Decision Theory: Part 1

Dr. Qiuzhuang Sun

STAT3023

Decision theory

- Many statistical procedures can be analysed through the general framework of statistical decision theory.
- We will begin with a special case: simple prediction problems.
- Suppose Y is a random variable from a known distribution and $\mathcal{D}$ is an arbitrary set which is called the decision space. For each possible value y of Y and a decision $d \in \mathcal{D}$, we measure the performance of d using a loss function $L(d|y) \geq 0$.
- Goal: choose d to minimise the expected loss (risk):

$$R(d) = E[L(d|Y)].$$

Decision theory

Example: **Squared error loss**.

$\mathcal{D} = \mathbb{R}$, $L(d|y) = C(d-y)^2$ for some $C > 0$. Which d minimises the risk?

Decision theory

Example: **Absolute error loss**.

$\mathcal{D} = \mathbb{R}$, $L(d|y) = C|d - y|$ for some $C > 0$. Y is continuous with cdf $F(\cdot)$ and density $f(\cdot)$. Which d minimises the risk?

Decision theory

Example: **Absolute error loss**.

$\mathcal{D} = \mathbb{R}$, $L(d|y) = C|d - y|$ for some $C > 0$. Y is continuous with cdf $F(\cdot)$ and density $f(\cdot)$. Which d minimises the risk?

Decision theory

Example: **0-1 (zero-one) loss**.

$\mathcal{D} = \mathbb{R}$, $L(d|y) = 1\{|d - y| > c\}$ for some $c > 0$. Y is continuous with density $f(\cdot)$, where $f(\cdot)$ is unimodal. That is, $f(y)$ is strictly increasing for $y < m$ and strictly decreasing for $y > m$ for some mode m. Further assume $f(y) > 0$ over an interval I with length at least $2c$. Which d minimises the risk?

Example: **0-1 (zero-one) loss**.

$\mathcal{D} = \mathbb{R}$, $L(d|y) = 1\{|d - y| > c\}$ for some $c > 0$. Y is continuous with density $f(\cdot)$, where $f(\cdot)$ is unimodal. That is, $f(y)$ is strictly increasing for $y < m$ and strictly decreasing for $y > m$ for some mode m. Further assume $f(y) > 0$ over an interval I with length at least $2c$. Which d minimises the risk?

Decision theory

Example: **Discrete selection**.

Suppose $\mathbb{R}$ is partitioned into disjoint sets $S_1, S_2, \ldots, S_k$, $\mathcal{D} = \{1, 2, \ldots, k\}$, and the loss is $L(d|y) = \sum_{j=1}^{k} L_{d,j} 1\{y \in S_j\}$, where $L_{i,j}$ is the (i,j)th entry of a $k \times k$ loss matrix, such that $L_{i,i} = 0$ for $i = 1, \ldots, k$, and $L_{i,j} = L_j$ for $i \neq j$.

Full decision theory framework

In the full framework, we have

- A family of distributions $\mathcal{F} = \{f_\theta(\cdot) : \theta \in \Theta\}$ for a random vector $\mathbf{X}$ taking values in $\mathcal{X}$.
- A decision space $\mathcal{D}$, where each decision $d(\cdot)$ is a **function** mapping a possible value $\mathbf{x} \in \mathcal{X}$ into $\mathcal{D}$.
- A non-negative-valued loss function such that when a decision d is made and the true distribution generating $\mathbf{X}$ is $f_\theta(\cdot)$, a loss of $L(d|\theta) = L(d(\mathbf{X})|\theta)$ is suffered.
- The risk function associated with decision function $d(\cdot)$ is:

$$R(\theta|d(\cdot)) = E_\theta[L(d(\mathbf{X})|\theta)], \quad \mathbf{X} \sim f_\theta(\cdot).$$

Full decision theory framework

Example: Suppose we have 2 independent observations X_1, X_2 from an exponential distribution with mean θ, which has PDF $f_\theta(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}}$ for $x > 0$. Take the loss as $L(d|\theta) = (d-\theta)^2$.

