Statistical Inference

https://people.bath.ac.uk/masss/APTS/lecture4.pdf

Simon Shaw

University of Bath

Lecture Four, 18 December 2019

Yesterday's lecture

- 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 $\text{Ev}(\mathcal{E}, x)$ is $\delta^*(x)$
- The classical risk for the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\}$ is

$$R(\theta, \delta) \;\; = \;\; \int_X L(\theta, \delta(x)) f_X(x \,|\, \theta)\, dx.$$

Example

Let $X = (X_1, \ldots, X_n)$ where $X_i \sim N(\theta, \sigma^2)$ and σ^2 is known. Suppose that $L(\theta, d) = (\theta - d)^2$ and consider a conjugate prior $\theta \sim N(\mu_0, \sigma_0^2)$. Possible decision functions include:

1. $\delta_1(x) = \overline{x}$, the sample mean.
2. $\delta_2(x) = \text{med}\{x_1, \ldots, x_n\} = \tilde{x}$, the sample median.
3. $\delta_3(x) = \mu_0$, the prior mean.
4. $\delta_4(x) = \mu_n$, the posterior mean where

$$\mu_n = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2}\right),$$

the weighted average of the prior and sample mean accorded to their respective precisions.

Example - continued

The respective classical risks are

1. $R(\theta, \delta_1) = \frac{\sigma^2}{n}$, a constant for θ, since $\overline{X} \sim N(\theta, \sigma^2/n)$.
2. $R(\theta, \delta_2) = \frac{\pi\sigma^2}{2n}$, a constant for θ, since $\tilde{X} \sim N(\theta, \pi\sigma^2/2n)$ (approximately).
3. $R(\theta, \delta_3) = (\theta - \mu_0)^2 = \frac{\sigma^2}{n}\left(\frac{\theta-\mu_0}{\sigma/\sqrt{n}}\right)^2$.
4. $R(\theta, \delta_4) = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-2}\left\{\frac{1}{\sigma_0^2}\left(\frac{\theta-\mu_0}{\sigma_0}\right)^2 + \frac{n}{\sigma^2}\right\}$.

Which decision do we choose? We observe that $R(\theta, \delta_1) < R(\theta, \delta_2)$ for all $\theta \in \Theta$ but other comparisons depend upon θ.

- The accepted approach for classical statisticians is to narrow the set of possible decision rules by ruling out those that are obviously bad.

Definition (Admissible decision rule)

A decision rule δ_0 is inadmissible if there exists a decision rule δ_1 which dominates it, that is

$$R(\theta, \delta_1) \ \leq \ R(\theta, \delta_0)$$

for all $\theta \in \Theta$ with $R(\theta, \delta_1) < R(\theta, \delta_0)$ for at least one value $\theta_0 \in \Theta$. If no such δ_1 exists then δ_0 is admissible.

- If δ_0 is dominated by δ_1 then the classical risk of δ_0 is never smaller than that of δ_1 and δ_1 has a smaller risk for θ_0.
- Thus, you would never want to use δ_0.[1]
- The accepted approach is to reduce the set of possible decision rules under consideration by only using admissible rules.

[1]Here I am assuming that all other considerations are the same in the two cases: e.g. for all $x \in \mathcal{X}$, $\delta_1(x)$ and $\delta_0(x)$ take about the same amount of resource to compute.

- We now show that admissible rules can be related to a Bayes rule δ^* for a prior distribution $\pi(\theta)$.

Theorem

If a prior distribution $\pi(\theta)$ is strictly positive for all Θ with finite Bayes risk and the classical risk, $R(\theta, \delta)$, is a continuous function of θ for all δ, then the Bayes rule δ^* is admissible.

Proof (Robert, 2007)

Letting $f(\theta, x) = f_X(x \mid \theta)\pi(\theta)$ we have

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

Proof continued

- Suppose that the Bayes rule δ^* is inadmissible and dominated by δ_1.
- Thus, in an open set C of θ, $R(\theta, \delta_1) < R(\theta, \delta^*)$ with $R(\theta, \delta_1) \leq R(\theta, \delta^*)$ elsewhere.
- Consequently, $\mathbb{E}\{L(\theta, \delta_1(X))\} < \mathbb{E}\{L(\theta, \delta^*(X))\}$ which is a contradiction to δ^* being the Bayes rule. □

- The relationship between a Bayes rule with prior $\pi(\theta)$ and an admissible decision rule is even stronger.
- The following result was derived by Abraham Wald (1902-1950)

Wald's Complete Class Theorem, CCT

In the case where the parameter space Θ and sample space $\mathcal{X}$ are finite, a decision rule δ is admissible if and only if it is a Bayes rule for some prior distribution $\pi(\theta)$ with strictly positive values.

