STAT 4010 – Bayesian Learning

Tutorial 4

Spring 2022

Cheuk Hin (Andy) CHENG (Email | Homepage)

Di SU (Email | Homepage)

1 Decision Theory

Definition 1. *(Decision Theory) Define the following settings,*

$$\mathscr{X} = \textit{Observation space}, \quad \Theta = \textit{Parameter space}, \quad \mathscr{D} = \textit{Decision space}.$$

Statistical inference is performed by using the observation $x \in \mathscr{X}$ to construct a decision $d(x) = d \in \mathscr{D}$ to understand the parameter $\theta \in \Theta$. The decision is made under the following systematic procedure.

1. *(Everyone) Fix a user-defined function, **utility function** for evaluating the goodness of a decision, defined as*

$$U : \Theta \times \mathscr{D} \to (-\infty, 0].$$

2. *Frequentist procedure*

   (a) *Compute the **loss function** to evaluate the loss of the decision, defined as*

$$L(\theta, d) = -U(\theta, d).$$

   (b) *Compute the **(frequentist) risk** to average out x and the effect of sampling error, defined as*

$$R(\theta, d) = \mathsf{E}_{x|\theta}[L(\theta, d) \mid \theta] = \int_{\mathscr{X}} L(\theta, d(x)) f(x \mid \theta) \mathrm{d}x.$$

   The best decision is chosen by minimizing the risk.

3. *Bayesian procedure (Keep in mind θ is no longer a constant):*

   (a) *Compute the **posterior loss function** to evaluate the loss of the decision, defined as*

$$L(\pi, d) = \mathsf{E}[L(\theta, d) \mid x] = \int_{\Theta} L(\theta, d) \pi(\theta \mid x) \mathrm{d}\theta.$$

   (b) *Compute the **Bayesian/average risk** to average out x as well as θ, defined as*

$$R(\pi, d) = \mathsf{E}_{x,\theta}[L(\theta, d)] = \int_{\Theta} \int_{\mathscr{X}} L(\theta, d(x)) f(x \mid \theta) \pi(\theta) \mathrm{d}x \mathrm{d}\theta.$$

   The best decision is chosen by minimizing the bayesian risk.

Remark 1.1. Recall by Bayes' formula, we have

$$f(x, \theta) = f(x \mid \theta) f(\theta) = f(\theta \mid x) f(x).$$

And by tower property, we have

$$\mathsf{E}(L(\theta, d)) = \mathsf{E}_\theta \mathsf{E}_{x|\theta}(L(\theta, d)|\theta) = \mathsf{E}_x \mathsf{E}_{\theta|x}(L(\theta, d)|x).$$

The dependency on the data and the parameter is summarized below.

	Depend on x	Not depend on x
Depend on θ	$L(\theta, d)$	$R(\theta, d)$
Not depend on θ	$L(\pi, d)$	$R(\pi, d)$

Example 1.1. (Example of Frequentist and Bayesian procedure) Consider in a coin flipping experiment, we want to know the the probability of getting a head or θ. We have two observations from the following model,

$$x_{1:2} \overset{\text{IID}}{\sim} \text{Bern}(\theta).$$

We agree to pick the best estimator according to the absolute difference. There are two candidates of estimator to choose,

$$d_1(x) = \frac{1}{2}, \quad \text{and} \quad d_2(x) = \bar{X} = n^{-1} \sum_{i=1}^{n} x_i.$$

1. Describe the observation, parameter and decision space.
2. Describe the utility function.
3. Describe the frequentist procedure when θ is 0.3 and 0.5.
4. Describe the bayesian procedure, choose a conjugate prior for θ.

