3) Bayesian Decision Theory

- ingredients: model $\{f_\theta : \theta \in \Theta\}$
  loss $L : \Theta \times \mathcal{A} \to [0, \infty)$
  prior π

- for any decision function δ the principle of minimizing expected risk leads to the prior risk of each $\delta \in \mathcal{D}$

$$r(\delta) = E_\pi(R(\theta, \delta))$$

$$= \int_\Theta \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a)\, \delta(x, da)\, P_\theta(dx)\, \pi(d\theta)$$

$$= \int_{\mathcal{X}} \int_\Theta \int_{\mathcal{A}} L(\theta, a)\, \delta(x, da)\, \pi(d\theta|x)\, M(dx)$$

where $P_\theta(dx)\, \pi(d\theta) = \pi(d\theta|x)\, M(dx)$

with $\pi(\theta|x) = \dfrac{\pi(\theta) f_\theta(x)}{m(x)}$ = conditional of θ given x

= posterior density of θ

$m(x) = \int_\Theta f_\theta(x)\, \pi(d\theta)$

= prior predictive density of x.

$$= \int_{\mathcal{X}} r(\delta|x)\, M(dx)$$

where $r(\delta|x) = \int_{\Theta}\int_{\Psi} L(\theta,\psi)\,\delta(x,d\psi)\,\Pi(d\theta|x)$

$= $ posterior risk of δ

- <u>note</u> - a nice consistency property for Bayes

- any parameter $\lambda = \Lambda(\theta)$ s.t. $\theta \leftrightarrow (\Lambda(\theta), \Psi(\theta)) = (\lambda, \psi)$ is called a <u>nuisance parameter</u> because interest is in ψ

- suppose instead we start with the model $\{m_\psi : \psi \in \Psi\}$ where

$$m_\psi(x) = \int_{\Psi^{-1}\{\psi\}} f_\theta(x)\,\Pi(d\theta\,|\,\Psi)(\psi)$$

$\Psi^{-1}\{\psi\} = \{\theta : \Psi(\theta) = \psi\}$ = conditional prior predictive of x given $\Psi(\theta) = \psi$.

- the marginal prior for ψ is

$$\pi_\Psi(\psi) = \int_{\Psi^{-1}\{\psi\}} \Pi(d\theta)$$

- then based on model $\{m_\psi : \psi \in \Psi\}$, prior π_Ψ the posterior of ψ is

$$\frac{\pi_\Psi(\psi)\, m_\psi(x)}{\int_\Psi m_\psi(x)\,\Pi_\Psi(d\psi)}$$ where

$$\int_{\Psi} m_\psi(x)\,\pi_\Psi(d\psi) = \int_\Psi \int_{\Psi^{-1}\{\psi\}} f_\theta(x)\,\pi(d\theta\,|\,\Psi)(\psi)\,\pi_\Psi(d\psi)$$

$$\overset{TTE}{=} \int_\Theta f_\theta(x)\,\pi(d\theta) = m(x)$$

$$\pi_\Psi(\psi)\, m_\psi(x) = \pi_\Psi(\psi)\int_{\Psi^{-1}\{\psi\}} f_\theta(x)\,\pi(d\theta\,|\,\Psi)(\psi)$$

$$\overset{\text{discrete case}}{=} \int_{\Psi^{-1}\{\psi\}} f_\theta(x)\,\pi(d\theta)$$

$$= m(x)\int_{\Psi^{-1}\{\psi\}} \pi(d\theta\,|\,x)$$

and $\int_{\Psi^{-1}\{\psi\}} \pi(d\theta\,|\,x)$ is the marginal posterior for ψ obtained from $\{f_\theta : \theta \in \Theta\}, \pi$

∴ we obtain the same posterior for ψ whether we start from $\{f_\theta : \theta \in \Theta\}, \pi$ or $\{m_\psi : \psi \in \Psi\}, \pi_\Psi$

∴ nuisance parameters are dealt with in a consistent way

- furthermore: if $L(\theta, \psi) = L^*(\Psi(\theta), \psi)$ for all $\theta \in \Theta$, $\psi \in \Psi$ (loss is really expressed on $\Psi \times \Psi$) then

$$r(\delta\,|\,x) = \int_\Theta \int_\Psi L(\theta, \psi)\, \delta(x, d\psi)\, \pi(d\theta\,|\,x)$$

$$= \int_\Theta \int_\Psi L^*(\Psi(\theta), \psi)\, \delta(x, d\psi)\, \pi(d\theta\,|\,x)$$

$= \int_{\mathcal{A}} \int_{\mathcal{A}} L^*(a', a)\, \delta(x, da)\, \pi_{\mathcal{A}}(da' \mid x)$

- note - quadratic loss, L^1 loss, 0-1 loss satisfy this condition.

