1 Decision Theoretic Setup: Loss, Posterior Risk, Bayes Action

Let $\mathcal{A}$ be action space and $a \in \mathcal{A}$ be an action. For example, in estimation problems, $\mathcal{A}$ is the set of real numbers and a is a number, say $a = 2$ is adopted as an estimator of $\theta \in \Theta$. In other words, the inference maker "took" the action $a = 2$ in estimating θ. In testing problems, the action space is $\mathcal{A} = \{\text{accept}, \text{reject}\}$. The action, as a function of observations is called a decision rule, or simply a *rule.* An example of a rule is $a(X_1, \ldots, X_n) = \bar{X}$. Often, the rules are denoted by $\delta(X)$.

No action can be taken without potential losses. Statisticians are pessimistic creatures that replaced nicely coined term *utility* to a more somber term *loss*, although, for all practical purposes, the loss is a negative utility. The loss function is denoted by $L(\theta, a)$ and represents the payoff by a decision maker (statistician) if he takes the action $a \in \mathcal{A}$, and the real state of nature is $\theta \in \Theta$.

The loss function usually satisfies the following properties, $L(a, a) = 0$ and $L(a, \theta)$ is nondecreasing function of $|a - \theta|$.

Examples are *squared error loss* (SEL) $L(\theta, a) = (\theta - a)^2$, *absolute loss*, $L(\theta, a) = |\theta - a|$, the *0-1 loss,* $L(\theta, a) = \mathbf{1}(|a - \theta| > m)$, etc.

The most common for estimation problems and mathematically easiest to work with is the SEL. The expected SEL (frequentist risk) is linked with variance and bias of an estimator,

$$E^{X|\theta}(\theta - \delta(X))^2 = Var(\delta(X)) + [\text{bias}(\delta(X))]^2.$$

where $\text{bias}(\delta(X) = E^{X|\theta}\delta(X)) - \theta$.

One criticism of the SEL is that it grows fast (quadratically) when the error increases, thus severely punishing the errors.

Example 1. The LINEX is defined as

$$L(\theta, a) = \exp\{c(a - \theta)\} - c(a - \theta) - 1, \; c \in R.$$

For $c > 0$, the loss function $L(\theta, a)$ is quite asymmetric about 0 with overestimation being more costly than under-estimation. As $|a - \theta| \to \infty$, the loss $L(\theta, a)$ increases almost exponentially when $a - \theta > 0$ and almost linearly when $a - \theta < 0$. For $c < 0$, the linearity-exponentiality phenomenon is reversed. Also, when $|a - \theta|$ is very small, $L(\theta, a)$ is near $c(a - \theta)^2/2$.

Definition 1. Bayesian expected loss is the expectation of the loss function with respect to posterior measure, i.e.,

$$\rho(a, \pi) = E^{\theta|X} L(a, \theta) = \int_\Theta L(\theta, a)\pi(\theta|x)d\theta.$$

The Expected Loss Principle. In comparing two actions $a_1 = \delta_1(X)$ and $a_2 = \delta_2(X)$, after data X had been observed, preferred action is the one for which the posterior expected loss is smaller. An action a^* that minimizes the posterior expected loss is called *Bayes* action.

In the following example we find the Bayes actions (and Bayes rules) for several common loss functions.
Example 2. (i) If the loss is squared error, the Bayes action a^* is found by minimizing

$$\varphi(a) = E^{\theta|X}(\theta - a)^2 = a^2 + (2E^{\theta|X}\theta)a + E^{\theta|X}\theta^2.$$

Since $\varphi'(a) = 0$ for $a = E^{\theta|X}\theta$ and $\varphi''(a) = 2 < 0$, the posterior mean $a^* = E^{\theta|X}\theta$ is the Bayes action.

(ii) Recall that $\left(\int_{A(x)}^{B(x)} \phi(x,t)dt\right)' = \int_{A(x)}^{B(x)} \phi'(x,t)dt + \phi(x,B(x))B'(x) - \phi(x,A(x))A'(x)$, and that the *median*,$m$, of random variable X is defined as $P(X \geq m) \geq 1/2$ and $P(X \leq m) \geq 1/2$.

