

Lecture 2 Admissibility and Bayesian

Two criteria will be introduced in what follows to evaluate and compare decision rules.

2.1 Admissibility

An order can be defined between two decision rules d and d' based on their risk functions:

- d is said to be (at least) as good as (resp. better than) d' if $R(\theta, d) \leq R(\theta, d') \forall \theta \in \Theta$ [resp. $R(\theta, d) \leq R(\theta, d') \forall \theta \in \Theta$, and $R(\theta, d) < R(\theta, d')$ for some $\theta \in \Theta$].
- d and d' are said to be equivalent if $R(\theta, d) = R(\theta, d') \forall \theta \in \Theta$ (Note: d and d' need not be the same rule even if they are equivalent).

A collection of rules form a partially ordered set. In particular, d is said to be inadmissible if there is a rule d' such that d' is better than d . d is said to be admissible if it is not inadmissible.

Note: Inadmissible rules should be ruled out, but admissible rules are by no means optimal in any sense. Sometimes it may not even be reasonable. The following example illustrates a method to show admissibility — focusing on some special values of θ that minimize the risk function.

Example 2.1 Let $X \sim \text{Bernoulli}(\theta)$. Estimate θ with squared error loss.

- $d_1(X) = X$ is admissible. By contradiction, suppose d' is better than d_1 . It follows from $R(\theta, d_1) = \theta(1-\theta)$ that $R(0, d_1) = R(1, d_1) = 0$. Therefore, $R(0, d') = E_0(d'(X))^2 = 0$ which implies $d'(0) = 0$; by the same token, $R(1, d') = E_1(d'(X) - 1)^2 = 0$ which implies $d'(1) = 1$. Hence $d'(X) = X = d_1(X)$.
- $d_2(X) \equiv 1/2$ is also admissible. Again by contradiction, suppose d'' is better than d_2 . Hence $E_{1/2}(d''(X) - 1/2)^2 = R(1/2, d'') \leq R(1/2, d_2) = 0$ which implies $d''(X) \equiv 1/2 \equiv d_2(X)$.

Note: Plots of $R(\theta, d_1)$ and $R(\theta, d_2)$ seem to suggest that d_1 and d_2 are “comparable”. It is somewhat misleading in this special case with a single observation X . Consider the case with data X^n , $n > 1$ and revisit the plots of $R(\theta, d_1)$ and $R(\theta, d_2)$ (note that $d_1(X^n) = \bar{X}$). You will see the difference between $R(\theta, d_1) = \theta(1-\theta)/n$ and $R(\theta, d_2) = (1/2 - \theta)^2$. (Question: Is d_1 still admissible?)

Another recipe for finding admissible rules, through a Bayesian approach, will be discussed later.

2.2 Bayesian framework

More often than not, neither risk function dominates the other when comparing two decision rules, i.e. the order between values of the two risk functions changes when evaluated at different parameter values. One way to resolve this is to integrate $R(\theta, d)$ over Θ against a density $\pi(\theta)$, which results in a summary quantity $B(\pi, d)$ as a benchmark for comparing decision rules. Assume all probability models on Θ and \mathcal{X} are *regular*, i.e. a reference measure on Θ (resp. \mathcal{X}), denoted by ν_Θ (resp. $\nu_{\mathcal{X}}$), is either a Lebesgue measure (continuous model) or a counting measure (discrete model). Here is a short list of terms:

- *prior* $\pi(\theta)$: a density of $\theta \in \Theta$;
- *likelihood* $f(x|\theta)$: a conditional density of $x \in \mathcal{X}$ given θ ;
- *marginal* $m(x) = \int_{\Theta} f(x|\theta)\pi(\theta) \nu_\Theta(d\theta)$;
- *posterior* $\pi(\theta|x) = \pi(\theta)f(x|\theta)/m(x)$.