- when condition is satisfied the dec. problem $(\{f_\theta : \theta \in \Theta\}, \pi, L)$ will be seen to be equivalent to the decision problem $(\{m_a : a \in \mathcal{A}\}, \pi_{\mathcal{A}}, L^*)$

- principle of minimizing expected loss (maximizing expected utility) says to look for $\delta^* \in \mathcal{D}$ s.t.

$$r(\delta^*) = \inf_{\delta \in \mathcal{D}} r(\delta)$$

and δ^* is called a <u>Bayes rule</u> and $r(\delta^*)$ is the Bayes risk.

- typically a Bayes rule exists but it may not be unique.

Lemma 1 If $\delta(x,\cdot)$ minimizes $r(\delta|x)$ for each $x \in \mathcal{X}$ (a.e. M) then δ is a Bayes rule.

Proof: Consider $\delta_0 \in \mathcal{D}$. Then

$$r(\delta_0) = \int_{\mathcal{X}} r(\delta_0|x)\, M(dx) \geq \int_{\mathcal{X}} r(\delta|x)\, M(dx)$$

$$= r(\delta).$$

eg $\Psi \subseteq \mathbb{R}^k$ convex

- $L(\theta, \psi) = (\Psi(\theta)-\psi)' A (\Psi(\theta)-\psi)$

- we can restrict to nonrandomized d s.t. $E_\theta(d) \in \Psi$, and suppose $b(x) = E_{\pi_\Psi(\cdot|x)}(\psi)$ = posterior mean.

- $$r(d|x) = \int_{\Psi} (\psi - d(x))' A (\psi - d(x))\, \pi_\Psi(d\psi|x)$$

$$= \int_{\Psi} (\psi - b(x) + b(x) - d(x))' A (\ \ )\, \pi_\Psi(d\psi|x)$$

$$= \int_\Psi (\psi - b(x))' A (\ \ )\, \pi_\Psi(d\psi|x) + 2\left(\int_\Psi (\psi - b(x))\, \pi_\Psi(d\psi|x)\right)' A (b(x) - d(x)) + (b(x)-d(x))' A (b(x)-d(x))$$

$$= \int_\Psi (\psi - b(x))' A (\ )\, \pi_\Psi(d\psi|x) + (b(x)-d(x))' A (\ \ )$$

$$\geq \int_\Psi (\psi - b(x))' A (\ \ )\, \pi_\Psi(d\psi|x)$$ attaining

lower bound when $b(x) = d(x)$

$\therefore$ posterior mean is a Bayes rule wrt quadratic loss

eg hypothesis testing

- $R(\theta, \varphi) = c_0 E_\theta(\varphi) I_{H_0}(\theta) + c_a(1 - E_\theta(\varphi)) I_{H_a}(\theta)$

- $r(\varphi) = E_\pi(R(\theta, \varphi))$

$$= c_0 \int_{H_0} E_\theta(\varphi)\, \pi(d\theta) + c_a \int_{H_a} (1 - E_\theta(\varphi))\, \pi(d\theta)$$

$$= c_0 \pi(H_0) E_{m(\cdot|H_0)}(\varphi) + c_a \pi(H_a) E_{m(\cdot|H_a)}(1-\varphi)$$

= a priori prob of rejecting H_0 given that H_0 is true

- note - if $\pi(H_0) = 0$ then $r(\varphi) = c_a E_{m(\cdot|H_a)}(1-\varphi)$ which is mini'd by $\varphi(x) \equiv 1$, i.e. always reject H_0 is the Bayes rule, so not useful

- general reccomdation is to require $0 < \pi(H_0) < 1$

- also $r(\varphi|x) = \int_\Theta (c_0 \varphi(x) I_{H_0}(\theta) + c_a(1-\varphi(x)) I_{H_a}(\theta)) \pi(d\theta|x)$

$$= c_0 \varphi(x) \pi(H_0|x) + c_a (1-\varphi(x)) \pi(H_a|x)$$

- then Bayes rule is given by

$$\phi(x) = \begin{cases} 1 & c_0 \pi(H_0|x) \leq c_a \pi(H_a|x) \\ 0 & \text{otherwise} \end{cases}$$

actually can do anything with equality

- when $c_0 = c_a = 1$ this says reject H_0 when $\pi(H_0|x) \leq 1/2$.

- note - what to do when $\pi(H_0) = 0$? which can happen with continuous parameters.