- We already know that the CRLB for unbiased estimation of θ is $\theta^2/2$, attained by $\bar{X} = \frac{X_1+X_2}{2} = d_{\mathsf{MVU}}(\mathbf{X})$.
- Consider the family of decisions $\{d_c(\cdot) : c > 0\}$ given by $d_c(\mathbf{X}) = c\bar{X}$.
- The risk of $d_c(\cdot)$ is:

Full decision theory framework

(Example continued)

Full decision theory framework

(Example continued)

Full decision theory framework

- The above example shows neither d_0 nor $d_{2/3}$ is uniformly better than the other. Rather, there are ranges of θ for which each is better.

- It is not very useful to compare risk functions in a pointwise sense. In fact we need some "overall" measure of risk to encompass all θ values.

Overall risk measure

Bayes (or integrated) risk: For a given non-negative weight function (prior) $w(\cdot)$:

$$B_w(d) = \int_\Theta w(\theta) R(\theta|d) d\theta.$$

If $\tilde{d}$ satisfies $B_w(\tilde{d}) \leq B_w(d)$ for any other decision function $d(\cdot)$, then $\tilde{d}$ is said to be a Bayes procedure (or Bayes decision rule) w.r.t. weight/prior $w(\cdot)$.

Overall risk measure

Maximum risk: For a given subset $\Theta_0 \subseteq \Theta$, a decision rule $\hat{d}(\cdot)$ is said to be minimax (over Θ_0) if

$$\max_{\theta \in \Theta_0} R(\theta | \hat{d}) \leq \max_{\theta \in \Theta_0} R(\theta | d)$$

for any other decision function $d(\cdot)$.

It can be understood as the best decision in the worst scenario.

Finding Bayes procedures

- Bayes procedures can be found by reducing the problem to a simple prediction problem.
- Recall the **Bayes risk** of a decision rule $d(\cdot)$ (w.r.t. to a weight function/prior $w(\cdot)$) is:

Finding Bayes procedures

(Example continued)

Finding Bayes procedures

(Example continued)

Finding Bayes procedures

Example: Suppose $X_1, \ldots, X_n$ are iid $N(\theta, 1)$, $\theta \in \Theta = \mathbb{R}$ with decision space $\mathcal{D}$ and loss $L(d|\theta)$.

(a) $\mathcal{D} = \mathbb{R}$, $L(d|\theta) = (d-\theta)^2$
(b) $\mathcal{D} = \mathbb{R}$, $L(d|\theta) = |d-\theta|$
(c) $\mathcal{D} = \mathbb{R}$, $L(d|\theta) = 1\{|d-\theta| > 1.96/\sqrt{n}\}$
(d) $\mathcal{D} = \{0, 1\}$,

$$L(d|\theta) = \begin{cases} L_0 & \text{if } d = 1,\ \theta \in \Theta_0 \\ L_1 & \text{if } d = 0,\ \theta \in \Theta_1 \\ 0 & \text{otherwise.} \end{cases}$$

Find Bayes procedures of the above with $w(\theta) = 1$, the "flat prior".

Finding Bayes procedures

(Example continued)

Finding Bayes procedures

(Example continued)

Finding Bayes procedures

(Example continued)

Finding Bayes procedures

Let's consider a more concrete example. For (d) in the last example, we assume $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, \infty)$. Recall $X_1, \ldots, X_n$ are iid $N(\theta, 1)$, $\theta \in \Theta = \mathbb{R}$, $\mathcal{D} = \{0, 1\}$, and

$$L(d|\theta) = \begin{cases} L_0 & \text{if } d = 1,\ \theta \in \Theta_0 \\ L_1 & \text{if } d = 0,\ \theta \in \Theta_1 \\ 0 & \text{otherwise.} \end{cases}$$

Find Bayes procedures with $w(\theta) = 1$.

Finding Bayes procedures

(Example continued)

Finding Bayes procedures

(Example continued)

Finding Bayes procedures

Again, we assume $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, \infty)$. Recall $X_1, \ldots, X_n$ are iid $N(\theta, 1)$, $\theta \in \Theta = \mathbb{R}$, $\mathcal{D} = \{0, 1\}$, and

$$L(d|\theta) = \begin{cases} L_0 & \text{if } d = 1,\ \theta \in \Theta_0 \\ L_1 & \text{if } d = 0,\ \theta \in \Theta_1 \\ 0 & \text{otherwise.} \end{cases}$$

Find Bayes procedures with the normal prior: $w(\theta)$ is the density of $N(\mu_0, \sigma_0^2)$.

