MAST90125: Bayesian Statistical learning

Lecture 3: Bayes and Minimax Estimation

Guoqi Qian

Outline

§1 Loss, risk, minimax risk and Bayes risk

§2 Bayes estimator

§3 Example and Remark

Loss and risk

- Recall the MSE of an estimator T of $\tau(\theta)$, $\text{MSE}(T) = E[T - \tau(\theta)]^2$, measures the *expected squared error loss* associated with T. One can consider other types of loss function.
- **Loss function** $L(t;\theta)$ is a real-valued function of an estimator T and parameter θ such that $L(t;\theta) \geq 0$ for every value of T, and $L(t;\theta) = 0$ when $t = \tau(\theta)$.
  Note that by default the estimator $T = t(\mathbf{X}_n)$ is a function of the data $\mathbf{X} = (X_1, \cdots, X_n)^T$ and is for estimating $\tau(\theta)$.
- **Risk function** $R_T(\theta)$ is the expectation of the loss function w.r.t. the data, i.e.

$$R_T(\theta) = E[L(T;\theta)] = \int L(t(\mathbf{x}_n);\theta) f(\mathbf{x}_n|\theta) d\mathbf{x}_n$$

where $f(\mathbf{x}_n|\theta)$ is the joint pdf (or pmf) of $\mathbf{X}_n$.

Minimization of loss and risk?

- Adopting a specific risk function $R_T(\theta)$ as the criterion, naturally the best estimator of $\tau(\theta)$ would be a T^* which minimizes $R_T(\theta)$ for all possible values of θ.
- Unfortunately, such a T^* usually does not exist except in very few cases. In other words, it is mostly impossible to find an estimator which is better than any other estimator in terms of $R_T(\theta)$.
- **Admissibility:** An estimator T_1 **dominates** another estimator T_2 iff $R_{T_1}(\theta) \leq R_{T_2}(\theta)$ for all $\theta \in \Theta$, and $R_{T_1}(\theta) < R_{T_2}(\theta)$ for at least some $\theta \in \Theta$. T is an **admissible estimator** iff no other estimators dominate it.
- It is not worth to consider inadmissible estimators.

Mini-max or min-expectation?

If we really want to find a "best" estimator w.r.t. $R_T(\theta)$, there are two possible ways.

1. Find a *minimax* estimator from the class of admissible estimators.
   - An estimator T_1 is a **minimax estimator** of $\tau(\theta)$

     if $\max_\theta R_{T_1}(\theta) \leq \max_\theta R_T(\theta)$ for every estimator T of $\tau(\theta)$.

     Namely, $T_1 = \arg\min_T\{\max_\theta R_T(\theta)\}$.
   - The minimax approach is conservative in general. Not much is known about its performance, but will not be pursued in this subject.

Mini-max or min-expectation?

2. Use Bayes approach: The key is to regard the parameter θ as a random variable having a pdf $p(\theta)$, where $p(\theta)$ is called the **prior distribution** or **prior density**.
   - Then **Bayes risk** is defined to be

     $$A_T = E_\theta[R_T(\theta)] = \int_\Theta R_T(\theta)p(\theta)d\theta,$$

     which is just an expected risk w.r.t. the prior distribution.
   - **Bayes estimator** is defined to be the estimator T^* which minimizes the Bayes risk:

     $$E_\theta[R_{T^*}(\theta)] \leq E_\theta[R_T(\theta)] \quad \text{for every estimator } T \text{ of } \tau(\theta).$$

     Namely, $T^* = \arg\min_T A_T = \arg\min_T E_\theta[R_T(\theta)]$.