  - general recommendation is to replace π by $p_0 \pi_{H_0} + (1-p_0)\pi_{H_a} = \pi$ where $p_0 \in (0,1)$ and π_{H_0}, π_{H_a} are priors on H_0, H_a respectively

  - then $\pi(H_0) = p_0 \in (0,1)$

- often Bayes factor is used as a measure of the evidence that H_0 is true, rather than $\pi(H_0|x)$

posterior odds H_0 true

prior odds H_0 true

- $BF(H_0|x) = \dfrac{\pi(H_0|x)}{\pi(H_a|x)} \Big/ \dfrac{\pi(H_0)}{\pi(H_a)}$

for given $\pi(H_0)$ a 1-1 increasing fn of $\pi(H_0|x)$

- now $\pi(H_0|x) = \int_{H_0} \frac{f_\theta(x)}{m(x)} \pi(d\theta)$

$= \frac{m(x|H_0)}{m(x)} \pi(H_0) \int_{H_0} \pi_{H_0}(d\theta)$

$$\pi(H_a \mid x) = \frac{m(x \mid H_a)}{m(x)}\, \pi(H_a)$$

$$\therefore\ BF(H_0 \mid x) = \frac{m(x \mid H_0)}{m(x \mid H_a)} = \text{ratio of integrated likelihoods}$$

Lemma ② $BF(H_0 \mid x) > 1$ iff $\pi(H_0 \mid x) > \pi(H_0)$

Proof: $1 < BF(H_0 \mid x)$ iff $\dfrac{\pi(H_0)}{\pi(H_a)} < \dfrac{\pi(H_0 \mid x)}{\pi(H_a \mid x)}$ iff

$\dfrac{\pi(H_0)}{1-\pi(H_0)} < \dfrac{\pi(H_0 \mid x)}{1-\pi(H_0 \mid x)}$ iff

$\dfrac{1}{\pi(H_0)} > \dfrac{1}{\pi(H_0 \mid x)}$ iff $\pi(H_0) < \pi(H_0 \mid x)$.

- but why do I need to modify a prior just to test H_0 and in addition specify $\pi(\cdot \mid H_0)$, $\pi(\cdot \mid H_a)$?

eg location normal

- $\bar{x} \sim N(\mu, \sigma_0^2/n)$, $\mu \sim N(\mu_0, \tau_0^2)$ and suppose we want to test $H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$

- the above formalism forces a change of the prior

- rather, better to remember that continuous probability is approximating a discrete reality

- so consider $H_0^\varepsilon = (\mu_0 - \varepsilon, \mu_0 + \varepsilon)$ so $\pi(H_0^\varepsilon) = \int_{\mu_0 - \varepsilon}^{\mu_0 + \varepsilon} \pi(\mu)\, d\mu > 0 \quad \forall \varepsilon > 0$ and

$$BF(H_0^\varepsilon \mid x) = \frac{\pi(H_0^\varepsilon \mid x)}{1 - \pi(H_0^\varepsilon \mid x)} \Big/ \frac{\pi(H_0^\varepsilon)}{1 - \pi(H_0^\varepsilon)}$$

$$= \frac{\pi(H_0^\varepsilon \mid x)}{\mathrm{Vol}(\mu_0 - \varepsilon, \mu_0 + \varepsilon)} \frac{1}{1 - \pi(H_0^\varepsilon \mid x)} \Big/ \frac{\pi(H_0^\varepsilon)}{\mathrm{Vol}(\mu_0 - \varepsilon, \mu_0 + \varepsilon)} \frac{1}{1 - \pi(H_0^\varepsilon)}$$

$$\to \pi(\mu_0 \mid x) \cdot 1 \Big/ \pi(\mu_0) \cdot 1$$

$$= \frac{\pi(\mu_0 \mid x)}{\pi(\mu_0)} = RB(\mu_0 \mid x) = \text{relative belief ratio at } \mu_0$$

- there is evidence in favor of (against) μ_0 iff $RB(\mu_0 \mid x) > (<1)$

- note - the approach generalizes and there is no need to modify a prior

  - BF with p_0, $\pi(\cdot \mid H_0)$, $\pi(\cdot \mid H_1)$ can exhibit "information inconsistency" in the continuous case

- poses another problem for the decision theory approach.

- axioms for Bayesian decision theory due to Savage (1954) The Foundations of Statistics

- there are 7 axioms (see Evans (2015) Measuring Statistical Evidence Using Relative Belief, section 2.3.5) that lead to the principle of minimizing expected loss

- are the axioms compelling?

- for me the axioms are designed to give the desired answer (the principle) and some of the axioms are not compelling (Axiom 6) and some are paradoxical (Axiom 2, the sure thing principle and the Allais paradox)