Assume the absolute loss.

$$\begin{aligned}\varphi(a) &= E^{\theta|X}|\theta - a| = \int_{\theta \geq a} (\theta - a)\pi(\theta|X)d\theta + \int_{\theta \leq a} (a - \theta)\pi(\theta|X)d\theta \\ &= \int_a^\infty (\theta - a)\pi(\theta|X)d\theta + \int_{-\infty}^a (a - \theta)\pi(\theta|X)d\theta.\end{aligned}$$

Then,

$$\begin{aligned}\varphi'(a) &= -\int_a^\infty \pi(\theta|X)d\theta + 0 - 0 + \int_{-\infty}^a \pi(\theta|X)d\theta + 0 - 0 \\ &= -P^{\theta|X}(\theta \geq a) + P^{\theta|X}(\theta \leq a) = 0.\end{aligned}$$

The value of a for which $P^{\theta|X}(\theta \geq a) = P^{\theta|X}(\theta \leq a)$ is the median of the posterior distribution.

Since $\varphi''(a) = 2\phi(a|X) > 0$, the median minimizes the $\varphi(a)$.

(iii) Recall that a number b is said to be a pth percentile (quantile) of a distribution of a random variable X if $P(X \leq b) \geq p$ and $P(X \geq b) \geq 1 - p$.

Let the loss be

$$L(\theta, a) = \begin{cases} K_1(\theta - a), & \theta \geq a \\ K_2(a - \theta), & \theta < a \end{cases}$$

This is a slight generalization of absolute error loss ($K_1 = K_2 = 1$). By mimicking (ii) we arrive to condition $K_1 P^{\theta|X}(\theta \geq a) = K_2 P^{\theta|X}(\theta \leq a)$ Thus, $P^{\theta|X}(\theta \leq a) = K_1/K_2[1 - P^{\theta|X}(\theta \leq a)]$ By solving this equation, we obtain that $P^{\theta|X}(\theta \leq a) = K_1/(K_1 + K_2)$, i.e., the Bayes action a^* is $K_1/(K_1 + K_2)$-percentile of the posterior distribution.

(iv) An interval of length $2c$, say $(b - c, b + c)$, is said to be a modal interval of length $2c$ for the distribution of a random variable X, if $P(b - c \leq X \leq b + c)$ takes on its maximum value out of all such intervals. For the loss function

$$L(\theta, a) = \begin{cases} 0, & |\theta - a| \leq c \\ 1, & \text{else} \end{cases} = \begin{cases} 0, & a - c \leq \theta \leq a + c \\ 1, & \text{else} \end{cases}$$

the probability $P^{\theta|X}(a - c \leq \theta \leq a + c)$ is minimized if a is chosen to be the midpoint of the modal interval of length $2c$. Thus, the Bayes action a^* is the midpoint of the modal interval of length $2c$ of the posterior.

Assume that $c \to 0$. Then the limiting case of the above loss is *hit-or-miss* loss,

$$L(\theta, a) = \mathbf{1}(\theta \neq a).$$

If the posterior is unimodal, the limiting Bayes action is $a^* = \text{argmax}_\theta \pi(\theta|X)$, the MAP rule. Of course direct minimization of "hit-or-miss" loss is impossible for absolutely continuous posteriors since the integrals are taken over a singleton set and such sets have posterior measure 0.

2 Bayes Principle in the Frequentist Decision Theoretic Setup

Let X be a random variable whose distribution is in $\{P_\theta,\ \theta \in \Theta\}$, a family which is indexed by a parameter (random variable) θ. We put on frequentist hat and make an inference about the parameter θ, given an observation X. A solution is a *decision procedure (decision rule)* $\delta(x)$, that identifies particular inference for each value of x **that can be observed.** Let, as in the Bayesian setup, $\mathcal{A}$ be the class of all possible realizations of $\delta(x)$, i.e. *actions.* The *loss function* $L(\theta, a)$ maps $\Theta \times \mathcal{A}$ into the set of real numbers and defines a cost to the statistician when he takes the action a and the true value of the parameter is θ. A *risk function* $R(\theta, \delta)$ characterizes the performance of the rule δ for each value of parameter $\theta \in \Theta$. The risk is usually defined in terms of underlying loss function $L(\theta, a)$ as

$$R(\theta, \delta) = E^{X|\theta} L(\theta, \delta(X)) = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) dx.$$

Since the risk function is defined as an average loss with respect to a sample space, it is called the *frequentist risk*. Let $\mathcal{D}$ be the collection of all measurable decision rules. There are several principles for assigning the preference among the rules in $\mathcal{D}$. We give now only the Bayes principle, other include minimax principle, and Γ-minimax principle, minimax regret principle, etc., and we may be talking about them later in the course if time permits.

