Math 5062: Bayesian Models

Jimin Ding

1 Lecture 1: Decision Theory Framework

Goals of statistic analysis: estimation, prediction, testing hypothesis, ranking. Examples on page 16 of Bickel and Docksum (2001). These examples motivate the decision theoretic framework: we need to

1. clarify the objectives of a study,
2. point to what the different possible actions are,
3. provide assessments of risk, accuracy, and reliability of statistical procedures,
4. provide guidance in the choice of procedures for analyzing outcomes of experiments.

We begin with a statistical model with observations X whose distribution $P_\theta \in \mathcal{P}$. Now we define the four components of the decision theory framework.

Action space: the space of actions or decisions or claims that we can contemplate making, often denoted by $\mathbb{A}$.

Examples:

1. Estimation: $\mathbb{A} = \mathbb{R}$
2. Testing: $\mathbb{A} = 0, 1$
3. Ranking: $\mathbb{A} = \{\text{permutations } (i_1, \cdots, i_k) \text{ of } (1, \cdots, k)\}$
4. Prediction: Here $\mathbb{A}$ is usually much larger. If the response variable Y is real, and the covariate $X \in \mathcal{X}$, then $\mathbb{A} = \{a : a \text{ is a function from } \mathcal{X} \text{ to } \mathbb{R}\}$ with $a(x)$ representing the prediction we would make if the new unobserved Y had covariate value x.

Loss function: a function that quantifies the loss for a given target and an action.

$$L : \mathcal{P} \times \mathbb{A} \to \mathbb{R}^+.$$

Examples: (In parametric models, $\mathcal{P}$ can be indexed by $\theta, \theta \in \Theta$)

$L(\theta, a) = (g(\theta) - a)^2$ Square/quadratic loss function.

$L(\theta, a) = |g(\theta) - a|$ Absolute loss function.

$L(\theta, a) = \min\{(g(\theta) - a)^2, d^2\}$ Truncated quadratic loss function.

$L(\theta, a) = 1$, if $|g(\theta) - a| > d$, 0 otherwise. Confidence interval loss function.

$L(\theta, a) = \mathbf{1}(g(\theta) < a)$, Asymmetric loss function.

For $\mathbb{A} \in \mathbb{R}^k, k > 1$:

$L(\theta, a) = \frac{1}{k} \sum_j (g_j(\theta) - a_j)^2$ Squared Euclidean distance.

$L(\theta, a) = \frac{1}{k} \sum_j |g_j(\theta) - a_j|$ Absolute distance

$L(\theta, a) = \max\{|g_j(\theta) - a_j|, j = 1, \cdots, k\}$ Supremum distance

Decision procedures: any function from the sample space taking its values in $\mathbb{A}$, often denoted by δ. If data $X = x$ is observed, the statistican takes action $\delta(x)$.
Question: How to choose an "optimal" decision/action/procedure?

Risk function: the average loss of the sample space. If δ action is used, L is the loss function, θ is the true value of the parameter, and $X = x$ is the outcome of the experiment. Then $L(\theta, \delta(x))$ depends on the particular outcome of the experiment. We may want to evaluate the properties of the action more based on overall or averaged loss.

$$R(\theta, \delta) = E[L(\theta, \delta(X))],$$

where the expectation is w.r.t X and may involve unknown parameter θ.
Example: The risk function defined by squared error loss function: MSE=Bias^2+Variance.

The optimal action will be the action that minimizes the risk function. However, the risk function depends on the unknown parameter θ, and it is usually very hard to find an action that uniformaly minimizes the risk function. One way to solve the problem is to restrict our attention to "unbiased" actions and then find the action uniformaly minimizes the risk function among the unbiased actions. For example, UMVUE. Another way is to aggregate the risk function over all possible values of $\theta \in \Theta$ with some weight function $\Pi(\theta)$ and then minimizes the aggregated risk. That is,

$$r(\Pi, \delta) = \int_\Theta R(\theta, \delta) d\Pi,$$

where Π is a known probability measure that gives the weight of θ. This $r(\Pi, \delta)$ is called the **Bayes risk**, and the optimal action that minimizes the Bayes risk is called **Bayes action/rule**, denoted by δ^*. A third method is to consider the worst situation, i.e. $\sup_{\theta\in\Theta} R(\theta, \delta)$. The resulting action that minimizes $\sup_{\theta\in\Theta} R(\theta, \delta)$ is called the **minimax rule**.

