

1 Basic Elements

Let Θ denote a set of possible hypotheses about the world, called the *parameter space*, and suppose we observe a random variable Z , taking values in a space \mathcal{Z} , whose distribution P_θ depends in some way on $\theta \in \Theta$. The collection of probability distributions $\{P_\theta : \theta \in \Theta\}$ is called an *experiment*.

The decision-maker observes Z , and chooses some *action* $a \in \mathcal{A}$. A (nonrandomized) *decision rule* is a function $d : \mathcal{Z} \rightarrow \mathcal{A}$. Let \mathcal{D} denote the set of all such rules. The decision-maker's objectives are embodied in a *loss function* $L(\theta, a)$, a real-valued function defined on $\Theta \times \mathcal{A}$.

Since the decision rule is a function of random variable Z , it seems reasonable to examine its average loss, which is given by the (frequentist) *risk function*:

$$R(\theta, d) = \mathbb{E}[L(\theta, d(Z))].$$

Here the expectation is with respect to the distribution P_θ of Z . In other words,

$$R(\theta, d) = \int L(\theta, d(z))dP_\theta(z).$$

Throughout we will assume that L and d are such that the integral in the previous expression is well-defined and finite. The $dP_\theta(z)$ notation of the integral indicates that it is a general (Lebesgue) integral—so we can include discrete, continuous, and multivariate probability distributions in this formulation.

Example 1 As an example, consider the simple normal model of Lecture Note 1. This has $Z = (Y_1, \dots, Y_n)$. Recall that we assumed that the Y_i were i.i.d. with distribution $\mathcal{N}(\mu, 1)$. So the unknown parameter θ is just the mean, μ . The parameter space Θ is the real line \mathbb{R} , and P_θ is the distribution that has the elements of Z i.i.d. normal with mean θ , variance 1. The decision problem was to choose a point estimate for θ , so the action space \mathcal{A} is just the set of possible guesses for θ , i.e. also \mathbb{R} , and a decision rule $d(Z)$ is an *estimator* of θ . The loss function is quadratic:

$$L(\theta, a) = (\theta - a)^2,$$

and taking the expectation of a decision rule (estimator) to get the risk gives what is usually called the *mean squared error* of the estimator:

$$R(\theta, d) = \mathbb{E}[(\theta - d(Z))^2].$$

2 Optimality of Decision Rules

The fundamental problem in decision theory is to choose the decision rule d . Since we have a well-defined risk function, and since we interpret greater risk as “bad,” we could try to choose d to minimize risk. However, the best choice will generally depend on θ . For example, continuing with the previous example, suppose we chose the constant decision rule $d(z) = c$. This has 0 risk if $\theta = c$, does quite well for θ close to c , but does very poorly for θ far away from c . So different choices for c lead to estimators that do well in different parts of the parameter space. This turns out to be true for many interesting classes of decision rules: their risk functions $R(\theta, d)$ cross.

In general, there is not a unique solution to the fundamental problem. However, it is possible that some decision rules can be eliminated from consideration on the basis that there is some other decision rule that does better regardless of θ .

Admissibility: a decision rule d is *admissible* if there exists no other rule d' with

$$R(\theta, d') \leq R(\theta, d) \quad \text{for all } \theta \in \Theta$$

and

$$R(\theta, d') < R(\theta, d) \quad \text{for some } \theta \in \Theta.$$

Our first goal is to characterize the set of admissible rules.

3 Bayes Decision Rules

Recall that in the previous lecture note we placed a distribution on θ , and this led to a unique solution for the estimator. Specifically, we placed a $\mathcal{N}(a, \tau^{-1})$ prior distribution on the mean, and this led to the estimate μ^* , a weighted average of a and \bar{y} . This idea can be extended to general decision problems.

