Decision theory[1]

[1]Based on Migon and Gamerman's (1999) *Statistical Inference: An Integrated Approach*.

Outline

Loss function and decision rule

A decision problem is completely specified by the description of three spaces:

Θ: Parameter (or states of the nature) space;

Ω: Space of possible results of an experiment;

$\mathcal{A}$: Space of possible actions.

A loss function associates losses to pairs of actions and states of the nature. Or, $L(\theta, a)$ has values in R^+ for $(\theta, a) \in \Theta \times \mathcal{A}$.

A decision rule $\delta(x)$ is a function from Ω into $\mathcal{A}$.

Bayesian Expected Loss and risk function

If $\pi^*(\theta)$ is the believed probability distribution of θ at the time of decision making, the Bayesian expected loss of an action a is

$$\rho(\pi^*, a) = E_{\pi^*}(L(\theta, a)) = \int_\Theta L(\theta, a) dF^{\pi^*}(\theta)$$

The risk function of a decision rule $\delta(x)$ is defined by

$$R(\theta, \delta) = E_{x|\theta}[L(\theta, \delta)] = \int_\Omega L(\theta, \delta(x)) dF(x|\theta)$$

Admissibility

A decision rule δ_1 is *R*-better than a decision rule δ_2 if

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

with restrict inequality for some θ. A rule δ_1 is *R*-equivalent to δ_2 if $R(\theta, \delta_1) = R(\theta, \delta_2)$ for all θ.

A decision rule is admissible if there exists no R-better decision rule.

Bayes risk

The Bayes risk of a decision rule δ with respect to a prior distribuiton π on Θ, is defined as

$$\begin{aligned} r(\pi, \delta) &= E_\theta[R(\theta, \delta)] = \int_\Theta R(\theta, \delta) dF^\pi(\theta) \\ &= \int_\Theta \int_\Omega L(\theta, \delta(x)) dF(x|\theta) dF^\pi(\theta) \\ &= \int_\Omega \left\{ \int_\Theta L(\theta, \delta(x)) dF^{\pi^*}(\theta) \right\} dF(x) \\ &= \int_\Omega \underbrace{\rho(\pi, \delta(x))}_{\text{Expected loss}} \; dF(x), \end{aligned}$$

where, by Bayes' Theorem,

$$dF(x|\theta) dF^\pi(\theta) = dF(x) dF^{\pi^*}(\theta).$$

Bayes risk is expectation of the Bayesian expected loss w.r.t. to the predictive $dF(x)$.

Conditional Bayes Principle

Choose an action $a \in \mathcal{A}$ which minimizes $\rho(\pi^*, a)$. Such an action will be called a Bayes rule or action.

Bayes Risk Principle: A decision rule δ_1 is preferred to a rule δ_2 if

$$r(\pi, \delta_1) < r(\pi, \delta_2)$$

Minimax Principle: A decision rule δ_1^* is preferred to a rule δ_2^* if

$$\sup_{\theta \in \Theta} R(\theta, \delta_1^*) < \sup_{\theta \in \Theta} R(\theta, \delta_2^*)$$

Definition: A rule δ^{*M} is a minimax decision rule if it minimizes $\sup_\theta R(\theta, \delta^*)$ among all rules in $\mathcal{D}^*$, i.e., if

$$\sup_{\theta \in \Theta} R(\theta, \delta^{*M}) = \inf_{\delta^* \in \mathcal{D}^*} \sup_{\theta \in \Theta} R(\theta, \delta^*)$$

Example

A doctor must decide whether John must undergo surgery or not.

States of the nature: $\Theta = \{\theta_0, \theta_1\}$
John is not sick (θ_0)
John is sick (θ_1)

Space of actions: $\mathcal{A} = \{a_0, a_1\}$
John should not undergo a surgery (a_0)
John should undergo a surgery (a_1)

Decisions and losses:

Θ	$\mathcal{A}$	
	No surgery (a_0)	Surgery (a_1)
Not sick (θ_0)	0	500
Sick (θ_1)	1000	100

Example

Losses represent the subjective evaluation of the decisor with respect to the combinations of actions and states of the nature.

