Foundations of Statistical Inference

Julien Berestycki

Department of Statistics
University of Oxford

MT 2016

Lecture 10 : Decision theory

Decision theory

Example You have been exposed to a deadly virus. About 1/3 of people who are exposed to the virus are infected by it, and all those infected by it die unless they receive a vaccine. By the time any symptoms of the virus show up, it is too late for the vaccine to work. You are offered a vaccine for £500. Do you take it or not?

The most likely scenario is that you don't have the virus but basing a decision on this ignores the *costs* (or *loss*) associated with the decisions. We would put a very high loss on dying!

Decision theory

Example You are in a gun fight with a trigger-happy tough cop. Unfortunately, he has his gun pointed at you and you must decide wether he still has some bullet in it. You estimate the probability that he does at less than 5%. You must decide whether or not to test this by reaching for your gun.

Decision Theory

- Decision Theory sits 'above' Bayesian and classical statistical inference and gives us a basis to compare different approaches to statistical inference.
- We make decisions by applying rules to data.
- Decisions are subject to risk.
- A risk function specifies the expected loss which follows from the application of a given rule, and this is a basis for comparing rules.
- We may choose a rule to minimize the maximum risk, or we may choose a rule to minimize the average risk.

Terminology

θ is the 'true state of nature', $X \sim f(x;\theta)$ is the data.

The **Decision rule** is δ. If $X = x$, adopt the action $\delta(x)$ given by the rule.

Example A single parameter θ is estimated from $X = x$ by $\delta(x) = \widehat{\theta}(x)$.

The rule $\widehat{\theta}$ is the functional form of the estimator. The action is the value of the estimator.

The **Loss function** $L_S(\theta, \delta(x))$ measures the loss from action $\delta(x)$ when θ holds.

Example

$L_S(\theta, \widehat{\theta}(x))$ is the loss function which increases for $\widehat{\theta}(x)$ being away from θ. Here are three common loss functions.

Zero-One loss

$$L_S(\theta, \widehat{\theta}(x)) = \begin{cases} 0 & |\widehat{\theta}(x) - \theta| < b \\ a & |\widehat{\theta}(x) - \theta| \geq b \end{cases}$$

where a, b are constants.

Absolute error loss

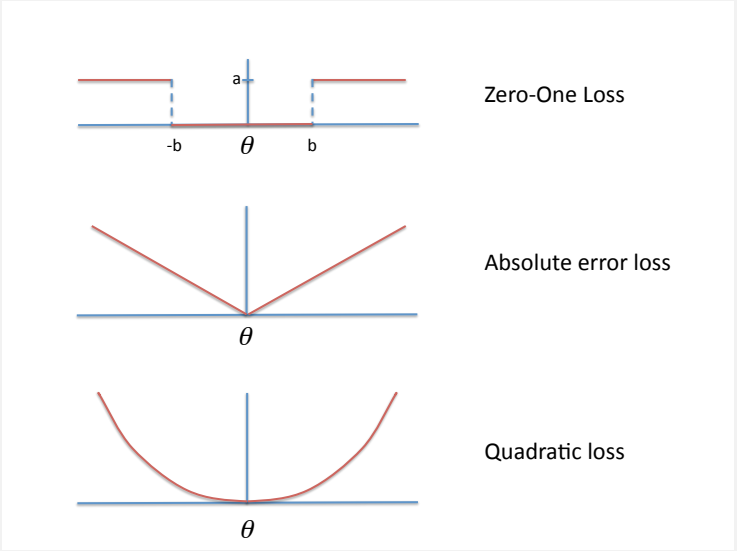
$$L_S(\theta, \widehat{\theta}(x)) = a|\widehat{\theta}(x) - \theta|$$

where $a > 0$.

Quadratic loss

$$L_S(\theta, \widehat{\theta}(x)) = (\widehat{\theta}(x) - \theta)^2.$$

Example



Risk function

Definition

The risk function $R(\theta, \delta)$ is defined as

$$R(\theta, \delta) = \int L_S(\theta, \delta(x)) f(x; \theta) dx,$$

This is the expected value of the loss (aka expected loss).