A *prior distribution* is any probability distribution $p(\theta)$ on the parameter space Θ . Define the *Bayes Risk* as the expectation of the risk function with respect to the prior distribution on θ :

$$r(p, d) = \mathbb{E}[R(\theta, d)] = \int R(\theta, d)p(\theta)d\theta.$$

A decision rule d_0 is a *Bayes decision rule* with respect to the prior distribution p if

$$r(p, d_0) = \inf_{d \in \mathcal{D}} r(p, d). \tag{1}$$

In other words, the Bayes decision rule minimizes the Bayes risk with respect to the prior distribution. Notice that if we regard utility as the negative of loss, $U = -L(\theta, a)$, then the Bayes decision rule is the rule that maximizes expected utility.

Since we are looking for admissible rules, one may ask whether Bayes rules are admissible. Fairly generally, the answer is yes. We will focus on the case where the parameter space is finite: $\Theta = \{\theta_1, \dots, \theta_k\}$. Then a prior distribution p can be represented as a k -tuple (p_1, \dots, p_k) , where $p_j = p(\theta_j)$.

Theorem 1 (*Admissibility of Bayes Decision Rules*) *Assume that $\Theta = \{\theta_1, \dots, \theta_k\}$, and that a decision rule d_0 is a Bayes rule with respect to the prior distribution (p_1, \dots, p_k) . If $p_j > 0$ for $j = 1, \dots, k$, then d_0 is admissible.*

Proof: Suppose that d_0 is inadmissible. Then there exists a \tilde{d} which is better than d_0 :

$$R(\theta_j, \tilde{d}) \leq R(\theta_j, d_0) \quad \text{for all } j$$

and

$$R(\theta_j, \tilde{d}) < R(\theta_j, d_0) \quad \text{for some } j.$$

Since all $p_j > 0$, this implies

$$\sum_{j=1}^k p_j R(\theta_j, \tilde{d}) < \sum_{j=1}^k p_j R(\theta_j, d_0).$$

But then d_0 cannot be Bayes with respect to (p_1, \dots, p_k) , a contradiction. ■

This result can be generalized to larger parameter spaces; see for example Ferguson, p. 62.

We have not completely characterized the class of admissible rules, but already we have a useful result. Theorem 1 implies that if we can show that a statistical procedure is equivalent to a Bayes rule with respect to some prior, then we know that that procedure is admissible.

4 The Complete Class Theorem

We will continue to assume that the parameter space Θ is finite. It will be convenient to expand the set of possible decision rules to allow for randomization. Let $\{d_1, \dots, d_m\} \subset \mathcal{D}$ be a finite set of nonrandomized decision rules, and let $(\alpha_1, \dots, \alpha_m)$ be a vector of probabilities (so $\alpha_i \geq 0$ for $i = 1, \dots, m$ and $\sum_{i=1}^m \alpha_i = 1$). Then we can define a *randomized decision rule* δ that employs d_i with probability α_i , and define its risk as

$$R(\theta, \delta) := \sum_{i=1}^m \alpha_i R(\theta, d_i).$$

Let \mathcal{D}^* denote the set of all such rules (for all finite m).

Definition: For a finite parameter space $\Theta = \{\theta_1, \dots, \theta_k\}$, the *risk set* consists of the risk vectors that correspond to some decision rule:

$$S = \left\{ (r_1, \dots, r_k) \in \mathbb{R}^k : \text{for some } \delta \in \mathcal{D}^*, r_j = R(\theta_j, \delta) \text{ for } j = 1, \dots, k \right\}.$$

It is not too hard to show that S is convex.