The decision must be guided by taking into consideration be uncertainty about the unknowns involved in the problem:

$$\begin{aligned} Pr(\theta_1) &= \pi \\ Pr(\theta_0) &= 1-\pi \end{aligned}$$

Risk Analysis

$$\rho(\pi, a) = E_\pi(L(\theta, a)) = \begin{cases} 0(1-\pi) + 1000\pi & \text{for } a_0 \\ 500(1-\pi) + 100\pi & \text{for } a_1 \end{cases}$$

Therefore,

$$\begin{aligned} \rho(\pi, a_0) &= 1000\pi \\ \rho(\pi, a_1) &= 500 - 400\pi \end{aligned}$$

Risk analysis

The two actions have equal risk if

$$\rho(\pi, a_0) = \rho(\pi, a_1)$$

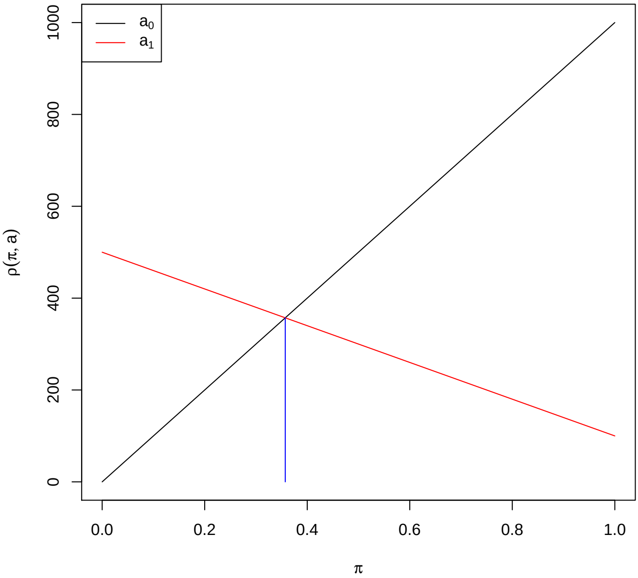
or when $\pi = 5/14 \approx 35.7\%$

$\pi < 5/14$
The risk of a_0, $\rho(\pi, a_0)$, is smaller than the risk of a_1, $\rho(\pi, a_1)$
$\Rightarrow a_0$ is the Bayes rule and the Bayes risk is 1000π.

$\pi > 5/14$
The risk of a_0, $\rho(\pi, a_0)$, is greater than the risk of a_1, $\rho(\pi, a_1)$
a_1 is the Bayes rule and the Bayes risk is $500 - 400\pi$.

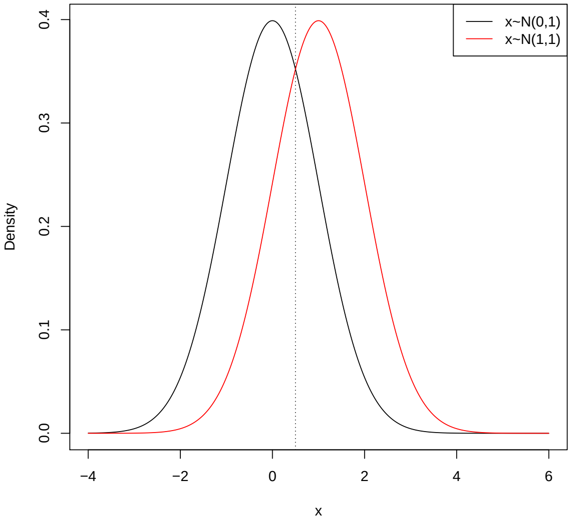
John should undergo a surgery if and only if $\pi > 5/14$.

Risk analysis



Adding some data to the mix

Suppose that $X|\theta_0 \sim N(0,1)$ and $X|\theta_1 \sim N(1,1)$.



