Decision Theory and Bayesian Inference

· It's a unifying framework for theory of statistic including estimation and testing.

(1) <u>Definition:</u>

· Suppose "a" is an action, $a \in A$. the action space.

$\Rightarrow$ The choice of action "a" depends on:

i) Observations of r.v.: Data X where $X \in S$. sample space.
ii) d, the decision function, i.e. $d: S \longrightarrow A$. $d(\vec{X}) = a$

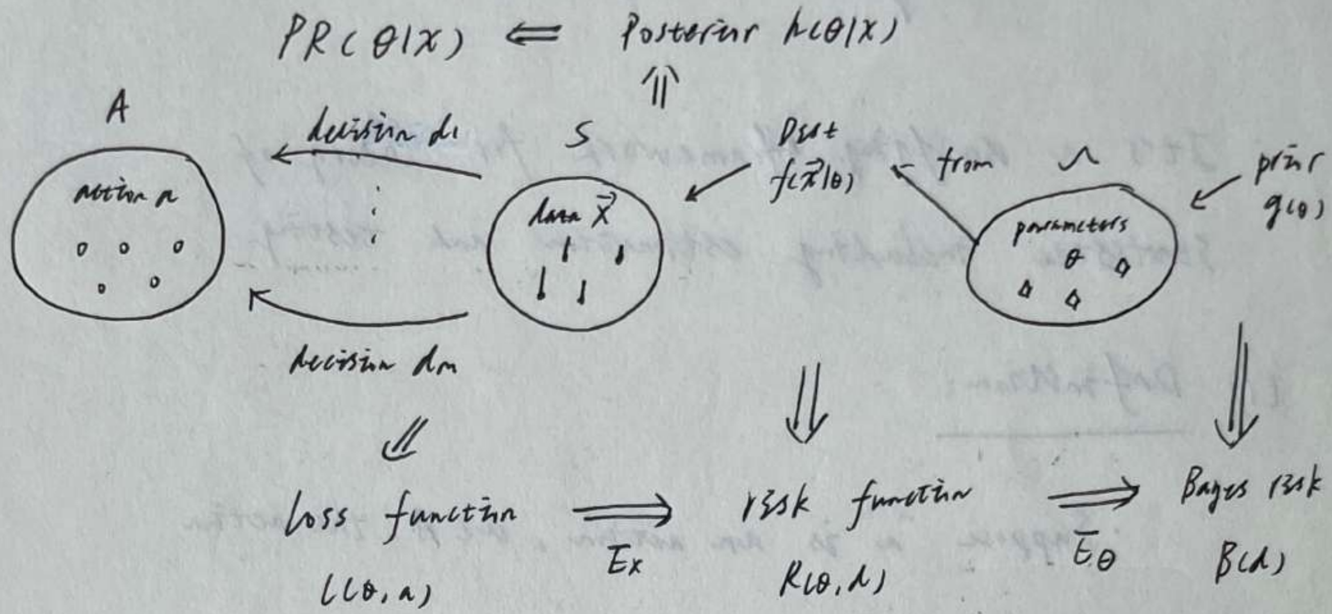
$\Rightarrow$ The prob dist of $\vec{X}$ depends on the parameter θ, called state of nature. $\theta \in \Lambda$. the space of parameters

$\Rightarrow$ Then we can define a loss function $l(\theta, a)$ on $\Lambda \times A$. Since $a = d(\vec{X})$ $\therefore l(\theta, a) = l(\theta, d(\vec{X}))$

The expected loss of $d(\vec{X})$ is the risk function:

$R(\theta, d) = E_X(l(\theta, d(\vec{X})))$ (It depends on θ)

⇒ Our aim is to minimize $R(\theta, d)$ by choosing a good decision $d(\vec{x})$!



e.g. estimate $\gamma(\theta)$, where $X_k \sim f(x|\theta)$, i.i.d. get data $\vec{X}$.