How to find the Bayes estimator? (1)

$$
\begin{aligned}
E_\theta[R_T(\theta)] &= \int_\Theta R_T(\theta)p(\theta)d\theta = \int_\Theta \left[\int_{\mathbf{x}_n} L(t(\mathbf{x}_n);\theta)f(\mathbf{x}_n|\theta)d\mathbf{x}_n\right] p(\theta)d\theta \\
&= \int_\Theta \int_{\mathbf{x}_n} L(t(\mathbf{x}_n);\theta)f(\mathbf{x}_n|\theta)p(\theta)d\mathbf{x}_n d\theta \\
&= \int_{\mathbf{x}_n} \left[\int_\Theta L(t(\mathbf{x}_n);\theta)\frac{f(\mathbf{x}_n|\theta)p(\theta)}{\int_\Theta f(\mathbf{x}_n|\theta)p(\theta)d\theta}d\theta\right]\left[\int_\Theta f(\mathbf{x}_n|\theta)p(\theta)d\theta\right]d\mathbf{x}_n \\
&= \int_{\mathbf{x}_n} \left[\int_\Theta L(t(\mathbf{x}_n);\theta)p(\theta|\mathbf{x}_n)d\theta\right] f(\mathbf{x}_n)d\mathbf{x}_n \\
&= \int_{\mathbf{x}_n} E_\theta[L(T;\theta)|\mathbf{x}_n]f(\mathbf{x}_n)d\mathbf{x}_n = E_{\mathbf{x}_n}\left(E_\theta[L(T;\theta)|\mathbf{x}_n]\right)
\end{aligned}
$$

Thus if an estimator $\tilde{T}$ minimizes $E_\theta[L(T;\theta)|\mathbf{x}_n]$ for any given $\mathbf{x}_n$, it must also minimize $E_\theta[R_T(\theta)]$.

How to find the Bayes estimator? (2)

Therefore, finding the Bayes estimator is equivalent to finding the estimator that minimizes $E_\theta[L(T;\theta)|\mathbf{x}_n]$ for any given $\mathbf{x}_n$.

Theorem (Bayes estimator under squared loss)

Suppose we choose to use the squared loss function $L(T;\theta) = [T - \tau(\theta)]^2$, *then*

$$T^* = E_\theta[\tau(\theta)|\mathbf{x}_n] = \int_\Theta \tau(\theta) p(\theta|\mathbf{x}_n) d\theta$$

is the Bayes estimator of $\tau(\theta)$ *that minimizes* $E_\theta([T - \tau(\theta)]^2|\mathbf{x}_n)$.

Proof: $E_\theta([T - \tau(\theta)]^2|\mathbf{x}_n) = T^2 - 2TE_\theta[\tau(\theta)|\mathbf{x}_n] + E_\theta[\tau(\theta)^2|\mathbf{x}_n]$ is a convex quadratic function of T, it follows that $\arg\min_T E_\theta([T - \tau(\theta)]^2|\mathbf{x}_n) = E_\theta[\tau(\theta)|\mathbf{x}_n]$. □

How to find the Bayes estimator? (3)

Remarks

1. Unless stated otherwise, we will use the squared loss function $L(T;\theta) = [T - \tau(\theta)]^2$.
2. $f(\mathbf{x}_n) = \int_\Theta f(\mathbf{x}_n|\theta)p(\theta)d\theta$ is the marginal pdf of $\mathbf{X}_n$.
3. The conditional pdf $p(\theta|\mathbf{x}_n) = \frac{f(\mathbf{x}_n|\theta)p(\theta)}{\int_\Theta f(\mathbf{x}_n|\theta)p(\theta)d\theta}$ is called the **posterior pdf** of θ.
4. The Bayes estimator T^* is interpreted as the posterior mean of $\tau(\theta)$ (provided that the squared loss is used).

Example 2.1 (1)

Example 2.1 Consider a random sample $X_n = (X_1, \cdots, X_n)$ from a Bernoulli distribution with pdf $f(x|\theta) = \theta^x(1-\theta)^{1-x};\ \ x = 0, 1.$ Let the prior pdf of θ be Uniform$(0, 1)$, i.e. $p(\theta) = I(0 < \theta < 1)$.

1. Find the Bayes estimator of θ.
2. Find the risk of the Bayes estimator of θ.
3. Find the Bayes risk of the Bayes estimator of θ.
4. Find the Bayes estimator of θ^2.
5. Formulate the risk of the Bayes estimator of θ^2.