It is easy to see that

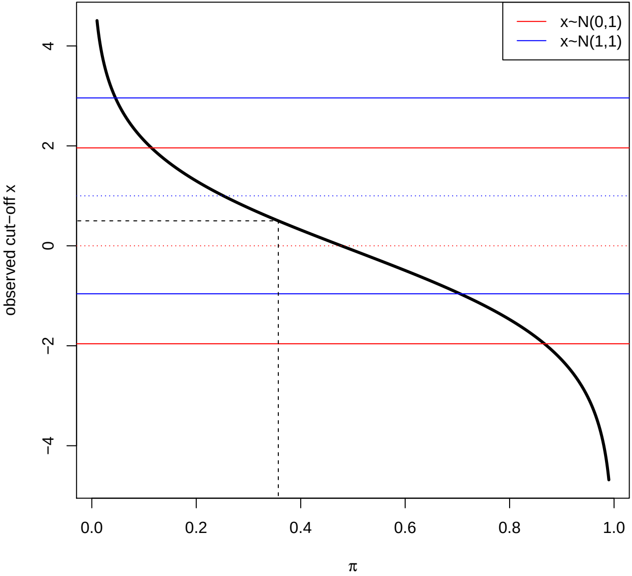
$$Pr(\theta_1|x) = \frac{1}{1 + \left(\frac{1-\pi}{\pi}\right)\exp\{1/2 - x\}}$$

a_0 is the Bayes rule when $Pr(\theta_1|x) < 5/14$, or when

$$x < \frac{1}{2} - \log\left(\frac{9\pi}{5(1-\pi)}\right)$$

If $\pi = 0.1$ $(0.2, 5/14, 0.5, 0.8)$, then a_0 is the Bayes rule when $x < 2.11$ $(1.30, 0.5, -0.09, -1.47)$.

Risk analysis



Comparing two decision rules

Let us assume two decision rules

$$\begin{aligned}\delta_1(x) &= a_0 1_{\{x<0.76\}}(x) + a_1 1_{\{x>0.76\}}(x)\\ \delta_2(x) &= a_0 1_{\{x<0.50\}}(x) + a_1 1_{\{x>0.50\}}(x),\end{aligned}$$

such that

$$\begin{aligned}R(\theta_0,\delta_1) &= \int L(\theta_0,\delta_1(x)) f_n(x;0,1)dx\\ &= \int_{-\infty}^{0.76} L(\theta_0,a_0) f_n(x;0,1)dx + \int_{0.76}^{\infty} L(\theta_0,a_1) f_n(x;0,1)dx\\ &= L(\theta_0,a_0)\Phi(0.76) + L(\theta_0,a_1)(1-\Phi(0.76)),\end{aligned}$$

and

$$\begin{aligned}R(\theta_1,\delta_1) &= L(\theta_1,a_0)\Phi(-0.24) + L(\theta_1,a_1)(1-\Phi(-0.24))\\ R(\theta_0,\delta_2) &= L(\theta_0,a_0)\Phi(0.50) + L(\theta_0,a_1)(1-\Phi(0.50))\\ R(\theta_1,\delta_2) &= L(\theta_1,a_0)\Phi(-0.50) + L(\theta_1,a_1)(1-\Phi(-0.50))\end{aligned}$$

Minimax Principle

It is easy to see that δ_2 is preferred to δ_1:

$$
\begin{aligned}
R(\theta_0, \delta_1) &= 500(1 - \Phi(0.76)) = 111.8136 \\
R(\theta_1, \delta_1) &= 1000\Phi(-0.24) + 100(1 - \Phi(-0.24)) = 464.6486 \\
R(\theta_0, \delta_2) &= 500(1 - \Phi(0.50)) = 154.2688 \\
R(\theta_1, \delta_2) &= 1000\Phi(-0.50) + 100(1 - \Phi(-0.50)) = 377.6838
\end{aligned}
$$

Bayes Risk Principle

Let $\pi = 0.3$, such that $1/2 - \log(9\pi/(5(1-\pi)) = 0.76$.

Then, the Bayes risks are

$$\begin{aligned}
r(\pi, \delta_1) &= (1-\pi)R(\theta_0, \delta_1) + \pi R(\theta_1, \delta_1) \\
&= (0.7)(111.8136) + (0.3)(464.6486) = 217.6641 \\
r(\pi, \delta_2) &= (1-\pi)R(\theta_0, \delta_2) + \pi R(\theta_1, \delta_2) \\
&= (0.7)(154.2688) + (0.3)(377.6838) = 221.2933
\end{aligned}$$

and δ_1 is preferred to δ_2.

In fact, δ_1 is preferred to δ_2 for all $\pi < 0.328042$.

Estimator, estimate and square loss

An estimator is an optimal decision rule with respect to a given loss function. Its observed value is called estimate.