Example In the context of point estimation, with Quadratic Loss, the risk function is the mean square error,

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

Admissible and inadmissible rules.

Definition

A procedure δ_1 is inadmissible if there exists another procedure δ_2 such that

$$R(\theta, \delta_1) \geq R(\theta, \delta_2), \text{ for all } \theta \in \Theta$$

with $R(\theta, \delta_1) > R(\theta, \delta_2)$ for at least some θ.

A procedure which is not inadmissible is admissible.

Example

Suppose $X \sim U(0, \theta)$. Consider estimators of the form $\widehat{\theta}(x) = ax$ (this is a family of decisions rules indexed by a).

Show that $a = 3/2$ is a necessary condition for the rule $\widehat{\theta}$ to be admissible for quadratic loss.

$$\begin{aligned} R(\theta, \widehat{\theta}) &= \int_0^\theta (ax - \theta)^2 \frac{1}{\theta} dx \\ &= (a^2/3 - a + 1)\theta^2 \end{aligned}$$

and R is minimized at $a = 3/2$.

This does not show $\widehat{\theta}(x) = 3x/2$ is admissible here.

It only shows that all estimators with $a \neq 3/2$ are inadmissible. The estimator $\widehat{\theta}(x) = 3x/2$ maybe inadmissible relative to other estimators not in this class.

Minimax rules

Definition

A rule δ is a minimax rule if $\max_\theta R(\theta, \delta) \leq \max_\theta R(\theta, \delta')$ for any other rule δ'. It minimizes the maximum risk.

Since minimax minimizes the maximum risk (*ie*, the loss averaged over all possible data $X \sim f$) the choice of rule is not influenced by the actual data $X = x$ (though given the rule δ, the action $\delta(x)$ is data-dependent).

It makes sense when the maximum loss scenario must be avoided, but can can lead to poor performance on average.

Bayes risk, Bayes rule, expected posterior loss

Definition

Suppose we have a prior probability $\pi = \pi(\theta)$ for θ. Denote by

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

the Bayes risk of rule δ. A Bayes rule is a rule that minimizes the Bayes risk. A Bayes rule is sometimes called a Bayes procedure.

Let $\pi(\theta|x) = \dfrac{L(x;\theta)\pi(\theta)}{h(x)}$ denote the posterior following from likelihood L and prior π.

Definition

The expected posterior loss is defined as

$$\int L_S(\theta, \delta(x))\pi(\theta|x)d\theta$$

Lemma

A Bayes rule minimizes the expected posterior loss.

Proof

$$\begin{aligned}\int R(\theta, \delta)\pi(\theta)d\theta &= \int\int L_S(\theta, \delta(x))L(\theta; x)\pi(\theta)dx\, d\theta \\ &= \int\int L_S(\theta, \delta(x))\pi(\theta|x)h(x)dx\, d\theta \\ &= \int h(x)\left(\int L_S(\theta, \delta(x))\pi(\theta|x)d\theta\right)dx\end{aligned}$$

That is for each x we choose $\delta(x)$ to minimize the integral

$$\int L_S(\theta, \delta(x))\pi(\theta|x)d\theta$$

Bayes rules for Point estimation

Zero-One loss

$$L_S(\theta, \widehat{\theta}(x)) = \begin{cases} 0 & |\widehat{\theta}(x) - \theta| < b \\ a & |\widehat{\theta}(x) - \theta| \geq b \end{cases}$$

where a, b are constants.

