Bayes Estimation

(for Frequentists!)

Outline

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

Frequentist Motivation

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

Loss $L(\theta, d)$, Risk $R(\theta; \delta) = \mathbb{E}_\theta[L(\theta, \delta(X))]$

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

For now, assume $\Lambda(\Theta) = 1$ (prob. meas.)
Later we will allow to be improper ($\Lambda(\Theta) = \infty$)

Note: • Λ and $c\Lambda$ for $c > 0$ functionally equiv.
• avg risk makes sense even if we don't "believe" $\theta \sim \Lambda$

$$r_\Lambda(\delta) = \int_\Theta R(\theta, \delta)\, d\Lambda(\theta)$$

$$= \mathbb{E}_{\theta \sim \Lambda}\big[\overbrace{R(\theta; \delta)}^{\mathbb{E}[L(\theta, \delta(X)) \mid \theta]}\big] \quad \text{where } \theta \sim \Lambda$$

$$= \mathbb{E}[L(\theta, \delta(X))] \quad \underbrace{\text{where } \theta \sim \Lambda, \; X \mid \theta \sim P_\theta}$$

$\mathbb{E}$ now means wrt joint distr. of (θ, X)

An estimator δ minimizing $r_\Lambda(\cdot)$ is called Bayes (a Bayes estimator). Dep. on $\mathcal{P}, \Lambda, L$

$$r_\Lambda(\delta) = \mathbb{E}\big[\,\mathbb{E}[L(\theta, \underbrace{\delta(X)}_{})) \mid X]\,\big]$$

Note: we choose this after seeing X

Prior, Posterior

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

Conditional dist. $\Lambda(\theta \mid X]$ called posterior dist.

"belief after seeing the data"

Epistemic uncertainty:

"I think there is a 50% chance that..."

More on this next time

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

Joint density $\lambda(\theta) p_\theta(x)$

Marginal density $q(x) = \int_\Theta \lambda(\theta) p_\theta(x) d\theta$

Posterior density $\lambda(\theta \mid x) = \dfrac{\lambda(\theta) p_\theta(x)}{q(x)}$

Bayes estimator depends on posterior:

$$\delta_\Lambda(x) = \arg\min_d \mathbb{E}[L(\theta, d) \mid X]$$

$$= \arg\min_d \int_\Theta L(\theta, d) \lambda(\theta \mid x) d\theta$$

Solve for Bayes estimator "one x at a time"

Bayes Estimator

Suppose $X|\theta \sim P_\theta$, $L(\theta, d) \geq 0$

$r_\Lambda(\delta_0) < \infty$ for some $\delta_0(x)$

Then $\delta_\Lambda(x)$ is Bayes with $r_\Lambda(\delta_\Lambda) < \infty$

iff $\delta_\Lambda(x) \in \underset{d}{\operatorname{argmin}}\ \mathbb{E}[L(\theta, d) \mid X = x]$ a.e. x

$\mathbb{P}(\delta_\Lambda(x) \notin \operatorname{argmin}) = 0$

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

$$\begin{aligned}
r_\Lambda(\delta) &= \mathbb{E}\big[\,\mathbb{E}[L(\theta, \delta(x)) \mid X = x]\,\big] \\
&\geq \mathbb{E}\big[\,\mathbb{E}[L(\theta, \delta_\Lambda(x)) \mid X = x]\,\big] \\
&= r_\Lambda(\delta_\Lambda) \\
&< \infty \quad (\text{take } \delta = \delta_0)
\end{aligned}$$

$(\Longleftarrow)$ Define $E_x(d) = \mathbb{E}[L(\theta, d) \mid X = x]$

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

Then $E_x(\delta^*(x)) \leq \min\big(E_x(\delta_0(X)), E_x(\delta_\Lambda(x))\big)$ $\forall x$

with ineq. strict on a set of measure > 0. $\boxtimes$

Posterior Mean

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

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

(why is the cross-term 0?)

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

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

$$\mathbb{E}[(d - g(\theta))^2 w(\theta) | X]$$
$$= d^2 \mathbb{E}[w(\theta)|X] - 2d\,\mathbb{E}[w(\theta)g(\theta)|X]$$
$$+ \underbrace{\mathbb{E}[w(\theta)g(\theta)^2 | X]}_{\text{no dep. on } d}$$

min at $d = \dfrac{\mathbb{E}[w(\theta)g(\theta)|X]}{\mathbb{E}[w(\theta)|X]} \quad (= \delta_\Lambda(x))$

Example: Beta-Binomial

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

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

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 $\frac{x}{n}$, $\frac{\alpha}{\alpha+\beta}$

$$= \overset{\text{UMVU}}{\frac{x}{n}} \cdot \frac{n}{n+\alpha+\beta} + \overset{\text{Prior Expectation}}{\frac{\alpha}{\alpha+\beta}} \cdot \overbrace{\frac{\alpha+\beta}{n+\alpha+\beta}}^{\left(1-\frac{n}{n+\alpha+\beta}\right)}$$

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

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

Example: Normal mean

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

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

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

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

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

Complete square:

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

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

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

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

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

Gaussian iid sample

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

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

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

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

<u>Interp</u>: $k = \sigma^2/\tau^2$ 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]

If the posterior is from the same family as the prior, we say the prior is <u>conjugate</u> to the likelihood.