Lemma: Let $L_1(\delta, \theta) = (\delta - \theta)^2$ be the loss associated with the estimation of θ by δ. The estimator of θ is $\delta_1 = E(\theta)$, the mean of the updated distribution of θ.

Proof: The risk function can be written as

$$\begin{aligned}
R(\theta, \delta) &= E[(\delta - \theta)^2] = E\{[(\delta - \delta_1) + (\delta_1 - \theta)]^2\} \\
&= E_\theta[(\delta - \delta_1)^2] + E_\theta[(\delta_1 - \theta)^2] + 2E_\theta[(\delta - \delta_1)(\delta_1 - \theta)] \\
&= (\delta - \delta_1)^2 + E_\theta[(\delta_1 - \theta)^2] + 2(\delta - \delta_1)E_\theta[\delta_1 - \theta] \\
&= (\delta - \delta_1)^2 + E_\theta[(\delta_1 - \theta)^2] = (\delta - \delta_1)^2 + V(\theta)
\end{aligned}$$

which is minimized for $\delta = \delta_1$.

The Bayes risk is $R(\delta_1) = V(\theta)$ and $R(\delta_1) \leq R(\delta)$, $\forall \delta$, with equality iff $\delta_1 = \delta$.

Absolute value loss

The quadratic loss is sometimes criticized for introducing a penalty that increases very strongly with the estimation error $\delta - \theta$.

In many cases, it is desirable to have a loss function that does not overly emphasize large estimation errors.

Lemma: Let $L_2(\delta, \theta) = |\delta - \theta|$ be the loss associated with the estimation of θ. The estimator of θ is $\delta_2 = \text{med}(\theta)$, the median of the updated distribution of θ.

0-1 loss

Another form to reduce the effect of large estimation errors is to consider loss functions that remain constant whenever $|\delta - \theta| > k$ for some k arbitrary. The most common choice is the limiting value as $k \to 0$. This loss function associates a fixed loss when an error is committed, irrespective of its magnitude.

Lemma: Let $L_3(\delta, \theta) = \lim_{\varepsilon \to 0} I_{|\theta - \delta|}([\varepsilon, \infty))$. The estimator of θ is $\delta_3 = \text{mode}(\theta)$, the mode of the updated distribution of θ.

Proof (for the θ continuous case):

$$
\begin{aligned}
E[L_3(\delta,\theta)] &= \lim_{\varepsilon \to 0} \int_{-\infty}^{\delta-\varepsilon} 1 \cdot p(\theta)\, d\theta + \int_{\delta-\varepsilon}^{\delta+\varepsilon} 0 \cdot p(\theta)\, d\theta + \int_{\delta+\varepsilon}^{\infty} 1 \cdot p(\theta) d\theta \\
&= \lim_{\varepsilon \to 0} 1 - \int_{\delta-\varepsilon}^{\delta+\varepsilon} p(\theta)\, d\theta = 1 - \lim_{\varepsilon \to 0} P(\delta - \varepsilon < \theta < \delta + \varepsilon) = p(\delta),
\end{aligned}
$$

which is minimized when $p(\delta)$ is maximized $\Rightarrow \delta_3 = \text{mode}(\theta)$. $\square$

GMLE

When the updated distribution is the posterior, the estimator associated with the 0-1 loss is the posterior mode. This is also referred to as the generalized maximum likelihood estimator (GMLE).

In the continuous case, it involves finding the solution to the equation

$$\frac{\partial p(\theta|\mathbf{x})}{\partial\theta} = 0.$$

Example

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample from a normal distribution with mean θ and variance σ^2, $N(\theta, \sigma^2)$, where $\phi = \sigma^{-2}$.

If a joint conjugate prior for (θ, ϕ) is used

$$\theta|\phi \sim N(\mu_0, (c_0\phi)^{-1}) \quad \text{and} \quad (n_0\sigma_0^2)\phi \sim \chi^2_{n_0},$$

then the joint posterior is also in the same family:

$$\theta|\phi, \mathbf{x} \sim N(\mu_1, (c_1\phi)^{-1}) \quad \text{and} \quad (n_1\sigma_1^2)\phi|\mathbf{x} \sim \chi^2_{n_1}.$$

Therefore, the logarithm of the posterior $p(\theta, \phi|\mathbf{x})$,

$$\log p(\theta, \phi|\mathbf{x}) = \kappa - \frac{\phi}{2}\left[c_1(\theta - \mu_1)^2 + n_1\sigma_1^2\right] + \left(\frac{n_1 + 1}{2} - 1\right)\log\phi$$