Solution:

1. We have
$$\mathscr{X} = \{0, 1\}^2 = \{00, 10, 01, 11\}, \quad \Theta = [0, 1], \quad \mathscr{D} = [0, 1].$$
2. The utility function is $U(d, \theta) = -|d - \theta|$, the closer d to θ, the larger the utility, which means the better d is.
3. The loss function is $L(\theta, d) = -U(d, \theta) = |d - \theta|$. The risk is
$$R(\theta, d) = \mathsf{E}[L(\theta, d) \mid \theta] = \sum_{x_{1:2} \in \mathscr{X}} |d - \theta| f(x_{1:2} \mid \theta) = \sum_{x_{1:2} \in \mathscr{X}} |d - \theta| \theta^{S_2} (1 - \theta)^{n - S_2}, \quad (1.1)$$
where $n = 2$ and $S_n = \sum_{i=1}^{n} x_i$. For d_1, we have
$$R(\theta = 0.5, d_1) = 0, \quad \text{and} \quad R(\theta = 0.3, d_1) = 0.2.$$

For d_2, we have

$$R(\theta = 0.5, d_2) = 0.25, \quad \text{and} \quad R(\theta = 0.3, d_2) = 0.294.$$

For both values of θ, we choose d_1.

```
##compute frequentist risk
get_joint_density <- function(theta=0.5){ ##c(f(00),f(10),f(01),f(11))
  sx = c(0,1,1,2)
  den = (theta^sx)*(1-theta)^(2-sx)
  return(den)
}
get_risk = function(d,theta=0.5){ ##d = c(d(00),d(10),d(01),d(11))
  den = get_joint_density(theta)
  risk = sum(abs(theta-d)*den)
  return(risk)
}
##compute risk of d_1
d1_risk_05 = get_risk(c(0.5,0.5,0.5,0.5),theta = 0.5)
d1_risk_03 = get_risk(c(0.5,0.5,0.5,0.5),theta = 0.3)
d1_risk_05;d1_risk_03
##compute risk of d_2
d2_risk_05 = get_risk(c(0,0.5,0.5,1),theta = 0.5)
d2_risk_03 = get_risk(c(0,0.5,0.5,1),theta = 0.3)
d2_risk_05;d2_risk_03
```

4. The conjugate prior for θ is Beta(α, β). Choose $\alpha = 1/2$ and $\beta = 1/2$. Then, the posterior is Beta$(\alpha + n\bar{x}, \beta + n(1-\bar{x}))$. We can simulate the posterior loss and bayesian risk.

x	00	10	01	11
$L(\pi, d_1)$	0.353	0.207	0.210	0.354
$L(\pi, d_2)$	0.161	0.207	0.210	0.170

Also, $R(\pi, d_1) = 0.317$ and $R(\pi, d_2) = 0.182$. We choose d_2 accordingly.

```
##compute posterior loss
get_posterior_loss <- function(x,seed = 4010,a=0.5,b=0.5,d1 = 0.5,nRep =
    2^10){
  bar_x = mean(x) ##d2 is bar_x
  set.seed(seed)
  theta = rbeta(nRep,a+n*bar_x,b+n*(1-bar_x))
  p_loss = array(NA,2)
  p_loss[1] = mean(abs(theta-d1)) ##posterior loss for d_1
  p_loss[2] = mean(abs(theta-bar_x)) ##posterior loss for d_2
  return(p_loss)
}

get_posterior_loss(x = c(0,0),seed=4010) ##posterior loss when x = 00
get_posterior_loss(x = c(1,0),seed=4011) ##posterior loss when x = 10
get_posterior_loss(x = c(0,1),seed=4012) ##posterior loss when x = 01
get_posterior_loss(x = c(1,1),seed=4013) ##posterior loss when x = 11

##compute bayesian risk by MC
```

```r
a=0.5
b=0.5
d1 = 0.5
nRep = 2^10

set.seed(4010)
theta = rbeta(nRep,a,b)
x = array(NA,c(nRep,2))
for (i in 1:nRep) {
  x[i,] = rbinom(n=2,size=1,theta[i])
}
d2 = apply(x,1,mean)
mean(abs(d1-theta)) ##bayesian risk for d1
mean(abs(d2-theta)) ##bayesian risk for d2
```

