Lecture 8. Bayes Estimators and Average Risk Optimality.

9th November 2020.

---

* Explore an alternative approach to achieve optimality.

* X (data). $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, $L(\theta, d)$ loss-function, $R(\theta, \delta)$ $= E_\theta(L(\theta, \delta(X)))$

* Average risk Optimality.

We need to introduce a measure Λ over the parameter space Ω. This measure Λ can be viewed as an assignment of weights to each of the parameter values $\theta \in \Omega$ a priori. [i.e. before any data is observed]. ← Prior distribution.

Θ: θ fixed / unknown constant

Given a measure Λ, our objective is to find an estimator δ_Λ which minimises the average risk, which is given by

$$r(\Lambda, \delta) = \int R(\theta, \delta)\, d\Lambda(\theta) = E_\Theta(R(\Theta, \delta))$$

If Λ is a probability distribution on Ω, we call Λ the prior distribution. Correspondingly, the estimator δ_Λ, if exists, is called the Bayes estimator with respect to Λ, and the minimised average risk is called the Bayes risk.

$$\begin{aligned} r(\Lambda, \delta) &= E_{(X,\Theta)}\left(L(\Theta, \delta(X))\right) \\ &= E_\Theta\left(E_X\left(L(\Theta, \delta(X)) \mid \Theta\right)\right) \\ &= E_\Theta\left(R(\Theta, \delta)\right). \end{aligned}$$

We shall pay attention to $E\left(L(\Theta, \delta(X)) \mid X = x\right)$, the conditional risk at (almost) every value of X. Notice that the expectation here is taken with respect to the cond. distribution of Θ given X. $(\Theta \mid X = x)$. ← data.

[THEOREM]. Suppose $\Theta \sim \Lambda$ and $X \mid \Theta = \theta \sim P_\theta$. If

(a) there exists δ_0, an estimator of $g(\theta)$ with finite risk for all θ, and

(b) there exists a value $\delta_\Lambda(x)$ that minimises

→ $E\left(L(\Theta, \delta_\Lambda(X)) \mid X = x\right)$ for almost every x, (*)

then δ_Λ is a Bayes estimator with respect to Λ.

Note: The almost sure statement (*) is defined with respect to the marginal distribution of X, which is given by

$$P(X \in A) = \int P_\theta(X \in A)\, d\Lambda(\theta).$$

Proof. Under the assumptions of the theorem (a) & (b), for any other estimator δ', say, and for almost every x,

$$E\left(L(\Theta, \delta_\Lambda(X)) \,\middle|\, X=x\right) \leq E\left(L(\Theta, \delta'(X)) \,\middle|\, X=x\right).$$

Taking expectation over X, we obtain

$$E\left(L(\Theta, \delta_\Lambda(X))\right) \leq E\left(L(\Theta, \delta'(X))\right)$$

for all δ'.

(Example). If we consider the squared loss function $L(\theta, d) = (\theta - d)^2$, to find the Bayes estimator, we need to minimise $E\left(\{g(\Theta) - \delta(X)\}^2 \mid X=x\right)$ and in this case, the Bayes estimator is $\delta_\Lambda(X) = E(g(\Theta)|X)$, the posterior mean of $g(\Theta)$ given $X=x$.

$$= E\left(\{g(\Theta) - E(g(\Theta)|X) + E(g(\Theta)|X) - \delta(X)\}^2 \mid X=x\right)$$

$$= E\left(\{g(\Theta) - E(g(\Theta)|X)\}^2 \mid X=x\right) + E\left(\{E(g(\Theta)|X) - \delta(X)\}^2 \mid X=x\right)$$

posterior mean $E_{\Theta|X=x}$

$$\text{posterior} = \frac{\text{joint}}{\text{marginal}} = \frac{\text{prior} \times \text{likelihood}}{\text{marginal}}.$$

Bayes' Theorem:

$$p(\theta|x) = \frac{p(\theta, x)}{\int p(\theta', x)\, d\theta'} = \frac{p(x|\theta)\,\pi(\theta)}{\int p(\theta', x)\, d\theta'}$$

(likelihood, prior)

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

(Example) Suppose $X \sim \text{Binomial}(n, \theta)$ given $\Theta = \theta$ and that Θ has a prior distribution Beta (α, β). The prior density is given by

hyperparameters.

non-informative / flat prior. Conjugate prior (共軛先驗)

$$\pi(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}\, I(0<\theta<1).$$

prior; normalising const.

Obviously, the model density is $f(x;\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$, in which case the posterior distribution of Θ given X is

$$\pi(\theta|x) \propto \left\{\binom{n}{x} \theta^x (1-\theta)^{n-x} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}\right\}$$

kernel

likelihood prior

$\int$ ← NUMERATOR(θ') → $d\theta'$

normalising constant.

If $X \sim \text{Beta}(\alpha, \beta)$

$E(X) = \frac{\alpha}{\alpha+\beta}$.

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

$$\sim \text{Beta}(x+\alpha,\ n-x+\beta) \qquad \frac{x+\alpha}{x+\alpha+n-x+\beta}$$

meaning that the posterior mean of $\Theta|X = \frac{x+\alpha}{n+\alpha+\beta}$.

