14.381 Recitation 4 : Short Additional Note on Bayesian Statistics

Joonhwan Lee

October 6, 2012

Bayesian Decision Theory

We have a state space Θ where each state $\theta \in \Theta$. We make action $a \in A$, which has different consequences depending on the state θ. Our consequence is denoted by 'loss function'

$$l(a, \theta) : A \times \Theta \to \mathbb{R}$$

This can be also interpreted as 'utility function' with opposite sign. We observe a signal $X \in \mathcal{X}_\theta$ associated with the state that we know the likelihood $f(X|\theta)$. We have a prior on the state, which is a distribution over Θ, denoted by $p(\theta)$. Our objective is to find a decision function $d(X) : X \to A$ that minimizes the following object.

$$\int_\Theta [\int_{\mathcal{X}_\theta} l(d(X), \theta) f(X|\theta) dX] p(\theta) d\theta = \int_\Theta E_\theta[l(d(X), \theta)] dp(\theta) = \int_\Theta R(\theta, d(X)) dp(\theta)$$

The object is often referred as 'expected utility' in economic theory. We call $R(\theta, d(X)) \equiv E_\theta[l(d(X), \theta)]$ the risk function. Note that the object can be also written as

$$\int_{\mathcal{X}} \int_\Theta l(d(X), \theta) \frac{f(X|\theta)p(\theta)}{m(X)} d\theta m(X) dX = \int_{\mathcal{X}} \int_\Theta l(d(X), \theta) f(\theta|X) d\theta m(X) dX$$

where $m(X)$ is the marginal density of X. Then the decision making problem reduces to

$$\min_d \int_\Theta l(d, \theta) f(\theta|X) d\theta = \min_d E[l(d, \theta)|X]$$

for each X. That is, the Bayesian decision is to minimize the posterior loss.

Let's take this process for the example of estimation. Our decision d is to estimate θ and some natural loss functions would be

$$\begin{aligned} l_1(d, \theta) &= (\theta - d)^2 \\ l_2(d, \theta) &= |\theta - d| \\ l_3(d, \theta) &= \tau|\theta - d|_+ + (1 - \tau)|\theta - d|_- \\ l_4(d, \theta) &= L \cdot 1_{\{|\theta - d| > K\}} \end{aligned}$$

It is not hard to find that minimizing posterior loss leads to posterior mean, median, τ-quantile and mode respectively for the above loss functions. Note that the risk function associated with $l_1(d, \theta)$ is MSE, which is the key concept of performance in frequentist view. Frequentist way of 'optimal' estimation is mostly described as

$$\min_{d(X)} \max_\theta R(\theta, d(X))$$

i.e. to minimize the maximal risk on Θ. Their view of the problem is as follows. Since we don't have any idea which state we are on, we should minimize the worst risk. The above problem is often approached by using Bayesian tools. For example, an estimator $\hat{\theta}(X)$ is minimax if it has constant risk and it is a Bayes estimator for some prior.

Now let's look at the problem of testing. Let us partition our parameter space by Θ_1 and Θ_0. We have two possible actions, namely reject or accept the null hypothesis of $\theta \in \Theta_0$. A natural loss function would be

$$\begin{aligned} l(1,\theta) &= k \quad \text{if } \theta \in \Theta_0 \\ l(0,\theta) &= 0 \quad \text{if } \theta \in \Theta_0 \\ l(1,\theta) &= 0 \quad \text{if } \theta \in \Theta_1 \\ l(0,\theta) &= 1 \quad \text{if } \theta \in \Theta_1 \end{aligned}$$

where 0 and 1 denote accpet and reject respectively. Minimizing posterior loss, we have our decision rule as follows.

$$\text{Reject if } \frac{P(\theta \in \Theta_1|X)}{P(\theta \in \Theta_0|X)} \geq k$$

and we get posterior odd-ratio test as the Bayesian optimal test. If we have simple hypotheses, then the test becomes likelihood ratio test which is known as most powerful test by the Neyman-Pearson lemma.

Some Principles behind Bayesian Statistics

Sufficiency Principle

Two observations x and y factorizing through the same value of a sufficient statistic T, that is, such that $T(x) = T(y)$, must lead to the same inference on θ.

Likelihood Principle

The information brought by and observation x about θ is entirely contained in the likelihjood function $l(\theta|x)$. Moreover, if $x_1$and x_2 are two observations depending on the same parameter θ, such that there exists a constant c satisfying

$$l_1(\theta|x_1) = cl_2(\theta|x_2)$$

for every θ, they then bring the same information about θ and must lead to identical inferences.

It is known that Bayesian inference satisfies both principles. Both principles may look natural. However, the latter principle is controversial. Frequentist testing or confidence interval often violates the likelihood principle. Most apparent cases are when we have stopping rule for experiments. For example, consider a simple model of coin tossing. We want to know p, the probability of getting head of the coin. We consider two different experiments.

1. We toss the coin 12 times. Let X be the number of times that head showed up.

2. We toss the coin until we obtain 3 heads. Let Y be the number of tails during the experiment.

Suppose we observed $X = 3$ for the first experiment and $Y = 9$ for the second experiment. Then both likelihoods are proportional to $p^3(1-p)^9$, and the principle dictates that inference on p should be the same for both experiments. However, note that the MLEs for both experiments are

$$\begin{aligned} \hat{p}_1 &= \frac{X}{12} \\ \hat{p}_2 &= \frac{3}{Y+3} \end{aligned}$$

and obviously $Var(\hat{p}_1) \neq Var(\hat{p}_2)$. That is, in frequentist's view, they have different sampling distributions and therefore leads to different inferences on p. Bayesian procedure, on the other hand, yields the same posterior of p for both experiments.