- An illuminating blackboard proof of this result can be found in Cox and Hinkley (1974, Section 11.6).
- There are generalisations of this theorem to non-finite decision sets, parameter spaces, and sample spaces but the results are highly technical.
- We'll proceed assuming the more general result, which is that a decision rule is admissible if and only if it is a Bayes rule for some prior distribution $\pi(\theta)$, which holds for practical purposes.

So what does the CCT say?

1. Admissible decision rules respect the SLP. This follows from the fact that admissible rules are Bayes rules which respect the SLP. This provides support for using admissible decision rules.
2. If you select a Bayes rule according to some positive prior distribution $\pi(\theta)$ then you cannot ever choose an inadmissible decision rule.

Point estimation

- We now look at possible choices of loss functions for different types of inference.
- For point estimation the decision space is $\mathcal{D} = \Theta$, and the loss function $L(\theta, d)$ represents the (negative) consequence of choosing d as a point estimate of θ.
- It will not be often that an obvious loss function $L : \Theta \times \Theta \to \mathbb{R}$ presents itself. There is a need for a generic loss function which is acceptable over a wide range of applications.

Suppose that Θ is a convex subset of $\mathbb{R}^p$. A natural choice is a convex loss function,

$$L(\theta, d) \;=\; h(d - \theta)$$

where $h : \mathbb{R}^p \to \mathbb{R}$ is a smooth non-negative convex function with $h(0) = 0$.

- This type of loss function asserts that small errors are much more tolerable than large ones.
- One possible further restriction is that h is an even function, $h(d-\theta) = h(\theta - d)$.
- In this case, $L(\theta, \theta+\epsilon) = L(\theta, \theta-\epsilon)$ so that under-estimation incurs the same loss as over-estimation.
- There are many situations where this is not appropriate and the loss function should be asymmetric and a generic loss function should be replaced by a more specific one.
- For $\Theta \subset \mathbb{R}$, the absolute loss function $L(\theta, d) = |\theta - d|$ gives a Bayes rule of the median of $\pi(\theta)$.
- We saw previously, that for quadratic loss $\Theta \subset \mathbb{R}$, $L(\theta, d) = (\theta - d)^2$, the Bayes rule was the expectation of $\pi(\theta)$. This attractive feature can be extended to more dimensions.

Example

If $\Theta \in \mathbb{R}^p$, the Bayes rule δ^* associated with the prior distribution $\pi(\theta)$ and the quadratic loss

$$L(\theta, d) \;=\; (d-\theta)^T Q\,(d-\theta)$$

is the posterior expectation $\mathbb{E}(\theta \,|\, X)$ for every positive-definite symmetric $p \times p$ matrix Q.

Example (Robert, 2007), $Q = \Sigma^{-1}$

Suppose $X \sim N_p(\theta, \Sigma)$ where the known variance matrix Σ is diagonal with elements σ_i^2 for each i. Then $\mathcal{D} = \mathbb{R}^p$. A possible loss function is

$$L(\theta, d) \;=\; \sum_{i=1}^{p} \left(\frac{d_i - \theta_i}{\sigma_i}\right)^2$$

so that the total loss is the sum of the squared componentwise errors.

- As the Bayes rule for $L(\theta, d) = (d - \theta)^T Q\,(d - \theta)$ does not depend upon Q, it is the same for an uncountably large class of loss functions.
- If we apply the Complete Class Theorem to this result we see that for quadratic loss, a point estimator for θ is admissible if and only if it is the conditional expectation with respect to some positive prior distribution $\pi(\theta)$.
- The value, and interpretability, of the quadratic loss can be further observed by noting that, from a Taylor series expansion, an even, differentiable and strictly convex loss function can be approximated by a quadratic loss function.