Examples: Normal-Normal model
Let $\boldsymbol{X} = (X_1, \cdots, X_n)^T$. Assume

$$X_1, \cdots, X_n | \mu \overset{\text{iid}}{\sim} N(\mu, \sigma^2),$$

with a known σ^2 and

$$\mu \sim N(\mu_0, \sigma_0^2).$$

Here μ_0 and σ_0^2 are hyperparameters. Then given the observed data,

$$\mu | \boldsymbol{X} = \boldsymbol{x} \sim N(\mu_*(\boldsymbol{x}), c^2),$$

where

$$\mu_*(\boldsymbol{x}) = w\mu_0 + (1-w)\bar{x}, \quad c^2 = w\sigma_0^2, \quad w = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}.$$

If we consider squared error loss function and estimation of $g(\theta) = \mu$, then

$$r(\Pi, \delta) = \int_\Theta R(\theta, \delta) d\Pi = E\left(E\{[\mu - \delta(\boldsymbol{X})]^2 | \mu\}\right) = E\left([\mu - \delta(\boldsymbol{X})]^2\right).$$

The problem of minimizing the Bayes risk can be hence viewed as the prediction of μ using $\boldsymbol{X}$. The best predictor is $E(\mu|\boldsymbol{X})$. So the Bayes decision is

$$\delta^*(\boldsymbol{x}) = \arg\min_{\delta \in \mathbb{A}} r(\Pi, \delta) = E(\mu|\boldsymbol{X} = \boldsymbol{x}),$$

where the expectation is w.r.t the conditional distribution of μ given $\boldsymbol{X} = \boldsymbol{x}$. In this case, $\delta^*(\boldsymbol{x})$ is the Bayes estimator of μ.

Now let's find the minimax rule in this example. For any decision rule δ,

$$\sup_{\theta \in \Theta} R(\theta, \delta) \geq \int_\Theta R(\theta, \delta) d\Pi \geq r(\Pi, \delta^*) = E\left([\mu - \delta^*(\boldsymbol{X})]^2\right) = E\{E\left([\mu - \delta^*(\boldsymbol{X})]^2 | X\right)\} = E(c^2) = c^2.$$

Since this result is true for any $\sigma_0^2 > 0$, let $\sigma_0^2 \to \infty$,

$$\sup_{\theta \in \Theta} R(\theta, \delta) \geq \sigma^2/n = \sup_{\theta \in \Theta} R(\theta, \bar{X}).$$

Hence $\bar{X}$ is the minimax estimator of μ.

Related Reading

1. Sh P113-116
2. BD chapter 1.2-1.3

2 Lecture 2: Bayesian Model and Estimation

Key: view model parameter θ as a realization of a random variable from some prior distribution Π. Two important distributions:

- Prior distribution: based on history/prior information/subjective belief. $\Pi(\theta, \xi)$, where ξ are **hyperparameters** which might be further modeled in a hierarchical model.
- Posterior distribution: conditional distribution of the model parameters given observed data. $P(\theta|\boldsymbol{X})$.

Theorem 2.1 (**Bayes Formula**). *Assume $\mathcal{P} = \{P(x|\theta) : \theta \in \Theta\}$ is dominated by a σ-finite measure μ with pdf $f(x|\theta) = \frac{dP(x|\theta)}{d\mu}$. Let Π be a prior distribution on Θ. Then the marginal pdf of X w.r.t. μ is $m(x) = \int_\Theta f(x|\theta)d\Pi$. Suppose $m(x) > 0$. Then*

1. *The posterior $P(\theta|x) \ll \Pi$ with pdf $\frac{dP(\theta|X)}{d\Pi} = \frac{f(x|\theta)}{m(x)}$.*
2. *If $\Pi \ll \lambda$ with pdf $\frac{d\Pi}{d\lambda} = \pi(\theta)$, where λ is a σ-finite measure, then the posterior $P(\theta|x) \ll \lambda$ with pdf $\frac{dP(\theta|x)}{d\lambda} = \frac{f(x|\theta)\pi(\theta)}{m(x)}$.*

Proof. Use Fubini's theorem. Detail see page 232 Shao (2003) □

If both X and θ are discrete and μ and λ are the counting measures, then above theorem becomes the Bayes formula in elementary statistics:

$$P(\theta = i|X = j) = \frac{P(X = j|\theta = i)P(\theta = i)}{\sum_i P(X = j|\theta = i)P(\theta = i)}.$$

Bayes risk:

$$r(\Pi, \delta) = \int_\Theta R(\theta, \delta)d\Pi(\theta), \quad \text{where } R(\theta, \delta) = E(L(\theta, \delta)).$$

Bayes rule:

$$\delta^*(x) = \arg\min_{\delta\in\mathbb{A}} r(\Pi, \delta) = \arg\min_{\delta\in\mathbb{A}} E(L(\theta, \delta(x))|X = x),$$

where the expectation is w.r.t. $P(\theta|x)$. We call δ^* the **Bayes estimator**, if the action is an estimation.