Example 2.1 (2)

- First the posterior pdf of θ is

$$
\begin{aligned}
p(\theta|\mathbf{x}_n) &= \frac{f(\mathbf{x}_n|\theta)p(\theta)}{\int_\Theta f(\mathbf{x}_n|\theta)p(\theta)d\theta} = \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} I(0<\theta<1)}{\int \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} I(0<\theta<1)d\theta} \\
&= \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} d\theta} \\
&= \frac{\Gamma(n+2)}{\Gamma(\sum_{i=1}^n x_i+1)\Gamma(n-\sum_{i=1}^n+1)} \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}
\end{aligned}
$$

which is a beta$\left(a = \sum_{i=1}^n x_i + 1, \quad b = n - \sum_{i=1}^n x_i + 1\right)$ pdf.

Example 2.1 (3)

1. Then the Bayes estimator of θ is

$$T_1^* = E(\theta|\mathbf{x}_n) = \frac{a}{a+b} = \frac{\sum_{i=1}^n x_i + 1}{n+2}.$$

   Note the MLE of θ is $\hat{\theta} = \frac{\sum_{i=1}^n x_i}{n}$.

2. Risk $R_{T_1^*}(\theta) = E[T_1^* - \theta]^2 = \frac{(n-4)\theta(1-\theta)+1}{(n+2)^2}$.
   This result is obtained by using the fact that
   $$\sum_{i=1}^n X_i \stackrel{d}{=} \text{binomial}(n,\theta) \text{ conditional on } \theta.$$

3. Bayes risk
   $A_{T_1^*} = \int R_{T_1^*}(\theta)p(\theta)d\theta = \int_0^1 \frac{(n-4)\theta(1-\theta)+1}{(n+2)^2} \times 1 d\theta = \frac{1}{6(n+2)}$.

Example 2.1 (4)

4. The Bayes estimator of θ^2 is

$$T_2^* = E(\theta^2|\mathbf{x}_n) = \frac{a(a+1)}{(a+b+1)(a+b)} = \frac{(\sum_{i=1}^{n} x_i + 1)(\sum_{i=1}^{n} x_i + 2)}{(n+3)(n+2)}.$$

Note the MLE of θ^2 is $\hat{\theta}^2 = \left(\frac{\sum_{i=1}^n x_i}{n}\right)^2$.

5. The risk of T_2^* is

$$R_{T_2^*}(\theta) = E[T_2^* - \theta^2]^2 = E\left[\frac{(\sum_{i=1}^{n} x_i + 1)(\sum_{i=1}^{n} x_i + 2)}{(n+3)(n+2)} - \theta^2\right]^2$$

which can still be calculated using the fact $(\sum_{i=1}^n X_i)|\theta \overset{d}{=} b(n,\theta)$.

Remarks

1. The idea involved in Bayes estimation is very appealing. Without any information or with only prior information about θ, we would estimate $\tau(\theta)$ by its prior mean. Once the data are observed, new information about θ is available, we then would estimate $\tau(\theta)$ by its posterior mean.
2. The posterior mean may be analytically intractable if the posterior pdf is mathematically complicated. This difficulty may be overcome by using a Monte Carlo technique to approximate the posterior mean. Monte Carlo statistical approaches are more and more popular which will be addressed in next chapter.
3. Bayes confidence intervals and testing are studied elsewhere.

Example 2.2 (1)

Example 2.2 Consider a random sample $X_n = (X_1, \cdots, X_n)$ from a Poisson distribution with pdf

$$f(x|\theta) = \frac{\theta^x}{x!} e^{-\theta}; \;\; x = 0, 1, \cdots .$$

Let the prior pdf of θ be Gamma(β, κ) with mean $\kappa\beta$ and variance $\kappa\beta^2$, i.e.

$$p(\theta) = \frac{1}{\beta^\kappa \Gamma(\kappa)} \theta^{\kappa-1} e^{-\theta/\beta}; \; \theta > 0; \beta > 0, \kappa > 0.$$

Find the Bayes estimator of θ and the associated risk.