We choose $L(\theta, d(\vec{x})) = [\gamma(\theta) - d(\vec{x})]^2$: quadratic loss func

(2) Minimization:

· Difficulties of minimizing $R(\theta, d)$

i) $R(\theta, d)$ depend on unknown θ.

ii) For different θ_1, θ_2. It may happen: $R(\theta_1, d_1) > R(\theta_1, d_2)$, $R(\theta_2, d_1) < R(\theta_2, d_2)$. how to choose?

① Def of minimax rule:

. Consider the worst case: $\sup_{\theta \in \Lambda} R(\theta, d)$

d^* is the minimax rule, if $\sup_{\theta \in \Lambda} R(\theta, d^*) = \inf_{d} \sup_{\theta \in \Lambda} R(\theta, d)$, d^* may not exist!

Remark: It's very conservative to consider the worst case which isn't likely to occur.

② Bayesian Rule:

- We assume $\theta \in \Theta$ is random with a prior dist. Then the Bayesian risk of decision function d is $B(d) = E_\theta(R(\theta, d))$

Def: Bayesian rule is a decision func. d^{**} attain the $\min_d \{B(d)\}$

Remark: It can be interpreted as average of risk with weight form.

⇒ Posterior Analysis:

A method for finding Bayesian Rule:

- Suppose $g(\theta)$ prior dist of θ. $f_{X|\theta}(x)$ condition θ of X.

$\Rightarrow f_{X,\theta} = g(\theta) f_{X|\theta}$. ⇒ Sum/Integration: f_X.

We obtain: $h_{\theta|X} = f_{X,\theta}/f_X$. posterior dist of θ.

Def: Posterior risk: $E_{\theta|X}(L(\theta, d(x)))$. $= PR(\theta|X)$

Remark: The observed data $X = x$ updates the p-risk!

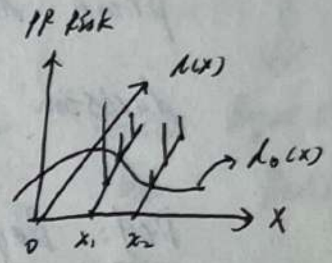
Thm. If $d_0(x)$ is a decision func. minimizes the posterior risk for each x. Then d_0 is Bayesian Rule.

Pf: $B(d) = E_\theta(R(\theta, d)) = E_\theta(E_x(l(\theta, d(x))|\theta))$

$$= \int \left[\int l(\theta, d(x)) f_{x|\theta}(x) dx \right] g_\theta(\theta) d\theta$$

$$= \int \left[\int l(\theta, d(x)) h_{\theta|x}(\theta) d\theta \right] f_x(x) dx$$

$$= \int E_{\theta|x}(l(\theta, d(x))) f_x(x) dx$$



Then $B(d)$ is minimized !

$\Rightarrow$ **Algorithm:**

1°) Calculate $h(\theta|x)$. for each x.

2°) Calculate $E_{\theta|x}(l(\theta, d(x)))$ for each x.

3°) Find $d(x_0) \in A$ minimizes every $PR(\theta|x_0)$. fixed x_0.

(3) Application of Decision Theory:

Estimation

① $\begin{cases} \text{Action space } A \longrightarrow \text{Parameter Space } \Omega. \\ \text{Decision Func. } d(X) \longrightarrow \text{estimator of } \theta. \\ l(\theta, d(X)) = [\theta - d(X)]^2 \text{ or } |\theta - d(X)| \end{cases}$

Thm. i) $E_{\theta|x}((\theta-\hat{\theta})^2|x) = Var_{\theta|x}(\theta|x) + [E_{\theta|x}(\theta|x) - \hat{\theta}]^2$

Then $\hat{\theta} = E_{\theta|x}(\theta|x)$ is the best predictor.

ii) $E_{\theta|x}(|\theta-\hat{\theta}||x)$ has the best predictor: median.