Theorem 2 (Complete Class) *If Θ is finite and $\delta \in \mathcal{D}^*$ is admissible, then δ is Bayes with respect to some prior distribution on Θ .*

Proof: Let $a = (R(\theta_1, \delta), \dots, R(\theta_k, \delta))$ denote the risk vector associated with δ , and let Q_a denote the set of risk vectors that are at least as good as a :

$$Q_a = \left\{ x \in \mathbb{R}^k : x_j \leq a_j \text{ for } j = 1, \dots, k \right\}.$$

Since δ is admissible, $Q_a \cap S = \{a\}$. Then $Q_a - \{a\}$ and S are disjoint convex sets, so we can use the separating hyperplane theorem to obtain a vector $p \in \mathbb{R}^k$ such that $p'x \leq p'y$ for all $x \in Q_a - \{a\}$ and $y \in S$. If some coordinate p_l of the vector p were negative, then $\sum_j p_j x_j > \sum_j p_j y_j$ by taking x_l sufficiently negative. Hence $p_j \geq 0$ for $j = 1, \dots, k$, and we can normalize p so that $\sum_j p_j = 1$. Now p is a probability distribution over Θ , and

$$r(p, d) = \sum_j p_j R(\theta_j, \delta) \leq p'y$$

for all $y \in S$ implies that δ is a Bayes rule with respect to p . ■

We now have a characterization of the class of admissible rules: it coincides with the class of Bayes decision rules. This result holds for a finite parameter set, but can be extended to more general spaces. It is important to note that this is a frequentist result in the sense that risk is a property of repeated sampling from P_θ . The prior distribution, in this formulation, could be viewed as simply a device with which to generate admissible decision rules. Another intriguing feature of this result is that it does not depend on the sample size growing to infinity.

Some qualifications are in order. Although we are now in a position to rule out many procedures as *inadmissible*, we are left with a rather large class of procedures which cannot be strictly ordered according the desiderata we have established so far. One possibility is to impose other restrictions (e.g. unbiasedness, invariance to reparametrization) on the space of decision rules. Sometimes, however, these other criteria leads to unappealing estimators (see for example Ferguson, pp. 132-136), or estimators which are inadmissible (see Ferguson, p. 152).

5 The Form of Bayes Decision Rules

The definition of the Bayes decision rule in Equation 1 seems to imply that we have to take an infimum with respect to the set of all decision functions \mathcal{D} . The purpose of this section is to show that there is usually a simpler way to calculate Bayes decision rules.

5.1 Finite Case

Definition: If the distribution P_θ has a probability density function $f(z|\theta)$, then when $Z = z$ is observed, $f(z|\theta)$, regarded as a function of θ , is called the *likelihood function*.

If Z is discrete, $f(z|\theta) = P_\theta(Z = z)$.

Theorem 3 Suppose that $\Theta = \{\theta_1, \dots, \theta_k\}$ and $\mathcal{Z} = \{z_1, \dots, z_J\}$. If for all $z \in \mathcal{Z}$

$$d_0(z) = \arg \min_{a \in \mathcal{A}} \sum_{i=1}^k L(\theta_i, a) f(z|\theta_i) p(\theta_i)$$

then d_0 is a Bayes decision rule with respect to the prior distribution p .

Proof: The Bayes risk is

$$\begin{aligned} r(p, d) &= \sum_{i=1}^k R(\theta_i, d) p(\theta_i) \\ &= \sum_{i=1}^k \left[\sum_{j=1}^J L(\theta_i, d(z_j)) f(z_j|\theta_i) \right] p(\theta_i) \\ &= \sum_{j=1}^J \left[\sum_{i=1}^k L(\theta_i, d(z_j)) f(z_j|\theta_i) p(\theta_i) \right]. \end{aligned}$$

To minimize $r(p, d)$, we cannot do better than to choose d to separately minimize each of the terms in the outer sum of the last line; i.e., choose $d(z_j)$ to minimize the j^{th} term for $j = 1, \dots, J$. ■

Definition: If $\Theta = \{\theta_1, \dots, \theta_k\}$, then the *conditional or posterior density of θ given $Z = z_j$* is

$$p(\theta_i|z_j) = \frac{f(z_j|\theta_i) p(\theta_i)}{\sum_{l=1}^k f(z_j|\theta_l) p(\theta_l)}$$