Set estimation

- For set estimation the decision space is a set of subsets of Θ so that each $d \subset \Theta$.
- There are two contradictory requirements for set estimators of Θ.
  1. We want the sets to be small.
  2. We also want them to contain θ.
- A simple way to represent these two requirements is to consider the loss function

$$L(\theta, d) \;=\; |d| + \kappa(1 - 1_{\theta \in d})$$

  for some $\kappa > 0$ where $|d|$ is the volume of d.
- The value of κ controls the trade-off between the two requirements.
  - If $\kappa \downarrow 0$ then minimising the expected loss will always produce the empty set.
  - If $\kappa \uparrow \infty$ then minimising the expected loss will always produce Θ.

- For loss functions of the form $L(\theta, d) = |d| + \kappa(1 - 1_{\theta \in d})$ we'll show there is a a simple necessary condition for a rule to be a Bayes rule.

Definition (Level set)

A set $d \subset \Theta$ is a level set of the posterior distribution exactly when $d = \{\theta : \pi(\theta \,|\, x) \geq k\}$ for some k.

Theorem (Level set property, LSP)

If δ^* is a Bayes rule for $L(\theta, d) = |d| + \kappa(1 - 1_{\theta \in d})$ then it is a level set of the posterior distribution.

Proof

Note that

$$\begin{aligned} \mathbb{E}\{L(\theta, d) \,|\, X\} &= |d| + \kappa(1 - \mathbb{E}(1_{\theta \in d} \,|\, X)) \\ &= |d| + \kappa\mathbb{P}(\theta \notin d \,|\, X). \end{aligned}$$

Proof continued

- For fixed x, we show that if d is not a level set of the posterior distribution then there is a $d' \neq d$ which has a smaller expected loss so that $\delta^*(x) \neq d$.
- Suppose that d is not a level set of $\pi(\theta \,|\, x)$. Then there is a $\theta \in d$ and $\theta' \notin d$ for which $\pi(\theta' \,|\, x) > \pi(\theta \,|\, x)$.
- Let $d' = d \cup d\theta' \setminus d\theta$ where $d\theta$ is the tiny region of Θ around θ and $d\theta'$ is the tiny region of Θ around θ' for which $|d\theta| = |d\theta'|$.
- Then $|d'| = |d|$ but

$$\mathbb{P}(\theta \notin d' \,|\, X) < \mathbb{P}(\theta \notin d \,|\, X)$$

Thus, $\mathbb{E}\{L(\theta, d') \,|\, X\} < \mathbb{E}\{L(\theta, d) \,|\, X\}$ showing that $\delta^*(x) \neq d$. □

- The Level Set Property Theorem states that δ having the level set property is necessary for δ to be a Bayes rule for loss functions of the form $L(\theta, d) = |d| + \kappa(1 - 1_{\theta \in d})$.
- The Complete Class Theorem states that being a Bayes rule is a necessary condition for δ to be admissible.
- Being a level set of a posterior distribution for some prior distribution $\pi(\theta)$ is a necessary condition for being admissible for loss functions of this form.
- Bayesian HPD regions satisfy the necessary condition for being a set estimator.
- Classical set estimators achieve a similar outcome if they are level sets of the likelihood function, because the posterior is proportional to the likelihood under a uniform prior distribution.[2]

[2] In the case where Θ is unbounded, this prior distribution may have to be truncated to be proper.

Hypothesis tests

- For hypothesis tests, the decision space is a partition of Θ, denoted

$$\mathcal{H} \;:=\; \{H_0, H_1, \ldots, H_d\}.$$

- Each element of $\mathcal{H}$ is termed a hypothesis.
- The loss function $L(\theta, H_i)$ represents the (negative) consequences of choosing element H_i, when the true value of the parameter is θ.
- It would be usual for the loss function to satisfy

$$\theta \in H_i \implies L(\theta, H_i) \;=\; \min_j L(\theta, H_j)$$

on the grounds that an incorrect choice of element should never incur a smaller loss than the correct choice.

- A generic loss function for hypothesis tests is the 0-1 ('zero-one') loss function

$$L(\theta, H_i) = 1 - 1_{\{\theta \in H_i\}}.$$

i.e., zero if $\theta \in H_i$, and one if it is not.

- The corresponding Bayes rule is to select the hypothesis with the largest posterior probability.
- The drawback is that this loss function is hard to defend as being realistic.
- An alternative approach is to co-opt the theory of set estimators.
- The statistician can use her set estimator δ to make at least some distinctions between the members of $\mathcal{H}$:
  - Accept H_i exactly when $\delta(x) \subset H_i$,
  - Reject H_i exactly when $\delta(x) \cap H_i = \emptyset$,
  - Undecided about H_i otherwise.