Pf: For ii) $E_{\theta|x}(|\theta-\hat{\theta}||x) = \int |\theta-\hat{\theta}| h_{\theta|x}(\theta) d\theta$

$$= \int_{\hat{\theta}}^{\infty} (\theta-\hat{\theta}) h_{\theta|x} d\theta + \int_{-\infty}^{\hat{\theta}} (\hat{\theta}-\theta) h_{\theta|x} d\theta \triangleq f(\hat{\theta})$$

$$\frac{\partial f}{\partial \hat{\theta}} = 0 \Rightarrow -\int_{\hat{\theta}}^{\infty} h_{\theta|x}(\theta) d\theta + \int_{-\infty}^{\hat{\theta}} h_{\theta|x} d\theta = 0$$

$\therefore \hat{\theta}$ = median of $h_{\theta|x}$!

② Def: i) d_1, d_2 two decision functions $\in A$.

says d_1 dominates d_2: $R(\theta, d_1) < R(\theta, d_2)$. $\forall \theta \in$ ⑪

ii) d. a decision func. d is admissable. if d is not strictly dominated by any other decision func.

Thm. i) If Λ is discrete. d^* is Bayesian rule w.r.t. prior pmf $g(\theta)$. where $g(\theta) > 0$. $\forall \theta$.

ii) If Λ is dense. d^* is Bayesian rule w.r.t prior pdf $g(\theta)$. $g(\theta) > 0$. $\forall \theta$. $R(\theta, d)$ is conti. of θ. $\forall d$.

Then d^* is admissible.

Remark: It claims the relation between Bayesian rule and admissible.

Pf: If $\exists d'$. s.t. $R(\theta, d^*) \geq R(\theta, d')$, $\forall \theta$.

$\exists \theta_0\, h$. $\theta \in (\theta_0 - h, \theta_0 + h) \Rightarrow R(\theta, d^*) > R(\theta, d') + \varepsilon$.

Then check: $B(d^*) - B(d') > 0$. contradict!

(4) Bayesian View for prob.:

(Personal Opinion)

① Bayesian prob is personal. It varies from person to person, embodying the beliefs of person. (Subjective)

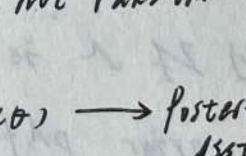
② Bayesian rule describes the prob. evolves with experience.

⇒ Difference between "B" and "F" Approach:

① Point Estimation:

$$\begin{cases} F: \theta \text{ is fixed. unknown. not random} \rightarrow \text{likelihood } L(\theta|X) \rightarrow \text{maximal: Find MLE.} \\ B: \theta \text{ has prior dist } g(\theta) \rightarrow \text{Posterior dist } h(\theta|x) \rightarrow \text{Centering: Find } E(\theta|X) \end{cases}$$

Remark: From $h(\theta|X) = \dfrac{f(x|\theta)g(\theta)}{\int f(x|\theta)g(\theta)d\theta}$ $\therefore h(\theta|x) \propto f(x|\theta)g(\theta)$

⇒ If θ is uniform:  almost const.

Then $g(\theta)$ has little effect. $h(\theta|x) \propto f(x|\theta) = L(\theta|x)$

③ Interval Estimation:

Frequentist: $p(\theta \in [\theta_L(\vec{X}), \theta_U(\vec{X})] \mid \theta) = 1-\alpha$. $\vec{X}$ is random, θ fixed.

If $\vec{x}$ is observed data, then $p(\theta \in [\theta_L(\vec{x}), \theta_U(\vec{x})]) = 0$ or 1.

Bayesian: $p(\theta \in [\theta_L(\vec{x}), \theta_U(\vec{x})] \mid \vec{X} = \vec{x}) = 1-\alpha$. θ is random.

$\vec{x}$ is the observed data. fixed.

③ Testing:

Frequentist: prob of Type I. II. error.

Bayesian: After observing data, compare posterior prob.

(5) Bayesian Inference

for Normal Dist:

- Suppose $X \mid \mu \sim N(\mu, \sigma^2)$. $\mu \sim N(\mu_0, \sigma_0^2)$. σ is known.

$\Rightarrow$ Posterior dist of μ is $N(\mu_1, \sigma_1^2)$. Where

$$\mu_1 = \frac{\xi_0}{\xi + \xi_0}\mu_0 + \frac{\xi}{\xi+\xi_0}x, \quad \sigma_1^2 = \frac{1}{\xi + \xi_0}$$

$$\xi_0 = \frac{1}{\sigma_0^2}, \quad \xi = \frac{1}{\sigma^2}.$$

Pf: Since the const. is for normalization.

