Notes on Bayesian Decision Theory

Adam N. Smith
July 31, 2021

This note reviews some of the key results in Bayesian decision theory. The motivation is to understand how and why Bayes estimators are "good" estimators. We outline conditions of Bayesian optimality and frequentist optimality, and then present a key result (the complete class theorem) connecting these two criteria which shows that all "good" estimators in the frequentist sense must be Bayes with respect to some prior.

1 Optimality

We first define a few concepts that are at the center of decision theory. Let $L(\theta, a)$ denote the **loss function** which describes the penalty incurred from taking an action $a \in \mathcal{A}$ when the true state of nature is $\theta \in \Theta$. A **decision rule** is then a function $\delta : \mathcal{X} \to \mathcal{A}$ which selects an action $a \in \mathcal{A}$ given data $x \in \mathcal{X}$.

Bayesian Optimality The goal is to characterize optimal decision rules. We first consider the Bayesian version of optimality. Let $\pi(\theta|x)$ denote the posterior distribution induced by the likelihood function $p(x|\theta)$ and prior $\pi(\theta)$. A **Bayes rule** δ^π with respect to the prior π is defined as the action which, for every $x \in \mathcal{X}$, minimizes the posterior expected loss:

$$\delta^\pi(x) = \arg\min_{a \in \mathcal{A}} \int_\Theta L(\theta, a)\pi(\theta|x)d\theta = \arg\min_{a \in \mathcal{A}} \int_\Theta L(\theta, a)p(x|\theta)\pi(\theta)d\theta. \quad (1)$$

Frequentist Optimality The posterior expected loss integrates the loss function over a distribution of model parameters, which is awkward in a frequentist setting. Therefore, classical decision theory defines optimal rules based on a different type of expected loss. The **(frequentist) risk function** $R(\theta, \delta)$ is defined as the expected loss, where the expectation is taken with respect to the data $x \in \mathcal{X}$:

$$R(\theta, \delta) = \mathbb{E}_x[L(\theta, \delta(x))] = \int_\mathcal{X} L(\theta, \delta(x))p(x|\theta)dx. \quad (2)$$

Note that Bayes rules can also be defined in terms of the frequentist risk. In particular, minimizing the posterior expected loss at each $x \in \mathcal{X}$ is equivalent to minimizing the *expected* frequentist risk (i.e., Bayes risk), where the expectation is taken with respect to θ.

In the Bayesian view of optimality, good rules minimize posterior expected loss. In the frequentist view, good rules minimize risk. This is formalized with the concept of admissibility. A decision rule δ is **inadmissible** if there exists another decision rule δ' such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all $\theta \in \Theta$ (with strict inequality holding for some θ). If there is no such δ', then δ is **admissible**.

Admissibility is a useful criterion when searching for decision rules. For example, knowing that an estimator is inadmissible is clearly bad in that another estimator with lower risk is guaranteed to exist. One of the most popular and surprising examples of an inadmissible estimator is given by James and Stein (1961).

Example 1 (James-Stein)**.** *Suppose that* $Y_i \sim N(\theta_i, 1)$ *and define the loss* $L(\theta, a) = \sum_{i=1}^{p}(\theta_i - a_i)^2$*. Then if* $p > 2$*,* $\delta(y) = y$ *is inadmissible.*

What makes this result surprising is that $\delta(y) = y$ is both the least squares estimator and the MLE![1] Even in a familiar setting of estimating normal means under squared error loss, the workhorse estimators do not meet this frequentist version of optimality. James and Stein (1961) also propose a new estimator:

$$\delta_{\text{JS}}(y) = \delta(y)\left[1 - \frac{p-2}{\sum_{i=1}^{p} y_i^2}\right]$$

which is shown to dominate $\delta(y)$. Notice that $[\cdot]$ will *shrink* the MLE towards 0 whenever $(p-2) < \sum_{i=1}^{p} y_i^2$ which is why the James-Stein estimator is referred to as a shrinkage estimator.[2] James and Stein (1961) were one of the first to document and popularize the benefits of shrinkage, which is now an indispensable tool in the analysis of modern high-dimensional data.

[1] This estimator is also a Bayes estimator under a uniform (and improper) prior. This example also highlights one of the nuances that will arise when studying the admissibility of Bayes rules. Specifically, Bayes rules with proper priors can easily be shown to be admissible, but Bayes rules with improper priors – referred to as *generalized* Bayes rules – need not be admissible.

[2] In an empirical Bayes framework, it can be shown that $\sum_{i=1}^{p} y_i^2$ is an unbiased estimator of 1 + the variance of θ_i and so shrinkage will be heavier as the variance of θ_i gets small.

2 Complete Class Theorem

We now turn to the question of how Bayes rules perform under this frequentist criterion of optimality. We first define the notion of a "complete class". A class of decision rules $C \subset \mathcal{D}$ is said to be **complete** if, for all $\delta \in \mathcal{D} \backslash C$, there exists a rule $\delta' \in C$ such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all $\theta \in \Theta$ (with strict equality holding for some θ). In words, if a class C is complete, then for any decision rule outside of C, there will always be a better decision rule (in terms of having lower frequent risk) inside of C. In fact, a complete class always contains the set of admissible rules and, in some scenarios, *is* the class of admissible decision rules. The main result in this section is that Bayes rules form a complete class.