Remark: The posterior mean can be rewritten as:

$$\frac{X+\alpha}{n+\alpha+\beta} = \frac{n}{n+\alpha+\beta}\left(\frac{X}{n}\right) + \frac{\alpha+\beta}{n+\alpha+\beta}\left(\frac{\alpha}{\alpha+\beta}\right)$$

Contribution from the prior.

ω $\quad$ $1-\omega$

Shrink the estimate towards $\alpha/(\alpha+\beta)$

"weighted average of the sample mean $\bar{X}_n$ and the prior mean $\alpha/(\alpha+\beta)$"

As $n \to \infty$ $E(\Theta|X) \to \bar{X}_n$. (Let the data "speak for themselves")

empirical evidence/observations...

(Example). (Normal Mean Estimation)

Let $X_1, \dots, X_n \overset{iid}{\sim} N(\Theta, \sigma^2)$, with σ^2 known. Furthermore, let $\Theta \sim N(\mu, b^2)$ where μ and b^2 are two fixed hyperparameters (prior). Then the posterior dist^n of $\Theta|X$ is

$$\pi(\theta|x) \propto \underbrace{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\theta)^2}{2\sigma^2}\right\}}_{\text{likelihood}} \cdot \underbrace{\frac{1}{\sqrt{2\pi b^2}} \exp\left\{-\frac{(\theta-\mu)^2}{2b^2}\right\}}_{\text{prior}}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2 - \frac{1}{2b^2}(\theta-\mu)^2\right\}$$

$$= \exp\left\{\frac{1}{\sigma^2}\sum_{i=1}^n x_i\theta - \frac{n\theta^2}{2\sigma^2} - \frac{1}{2b^2}\theta^2 + \frac{\mu}{b^2}\theta\right\}$$

$(a+b)^2 = a^2 + 2ab + b^2$

$$\propto \dots$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)\theta^2 + \left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{b^2}\right)\theta\right\} \dots \exp\left(-\frac{1}{2\sigma^2}(\theta-\dots\right.$$

The posterior distribution of Θ given X is normal, with $(\tilde{\mu}, \tilde{\sigma}^2)$ as

$$\exp\left\{-\frac{1}{2\tilde{\sigma}^2}(\theta-\tilde{\mu})^2\right\}, \text{ where } \begin{cases} \tilde{\mu} = \dfrac{n\bar{x}/\sigma^2 + \mu/b^2}{n/\sigma^2 + 1/b^2} \\ \tilde{\sigma}^2 = \dfrac{1}{n/\sigma^2 + 1/b^2} \end{cases}.$$

Hence, the posterior mean of $\Theta|X$ is $\dfrac{n\bar{x}/\sigma^2 + \mu/b^2}{n/\sigma^2 + 1/b^2}$

$$= \frac{n/\sigma^2}{\dots}\bar{x} + \frac{1/b^2}{\dots}\mu$$

data. $\quad$ prior mean

improper prior.

$N(\mu, \infty)$

Bayes estimator δ_Λ B $\tilde{\mu}$. If we adopt the squared loss function...

$n/\sigma^2 + 1/b^2$ ... $n/\sigma^2 + 1/b^2$ ...

1 as $n \to \infty$ ... 0 as $n \to \infty$.

prior: $N(\mu; b^2)$

Suppose $b^2 \to \infty$...

$$\delta_{\Lambda(\mu,\infty)} = \frac{n/\sigma^2}{n/\sigma^2 + 0}\bar{x} + \frac{1/\infty}{n/\sigma^2 + 1/\infty} = \bar{x}$$

(Example). Assume that we consider $L(\theta, d) = \omega(\theta)\{d - g(\theta)\}^2$, where $\omega(\theta) \geq 0$, which can be interpreted as a weight function. Our goal is to find the corresponding Bayes estimator, which minimizes $E\left(\omega(\Theta)\{g(\Theta) - d\}^2 \mid X = x\right)$ with respect to d. (*)

(*) can be rewritten as

$$d^2 E\left(\omega(\Theta) \mid X = x\right) - 2d E\left(\omega(\Theta) g(\Theta) \mid X = x\right) + E\left(\omega(\Theta) g(\Theta)^2 \mid X = x\right). \quad (\dagger)$$

Taking derivative of $(\dagger)$ w.r.t. d, we obtain

$$2d^* E\left(\omega(\Theta) \mid X = x\right) - 2E\left(\omega(\Theta) g(\Theta) \mid X = x\right) = 0$$

$$\Rightarrow \quad \underset{\Lambda}{\delta}(X) = d^* = \frac{E\left(\omega(\Theta) g(\Theta) \mid X = x\right)}{E\left(\omega(\Theta) \mid X = x\right)}.$$

In particular, if $\omega(\cdot) \equiv 1$, $\delta_\Lambda(X)$ [with $\omega(\cdot) \equiv 1$] $= E\left(g(\Theta) \mid X = x\right)$.