We just care the ratio (about "μ"):

$$h(\mu \mid x) \propto f(x \mid \mu) g(\mu) \propto e^{-\frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$$

$$= e^{-\frac{1}{2}(\mu^2 \underbrace{[\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}]}_{=a} - 2\mu \underbrace{(\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})}_{=b} + \frac{x^2}{\sigma^2} + \frac{\mu_0^2}{\sigma_0^2})}$$

$$\propto e^{-\frac{1}{2a^{-1}}(\mu - \frac{b}{a})^2}$$

Cor. For n samples $\vec{X} = (x_1 \cdots x_n)$. i.i.d.

$$\mu | \vec{X} \sim N\left(\frac{\xi_0}{n\xi + \xi_0} \mu_0 + \frac{n\xi}{n\xi+\xi_0} \bar{x}, \ \frac{1}{n\xi+\xi_0} \right)$$

Remark: i) Note that $\frac{\xi_0}{n\xi+\xi_0} + \frac{n\xi}{n\xi+\xi_0} = 1$. We mix up the prior and the data to generate the posterior. (Weighted Average!)

ii) $\frac{1}{n\xi+\xi_0} < \sigma_0^2$. Which means the dist of $\mu|\vec{x}$ is more concentrated. It carries more information (informative)

iii) If n is large enough. Then the data will dominate the prior dist!

iv) For objectiveness. the prior needs to be vague, non informative!

(6) Bayesian Inference for Binomial Dist:

. $X|p \sim B(n.p)$. $p \sim Beta(a,b)$

$\Rightarrow p|x \sim Beta(a+x, n+b-x)$

Similarly, $\mu_{post} = \frac{a+b}{a+b+n} \cdot \frac{a}{a+b} + \frac{n}{a+b+n} \cdot \bar{x}$

If $n \to \infty$. $\mu_{post} \longrightarrow \bar{x}$!

Remark: Define:

$$\begin{cases} G: \text{family of prior dist given for } \Theta \\ H: \text{family of conditional dist } f(x|\theta) \end{cases}$$

$\Rightarrow$ G is called conjugate prior to H:

if the posterior of G under H also

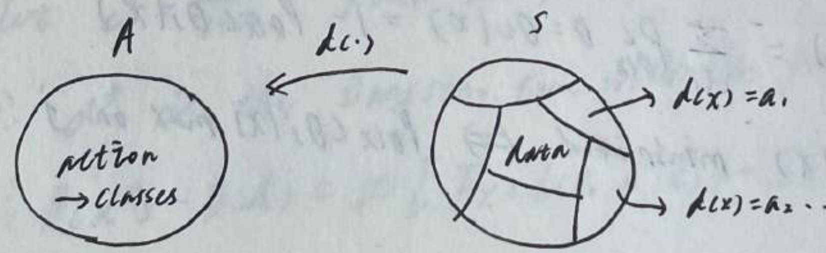
belongs to G:

e.g. G: Normal dist. H: Normal dist $\longrightarrow G|H = G$

H: Binomial dist G: Beta dist $\longrightarrow G|H = G$

(7) Application of Decision Theory:

① Classification:



Let $A = \{a_1 : \in \text{Class } \theta_1,\ a_2 : \in \text{Class } \theta_2, \cdots\ a_m : \in \text{Class } \theta_m\}$.

Then $d(\cdot)$ is function to classify datas in S.

$\Rightarrow$ For parameter θ, let $A = \{\theta_i\}_1^m$.

$\lambda_i = P(\Theta = \theta_i)$, s.t. $\sum_1^m \lambda_i = 1$. $f(x|\theta_i)$ is known. $1 \le i \le m$

If l_{ij} is the loss in classifying class i to class j.