We need to minimize

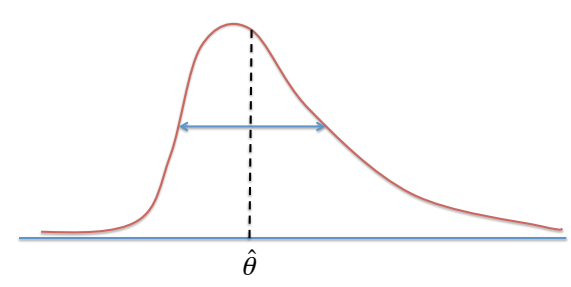
$$\begin{aligned}\int_{-\infty}^{\infty} \pi(\theta|x)L_S(\theta, \widehat{\theta})d\theta &= a\int_{\widehat{\theta}+b}^{\infty} \pi(\theta|x)d\theta + a\int_{-\infty}^{\widehat{\theta}-b} \pi(\theta|x)d\theta \\ &\propto 1 - \int_{\widehat{\theta}-b}^{\widehat{\theta}+b} \pi(\theta|x)d\theta\end{aligned}$$

That is we want to maximize

$$\int_{\widehat{\theta}-b}^{\widehat{\theta}+b} \pi(\theta|x)d\theta$$

Zero-one loss

If $\pi(\theta|x)$ is unimodal the maximum is attained by choosing $\widehat{\theta}$ to be the mid-point of the interval of length $2b$ for which $\pi(\theta|b)$ has the same value at both ends.



As $b \to 0$, $\widehat{\theta} \to$ the global mode of the posterior distribution. If $\pi(\theta|x)$ is unimodal and symmetric, the optimal $\widehat{\theta}$ is the median (equal to the mean and mode) of the posterior distribution.

Absolute error loss

Absolute error loss

$$L_S(\theta, \widehat{\theta}(x)) = a|\widehat{\theta}(x) - \theta|$$

We need to minimise

$$\int |\widehat{\theta} - \theta|\pi(\theta|x)d\theta = \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta} - \theta)\pi(\theta|x)d\theta + \int_{\widehat{\theta}}^{\infty} (\theta - \widehat{\theta})\pi(\theta|x)d\theta.$$

Differentiate wrt $\widehat{\theta}$ and equate to zero.

$$\int_{-\infty}^{\widehat{\theta}} \pi(\theta|x)d\theta - \int_{\widehat{\theta}}^{\infty} \pi(\theta|x)d\theta = 0$$

That is, the optimal $\widehat{\theta}$ is the median of the posterior distribution.

Quadratic loss

Minimize

$$\begin{aligned} \mathbb{E}_{\theta|x}[(\widehat{\theta} - \theta)^2] &= \mathbb{E}_{\theta|x}[(\widehat{\theta} - \bar{\theta} + \bar{\theta} - \theta)^2] \\ &= (\widehat{\theta} - \bar{\theta})^2 + 2(\hat{\theta} - \bar{\theta})\mathbb{E}_{\theta|x}(\theta - \bar{\theta}) + \mathbb{E}_{\theta|x}[(\theta - \bar{\theta})^2] \end{aligned}$$

where $\bar{\theta}$ is the posterior mean of θ.

Note $\hat{\theta}$ and $\bar{\theta}$ are constants in the posterior distribution of θ so that $(\hat{\theta} - \bar{\theta})\mathbb{E}(\theta - \bar{\theta}) = 0$. So

$$\mathbb{E}_{\theta|x}[(\widehat{\theta} - \theta)^2] = (\widehat{\theta} - \bar{\theta})^2 + \mathrm{V}_{\theta|x}(\theta)$$

The Quadratic loss function is minimized when $\widehat{\theta} = \bar{\theta}$, the posterior mean.

Summary

The form of the Bayes rule depends upon the loss function in the following way

- Zero-one loss (as $b \to 0$) leads to the posterior mode.
- Absolute error loss leads to the posterior median.
- Quadratic loss leads to the posterior mean.

Note These are not the only loss functions one could use in a given situation, and other loss functions will lead to different Bayes rules.

Example

$X \sim$ Binomial (n, θ), and the prior $\pi(\theta)$ is a Beta (α, β) distribution.

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha + \beta}$$

The distribution is unimodal if $\alpha, \beta > 1$ with mode $(\alpha - 1)/(\alpha + \beta - 2)$.

The posterior distribution of $\theta \mid x$ is Beta $(\alpha + x, \beta + n - x)$.