[THEOREM]. (TPE 4.2.3). If δ is unbiased for $g(\theta)$ with $r(\Lambda, \delta) < \infty$ and $E\left(g(\Theta)^2\right) < \infty$, then δ is not Bayes under the squared loss function unless its average risk is zero, i.e. $E_{(X,\Theta)}\left(\{\delta(X) - g(\Theta)\}^2\right) = 0$.

Can $\bar{X}_n$ be a Bayes estimator?

Proof. Let δ be an unbiased estimator under the squared loss function. Then, we know that δ is the posterior mean, i.e.

$$\delta(X) = E\left(g(\Theta) \mid X\right) \quad \text{a.s.}. \quad \star$$

Thus, we have

$$E\left(\delta(X) g(\Theta)\right) = E\left(E(\delta(X) g(\Theta) \mid X)\right) = E\left(\delta(x) E(g(\Theta) \mid X)\right)$$
$$\overset{\star}{=} E\left(\delta^2(X)\right). \quad (1)$$

Also,

$$E\left(\delta(X) g(\Theta)\right) = E\left(E\left(\delta(X) g(\Theta) \mid \Theta\right)\right) = E\left(g(\Theta) E(\delta(X) \mid \Theta)\right)$$
$$\overset{\text{unbiasedness of } \delta}{=} E\left(g^2(\Theta)\right). \quad (2)$$

Observe that.

$$E\left(\{\delta(X) - g(\Theta)\}^2\right) = E\left(\delta^2(X)\right) - 2E\left(\delta(X) g(\Theta)\right) + E\left(g^2(\Theta)\right)$$
$$= E\left(\delta^2(X)\right) - E\left(\delta(X) g(\Theta)\right)$$
$$+ E\left(g^2(\Theta)\right) - E\left(\delta(X) g(\Theta)\right)$$

$$= E(\delta^2(X)) - E(\delta^2(X)) + E(g^2(\Theta)) - E(g^2(\Theta)) \quad \text{(due to ① and ②)}$$

$$= 0.$$

Thus, we have that $E\left(\{\delta(X) - g(\Theta)\}^2\right) = 0$, i.e. the average risk is zero. The claim is thus proved. ∎

(Example). Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\Theta, \sigma^2)$ ($N(\mu, b^2)$), with σ^2 known. Is $\bar{X}_n$ Bayes under the squared loss function for some choice of the prior distribution?

Observe that $E(\bar{X} \mid \theta) = \theta$, hence $\bar{X}_n$ is unbiased for θ. The corresponding average risk under the squared loss function is given by

$$E_{(X,\Theta)}\left(\{\bar{X} - \Theta\}^2\right) = \frac{\sigma^2}{n} \neq 0.$$

$\Rightarrow \bar{X}$ is not Bayes estimator under any prior distributions... ‖

guaranteed admissibility ←

[THEOREM] (TPE 5.2.4) A unique Bayes estimator (a.s. for all P_θ) is admissible.

An estimator is admissible if it is not uniformly dominated by some other estimator. δ is said to be inadmissible if and only if there exists δ' such that

$$\left\{ \begin{array}{l} R(\theta, \delta') \leq R(\theta, \delta) \quad \forall \theta \in \Omega \text{ and} \\ R(\theta, \delta') < R(\theta, \delta) \quad \text{for some } \theta \in \Omega. \end{array} \right\}$$

Proof. Suppose δ_Λ is Bayes for Λ, and for some δ', $R(\theta, \delta') \leq R(\theta, \delta_\Lambda)$ for all $\theta \in \Omega$. If we take expectations w.r.t Θ, the inequality above is preserved and we can write

$$\int_{\theta \in \Omega} R(\theta, \delta') \, d\Lambda(\theta) \leq \int_{\theta \in \Omega} R(\theta, \delta_\Lambda) \, d\Lambda(\theta)$$

This implies that δ' is also Bayes because δ' has less (or equal) risk than δ_Λ which minimizes the average risk. Hence $\delta' = \delta_\Lambda$ with probability one for all P_θ.

Question: When is a Bayes estimator unique?

[THEOREM]. Let Q be the marginal distribution of X, that is

$$Q(E) = \int P(X \in E \mid \theta) \, d\Lambda(\theta).$$

TPE 1-7.27 uniqueness of the minimizer of a strictly convex

Then, under a strictly convex loss function, δ_Λ is unique (a.s. for all P_θ) if

function (proved by contradiction)

(a) $r(\Lambda, \delta_\Lambda)$ is finite and ⇐ finiteness for comparison

(b) $P_\theta \ll Q$ (absolute continuity)

Ω is defined on the support Λ.

$P_\theta(X \in A)$ is a continuous function of θ for all meas. set of A.

Benefits of Bayes:
① Admissible
② Incorporate prior information (domain knowledge) → frequentist
③ ...

Next lecture: Minimax Estimator.

$\int R(\theta, \delta)\, d\Lambda(\theta)$ (average) $\times$ → $\sup_{\theta \in \Omega} R(\theta, \delta)$ (maximum) (most conservative)

Worst-case scenario / Optimality

Testing of Statistical Hypotheses, (UMP)