2 Bayes estimator

Definition 2. *(Bayes Estimator)*

1. *The Bayes estimator under the prior $\theta \sim \pi(\theta)$ and the loss $L(\theta, \widehat{\theta})$ is defined as*
$$\widehat{\theta}_\pi = \arg\min_{\widehat{\theta}} R(\pi, \widehat{\theta}).$$
2. *The value $R\left(\pi, \widehat{\theta}_\pi\right)$ is called the Bayes risk.*

Theorem 2.1. *(A simpler way of finding Bayes estimators). Assume there is an estimator $\widehat{\theta}_0$ such that $R\left(\theta, \widehat{\theta}_0\right) < \infty$. Then, for each $x \in X$, the Bayes estimator is*
$$\widehat{\theta}_\pi(x) = \arg\min_{\widehat{\theta}} L(\pi, \widehat{\theta})$$

Theorem 2.2. *(Bayes estimators under special losses). Suppose the conditions in Theorem 2.1 hold.*

1. *Quadratic loss. Let $L(\theta, \widehat{\theta}) = (\widehat{\theta} - \theta)^2$. (See Example 3.13 for a generalization.) The Bayes estimator is*
$$\widehat{\theta}_\pi = \mathrm{E}(\theta \mid x).$$
2. *Absolute loss. Let $L(\theta, \widehat{\theta}) = |\widehat{\theta} - \theta|$. (See Example 3.14 for a generalization.) The Bayes estimator is*
$$\widehat{\theta}_\pi = Q_{1/2}(\theta \mid x).$$
3. *Zero-one loss. Let $L(\theta, \widehat{\theta}) = \mathbb{1}(|\widehat{\theta} - \theta| > \varepsilon)$ for some $\varepsilon > 0$. The Bayes estimator is*
$$\widehat{\theta}_\pi = \arg\max_{\widehat{\theta}} \mathrm{P}(\theta \in [\widehat{\theta} - \varepsilon, \widehat{\theta} + \varepsilon] \mid x).$$

Example 2.1. (Q4 Midterm 2019 - Bayes Estimator (20%)). Let Θ be the parameter space. Suppose that the decision space is $\mathscr{D} = \Theta$. Consider the generalized squared-error loss function $L : \Theta \times \mathscr{D} \to \mathbb{R}$ defined as

$$L(\theta, \widehat{\theta}) = \{g(\widehat{\theta}) - g(\theta)\}^2 \tag{2.1}$$

where $g : \Theta \to \mathbb{R}$ is a continuous and strictly increasing function whose inverse function g^{-1} exists. (Note: g may NOT be differentiable.) Let $\theta \sim \pi(\theta)$ such that the Bayesian risk exists.

1. Find the Bayes estimator $\widehat{\theta}_\pi$ of θ.
2. Let $W \sim \mathrm{N}(\mu, \sigma^2)$. Prove that
$$\mathrm{E}\left(e^W\right) = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}$$
3. For parts (3) and (4), consider $g(t) = e^t, [x_1, \ldots, x_n \mid \theta] \overset{\text{ID}}{\sim} \mathrm{N}(\theta, \sigma^2)$ and $\theta \sim \mathrm{N}(\theta_0, \tau_0^2)$. Prove that the Bayes estimator of θ under the loss (2.1) is $\widehat{\theta}_\pi = \theta_n + \frac{\tau_n^2}{2}$, where
$$\theta_n = \frac{\tau_0^2 \bar{x}_n + \sigma^2 \theta_0 / n}{\tau_0^2 + \sigma^2/n}, \quad \tau_n^2 = \frac{\tau_0^2 \sigma^2 / n}{\tau_0^2 + \sigma^2/n}, \quad \bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i. \tag{2.2}$$
4. Comment the following statement: "Under the squared-error loss, the Bayes estimator of θ is $\widehat{\theta}'_\pi = \theta_n$, which is different from the $\widehat{\theta}_\pi$ in (2.2). In practice, we should choose the estimator which has a smaller mean squared error."

