STATS310A - Lecture 8

Dominik Rothenhaeusler
Scribed by Michael Howes

10/07/21

Contents

1 Announcement **1**

2 Recap **1**

3 Examples **2**
3.1 Absolute error 2
3.2 Weighted squared error loss 2
3.3 Binary classification 2
3.4 Normal likelihood and normal prior 3

4 Evaluating Bayes estimators **3**
4.1 Bias 3

1 Announcement

- Recommending reading on Bayesian vs Frequentist modeling in the scribed notes.
- Andrew Gelman's blog is highly recommended.
- "Why isn't everyone a Bayesian?" by Brad Efron is also recommended.
- Midterm is in two weeks in class on Oct 27th.
- Last year's midterm will be posted to Canvas.
- There will be a section for midterm revision.
- The exam is open book.

2 Recap

We have a new optimality goal: minimize the average risk. That is minimize

$$r(\Lambda, \delta) = \mathbb{E}[r(\Theta, \delta(X))],$$

where $\Theta \sim \Lambda$ and $X|\Theta = \theta \sim \mathbb{P}_\theta$. We saw that it suffices to minimize the posterior risk. That is the Bayes estimator is given by

$$\delta_\Lambda(x) = \arg\min_\delta \mathbb{E}[L(\Theta, \delta)|X = x].$$

We saw an example yesterday when we had

- Prior: Beta distribution.
- Likelihood: Binomial.
- Posterior: Beta distribution.

This is an example of *conjugate priors*.

Definition 1. Given a family of distribution $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$, a family of priors $\{\Lambda_\xi : \xi \in \Xi\}$ are *conjugate priors to* $\mathcal{P}$, if for every $x \in \mathcal{X}$ and $\xi \in \Xi$, there exists $\xi' \in \Xi$, such that $\Lambda_\xi | X = x$ has distribution $\Lambda_{\xi'}$.

Such priors are computational convenient.

3 Examples

3.1 Absolute error

Suppose that $L(\theta, d) = |\theta - d|$, then δ_Λ is the median of the posterior distribution.

3.2 Weighted squared error loss

Suppose that $L(\theta, d) = w(\theta)(\theta - d)^2$ for some weight function $w(\theta) \geq 0$. The Bayes estimator for this loss function minimises

$$\mathbb{E}[w(\theta)(\theta - \delta)^2 | X = x].$$

Note that

$$\begin{aligned}\mathbb{E}[w(\theta)(\theta - \delta)^2 | X = x] =& \mathbb{E}[w(\theta)\theta^2 | X = x] - 2\mathbb{E}[\delta w(\theta)\theta | X = x] + \mathbb{E}[w(\theta)\delta^2 | X = x] \\ =& \mathbb{E}[w(\theta)\theta^2 | X = x] - 2\delta\mathbb{E}[w(\theta)\theta | X = x] + \delta^2\mathbb{E}[w(\theta)\delta^2 | X = x].\end{aligned}$$

Thus the Bayes estimator is the value of δ that minimizes $-2\delta\mathbb{E}[w(\theta)\theta | X = x] + \delta^2\mathbb{E}[w(\theta) | X = x]$. Differentiating this with respect to δ we see that $-2[w(\theta)\theta | X = x] + 2\delta\mathbb{E}[w(\theta) | X = x]$ and so

$$\delta_\Lambda(x) = \frac{\mathbb{E}[\theta w(\theta) | X = x]}{\mathbb{E}[w(\theta) | X = x]}.$$

An alternative solution to this problem is the define a new measure

$$\Theta \sim w(\theta) d\Lambda(\theta).$$

Then δ_Λ is the conditional expectation under the new measure. Thinking of $w(\theta)d\Lambda(\theta)$ as a new measure is a useful computational trick.

3.3 Binary classification

Suppose $\Omega = \{0, 1\}$. For example we may be classifying emails as either spam or not span. $X \sim f_0$ or $X \sim f_1$, $D = \{0, 1\}$ and $L(\theta, d) = \mathbf{1}(\theta = d)$, where $\mathbf{1}(A)$ is the indicator function of A. Our prior Λ is specified by $\pi(1) = p$, $\pi(0) = 1 - p$. The average risk of an estimator is $P(\delta(X) \neq \Theta)$. We want to minimise

$$P(\delta \neq \Theta | X = x) = P(\Theta = 1 | X = x)\mathbf{1}(\delta = 0) + P(\Theta = 0 | X = x)\mathbf{1}(\delta = 1).$$

Thus we can see that if $P(\Theta = 1 | X = x) > P(\Theta = 0 | X = x)$, then we will pick $\delta = 1$ and otherwise we will pick $\delta = 0$. Note that $P(\Theta = 1 | X = x) \propto p f_1(x)$ and $P(\Theta = 0 | X = x) \propto (1 - p) f_2(x)$. Thus we will set $\delta = 1$ if and only if

$$\frac{f_1(x)}{f_0(x)} > \frac{1 - p}{p}.$$

Continuing this example suppose we have $\lambda_0, \lambda_1 > 0$ and we know $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda_\theta)$ where $\theta \in \{0, 1\}$. With the same loss and prior as before we know that we will set $\delta = 1$ if and only if

$$\frac{1-p}{p} < \frac{f_1(x)}{f_0(x)} = \frac{\lambda_1}{\lambda_0} \exp\left\{-(\lambda_1 - \lambda_0)\sum_{i=1}^n x_i\right\}.$$