So we can restate Theorem 3 as saying that a Bayes rule minimizes the *posterior expected loss*, given z :

$$d_0(z) = \arg \min_{a \in \mathcal{A}} \sum_{i=1}^k L(\theta_i, a) p(\theta_i|z) \quad \text{for } z \in \mathcal{Z}.$$

5.2 Continuous Case

Essentially the same argument works in the continuous case, where $f(z|\theta)$ and $p(\theta)$ are interpreted as densities with respect to Lebesgue measure. Then we can write the Bayes risk as

$$\begin{aligned} r(p, d) &= \int_{\Theta} R(\theta, d) p(\theta) d\theta \\ &= \int_{\Theta} \left[\int_{\mathcal{Z}} L(\theta, d(z)) f(z|\theta) dz \right] p(\theta) d\theta \\ &= \int_{\mathcal{Z}} \left[\int_{\Theta} L(\theta, d(z)) f(z|\theta) p(\theta) d\theta \right] dz, \end{aligned}$$

and choosing d to minimize the inner integral separately at each value of z will be optimal.

Definition: The *conditional* or *posterior distribution of θ given z* has density

$$p(\theta|z) = \frac{f(z|\theta)p(\theta)}{\int f(z|t)p(t)dt}.$$

As in the finite case, the Bayes decision rule minimizes posterior expected loss, conditional on the observed value z :

$$d_0(z) = \arg \min_{a \in \mathcal{A}} \int L(\theta, a) p(\theta|z) d\theta.$$

The thing to take from these calculations is that regardless of the loss function, we always want to calculate the posterior distribution, or at least have a way to take expectations with respect to it. If the likelihood function and the prior distribution are easy to evaluate at a given value for θ , then the main difficulty will be in evaluating the integral in the denominator of the expression for the posterior density. Sometimes there will be analytic results that allow us to express the posterior density in a closed form, but in a large number of cases, the only way to evaluate the posterior distribution will be to use numerical methods.

6 Optimality: Bayes and Minmax Rules

We have shown that under some conditions, the class of admissible rules coincides with the class of Bayes rules. This does not “pin down” a best rule. The problem, as we have seen, is that different rules may do better in different parts of the parameter space.

In order to have some more concrete notion of optimality, we would need to take a stand on how to rank rules that do not strictly dominate each other. One way that we have already worked with is to select a weighting of the parameter space. We can think of a prior $p(\theta)$ as a weighting of the parameter space, and use the integrated (Bayes) risk:

$$r(p, d) = \int R(\theta, d) p(\theta) d\theta.$$

This weights the parameter space according to $p(\theta)$, and the Bayes rule by definition minimizes weighted average risk. Of course, different weightings will lead to different optimal rules.

Another approach is to look at the worst-case performance of the decision rule over the parameter space. The worst-case risk of a decision rule d is:

$$\sup_{\theta \in \Theta} R(\theta, d)$$

A rule d that minimizes this quantity is said to be *minmax*. In other words, a minmax rule d_m satisfies:

$$\sup_{\theta \in \Theta} R(\theta, d_m) = \inf_d \sup_{\theta \in \Theta} R(\theta, d).$$

Minmax rules may not always exist, or there may be multiple minmax rules. In some cases, they are difficult to calculate.

The minmax criterion can lead to very conservative decision-making, especially if the parameter space is “large,” because the emphasis is on avoiding poor performance for *any* value of the parameter. A related criterion that is sometimes used is the minmax regret criterion. The regret loss of a rule is defined as the difference between the loss, and the loss of the best possible action under θ :

$$L_r(\theta, a) = L(\theta, a) - \inf_{a \in \mathcal{A}} L(\theta, a).$$

The regret risk is then:

$$R_r(\theta, d) = \int L_r(\theta, d(z)) dP_\theta(z).$$

We can then look for a rule that is minmax with respect to this modified version of the risk.