Theorem 2.2 (**Existence and Uniqueness of Bayes Estimator**). *Assume the conditions in the previous theorem hold. Assume $L(\theta, a)$ is convex in a for each given $x \in \mathcal{X}$ and $E(L(\theta, \delta(x))|X = x) < \infty$ for some δ (there exists an estimator with finite risk.) Furthermore, assume $\mathbb{A}$ is compact or $L(\theta, a) \to +\infty$ as $\|a\| \to +\infty$ uniformaly in θ. Then*

1. *Bayes estimator*
$$\delta^* = \arg\min_{\delta\in\mathbb{A}} E(L(\theta, \delta(x))|X = x)$$
exists.

2. If L is strictly convex, then δ^* is unique.

Proof. See page 228 in Lehmann and Casella (1998). □

Examples:

1. $L(\theta, a) = [g(\theta) - a]^2$, then

$$\delta^*(x) = E(g(\theta)|X = x) = \frac{\int_\Theta g(\theta) f(x|\theta) d\Pi}{\int_\Theta f(x|\theta) d\Pi}$$

2. $L(\theta, a) = w(\theta)[g(\theta) - a]^2$ with some known weight function $w(\theta)$, then

$$\delta^*(x) = \frac{\int_\Theta w(\theta) g(\theta) f(x|\theta) d\Pi}{\int_\Theta w(\theta) f(x|\theta) d\Pi}$$

3. $L(\theta, a) = |g(\theta) - a|^2$, then $\delta^*(x)$ is any median of the posterior.

4. $L(\theta, a) = \mathbf{1}(|\theta - a| > c)$ with some known constant c, then $\delta^*(x)$ is the interval I of length $2c$ which maximizes $P(\theta \in I|X = x)$.

Conjugate prior: the distributions of the posterior and prior belong to the same family. Depends on both the prior and the model distribution. For example,

1. Normal-Normal model (midterm of math 5061)
2. Poisson-Gamma model (Ex.)
3. Multinomial-Dirichlet model (final of math 5061)

Improper prior: note that Bayes risk $r(\Pi, \delta)$ is well defined even if Π is not a probability measure but a σ-finite measure. (Then $m(x)$ may not be finite.)

- $\Pi(\Theta) = 1$: proper prior
- $\Pi(\Theta) \neq 1$: improper prior $\Rightarrow \delta^*$ is referred to as **a generalized Bayes estimator**.
  Noninformative prior is usually an improper prior.
  The resulting generalized Bayes estimator is usally the limit of some Bayes estimators from a proper prior.
  Example 4.3 on page 236 Shao 2003.

More Bayesian Models:

1. **Normal-Gamma model.** Let

$$X_i \overset{\text{iid}}{\sim} N(0,\sigma^2), \quad \tau = \frac{1}{2\sigma^2} \sim Gamma(\alpha,\nu).$$

(σ^2 follows inverse Gamma distribution.) Then

$$\tau|\boldsymbol{X} \sim Gamma(\alpha^*,\nu^*), \quad \alpha^* = n/2+\alpha, \nu^* = (\sum X_i^2 + \nu^{-1})^{-1}.$$

Under square loss, the Bayes esimator for $\sigma^2 = 1/(2\tau^2)$ is

$$E[1/(2\tau^2)|\boldsymbol{X}] = \frac{1}{2}E[1/(\tau^2)|\boldsymbol{X}] = \frac{1}{2}\frac{1}{(\alpha^*-1)\nu^*} = \frac{\sum X_i^2+\nu^{-1}}{n/2-1+\alpha}.$$

2. If we consider **scale-invariant loss function**

$$L(\theta,a) = (\frac{\theta-a}{\theta})^2 = (\frac{\sigma^2-a}{\sigma^2})^2,$$

then the Bayes estimator for σ^2 is

$$\frac{E[w(\theta)g(\theta)]}{E[w(\theta)]} = \frac{E[\frac{1}{\sigma^4}\sigma^2]}{E\frac{1}{\sigma^4}} = \frac{E[\tau]}{2E[\tau^2]} = \frac{\sum X_i^2+\nu^{-1}}{2\alpha+n+2}, \quad \alpha^* > 1$$