Most common in exp. families

Conjugate Priors

Suppose
$$X_i \mid \eta \overset{iid}{\sim} p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x) \qquad \eta \in \Xi \subseteq \mathbb{R}^s, \quad i = 1, \dots, n$$

For carrier $\lambda_0(\eta)$, define $s+1$-dim family:
$$\lambda_{\mu,k}(\eta) = e^{k\mu'\eta - kA(\eta) - B(k\mu, k)} \lambda_0(\eta)$$

Suff. stat $\begin{pmatrix} \eta \\ -A(\eta) \end{pmatrix} \in \mathbb{R}^{s+1}$ Nat. param. $\begin{pmatrix} k\mu \\ k \end{pmatrix}$

$$\Rightarrow \lambda(\eta \mid x_1, \dots, x_n) \propto_\eta \left(\prod_{i=1}^n e^{\eta' T(x_i) - A(\eta)} h(x_i) \right) \cdot e^{k\mu'\eta - kA(\eta) - B(k\mu, k)} \lambda_0(\eta)$$
$$\propto_\eta e^{(k\mu + \Sigma T(x_i))'\eta - (k+n) A(\eta)} \lambda_0(\eta)$$
$$= \lambda_{\mu_{post},\, k+n}(\eta)$$

where
$$\mu_{post} = \frac{k\mu + n\bar{T}}{k+n}, \qquad \bar{T}(X) = \frac{1}{n} \sum_{i=1}^n T(x_i)$$

often Bayes est. for $\mathbb{E}_\eta T$

then
$$\mu_{post} = \bar{T} \cdot \frac{n}{k+n} + \mu \cdot \frac{k}{k+n}$$

UMVUE from data — "UMVUE" from "pseudo data"

Conjugate Prior Examples

Likelihood	Prior
$X_i \mid \theta \sim \text{Binom}(n,\theta)$ $= \theta^{x}(1-\theta)^{n-x}\binom{n}{x}$	$\theta \sim \text{Beta}(\alpha,\beta)$ $= \theta^{\alpha-1}(1-\theta)^{\beta-1}\dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
$X_i \mid \theta \sim N(\theta,\sigma^2)$ (σ^2 known) $= \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-(\theta-x)^2/2\sigma^2}$	$\theta \sim N(\mu,\tau^2)$ $= \dfrac{1}{\sqrt{2\pi\tau^2}} e^{(\theta-\mu)^2/2\tau^2}$
$X_i \mid \theta \sim \text{Pois}(\theta)$ $x=0,1,\ldots$ $= \dfrac{\theta^{x} e^{-\theta}}{x!}$	$\theta \sim \text{Gamma}(v,s)$ $\theta > 0$ $= \dfrac{1}{\Gamma(v)s^{2}}\theta^{v-1}e^{-\theta/s}$

Gamma / Poisson:

$$\lambda(\theta \mid x) \propto_{\theta} \theta^{v-1+\Sigma x_i} e^{-(s^{-1}+n)\theta}$$

$$= \text{Gamma}\left(v+\Sigma x_i,\ (s^{-1}+n)^{-1}\right)$$

$\Rightarrow k = s^{-1}, \quad \mu = vs$

$\lambda_0(\theta) = \theta^{-1}$ (not normalizable)

Flexibility of Bayes

Any Λ, $\mathcal{P}$, L, $g(\theta)$: δ_Λ defined straightforwardly

$$\delta_\Lambda(x) = \arg\min_d \int L(\theta, d)\, \lambda(\theta \mid x)\, d\theta$$

Problem reduced to (possibly hard) computation

Posterior is "one stop shop" for all answers

No need for:

- special family structure (exp. fam. / complete s.s.)
- special estimator (U-estimable)
- convex or nice L

$\Rightarrow$ Highly expressive modeling & estimation

Caveat: Limited by ability to do computations

Source #2: "Objective" or "vague" prior

Using default prior removes subjectivity
(But then what does the posterior mean?)

<u>Flat prior</u> $\lambda(\theta) \propto_\theta 1$ on Θ

"Indifference" (in θ parameterization)

Often <u>improper</u> ($\Lambda(\Theta) = \infty$) but usually ok

<u>Ex</u>: $\theta \sim$ flat prior on $\mathbb{R}$

$X \mid \theta \sim N(\theta, \sigma^2)$

$$\lambda(\theta \mid x) \propto_\theta p_\theta(x)$$
$$= \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2\sigma^2}$$
$$\propto_\theta N(x, \sigma^2)$$

<u>Jeffreys prior</u> $\lambda(\theta) \propto_\theta |J(\theta)|^{1/2}$

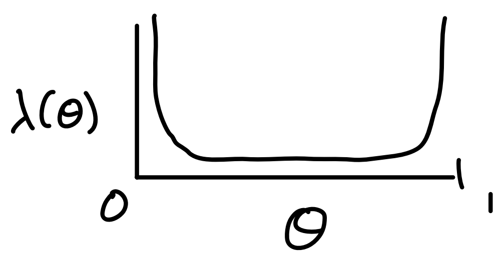
Higher density where P_θ "changing faster"

Invariant to parameterization

<u>Ex</u>. $X \mid \theta \sim \text{Binom}(n, \theta)$

$$\lambda(\theta) \propto_\theta J(\theta)^{1/2} = \left(\frac{n}{\theta(1-\theta)}\right)^{1/2} \propto_\theta \text{Beta}\left(\tfrac{1}{2}, \tfrac{1}{2}\right)$$

$\lambda(\theta) \to \infty$ as $\theta \to 0$ or 1:



(33x)

$D_{KL}(0.001 \,\|\, 0.01) \gg D_{KL}(0.49 \,\|\, 0.5)$

$7n \cdot 10^{-3}$ $\quad$ $2n \cdot 10^{-4}$