Bayes Estimation

Outline

1) Bayes risk, Bayes estimator
2) Examples
3) Conjugate priors

Frequentist Motivation

Model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ for data X

Notation change: for now, reserving Θ to denote random variable

Loss $L(\theta, d)$, Risk $R(\theta; \delta)$

The Bayes risk is the average-case risk, integrated wrt some measure Λ, called prior

For now, assume $\Lambda(\Omega) = 1$ (prob. meas.)
Later we will allow to be improper ($\Lambda(\Omega) = \infty$)
(Note Λ and $c\Lambda$ for $c > 0$ functionally equiv.)

$$R_{\text{Bayes}}(\Lambda, \delta) = \int_\Omega R(\theta, \delta)\, d\Lambda(\theta)$$

$$= \mathbb{E}\, R(\Theta; \delta) \quad \text{where } \Theta \sim \Lambda$$

$$= \mathbb{E}\, L(\Theta, \delta(X)) \quad \begin{array}{l} \Theta \sim \Lambda \\ X \mid \Theta = \theta \sim P_\theta \end{array}$$

An estimator δ minimizing $R_{\text{Bayes}}(\Lambda; \cdot)$ is called Bayes (a Bayes estimator).
Depends on $\mathcal{P}$, Λ, L

Note avg-case loss makes sense even if we don't "believe" the parameter is random.

Bayes Estimator

Thm Suppose $\Theta \sim \Lambda$

$$X \mid \Theta = \theta \sim P_\theta$$

$$L(\theta, d) \geq 0 \quad \forall \theta, d$$

$$R_{Bayes}(\Lambda; \delta_0) < \infty \quad \text{for some } \delta_0(x)$$

Then $\delta_\Lambda(x) \in \underset{d}{\text{argmin}}\ \mathbb{E}\left[L(\Theta, d) \mid X = x\right]$ a.e. x

$\iff$ $\delta_\Lambda(x)$ is Bayes with $R_{Bayes}(\Lambda; \delta_\Lambda) < \infty$

Interpretation: we can find the Bayes estimator by minimizing $\mathbb{E}[L(\Theta, d) \mid X = x]$ "one x at a time."

<u>Proof</u> $(\Longrightarrow)$ Let δ be any other estimator

$$R_{Bayes}(\Lambda; \delta) = \mathbb{E}\, L(\Theta, \delta(X))$$

$$= \mathbb{E}\Big[\, \mathbb{E}\big[L(\Theta, \delta(x)) \mid X = x\big]\Big]$$

$$\geq \mathbb{E}\Big[\, \mathbb{E}\big[L(\Theta, \delta_\Lambda(x)) \mid X = x\big]\Big]$$

$$= R_{Bayes}(\Lambda, \delta_\Lambda)$$

$$(< \infty \quad \text{if} \quad \delta = \delta_0)$$

$(\Longleftarrow)$ Let $E_x(d) = \mathbb{E}\big[L(\Theta, d) \mid X = x\big]$

Define

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

where $E_x(d^*) < E_x(\delta_\Lambda(x))$

Then $E_x(\delta^*(x)) \overset{a.s.}{\leq} E_x(\delta_0(x))$

and $E_x(\delta^*(x)) \overset{a.s.}{\leq} E_x(\delta_\Lambda(x))$

with ineq. strict on a set of measure > 0.

$\boxtimes$

Prior, Posterior

Usual interp. of Λ is prior belief about Θ before seeing the data

Epistemic uncertainty: "I think there is a 50% chance that..."

[Mathematically, it makes no more or less sense to take θ as fixed or random, but a matter of philosophy whether this is scientifically appropriate] More on this later...

Conditional dist. of Θ given X, which we will write $\mathcal{L}(\Theta \mid X)$ (law), is called the posterior distribution, aka our beliefs after seeing the data.