3. **Normal-Gamma model with unknown mean and unknown variance.** Let

$$\begin{aligned} X_i &\overset{\text{iid}}{\sim} N(\mu,\sigma^2), \\ \mu &\sim N(\mu_0, \sigma_0^{/}(2\sigma^2)) \\ \tau = \frac{1}{2\sigma^2} &\sim Gamma(\alpha,\nu). \end{aligned}$$

Consider $\theta = (\mu,\sigma^2)$. Posterior of $\theta|\boldsymbol{X}$ is proportional to

$$\tau^{n/2}\exp\{-\sum(x_i-\mu)^2\tau\}(\sigma_0^2\tau)^{1/2}\exp\{-(\mu-\mu_0)^2\tau/(2\sigma_0^2)\}\tau^{\alpha-1}\exp\{-\tau/\nu\}$$

$$\begin{aligned} \Rightarrow \mu|\boldsymbol{X},\tau &\sim N(\mu*,c^2), \quad \mu* = \frac{n\bar{x}+\mu_0/(2\sigma_0^2)}{n+1/(2\sigma_0^2)}, \quad c^2 = [(2n+\sigma_0)^2\tau]^{-1}, \\ \tau|\boldsymbol{X} &\sim Gamma(\alpha^*,\nu^*), \quad \alpha^* = n/2+\alpha, \quad \nu^* = (\sum X_i^2+\nu^{-1}-(n+\frac{1}{2\sigma_0^2})\mu*^2)^{-1}. \end{aligned}$$

Hence Bayes estimator for θ under square loss is

$$\hat{\mu} = \mu*, \quad \hat{\sigma}^2 = \frac{1}{2(\alpha^*-1)\nu^*}, \text{ if } \alpha^* > 1.$$

Note these are biased estimators.

4. **ANOVA with hetergeneous variance.** (Example 4.9 on page 244 in Shao (2003).)

3 Lecture 3: More on Baysian Estimation

3.1 Empirical and Hierarchical Bayes Method for Priors

If hyperparameters, ξ, are estimated by historical or observed data, the resulting estimation is called **empirical Bayes estimation**. If the hyperparameters are viewed as random variables and modeled by a second-stage prior (**hyper-prior**), the resulting estimation is called **hierarchical Bayes estimation**.

Examples: Normal-Normal Model
Let $X_1, \cdots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with unknown $\mu \in \mathbb{R}$ and a known σ^2. Assume $\mu \sim N(\mu_0, \sigma_0^2)$.
Empirical Bayes Estimation: the simplest empirical Bayes method is to view the observed data as a "sample" from the marginal distribution $P(x|\xi) = \int P(x|\theta) d\Pi(\theta|\xi)$.
In this case, the pdf of $\boldsymbol{X}|\xi$ is

$$m(\boldsymbol{x}|\xi) = \int_{\mathbb{R}} (\sqrt{2\pi\sigma^2})^{-n/2} \exp\{-\frac{\sum_i (x_i - \mu)^2}{2\sigma^2}\} (\sqrt{2\pi\sigma_0^2})^{-n/2} \exp\{-\frac{\sum_i (\mu - \mu_0)^2}{2\sigma_0^2}\} d\mu,$$

where $\xi = (\mu_0, \sigma_0^2)$. One can find the conditional expectation and variance of $X|\xi$ without calculating $m(\boldsymbol{x}|\xi)$:

$$E(X|\xi) = E[E(X|\theta)|\xi] = E[\mu|\xi] = \mu_0,$$

$$Var(X|\xi) = Var[E(X|\theta)|\xi] + E[Var(X|\theta)|\xi] = Var[\mu|\xi] + E[\sigma^2|\xi] = \sigma_0^2 + \sigma^2.$$

Then we may estimate prior parameters using the moment method

$$\hat{\mu_0} = \bar{x}, \quad \hat{\sigma_0^2} = \frac{1}{n}\sum_i x_i^2 - \bar{x} - \sigma^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 - \sigma^2.$$

With respect to squared error loss function, the Bayes estimator for μ is hence the sample mean $\bar{x}$.
Hierarchical Bayes Estimation: instead of estimating hyperparameters, in hierarchical Bayes approach we put a second-stage prior on hyperparameters, $\Lambda(\theta|\xi)$. If the second-stage prior also depends on some unknown parameters, one may go on to consider a third-stage prior. In most applications, however, two-stage priors are sufficient, since misspecifying a second-stage prior is much less serious than misspecifying a first-stage prior. In addition, the second-stage prior can be chosen to be noninformative. The hierarchical Bayes estimation is

$$\delta(x) = \int \delta(x, \xi) dP(\xi|x),$$

where $\delta(x, \xi)$ is the Bayes estimation when ξ is known.
In this case, we consider $\Lambda(\xi)$ be the Lebesgue measure. From the previous example, we see

$$\delta(x, \xi) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\xi + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x}.$$