Solution:

1. There are two methods.
   - (Method 1) The posterior loss is
$$\begin{aligned} L(\pi, \widehat{\theta}) &= \mathrm{E}\{L(\theta, \widehat{\theta}) \mid x\} \\ &= \mathrm{E}\left[\{g(\widehat{\theta}) - \mathrm{E}\{g(\theta) \mid x\} + \mathrm{E}\{g(\theta) \mid x\} - g(\theta)\}^2 \mid x\right] \\ &= \left[g(\widehat{\theta}) - \mathrm{E}\{g(\theta) \mid x\}\right]^2 + \mathrm{E}\left([\mathrm{E}\{g(\theta) \mid x\} - g(\theta)]^2 \mid x\right) \\ &\quad + 2\left[g(\widehat{\theta}) - \mathrm{E}\{g(\theta) \mid x\}\right] \underbrace{\mathrm{E}\left[\mathrm{E}\{g(\theta) \mid x\} - g(\theta) \mid x\right]}_{=0} \\ &= \underbrace{[g(\widehat{\theta}) - \mathrm{E}\{g(\theta) \mid x\}]^2}_{\star} + \underbrace{\mathrm{E}\left([\mathrm{E}\{g(\theta) \mid x\} - g(\theta)]^2 \mid x\right)}_{\text{does not depends on } \hat{\theta}} \end{aligned}$$

     Hence, $L(\pi, \widehat{\theta})$ is minimized (with respect to $\widehat{\theta}$) iff $\star$ is minimized. Clearly, the minimum is achieved iff $g(\widehat{\theta}) - \mathrm{E}\{g(\theta) \mid x\} = 0$. Solving for $\widehat{\theta}$, we know that the Bayes estimator is
$$\widehat{\theta}_\pi = g^{-1}(\mathrm{E}\{g(\theta) \mid x\})$$

- (Method 2) We first find the minimizer with respect to $\widehat{\phi} = g(\widehat{\theta})$. Note that the posterior loss and its derivatives are

$$L(\pi, \widehat{\theta}) = \widehat{\phi}^2 - 2\widehat{\phi}\mathrm{E}\{g(\theta) \mid x\} + \mathrm{E}\left\{g^2(\theta) \mid x\right\}$$
$$\frac{\mathrm{d}}{\mathrm{d}\widehat{\phi}} L(\pi, \widehat{\theta}) = 2\widehat{\phi} - 2\mathrm{E}\{g(\theta) \mid x\}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\widehat{\phi}^2} L(\pi, \widehat{\theta}) = 2 > 0$$

  Clearly, $\widehat{\phi} = \mathrm{E}\{g(\theta) \mid x\}$ minimizes $L(\pi, \widehat{\theta})$. Since g is a strictly increasing function, we know that $\widehat{\theta} = g^{-1}(\mathrm{E}\{g(\theta) \mid x\})$ minimizes $L(\pi, \widehat{\theta})$. Hence, the Bayes estimator is

$$\widehat{\theta}_\pi = g^{-1}(\mathrm{E}\{g(\theta) \mid x\}).$$

2. Represent $W = \mu + \sigma Z$, where $Z \sim \mathrm{N}(0, 1)$. So, by matching the Normal density, we have

$$\begin{aligned}
\mathrm{E}\left(e^W\right) &= \mathrm{E}\left(e^{\mu+\sigma Z}\right) = e^\mu \mathrm{E}\left(e^{\sigma Z}\right) \\
&= e^\mu \int_{-\infty}^{\infty} e^{\sigma z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, \mathrm{d}z \\
&= e^\mu e^{\sigma^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(z-\sigma)^2\right\} \mathrm{d}z \\
&= \exp\left\{\mu + \frac{\sigma^2}{2}\right\}.
\end{aligned}$$

3. Recall that $[\theta \mid x_{1:n}] \sim \mathrm{N}\left(\theta_n, \tau_n^2\right)$, where θ_n and τ_n^2 are given by (2.2). By the result of parts (1) and (2), we have

$$\widehat{\theta}_\pi = \log \mathrm{E}\left(e^\theta \mid x_{1:n}\right) = \log \exp\left\{\theta_n + \frac{\tau_n^2}{2}\right\} = \theta_n + \frac{\tau_n^2}{2}$$

4. It is true that these two Bayes estimators $\left(\widehat{\theta}_\pi \text{ and } \widehat{\theta}'_\pi\right)$ are different because they are evaluated under different loss functions. However, there is no absolute rule for comparing these two estimators.