Then, we can find Bayesian Rule:

$$p(\theta_i|x) = \frac{\pi_i f(x|\theta_i)}{\sum \pi_k f(x|\theta_k)} = P(\Theta = \theta_i \mid X = x). \text{ let } d(x) = a_j$$

We obtain $PR(j|x) = \sum l_{ij} p(\theta_i|x)$. Posterior risk

$\Rightarrow$ Choose j to minimize $PR(i|x)$. for $\forall\ 1 \le i \le m$

Then $d: x \mapsto a_j$ is a Bayesian Rule.

e.g. (0-1 loss)

$$l_{ij} = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases}$$ (Because the classification is "qualitative", not "quantative")

$$R(i,d) = E_X(l(\theta_i, d(X))) = \sum_j l_{ij} P_{\theta_i}(d(X) = j)$$

$$= \sum_{j \neq i} P_{\theta_i}(d(X) = j) = 1 - P_{\theta_i}(d(X) = i)$$

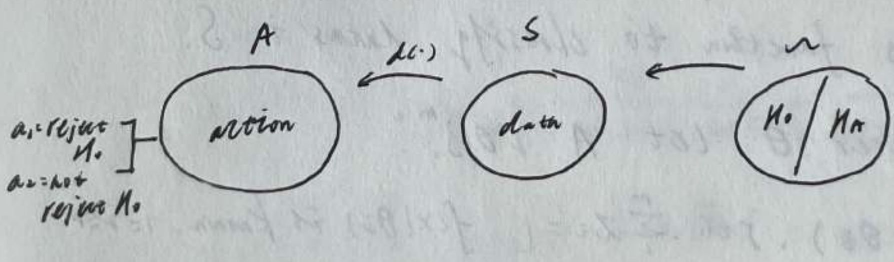
(From frequentist approach, minmax it)

$\Rightarrow$ For Bayesian approach:

$$PR(j|X) = \sum_{i \neq j} P_{\theta|x}(\theta = \theta_i|x) = 1 - P_{\theta|x}(\theta_j|x)$$

$\therefore$ $PR(j|x)$ minimized $\Leftrightarrow$ $P_{\theta|x}(\theta_j|x)$ max on j!

② Hypothesis Testing:



Then both type I and type II errors are misclassification errors.

$\Rightarrow$ <u>Bayesian approach:</u>

$p(H_0) = \pi$. $p(H_A) = 1-\pi$. then we obtain:

$$\frac{p(H_0|x)}{p(H_A|x)} = \frac{p(H_0)}{p(H_A)} \frac{p(x|H_0)}{p(x|H_A)} > 1 \text{. (LHS} = \frac{PR(H_0|x)}{PR(H_A|x)}\text{, under 0-1 loss)}$$

$$\Leftrightarrow \frac{p(x|H_0)}{p(x|H_A)} > \frac{1-\pi}{\pi} \triangleq c \text{. Which means:}$$

assign $x \xrightarrow{d} H_0$. when $\frac{p(x|H_0)}{p(x|H_A)} > c$, get

$d(x)$ is Bayesian Rule under 0-1 loss.

<u>Remark:</u> For protecting H_0. we can assign large "π" to $p(H_0)$.

<u>Alternative proof of N-P Lemma:</u>

d^*: test accept $H_0 \Leftrightarrow \frac{f_0(x)}{f_1(x)} > c$. with level α^*.

Then d^* is the most power test of level $\alpha \le \alpha^*$

<u>Pf:</u> let $c^1 = \frac{\pi}{1-\pi}$. $p(H_0) = \pi$. $p(H_A) = 1-\pi$.

$\because$ d^* is the Bayesian rule with this prior. with 0-1 loss

$$\therefore \beta(d^*) - \beta(d) = \pi \left[E_X(L(H_0, d^*(x)) - E_X(L(H_0, d(x))\right] + (1-\pi)\left[E_X(L(H_A, d^*(x))) - E_X(L(H_A, d(x)))\right]$$

$$= \pi(\alpha^* - \alpha) + (1-\pi)\left[E_X(L(H_A, d^*(x)) - E_X(L(H_A, d(x)))\right] \le 0$$

$$\therefore E_X(L(H_A, d^*(x))) \le E_X(L(H_A, d(x)))$$

i.e. $\beta_{d^*} \ge \beta_d$.