Differentiate it with respect to θ and ϕ:

$$\frac{\partial \log p(\theta, \phi | \mathbf{x})}{\partial \theta} = -\frac{\phi}{2} \left[2c_1(\theta - \mu_1) \right] = -\phi c_1(\theta - \mu_1),$$

so

$$\hat{\theta} = \mu_1 \text{ is a critical point.}$$

Similarly,

$$\frac{\partial \log p(\theta, \phi | \boldsymbol{x})}{\partial \phi} = -\frac{c_1(\theta - \mu_1)^2 + n_1 \sigma_1^2}{2} + \left(\frac{n_1 + 1}{2} - 1 \right) \frac{1}{\phi}$$

such that

$$\frac{\partial \, log \; p(\theta = \mu_1, \phi | \mathbf{x})}{\partial \phi} = 0$$

leads to

$$\hat{\phi} = \left(\frac{n_1 - 1}{n_1} \right) \frac{1}{\sigma_1^2}$$

The second order conditions are satisfied as

$$
\begin{aligned}
\frac{\partial^2 \log p(\theta = \mu_1, \phi = \hat{\phi}|\mathbf{x})}{\partial^2 \theta} &= -c_1 \hat{\phi} < 0 \\
\frac{\partial^2 \log p(\theta = \mu_1, \phi = \hat{\phi}|\mathbf{x})}{\partial^2 \phi} &= -\left(\frac{n_1 + 1}{2} - 1 \right) \frac{1}{\hat{\phi}^2} < 0 \\
\frac{\partial^2 \log p(\theta = \mu_1, \phi = \hat{\phi}|\mathbf{x})}{\partial \theta \partial \phi} &= 0
\end{aligned}
$$

Therefore, $(\mu_1, \hat{\phi})$ is the mode of the joint posterior distribution of (θ, ϕ).

Joint and marginal modes

The above calculations do not guarantee that μ_1 is the maximum of the marginal distribution of θ and $\hat{\phi}$ is the maximum of the marginal distribution of ϕ.

The marginal distribution of θ is a Student-t centered at μ_1:

$$\phi|\mathbf{x} \sim G(n_1/2, n_1\sigma_1^2/2)$$

that has posterior mode

$$\tilde{\phi} = \left(\frac{n_1 - 2}{n_1}\right)\frac{1}{\sigma_1^2} \neq \left(\frac{n_1 - 1}{n_1}\right)\frac{1}{\sigma_1^2} = \hat{\phi}$$

and posterior mean

$$E(\phi|\mathbf{x}) = \sigma_1^{-2}.$$

Mode and mean are not invariant under transformations.

Mode of σ^2

$\tilde{\phi}^{-1}$ is not the joint nor the marginal mode of σ^2.

To evaluate the mode of σ^2, $p(\sigma^2|\mathbf{x})$ must be obtained:

$$\log\, p(\sigma^2|\mathbf{x}) = k - \left(\frac{n_1}{2}+1\right)\log\sigma^2 - \frac{n_1\sigma_1^2}{2\sigma^2}.$$

so

$$\frac{\partial \log p(\sigma^2|\mathbf{x})}{\partial\sigma^2} = -\left(\frac{n_1}{2}+1\right)\frac{1}{\tilde{\sigma}^2} + \frac{n_1\sigma_1^2}{2\tilde{\sigma}^4} = 0$$

where

$$\tilde{\sigma}^2 = \left(\frac{n_1}{n_1+2}\right)\frac{1}{\sigma_1^2} \neq \left(\frac{n_1}{n_1-2}\right)\frac{1}{\sigma_1^2} = \tilde{\phi}^{-1}.$$

The second order condition guarantees the maximum as

$$\frac{\partial^2 \log p(\tilde{\sigma}^2|\mathbf{x})}{\partial(\sigma^2)^2} = \left(\frac{n_1}{2}+1\right)\frac{1}{\tilde{\sigma}^4} - 2\frac{n_1\sigma_1^2}{2\tilde{\sigma}^6} = -\frac{1}{2}\frac{(n_1+2)^3}{(n_1\sigma_1^2)^2} < 0.$$