With zero-one loss and $b \to 0$ the Bayes estimator is $(\alpha + x - 1)/(\alpha + \beta + n - 2)$.

For a quadratic loss function, the Bayes estimator is $(\alpha + x)/(\alpha + \beta + n)$.

For an absolute error loss function is the median of the posterior.

Randomized decision rules

Suppose we have a collection of l decision rules $d_1, \ldots, d_l$.For probability weights $p_1, \ldots, p_l$ define $d^* = \sum p_i d_i$ to be the rule *'select rule d_i with probability p_i and apply'*.

Definition

d^* is a **randomized** decision rule.

The risk function of a randomized decision rule is

$$R(\theta, d^*) = \sum_{i=1}^{l} p_i R(\theta, d_i).$$

Minimax rules may be of this form.

Finite decision problem

Definition

A decision problem is said to be finite when the parameter space $\Theta = \{\theta_1, \ldots, \theta_k\}$ is finte.

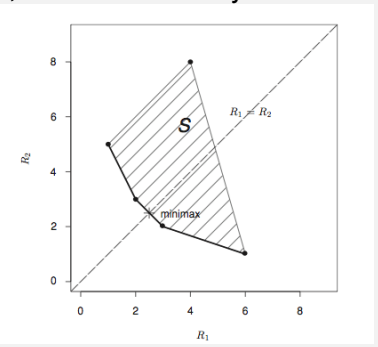
In this case the notions of admissible, minimax and Bayes can be given geometric interpretations.

Definition

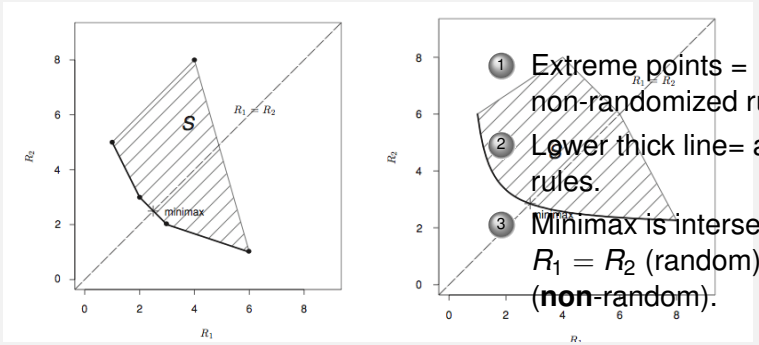
The risk set $S \subset \mathbb{R}^k$ is the set of points $(R(\theta_1, d), \ldots R(\theta_k, d))$ for some decision rule d.

Lemma

S is a convex set.



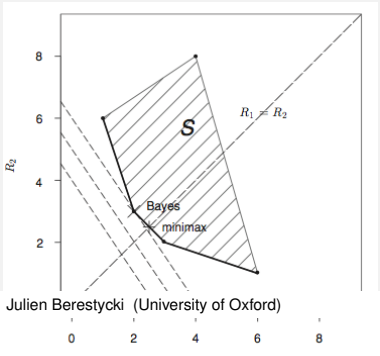
Finite decision problem

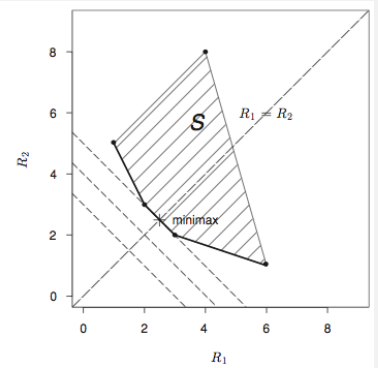


1. Extreme points = non-randomized rules.
2. Lower thick line= admissible rules.
3. Minimax is intersection with $R_1 = R_2$ (random) (**non**-random).

Finite decision problem

To find the Bayes rule; suppose prior is (π_1, π_2). For any c the line $\pi_1 R_1 + \pi_2 R_2 = c$ represents a class of decision rules with same Bayes risk c.





Bayes rule is not unique but can be chosen non-random.