Example 2.2 (2)

The posterior distribution or posterior pdf of θ is

$$
\begin{aligned}
p(\theta|\mathbf{x}_n) &= \frac{f(\mathbf{x}_n|\theta)p(\theta)}{\int_\Theta f(\mathbf{x}_n|\theta)p(\theta)d\theta} \\
&= \left[\frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}\theta^{\kappa-1}e^{-\theta/\beta}}{\prod_{i=1}^n (x_i!)\beta^\kappa\Gamma(\kappa)}\right] \Big/ \left[\int \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}\theta^{\kappa-1}e^{-\theta/\beta}}{\prod_{i=1}^n (x_i!)\beta^\kappa\Gamma(\kappa)} d\theta\right] \\
&= \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}\theta^{\kappa-1}e^{-\theta/\beta}}{\int e^{-n\theta}\theta^{\sum_{i=1}^n x_i}\theta^{\kappa-1}e^{-\theta/\beta}d\theta} = \frac{\theta^{\sum_{i=1}^n x_i+\kappa-1}e^{-\theta(n+1/\beta)}}{\int_0^\infty \theta^{\sum_{i=1}^n x_i+\kappa-1}e^{-\theta(n+1/\beta)}d\theta} \\
&= \frac{\theta^{\sum_{i=1}^n x_i+\kappa-1}e^{-\theta(n+1/\beta)}}{(n+1/\beta)^{-(\sum_{i=1}^n x_i+\kappa)}\Gamma(\sum_{i=1}^n x_i+\kappa)}
\end{aligned}
$$

which is a Gamma$((n+1/\beta)^{-1}, \sum_{i=1}^n x_i + \kappa)$ pdf.

Example 2.2 (3)

The Bayes estimator of θ is therefore

$$T = E(\theta|\mathbf{x}_n) = (n + 1/\beta)^{-1}(\sum_{i=1}^{n} x_i + \kappa)$$

which is very close to the MLE, $\hat{\theta} = \bar{x}_n$, if β is large and κ is small. The risk in this case is

$$\begin{aligned} R_T(\theta) &= E[T-\theta]^2 = \mathrm{Var}(T) + [E(T)-\theta]^2 \\ &= \frac{n\mathrm{Var}(X)}{(n+1/\beta)^2} + \left[\frac{n\theta+\kappa}{n+1/\beta} - \theta\right]^2 \\ &= \frac{n\theta + [\kappa - \theta/\beta]^2}{(n+1/\beta)^2} \end{aligned}$$

where we have used the fact that, given θ, $\sum_{i=1}^{n} X_i \stackrel{d}{=} \mathrm{Poisson}(n\theta)$.

Example 2.3

Example 2.3 Consider a random sample $Y_n = (Y_1, \cdots, Y_n)$ with $Y_i \stackrel{d}{=} \text{Poisson}(e^{\beta x_i})$ and x_i being given, $i = 1, \cdots, n$. Let the prior pdf of β be $N(0, 1)$.

Then the posterior pdf of β is

$$p(\beta|\mathbf{y}_n, \mathbf{x}_n) = \frac{f(\mathbf{y}_n|\beta, \mathbf{x}_n)p(\beta)}{\int_B f(\mathbf{y}_n|\beta, \mathbf{x}_n)p(\beta)d\beta}$$

$$= \frac{(\prod_{i=1}^n y_i!)^{-1} e^{-\sum_{i=1}^n e^{\beta x_i}} e^{\beta \sum_{i=1}^n x_i y_i} (\sqrt{2\pi})^{-1} e^{-\beta^2/2}}{\int_{-\infty}^{\infty} (\prod_{i=1}^n y_i!)^{-1} e^{-\sum_{i=1}^n e^{\beta x_i}} e^{\beta \sum_{i=1}^n x_i y_i} (\sqrt{2\pi})^{-1} e^{-\beta^2/2} d\beta}.$$

The Bayes estimator of β is

$T = E(\beta|\mathbf{y}_n, \mathbf{x}_n) = \int_{-\infty}^{\infty} \beta p(\beta|\mathbf{y}_n, \mathbf{x}_n) d\beta,$

which does not have a closed form and will have to be calculated using a Monte Carlo method.