau^2}{1/\sigma^2+1/\tau^2}, \frac{1}{1/\sigma^2+1/\tau^2}\right).$$

The posterior mean is

$$\mathbb{E}[\Theta \mid X] = X\frac{1/\sigma^2}{1/\sigma^2+1/\tau^2}+\mu\frac{1/\tau^2}{1/\sigma^2+1/\tau^2},$$

which is called a **precision-weighted average**.

These examples show that when calculating $\lambda(\theta \mid x)$, we should ignore the parts not depending on θ and try to recognize the resulting shape of the density as a distribution we know already.