Theorem 1 (Complete Class Theorem)**.** *If Θ is finite and the risk set is convex, then the class of all Bayes rules forms a complete class. Thus, a decision rule is admissible if and only if it is a Bayes rule δ^π with respect to some prior π.*[3]

Proof. Need to show that (i) all Bayes rules are admissible and (ii) all admissible rules are Bayes rules. Assume $\Theta = \{\theta_1, \ldots, \theta_K\}$ is finite and the prior puts positive mass everywhere $\pi(\theta) > 0$ for all $\theta \in \Theta$.

(i) Proof by contradiction. Assume a Bayes rule δ^π is inadmissible so that another decision rule δ has lower risk.

$$R(\theta, \delta) \leq R(\theta, \delta^\pi) \text{ for all } \theta \in \Theta$$

Therefore,

$$R(\pi, \delta) = \sum_{k=1}^{K} \pi(\theta_k) R(\theta_k, \delta) < \sum_{k=1}^{K} \pi(\theta_k) R(\theta_k, \delta^\pi) = R(\theta, \delta^\pi)$$

which contradicts the assumption that δ^π is a Bayes rule (i.e., minimizes expected risk).

(ii) This direction is much more technical. Assume the decision rule δ_0 is admissible. Then do the following:

[3] This theorem was first proven by Wald (1947) under the conditions that the parameter space Θ is finite, the sample space $\mathcal{X}$ is finite, and the prior $\pi(\theta)$ is proper and strictly positive. Generalizations to infinite dimensional parameter spaces and unbounded loss functions can be found in Le Cam (1955) and Brown et al. (1976). Generalizations to improper priors can be found in Sacks (1963), Stone (1967), and Berger and Srinivasan (1978).

- Define the risk set $\mathcal{S} = \{(R(\theta_1, \delta), \ldots, R(\theta_K, \delta))\}$ which is convex if Θ is finite
- Define the lower orthant set $\mathcal{Q}(\mathbf{s}) = \{\mathbf{s} \in \mathbb{R}^K : s_k \leq R(\theta_k, \delta)\ \forall k \leq K\}$
- Show $\mathbf{s}_0 = (R(\theta_1, \delta_0), \ldots, R(\theta_K, \delta_0))$ is a lower boundary point of $\mathcal{S}$
- Invoke the separating hyperplane theorem to prove existence of a vector $\mathbf{v}$ separating the risk set $\mathcal{S}$ from the lower orthant set $\mathcal{Q}(\mathbf{s}_0)$:

$$\mathbf{v} \cdot \mathbf{x} \leq \mathbf{v} \cdot \mathbf{s} \text{ for all } \mathbf{x} \in \tilde{\mathcal{Q}}(\mathbf{s}_0) \text{ and } \mathbf{s} \in \mathcal{S}$$

  where $\tilde{\mathcal{Q}}(\mathbf{s}_0) = \mathcal{Q}(\mathbf{s}_0) \backslash \{\mathbf{s}_0\}$ and $v_k \geq 0$ for all $k = 1, \ldots, K$
- Rescale $\mathbf{v}$ to be defined on the probability simplex: $\pi = \mathbf{v}/(\mathbf{v} \cdot \mathbf{1})$

$$\mathbf{v} \cdot \mathbf{x} \leq \mathbf{v} \cdot \mathbf{s} \implies \pi \cdot \mathbf{x} \leq \pi \cdot \mathbf{s} \implies \pi \cdot \mathbf{s}_0 \leq \pi \cdot \mathbf{s}$$

- Therefore δ_0 is a Bayes rule with respect to π

The Complete Class Theorem is profound because it *equates* admissible rules with Bayes rules. That is, any admissible frequentist decision rule can be derived from a Bayesian perspective. At a high level, this suggests that we only need to consider Bayes rules when searching for good estimators (regardless of the inference paradigm!).

References

Berger, J. O. and Srinivasan, C. (1978). Generalized Bayes estimators in multivariate problems. *The Annals of Statistics*, 6(4):783–801.

Brown, L. D., Cohen, A., and Strawderman, W. E. (1976). A complete class theorem for strict monotone likelihood ratio with applications. *The Annals of Statistics*, 4(4):712–722.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In Neyman, J., editor, *Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379.

Le Cam, L. (1955). An extension of wald's theory of statistical decision functions. *The Annals of Mathematical Statistics*.

Sacks, J. (1963). Generalized Bayes solutions in estimation problems. *The Annals of Mathematical Statistics*, 34(3):751–768.

Stone, M. (1967). Generalized Bayes decision functions, admissibility and the exponential family. *The Annals of Mathematical Statistics*, 38(3):818–822.

Wald, A. (1947). An essentially complete class of admissible decision functions. *The Annals of Mathematical Statistics*, 18(4):549–555.