The identity $\pi(\theta)f(x|\theta) = m(x)\pi(\theta|x)$ is called the Bayes formula; and

$$B(\pi, d) = E_\pi R(\theta, d) = \int_{\Theta} R(\theta, d)\pi(\theta) \nu_\Theta(d\theta)$$

is called the *Bayes risk* of d under the prior π . A decision rule d^π is called a *Bayes rule* (with respect to π) if

$$B(\pi, d^\pi) = \min_{d \in \mathcal{D}} B(\pi, d),$$

where \mathcal{D} is a class of rules of interest (we usually suppress “ $d \in \mathcal{D}$ ” unless necessary).

2.2.1 Posterior risk and Bayesian inference

As a minimization problem of finding a Bayes rule, there is a useful sufficient condition based on the posterior risk of action a given $x \in \mathcal{X}$:

$$r(x, a) = \int_{\Theta} L(\theta, a) \pi(\theta|x) \nu_\Theta(d\theta).$$

Note: In what follows, statements in theorems, propositions, etc. often start with “under certain regularity conditions ...”. Such an informal style may serve better than making a long list of specific regularity conditions. Hopefully, the proofs will make it clear what regularity conditions are required.

Theorem 1 Under certain regularity conditions, if d' satisfies

$$r(x, d'(x)) = \min_d r(x, d(x)) \quad \forall x \in \mathcal{S} = \{x \in \mathcal{X} : m(x) > 0\},$$

then d' is a Bayes rule d^π .

Proof: For any $d \in \mathcal{D}$,

$$\begin{aligned} B(\pi, d) &= \int_{\Theta} R(\theta, d) \pi(\theta) \nu_{\Theta}(d\theta) \\ &= \int_{\Theta} \int_{\mathcal{X}} L(\theta, d(x)) f(x|\theta) \nu_{\mathcal{X}}(dx) \pi(\theta) \nu_{\Theta}(d\theta) \\ &= \int_{\Theta} \int_{\mathcal{S}} L(\theta, d(x)) m(x) \pi(\theta|x) \nu_{\mathcal{X}}(dx) \nu_{\Theta}(d\theta) \\ &= \int_{\mathcal{S}} \int_{\Theta} L(\theta, d(x)) \pi(\theta|x) \nu_{\Theta}(d\theta) m(x) \nu_{\mathcal{X}}(dx) \quad (\text{Fubini Theorem}) \\ &= \int_{\mathcal{X}} r(x, d(x)) m(x) \nu_{\mathcal{X}}(dx). \end{aligned}$$

Hence Theorem 1 follows straightforwardly. *QED.*

Corollary 1 The following useful results in Bayesian inference are special cases of Theorem 1:

(i) For estimation of θ with the squared error loss, a Bayes estimator is the posterior mean

$$d^\pi(X) = E_\pi(\theta|X).$$

(ii) For estimation of θ with the absolute error loss, a Bayes estimator is a posterior median, i.e. a median of $\pi(\theta|X)$.

(iii) For testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ with the 0-1 loss, a Bayesian test rejects H_0 if $P(\Theta_0|X) < P(\Theta_1|X)$ (or equivalently, $P(\Theta_0|X) < 1/2$) and accepts H_0 otherwise.

Note: (i) and (iii) can be shown easily. Try to show (ii), and I will provide a proof later.

A basic idea in Bayesian statistics is summarized in the following

“No data principle”: With the data x , a Bayes rule makes a decision in the same way as what we would have made without data, except for the prior $\pi(\theta)$ replaced by the posterior $\pi(\theta|x)$. Using the empty set \emptyset to represent “no data”, the corresponding Bayes risk and the posterior risk will coincide:

$$r(\emptyset, d(\emptyset)) = \int_{\Theta} L(\theta, d(\emptyset)) \pi(\theta|\emptyset) \nu_{\Theta}(d\theta) = \int_{\Theta} R(\theta, d) \pi(\theta) \nu_{\Theta}(d\theta) = B(\pi, d).$$