Under the Bayes principle, the prior distribution π is specified on the parameter space Θ. Any rule δ is characterized by its *Bayes risk*

$$r(\pi, \delta) = \int R(\theta, \delta) \pi(d\theta) = E^\theta R(\theta, \delta).$$

The rule δ_π that minimizes Bayes risk is called *Bayes rule*, i.e.

$$\delta_\pi = arg \inf_{\delta \in \mathcal{D}} r(\pi, \delta).$$

The *Bayes risk of the prior distribution* π (*Bayes envelope function*) is

$$r(\pi) = r(\pi, \delta_\pi).$$

Name	Bayes	Name	Frequentist
action	$a \in \mathcal{A}$	rule	$\delta(x) \in \mathcal{D}$
$\odot\odot$		Risk	$R(\theta, \delta(X)) = E^{X\|\theta} L(\theta, \delta(X))$
posterior expected loss	$\rho(\pi, a) = E^{\theta\|X} L(\theta, a)$	Bayes risk	$r(\pi, \delta) = E^\theta E^{X\|\theta} L(\theta, \delta(X))$
Bayes action	$\text{argmin}_a \rho(\pi, a)$	Bayes rule	$\delta^*(x) = \text{argmin}_{\delta \in \mathcal{D}} r(\pi, \delta)$

Since, by Fubini's Theorem $r(\pi, \delta) = E^\theta E^{X|\theta} L(\theta, \delta(X)) = E^X E^{\theta|X} L(\theta, \delta(X)) = E^X \rho(\pi, a(X))$ is minimized when $\rho(\pi, a)$ is minimized, for any fixed x, $\delta_B(x) = a^*(x)$. This fact is true whenever $r(\pi)$ is finite.

This result links the conditional Bayesian and decision theoretic frequentist inference: the frequentist Bayes rule conditional on X is the Bayes action.

In the terms of Bayes rule, when the loss is squared error, the Bayes rule is the posterior expectation,

$$\delta_B(x) = \frac{\int_\Theta \theta f(x|\theta) \pi(\theta) d\theta}{\int_\Theta f(x|\theta) \pi(\theta) d\theta},$$

more generally, if the loss is *weighted squared error*, $L(\theta, a) = \omega(\theta)(\theta - a)^2$, the Bayes rule is

$$\delta_B(x) = \frac{\int_\Theta \omega(\theta)\theta f(x|\theta)\pi(\theta)d\theta}{\int_\Theta \omega(\theta) f(x|\theta)\pi(\theta)d\theta},$$

According to a Bayes principle the rule $\delta_1(X)$ is preferred to $\delta_2(X)$ if $r(\pi, \delta_1) < r(\pi, \delta_2)$. The frequentists use Bayes principle to compare frequentist risks of the rules, $R(\theta, \delta_1)$ and $R(\theta, \delta_2)$. Analysis of frequentist risk functions leads to various concepts of frequentist procedure choice: minimaxity, admissibility, unbiasedness, equivariance, etc. This will be revisited later in the course.

3 Exercises

1. Under the LINEX loss defined in Example 1 find the Bayes rule under the model $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1), \theta \propto 1$.

2. If $X|\theta \sim \mathcal{B}(n, \theta)$ and $\theta \sim \mathcal{B}e(\alpha, \beta)$, find the Bayes rule under the loss

$$L(\theta, a) = \frac{(\theta - a)^2}{\theta(1 - \theta)}.$$

Be careful about the treatment of $x = 0$ and $x = n$.
3. If $X|\theta \sim \mathcal{G}(n/2, 2\theta)$ and $\theta \sim \mathcal{IG}(\alpha, \beta)$, find the Bayes rule under the loss

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

References

[1] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, Second Edition, Springer Verlag.

[2] Robert, C. (2001). *Bayesian Choice*, Second Edition, Springer Verlag.