Thus we set $\delta = 1$ if and only if $\sum_{i=1}^n x_i$ exceeds some value which is a function of p, λ_0 and λ_1. Thus the Bayes estimator depends only on the sufficient statistic $\sum_{i=1}^n X_i$. This is a general result. It holds because

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta) \alpha g_\theta(T(x))h(x)\pi(\theta) \propto g_\theta(T(x))\pi(\theta),$$

by the NFFC.

3.4 Normal likelihood and normal prior

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$ where σ^2 is known and $\Theta \sim N(\mu, b^2)$ where both μ and b^2 are known. We then have

$$\begin{aligned}
\text{posterior } &\propto \text{prior} \times \text{likelihood} \\
&\propto e^{-\frac{1}{2b^2}(\theta-\mu)^2} \prod_{i=1}^n e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2} \\
&\propto \exp\left\{-\frac{1}{2}\theta^2\left(\frac{1}{b^2} + \frac{n}{\sigma^2}\right) + \theta\left(\frac{\mu}{b^2} + \sum_{i=1}^n \frac{x_i}{\sigma^2}\right)\right\}.
\end{aligned}$$

Thus the posterior has a normal distribution with parameters.

$$\begin{aligned}
\text{Variance } &= \frac{1}{\frac{1}{b^2} + \frac{n}{\sigma^2}}, \\
\text{Mean } &= \left(\frac{\mu}{b^2} + \frac{n\bar{x}}{\sigma^2}\right) \cdot \frac{1}{\frac{1}{b^2} + \frac{n}{\sigma^2}} \\
&= \bar{x}\frac{\frac{n}{\sigma^2}}{\frac{1}{b^2} + \frac{n}{\sigma^2}} + \mu\frac{\frac{1}{b^2}}{\frac{1}{b^2} + \frac{n}{\sigma^2}} \\
&= \bar{x}w_1 + \mu w_2.
\end{aligned}$$

Since $w_1 + w_2 = 1$, we see that the Bayes estimator under squared error loss is again a convex combination of the UMVUE and the prior mean. We see again that the Bayes estimator depends only on the sufficient statistic $\bar{x}$.

4 Evaluating Bayes estimators

4.1 Bias

Under squared error loss Bayes estimators are always biased. More precisely:

Theorem 1. *If δ is unbiased for $g(\theta)$ with $r(\Lambda, \delta) < \infty$ and $\mathbb{E}[g(\Theta)^2] < \infty$, then δ is not the Bayes estimator under squared error loss unles $r(\Lambda, \delta) = 0$ (ie $g(\Theta) = \delta(X)$ with probability 1).*

Proof. We will proceed by contradiction. Suppose that δ is both unbiased and the Bayes estimator. We know that the Bayes estimator under squared error loss is given by

$$\delta(x) = \mathbb{E}[g(\Theta)|X = x].$$

It follows that

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

Also since $\delta(X)$ is unbiased, we have $\mathbb{E}[\delta(X)|\Theta = \theta] = g(\theta)$. This implies

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

Hence

$$\begin{aligned} r(\Lambda, \delta) &= \mathbb{E}[(g(\Theta) - \delta(X))^2] \\ &= \mathbb{E}[g(\Theta)^2] - 2\mathbb{E}[g(\Theta)\delta(X)] - \mathbb{E}[\delta(X)^2] \\ &= \mathbb{E}[g(\Theta)\delta(X)] - 2\mathbb{E}[g(\Theta)\delta(X)] - \mathbb{E}[g(\Theta)\delta(X)] \\ &= 0, \end{aligned}$$

as required. □

Suppose that $X_i \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$. The estimator $\delta(X) = \frac{1}{n}\sum_{i=1}^n X_i$ is not the Bayes estimator since it has risk $\frac{\sigma^2}{n} > 0$ and δ is unbiased.

Theorem 2. *A Bayes estimator with finite risk is admissible on the support of Λ. That is there are no estimators δ' satifying*

(a) $R(\theta, \delta') \leq R(\theta, \delta_\Lambda)$ with probability 1 under Λ.

(b) $R(\theta, \delta') < R(\theta, \delta_\Lambda)$ for all $\theta \in \Omega'$ for some $\Omega' \subseteq \Omega$ with $\Lambda(\Omega') > 0$.

Proof. Suppose that there did exist such a δ'. Then

$$r(\Lambda, \delta') = \int R(\theta, \delta')d\Lambda(\theta) < \int R(\theta, \delta_\Lambda)d\Lambda(\theta) = r(\Lambda, \delta_\Lambda),$$

a contradiction. □