2.2.2 Conjugate families

It becomes clear based on Theorem 1 and Corollary 1 that in any Bayesian problem the posterior $\pi(\theta|x)$ is the key. A posterior is determined by a prior and a likelihood. The concept likelihood, representing a data model, is important in both Bayesian and frequentist statistics. A frequentist statistician does not believe a prior, which is also subject to different interpretations in a Bayesian framework. A subjective Bayesian view takes a prior as a belief — taken for granted without questioning. A practical Bayesian approach treats a prior as a summary of some relevant information before data are collected. In a decision-theoretical setting, a prior reflects a preference: we would like to have lower risk attached to more likely values of θ (under prior π); while we are not as concerned about the risk attached to less likely values of θ .

How should a prior be specified? Various considerations can be given: (a) a purely subjective choice; (b) a fully objective one, called a non-informative prior; (c) careful incorporation of subject matters; (d) driven by computational convenience. There is no unique way in general. Here we only elaborate on the approach (d).

Definition 1 Let $\mathcal{F} = \{f(\cdot|\theta), \theta \in \Theta\}$ be a parametric family. Π is called a conjugate family to \mathcal{F} if using any $\pi \in \Pi$ as a prior and $f(x|\theta)$ as the likelihood would end up with a posterior $\pi(\cdot|x) \in \Pi, \forall x \in \mathcal{X}$ and $\theta \in \Theta$.

A great advantage of using a conjugate family is that derivation of the posterior reduces to simply updating parameters in the prior.

Example 2.2 (several conjugate families)

- (i) Beta distributions form a conjugate family to Bernoulli distributions. Let X_1, \dots, X_n be iid random variables sharing a common Bernoulli distribution with parameter θ . Assume $\theta \sim \mathcal{B}(a, b)$ where $\mathcal{B}(a, b)$ denotes a beta distribution with hyper-parameters $a > 0$ and $b > 0$, i.e. the prior

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a, b)},$$

where $B(a, b) = \int_0^1 u^{a-1}(1-u)^{b-1} du$. Then $\theta|x^n \sim \mathcal{B}(a + s_n, b + n - s_n)$ with $s_n = x_1 + \dots + x_n$.

- (ii) Gamma distributions form a conjugate family to Poisson distributions. Let X_1, \dots, X_n be iid random variables sharing a common Poisson distribution with parameter θ . Assume $\theta \sim \mathcal{G}(\alpha, \lambda)$ where $\mathcal{G}(\alpha, \lambda)$ denotes a gamma distribution with hyper-parameters $\alpha > 0$ and $\lambda > 0$, i.e. the prior

$$\pi(\theta) = \frac{\lambda^\alpha \theta^{\alpha-1} e^{-\lambda\theta}}{\Gamma(\alpha)},$$

where $\Gamma(\alpha) = \int_0^\infty v^{\alpha-1} e^{-v} dv$. Then $\theta|x^n \sim \mathcal{G}(\alpha + s_n, \lambda + n)$.

- (iii) Normal distributions form a conjugate family to themselves when estimating the mean. Let X_1, \dots, X_n be iid random variables sharing a common distribution $N(\theta, \sigma^2)$. Assume $\sigma > 0$ is a known constant, and $\theta \sim N(\mu, \tau^2)$ with hyper-parameters $\mu \in \mathbb{R}$ and $\tau > 0$. Then $\theta|x^n \sim N(\mu_{\bar{x}}, \tau_{\bar{x}}^2)$ where

$$\begin{aligned}\mu_{\bar{x}} &= \frac{\tau^2}{\tau^2 + \sigma^2/n} \bar{x} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n} \mu; \\ \tau_{\bar{x}}^2 &= \frac{\tau^2 \sigma^2/n}{\tau^2 + \sigma^2/n}.\end{aligned}$$

Note: More examples for conjugate families are given in the references. In fact, (i) – (iii) are special cases of a more general result, i.e. an exponential family (to be introduced later) form a conjugate family to itself.