Statistical Inference
Lecture Four

https://people.bath.ac.uk/masss/APTS/apts.html

Simon Shaw

University of Bath

APTS, 14-18 December 2020

Overview of Lecture Four

- Bayesian statistical decision problem, $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$.
- The risk of decision $d \in \mathcal{D}$ under the distribution $\pi(\theta)$ is $\rho(\pi(\theta), d) = \int_\theta L(\theta, d)\pi(\theta)\, d\theta$.
- The Bayes risk $\rho^*(\pi)$ minimises the expected loss,

$$\rho^*(\pi) \;=\; \inf_{d \in \mathcal{D}} \rho(\pi, d)$$

  with respect to $\pi(\theta)$.
- A decision $d^* \in \mathcal{D}$ for which $\rho(\pi, d^*) = \rho^*(\pi)$ is a Bayes rule against $\pi(\theta)$.
- A decision rule $\delta(x)$ is a function from $\mathcal{X}$ into $\mathcal{D}$,
- We view the set of decision rules, to be our possible set of inferences about θ when the sample is observed so that $\mathrm{Ev}(\mathcal{E}, x)$ is $\delta^*(x)$
- The Bayes rule for the posterior decision respects the strong likelihood principle.

Introduction

- Statistical Decision Theory allows us to consider ways to construct the Ev function that reflects our needs, which will vary from application to application, and which assesses the consequences of making a good or bad inference.
- The set of possible inferences, or decisions, is termed the decision space, denoted $\mathcal{D}$.
- For each $d \in \mathcal{D}$, we want a way to assess the consequence of how good or bad the choice of decision d was under the event θ.

Definition (Loss function)

A loss function is any function L from $\Theta \times \mathcal{D}$ to $[0, \infty)$.

- The loss function measures the penalty or error, $L(\theta, d)$ of the decision d when the parameter takes the value θ.
- Thus, larger values indicate worse consequences.

The three main types of inference about θ are

1. point estimation,
2. set estimation,
3. hypothesis testing.

It is a great conceptual and practical simplification that Statistical Decision Theory distinguishes between these three types simply according to their decision spaces.

Type of inference	Decision space $\mathcal{D}$
Point estimation	The parameter space, Θ.
Set estimation	A set of subsets of Θ.
Hypothesis testing	A specified partition of Θ, denoted $\mathcal{H}$.

Bayesian statistical decision theory

In a Bayesian approach, a statistical decision problem $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$ has the following ingredients.

1. The possible values of the parameter: Θ, the parameter space.
2. The set of possible decisions: $\mathcal{D}$, the decision space.
3. The probability distribution on Θ, $\pi(\theta)$. For example,
    1. this could be a prior distribution, $\pi(\theta) = f(\theta)$.
    2. this could be a posterior distribution, $\pi(\theta) = f(\theta \,|\, x)$ following the receipt of some data x.
    3. this could be a posterior distribution $\pi(\theta) = f(\theta \,|\, x, y)$ following the receipt of some data x,y.
4. The loss function $L(\theta, d)$.

In this setting, only θ is random and we can calculate the expected loss, or risk.

Definition (Risk)

The risk of decision $d \in \mathcal{D}$ under the distribution $\pi(\theta)$ is

$$\rho(\pi(\theta), d) \;=\; \int_{\theta} L(\theta, d)\pi(\theta)\, d\theta.$$

We choose d to minimise this risk.

Definition (Bayes rule and Bayes risk)

The Bayes risk $\rho^*(\pi)$ minimises the expected loss,

$$\rho^*(\pi) \;=\; \inf_{d \in \mathcal{D}} \rho(\pi, d)$$

with respect to $\pi(\theta)$. A decision $d^* \in \mathcal{D}$ for which $\rho(\pi, d^*) = \rho^*(\pi)$ is a Bayes rule against $\pi(\theta)$.

The Bayes rule may not be unique, and in weird cases it might not exist. We solve $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$ by finding $\rho^*(\pi)$ and (at least one) d^*.

Example - quadratic loss

Suppose that $\Theta \subset \mathbb{R}$ and we wish to find a point estimate for θ. We consider the loss function $L(\theta, d) = (\theta - d)^2$.

- The risk of decision d is

$$\begin{aligned}\rho(\pi, d) = \mathbb{E}\{L(\theta, d) \,|\, \theta \sim \pi(\theta)\} &= \mathbb{E}_{(\pi)}\{(\theta - d)^2\} \\ &= \mathbb{E}_{(\pi)}(\theta^2) - 2d\mathbb{E}_{(\pi)}(\theta) + d^2,\end{aligned}$$

  where $\mathbb{E}_{(\pi)}(\cdot)$ denotes the expectation with respect to $\pi(\theta)$.
- Differentiating with respect to d we have

$$\frac{\partial}{\partial d}\rho(\pi, d) = -2\mathbb{E}_{(\pi)}(\theta) + 2d.$$

- So, the Bayes rule is $d^* = \mathbb{E}_{(\pi)}(\theta)$.