   - On one side, using the mean-squared error (MSE) is possible. Recall that MSE is the frequentist risk under the squared-error loss. In this case, we compare the frequentist's properties (i.e., MSE) of two Bayesian estimators. The MSE is useful for understanding, e.g., consistency and the L^2-convergence rate. So, it is not wrong to do so.
   - But, on the other side, comparing the estimators by MSE may not be relevant or fair from the Bayesian point of view. For Bayesian, once we fix our loss function (e.g., (4.2) or the sqaured-error loss), the goal is to minimize the Bayesian risk but not other frequentist risk (e.g., the MSE). Hence, "evaluating an estimator that optimizes criteria A by another criteria B " may not be fair. It is possible in practice because the loss function $L(\theta, \widehat{\theta})$ may be selected due to some problem-specific purposes, e.g., handling asymmetric loss.

Example 2.2. (Q4 Midterm 2021- Bayes Estimator). Let $y_i \in \mathbb{R}$ be the response value at time i, and $x_i \in \{0, 1\}$ be a binary covariate. (For example, y_i and x_i may denote the rate of return a stock and the market condition, respectively.) Suppose that $y_0 = 0$ and $x_{1:n}$ are non-random and always known. Consider a regime-switching AR (1) model:

$$[y_i \mid \theta_1, \theta_0, y_{i-1}] \sim \mathrm{N}\left(\theta_1^{x_i}\theta_0^{1-x_i} y_{i-1}, 1\right), \quad i = 1, \ldots, n,$$
$$\theta_1, \theta_0 \overset{\text{IID}}{\sim} \mathrm{Unif}(-1, 1).$$

1. Derive the posterior distribution of $\theta = (\theta_0, \theta_1)^{\mathrm{T}}$.
2. Consider the loss function:

$$L(\theta, \widehat{\theta}) = \sum_{i=1}^{n} \left(\theta_1^{x_i}\theta_0^{1-x_i} - \widehat{\theta}_1^{x_i}\widehat{\theta}_0^{1-x_i}\right)^2$$

where $\widehat{\theta} = \left(\widehat{\theta}_0, \widehat{\theta}_1\right)^{\mathrm{T}}$.

   (a) Interpret the loss function.

   (b) Derive and interpret the Bayes estimator.

   Use no more than about 50 words for the interpretations in (a) and (b).

Solution:

1. Let $n_1 = \sum_{i=1}^{n} x_i$ be the number of i such that $x_i = 1$. Then $n_0 = n - n_1$ is the number of i such that $x_i = 0$. Denote all the indexes $i \in \{1, 2, \ldots, n\}$ satisfying $x_i = 1$ to be $j_1, j_2, \ldots, j_{n_1}$. Similarly, denote all the indexes $i \in \{1, 2, \ldots, n\}$ satisfying $x_i = 0$ to be $k_1, k_2, \ldots, k_{n_0}$. The sampling distribution can be written as

$$\begin{aligned} f\left(y_{1:n} \mid \theta_1, \theta_0\right) &= f\left(y_{1:n} \mid \theta_1, \theta_0, y_0\right) \\ &= \prod_{i=1}^{n} f\left(y_i \mid \theta_1, \theta_0, y_{i-1}\right) \\ &= (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \theta_1^{x_i}\theta_0^{1-x_i} y_{i-1}\right)^2\right\} \\ &= (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{s=1}^{n_1}\left(y_{j_s} - \theta_1 y_{j_s-1}\right)^2\right\} \exp\left\{-\frac{1}{2}\sum_{t=1}^{n_0}\left(y_{k_t} - \theta_0 y_{k_t-1}\right)^2\right\} \end{aligned}$$