Finding Bayes procedures

(Example continued)

The Bayesian interpretation

- "Frequentists vs Bayesians" debate in statistics:
  - The frequentists approach to statistical modelling assumes that the data are generated from a **fixed** distribution in a known family:
    $$\{f_\theta(\cdot) : \theta \in \Theta\}$$
    Inference consists of hypothesis testing, point/interval estimation, etc.
  - The Bayesian approach is to specify a known weight function/prior distribution $w(\cdot)$ on Θ and assume the data is generated by (assuming $w(\cdot)$ is a "proper" distribution):
    (i) First draw a random value θ from $w(\cdot)$;
    (ii) Conditional on θ, data are generated from $f_\theta(\cdot)$.

    Inference is based on the posterior distribution $p(\theta|\mathbf{x})$, given the observed values $\mathbf{x}$.

The Bayesian interpretation

Assuming $\int_\Theta w(\theta)d\theta = 1$. In the Bayesian calculation:

- $f_\theta(\cdot)$ is the conditional density/PMF of $\mathbf{X}$ given θ
- $w(\theta)f_\theta(\mathbf{X})$ is the joint density/PMF of $(\theta, \mathbf{X})$
- $m(\mathbf{x}) = \int_\Theta w(\theta)f_\theta(\mathbf{x})d\theta$ is the marginal of $\mathbf{X}$ at $\mathbf{x}$
- posterior
$$p(\theta|\mathbf{x}) = \frac{w(\theta)f_\theta(\mathbf{x})}{m(\mathbf{x})}$$
is the conditional density of θ given $\mathbf{X} = \mathbf{x}$
- $B_w(d) = \int_\Theta \int_\mathbf{x} L(d(\mathbf{x})|\theta)f_\theta(x)w(\theta)d\mathbf{x}d\theta$ is the Bayes risk as the "overall" expected loss

The Bayesian interpretation

- We shall always take the frequentist point of view. That is, we assume there is a fixed, non-random but unknown true parameter value (even though there is a Bayesian interpretation of the overall risk).

- We do not want to restrict ourselves to integrable weight functions (so-called "proper" priors). Our examples have shown that even if $w(\cdot)$ is NOT integrable (i.e. is an "improper" prior), the resulting posterior may still be integrable.

Conjugate Priors

In many examples the weight function (or "prior") $w(\cdot)$ may be chosen in such a way that the posterior is of the same form or from the same family. When this happens, it is called a conjugate prior.

Example 1: normal prior for mean θ of a normal distribution (see tutorial questions).

Example 2: $X_1, \ldots, X_n$ are independent and $X_i \sim \text{Binomial}(m_i, \theta)$ with known $m_1, \ldots, m_n$. The prior of θ is $\text{Beta}(\alpha, \beta)$ distribution:

$$w(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

Conjugate Priors

(Example continued)

Conjugate Priors

(Example continued)

Conjugate Priors

Example 3: Let $X_1, \ldots, X_n$ be independent and $X_i \sim \text{Poisson}(k_i\theta)$ with known $k_1, \ldots, k_n$. (A real scenario can be the number of bugs on a leaf with area k_i.) Consider a gamma prior of θ with shape α and rate λ:

$$w(\theta) = \frac{\theta^{\alpha-1}e^{-\lambda\theta}\lambda^{\alpha}}{\Gamma(\alpha)}, \quad \theta > 0.$$

Conjugate Priors

(Example continued)

Table of conjugate Priors

Parameter θ	Conjugate prior $w(\theta)$
normal mean	normal
Binomial success probability	Beta
Negative binomial success probability	Beta
Poisson mean	Gamma
Gamma scale	inverse Gamma
normal variance	inverse Gamma
$U(0, \theta)$	Pareto

See more on Wikipedia