Densities: prior $\lambda(\theta)$, likelihood $p_\theta(x)$

$$\Rightarrow q(x) = \int_\Omega \lambda(\theta) p_\theta(x) d\theta \quad \text{marginal density of } X$$

$$\lambda(\theta \mid x) = \frac{\lambda(\theta) p_\theta(x)}{q(x)} \quad \text{posterior density}$$

Bayes est: $\delta_\Lambda(x) = \arg\min_d \int_\Omega L(\theta, d) \lambda(\theta \mid x) d\theta$

Posterior Mean

If $L(\theta, d) = (g(\theta) - d)^2$ then the Bayes estimator is the posterior mean:

$$\mathbb{E}[(g(\Theta) - d)^2 \mid X]$$
$$= \mathbb{E}\left[\left(g(\Theta) - \mathbb{E}[g(\Theta) \mid X] + \mathbb{E}[g(\Theta) \mid X] - d\right)^2 \mid X\right]$$
$$= \mathrm{Var}(g(\Theta) \mid X) + \left(\mathbb{E}[g(\Theta) \mid X] - d\right)^2$$

(why is the cross-term 0?)

$$\Rightarrow \delta_\Lambda(x) = \mathbb{E}[g(\Theta) \mid X = x]$$

Weighted sq. error:

$$L(\theta, d) = w(\theta)(g(\theta) - d)^2 \qquad \text{e.g. } \left(\frac{\theta - d}{\theta}\right)^2$$

sq. rel. error

$$\mathbb{E}[(d - g(\Theta))^2 w(\Theta) \mid X]$$
$$= d^2 \, \mathbb{E}[w(\Theta) \mid X] - 2d \, \mathbb{E}[w(\Theta) g(\Theta) \mid X]$$
$$+ \mathbb{E}[\cancel{w(\Theta) g(\Theta)^2} \mid X]$$

no dep. on d

$$\text{min at } d = \frac{\mathbb{E}[w(\Theta) g(\Theta) \mid X]}{\mathbb{E}[w(\Theta) \mid X]} \quad (= \delta_\Lambda(x))$$

Example: Beta-Binomial

$$X \mid \Theta = \theta \sim \text{Binom}(n, \theta) = \theta^x (1-\theta)^{n-x} \binom{n}{x}$$

$$\Theta \sim \text{Beta}(\alpha, \beta) = \theta^{\alpha-1}(1-\theta)^{\beta-1} \underbrace{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}}_{\text{normalizing const.}}$$

θ is r.v. here

Marginal dist. of X called <u>Beta-Binomial</u>

<u>Posterior</u>:

$$\lambda(\theta \mid x) = \lambda(\theta)\, p_\theta(x) / q(x)$$

we can drop factors that don't depend on θ

$$\propto_\theta \theta^{\alpha-1}(1-\theta)^{\beta-1}\, \theta^x (1-\theta)^{n-x}$$

$$= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

$$\Rightarrow \Theta \mid X = x \sim \text{Beta}(x+\alpha,\, n-x+\beta)$$

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

convex combo of $\frac{X}{n}$, $\frac{\alpha}{\alpha+\beta}$

$$= \underbrace{\frac{X}{n}}_{\text{UMVU}} \cdot \frac{n}{n+\alpha+\beta} + \underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{Prior Expectation}} \cdot \underbrace{\frac{\alpha+\beta}{n+\alpha+\beta}}_{\left(1-\frac{n}{n+\alpha+\beta}\right)}$$

<u>Interp.</u>: $k = \alpha+\beta$ "pseudo-trials," α successes

(Recall $\frac{X+3}{n+6}$ from Lec. 2)