Note the pdf of $\xi|\boldsymbol{X}$ is

$$f(\xi|\boldsymbol{x}) = \frac{\int f(\boldsymbol{x}|\theta)f(\theta|\xi)f(\xi)d\theta}{\int\int f(\boldsymbol{x}|\theta)f(\theta|\xi)f(\xi)d\theta d\boldsymbol{x}} \propto \int f(\boldsymbol{x}|\theta)f(\theta|\xi)f(\xi)d\theta.$$

Hence, in this example,

$$\begin{aligned} f(\xi|\boldsymbol{x}) &\propto \int_{\mathbb{R}} \exp\{-\frac{\sum_i (x_i-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\}d\mu \\ &\propto \int_{\mathbb{R}} \exp\{-\frac{n(\bar{x}-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\}d\mu \\ &= \int_{\mathbb{R}} \exp\{-[a\mu^2 - 2b\mu + c]\}d\mu \\ &\propto \exp c - (2b/a)^2, \end{aligned}$$

where $a = n/(2\sigma^2) + 1/(2\sigma_0^2)$,$b = n\bar{x}/\sigma^2 + \mu_0/\sigma_0^2$, and $c = n\bar{x}^2/(2\sigma^2) + \mu_0^2/(2\sigma_0^2)$. Hence

$$f(\xi|\boldsymbol{x}) \propto \exp -\frac{n(\xi-\bar{x})^2}{2(n\sigma_0^2+\sigma^2)}.$$

So $\xi|\boldsymbol{x}$ follows normal distribution with mean $\bar{x}$. We may estimate

$$\hat{\mu}_0 = \bar{x}.$$

The hierarchical generalized Bayes rule is then $\bar{x}$.

3.2 Properties on Bayesian Estiamtors

Biasness of Bayes estimators Theorem. *Let $\delta(X)$ be a Bayes estimator of $g(\theta)$ under squared error loss. Then $\delta(X)$ is biased unless the Bayes risk $r(\Pi, \delta) = 0$, i.e. $E[(g(\theta) - \delta(X))^2] = 0$, where E is w.r.t both X and θ.*

Proof. If $\delta(X)$ is unbiased, then $E[\delta(X)|\theta] = g(\theta), \forall\theta$. Then

$$\begin{aligned} E[\delta(X)g(\theta)] &= E[E(\delta(X)g(\theta)|\theta)] = E[g^2(\theta)], \\ E[\delta(X)g(\theta)] &= E[E(\delta(X)g(\theta)|X)] = E[\delta^2(X)]. \\ \therefore E[(g(\theta)-\delta(X))^2] &= E[g^2(\theta)] + E[\delta^2(X)] - 2E[\delta(X)g(\theta)] = 0. \end{aligned}$$

□

Examples: Sample mean is usually not Bayes estimator.

1.**Normal-Noraml model.**

Bias= $\frac{\sigma^2}{n\sigma_0^2+\sigma^2}(\mu_0 - \mu)$, which is not 0. For large n, the bias is the order of $1/n$.

2.**Bin-Beta model.**
Let $X_i \overset{\text{iid}}{\sim} Ber(p), p \sim Beta(a, b)$. Then $p|\boldsymbol{X} \sim Beta(a^*, b^*)$, where

$$a^* = a + \bar{x}, b^* = b + n(1 - \bar{X}).$$

Here $Var(\bar{X}) = p(1-p)/n$, which is not a constant as above, but depends on the mean p. But for $p \in (0, 1)$, $p(1-p) > 0, \Rightarrow r(\Pi, \bar{X}) = \int p(1-p)/n d\Pi > 0$ unless Π only have point mass on 0 and/or 1. Hence $\bar{X}$ cannot be Bayes estimator.

3.**Sample mean in general.**
Let $X_i \overset{\text{iid}}{\sim} E(X_i) = \theta, Var(X_i) = \sigma^2$, which is independent of θ. Note that

$$R(\theta, \bar{X}) = Var(\bar{X}) = \sigma^2/n \neq 0.$$

Then $r(\Pi, \bar{X}) \neq 0$ w.r.t. any prior Π, hence $\bar{X}$ cannot be Bayes estimator.

3.3 Related Reading

1. Sh P231-245
2. LC chapter 4