   The posterior is

$$\begin{aligned} f\left(\theta_1, \theta_0 \mid y_{1:n}\right) &\propto f\left(y_{1:n} \mid \theta_1, \theta_0\right) f\left(\theta_1, \theta_0\right) \\ &\propto \exp\left\{-\frac{1}{2}\sum_{s=1}^{n_1}\left(y_{j_s} - \theta_1 y_{j_s-1}\right)^2\right\} \exp\left\{-\frac{1}{2}\sum_{t=1}^{n_0}\left(y_{k_t} - \theta_0 y_{k_t-1}\right)^2\right\} \mathbb{1}\left(-1 < \theta_1, \theta_0 < 1\right) \\ &\propto \exp\left[-\frac{1}{2}\left\{\left(\sum_{s=1}^{n_1} y_{j_s-1}^2\right)\theta_1^2 - 2\left(\sum_{s=1}^{n_1} y_{j_s} y_{j_s-1}\right)\theta_1\right\}\right] \\ &\quad \times \exp\left[-\frac{1}{2}\left\{\left(\sum_{t=1}^{n_0} y_{k_t-1}^2\right)\theta_0^2 - 2\left(\sum_{t=1}^{n_0} y_{k_t} y_{k_t-1}\right)\theta_0\right\}\right] \mathbb{1}\left(-1 < \theta_1, \theta_0 < 1\right). \end{aligned}$$

Denote

$$\mu_0 = \frac{\sum_{k=1}^{n_0} y_{k_m} y_{k_m-1}}{\sum_{k=1}^{n_0} y_{k_m-1}^2}, \quad \sigma_0 = \frac{1}{\sqrt{\sum_{k=1}^{n_0} y_{k_m-1}^2}}, \quad \mu_1 = \frac{\sum_{l=1}^{n_1} y_{j_l} y_{j_l-1}}{\sum_{l=1}^{n_1} y_{j_l-1}^2} \quad \text{and} \quad \sigma_1 = \frac{1}{\sqrt{\sum_{l=1}^{n_1} y_{j_l-1}^2}}.$$

Then

$$\begin{aligned} f(\theta_1, \theta_0 \mid y_{1:n}) &\propto \texttt{dnorm}(\theta_1, \mu_1, \sigma_1)\, \texttt{dnorm}(\theta_0, \mu_0, \sigma_0)\, \mathbb{1}(-1 < \theta_1, \theta_0 < 1) \\ &= \texttt{dTN}(\theta_1, \mu_1, \sigma_1, -1, 1)\, \texttt{dTN}(\theta_0, \mu_0, \sigma_0, -1, 1). \end{aligned}$$

2. Observe that the loss function can be written as

$$\begin{aligned} L(\theta, \hat{\theta}) &= \sum_{i=1}^{n} \left(\theta_1^{x_i} \theta_0^{1-x_i} - \hat{\theta}_1^{x_i} \hat{\theta}_0^{1-x_i} \right)^2 \\ &= n_1 \left(\theta_1 - \hat{\theta}_1 \right)^2 + n_0 \left(\theta_0 - \hat{\theta}_0 \right)^2 \end{aligned}$$

Since n_1 and n_0 are given, we can find the minimizer of the $\mathscr{L}^2$ loss of θ_0 and of θ_1 respectively. By Theorem 3.2 and Example 3.13 from the lecture notes, the Bayes estimator of θ is

$$\hat{\theta} = \left(\hat{\theta}_1^{\pi}, \hat{\theta}_0^{\pi} \right)^T = \left(\mathrm{E}(\theta_1 \mid y_{1:n}), \mathrm{E}(\theta_0 \mid y_{1:n}) \right)^T.$$

Specifically,

$$\hat{\theta}_j^{\pi} = \mathrm{E}(\theta_j \mid y_{1:n}) = \mu_j - \sigma_j \frac{\texttt{dnorm}(u_j) - \texttt{dnorm}(l_j)}{\texttt{pnorm}(u_j) - \texttt{pnorm}(l_j)},$$

where $u_j = (1 - \mu_j)/\sigma_j$ and $l_j = (-1 - \mu_j)/\sigma_j$ for $j = 0, 1$. For the interpretations, any reasonable answers are accepted.