Example - quadratic loss (continued)

- The corresponding Bayes risk is

$$\begin{aligned}\rho^*(\pi) = \rho(\pi, d^*) &= \mathbb{E}_{(\pi)}(\theta^2) - 2d^*\mathbb{E}_{(\pi)}(\theta) + (d^*)^2 \\ &= Var_{(\pi)}(\theta) + (d^* - \mathbb{E}_{(\pi)}(\theta))^2 \\ &= Var_{(\pi)}(\theta)\end{aligned}$$

where $Var_{(\pi)}(\theta)$ is the variance of θ computed with respect to $\pi(\theta)$.

1. If $\pi(\theta) = f(\theta)$, a prior for θ, then the Bayes rule of an immediate decision is $d^* = \mathbb{E}(\theta)$ with corresponding Bayes risk $\rho^* = Var(\theta)$.
2. If we observe sample data x then the Bayes rule given this sample information is $d^* = \mathbb{E}(\theta \,|\, X)$ with corresponding Bayes risk $\rho^* = Var(\theta \,|\, X)$ as $\pi(\theta) = f(\theta \,|\, x)$.

- Typically we solve:
  1. $[\Theta, \mathcal{D}, f(\theta), L(\theta, d)]$, the immediate decision problem,
  2. $[\Theta, \mathcal{D}, f(\theta \mid x), L(\theta, d)]$, the decision problem after sample information.
- We may also want to consider the risk of the sampling procedure, before observing the sample, to decide whether or not to sample.
- We now consider both θ and X as random.
- For each possible sample, we need to specify which decision to make.

Definition (Decision rule)

A decision rule $\delta(x)$ is a function from $\mathcal{X}$ into $\mathcal{D}$,

$$\delta : \mathcal{X} \to \mathcal{D}.$$

If $X = x$ is the observed value of the sample information then $\delta(x)$ is the decision that will be taken. The collection of all decision rules is denoted by Δ so that $\delta \in \Delta \Rightarrow \delta(x) \in \mathcal{D} \ \forall x \in X$.

- We wish to solve the problem $[\Theta, \Delta, f(\theta, x), L(\theta, \delta(x))]$.

Definition (Bayes (decision) rule and risk of the sampling procedure)

The decision rule δ^* is a Bayes (decision) rule exactly when

$$\mathbb{E}\{L(\theta, \delta^*(X))\} \;\; \leq \;\; \mathbb{E}\{L(\theta, \delta(X))\}$$

for all $\delta(x) \in \mathcal{D}$. The corresponding risk $\rho^* = \mathbb{E}\{L(\theta, \delta^*(X))\}$ is termed the risk of the sampling procedure.

- If the sample information consists of $X = (X_1, \ldots, X_n)$ then ρ^* will be a function of n and so can be used to help determine sample size choice.

Bayes rule theorem, BRT

Suppose that a Bayes rule exists for $[\Theta, \mathcal{D}, f(\theta \,|\, x), L(\theta, d)]$. Then

$$\delta^*(x) = \arg\min_{d \in \mathcal{D}} \mathbb{E}(L(\theta, d) \,|\, X = x).$$

Proof

Let δ be arbitrary. Then

$$\begin{aligned}
\mathbb{E}\{L(\theta, \delta(X))\} &= \int_x \int_\theta L(\theta, \delta(x)) f(\theta, x)\, d\theta dx \\
&= \int_x \int_\theta L(\theta, \delta(x)) f(\theta \,|\, x) f(x)\, d\theta dx \\
&= \int_x \left\{ \int_\theta L(\theta, \delta(x)) f(\theta \,|\, x)\, d\theta \right\} f(x)\, dx \\
&= \int_x \mathbb{E}\{L(\theta, \delta(x)) \,|\, X\} f(x)\, dx
\end{aligned}$$

Proof continued

Now, as $f(x) > 0$, the $\delta^* \in \Delta$ which minimises $\mathbb{E}\{L(\theta, \delta(X))\}$ may equivalently be found as the δ^* which satisfies

$$\rho(f(\theta), \delta^*) = \inf_{\delta(x) \in \mathcal{D}} \mathbb{E}\{L(\theta, \delta(x)) \,|\, X\},$$

giving the result. $\square$

- The minimisation of expected loss over the space of all functions from $\mathcal{X}$ to $\mathcal{D}$ can be achieved by the pointwise minimisation over $\mathcal{D}$ of the expected loss conditional on $X = x$.
- The risk of the sampling procedure is $\rho^* = \mathbb{E}[\mathbb{E}\{L(\theta, \delta^*(x)) \,|\, X\}]$.

Example - quadratic loss

We have $\delta^* = \mathbb{E}(\theta \,|\, X)$ and $\rho^* = \mathbb{E}\{Var(\theta \,|\, X)\}$.

We could consider Δ, the set of decision rules, to be our possible set of inferences about θ when the sample is observed so that $\text{Ev}(\mathcal{E}, x)$ is $\delta^*(x)$. We thus have the following result.

Theorem

The Bayes rule for the posterior decision respects the strong likelihood principle.

Proof

If we have two Bayesian models with the same prior distribution then if $f_{X_1}(x_1 \mid \theta) = c(x_1, x_2) f_{X_2}(x_2 \mid \theta)$ the corresponding posterior distributions are the same and so the corresponding Bayes rule (and risk) is the same. $\square$