Example: Normal mean

$$X \mid \Theta = \theta \sim N(\theta, \sigma^2) \propto_\theta e^{-(x-\theta)^2/2\sigma^2}$$

$$\Theta \sim N(\mu, \tau^2) \propto_\theta e^{-(\theta-\mu)^2/2\tau^2}$$

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

$$\propto_\theta \exp\left\{ \frac{x\theta}{\sigma^2} - \frac{\theta^2}{2\sigma^2} - \frac{\theta^2}{2\tau^2} + \frac{\theta\mu}{\tau^2} \right\}$$

$$= \exp\left\{ \theta \underbrace{\left(\frac{x}{\sigma^2} + \frac{\mu}{\tau^2}\right)}_{b} - \theta^2 \underbrace{\left(\frac{\sigma^{-2}+\tau^{-2}}{2}\right)}_{a^2} \right\}$$

Complete square:

$$a^2\theta^2 - b\theta = \left(\theta a - \frac{b}{2a}\right)^2 - c(a,b)$$

$$= \left(\theta - \frac{b}{2a^2}\right)^2 \cdot a^2 - c$$

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

$$\propto_\theta N\left(\underbrace{\frac{x\sigma^{-2} + \mu\tau^{-2}}{\sigma^{-2}+\tau^{-2}}}_{\text{precision-weighted average of } x,\, \mu}, \underbrace{\frac{1}{\sigma^{-2}+\tau^{-2}}}_{\text{harmonic mean of } \sigma^2,\, \tau^2} \right)$$

$$\mathbb{E}[\Theta \mid x] = x \cdot \frac{\sigma^{-2}}{\sigma^{-2}+\tau^{-2}} + \mu \cdot \frac{\tau^{-2}}{\sigma^{-2}+\tau^{-2}}$$

Gaussian iid sample

$\theta \sim N(\mu, \tau^2)$ (dropping Θ for computations)

$X_i \mid \theta \overset{iid}{\sim} N(\theta, \sigma^2)$, $i = 1, \ldots, n$

$\bar{X} \mid \theta \sim N(\theta, \frac{\sigma^2}{n})$

$$\Rightarrow \mathbb{E}[\theta \mid X] = X \cdot \frac{n\sigma^{-2}}{n\sigma^{-2} + \tau^{-2}} + \mu \cdot \frac{\tau^{-2}}{n\sigma^{-2} + \tau^{-2}}$$

$$= X \cdot \frac{n}{n + \sigma^2/\tau^2} + \mu \cdot \frac{\sigma^2/\tau^2}{n + \sigma^2/\tau^2}$$

Interp: $k = \sigma^2/\tau^2$ pseudo-observations, mean μ

If $n \gg k$, "data swamps prior"

If $n \ll k$, "prior swamps data"

Note in both examples :

- Prior & Likelihood have similar fcn. form
- Posterior comes from same exp. fam. as prior
- Prior can be interp. as "pseudo-obs."

This is because we chose conjugate priors for the parameter; topic for next lecture.

Bias and Bayes

[Bayes estimators are almost always biased, especially when the parameter value is extreme relative to the prior]

Theorem The posterior mean is biased unless

$$\delta_\Lambda(x) \overset{a.s.}{=} g(\Theta)$$

Proof Suppose δ_Λ is unbiased. Then

$$g(\Theta) = \mathbb{E}[\delta(X) \mid \Theta]$$

$$\delta_\Lambda(X) = \mathbb{E}[g(\Theta) \mid X] \quad \text{(by def.)}$$

Condition on X:

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

$$= \delta_\Lambda(x)^2$$

Cond. on Θ:

$$\mathbb{E}[\delta_\Lambda(X) g(\Theta) \mid \Theta] = g(\Theta)^2$$

$$\Rightarrow \mathbb{E}[\delta_\Lambda g(\Theta)] = \mathbb{E}[\delta_\Lambda^2] = \mathbb{E}[g(\Theta)^2]$$

$$\mathbb{E}[(\delta_\Lambda - g(\Theta))^2] = \mathbb{E}\delta_\Lambda^2 + \mathbb{E}g(\Theta)^2 - 2\mathbb{E}[\delta_\Lambda g(\Theta)]$$

$$= 0 \qquad \boxtimes$$