Lecture - Bayesian Theory and estimation techniques

Dr. Arabin Kumar Dey

Assistant Professor
Department of Mathematics
Indian Institute of Technology Guwahati

July 31, 2013

Outline

1. Baysian Theory and other estimation techniques

Outline

1. Baysian Theory and other estimation techniques

The word “Bayesian” traces its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace was among the first thinkers to consider the laws of chance and randomness in a quantitative, scientific way. Both Bayes and Laplace were aware of a relation that is now known as Bayes Theorem:

Another motivation for the Bayesian approach is decision theory. In statistical decision theory, we formalize good and bad results with a loss function. In statistical decision theory, we formalize good and bad results with a loss function.

- A loss function $L(\theta, \delta(x))$ is a function of $\theta \in \Theta$ a parameter or index.
- $\delta(x)$ is a decision based on the data $x \in X$.
- For example, $\delta(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$ might be the sample mean, and θ might be the true parameter.
- The loss function determines the penalty for deciding $\delta(x)$ if θ is the true parameter.

To give some intuition, in the discrete case, we might use a 0-1 loss, which assigns

$$L(\theta, \delta(x)) = \begin{cases} 0 & \text{if } \delta(x) = \theta, \\ 1 & \text{if } \delta(x) \neq \theta \end{cases}$$

or in the continuous case, we might use the squared error loss

$$L(\theta, \delta(x)) = (\theta - \delta(x))^2$$

What are loss functions

Notice that in general, $\delta(x)$ does not necessarily have to be an estimate of θ.

- Loss functions provide a very good foundation for statistical decision theory.
- They are simply a function of the state of nature (θ) and a decision function $(\delta(\cdot))$.
- In order to compare procedures we need to calculate which procedure is best even though we cannot observe the true nature of the parameter space θ and data X.
- This is the main challenge of decision theory and the break between frequentists and Bayesians.

Frequentist Risk

Definition

The frequentist risk is

$$R(\theta, \delta(x)) = E_\theta(L(\theta, \delta(x))) = \int_X L(\theta, \delta) f(x|\theta) dx$$

where θ is held fixed and the expectation is taken over X.

Thus, the risk measures the long-term average loss resulting from using δ.

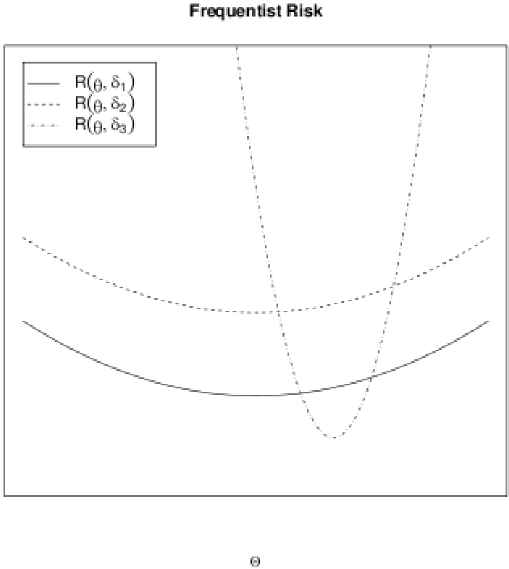

FIGURE 2.1: frequentist Risk

Often one decision does not dominate the other everywhere as is the case with decisions δ_1, δ_2. The challenge is in saying whether, for example, δ_1 or δ_3 is better. In other words, how should we aggregate over Θ?
Frequentists have a few answers for deciding which is better:

1 **Admissibility. A decision which is inadmissible is one that is dominated everywhere.** For example, in Figure 1, δ_2 dominates δ_1 for all values of θ. It would be easy to compare decisions if all but one were inadmissible. But usually the risk functions overlap, so this criterion fails.

Restricted classes of procedure.

- We say that an estimator θ is an unbiased estimator of θ if $E_\theta[\hat{\theta}] = \theta$ for all θ. If we restrict our attention to only unbiased estimators then we can often reduce the situation to only risk curves like δ_1 and δ_2 in Figure above, eliminating overlapping curves like δ_3.
- The existence of an optimal unbiased procedure is a nice frequentist theory, but many good procedures are biased-for example Bayesian procedures are typically biased.

- More surprisingly, some unbiased procedures are actually inadmissible. For example, James and Stein showed that the sample mean is an inadmissible estimate of the mean of a multivariate Gaussian in three or more dimensions.
- There are also some problems were no unbiased estimator exists-for example, when p is a binomial proportion and we wish to estimate $1/p$ (see Example 2.1.2 on page 83 of Lehmann and Casella).
- If we restrict our class of procedures to those which are equivariant, we also get nice properties. We do not go into detail here, but these are procedures with the same group theoretic properties as the data.

3. **Minimax.** In this approach we get around the problem by just looking at $\sup_\theta R(\theta, \delta(x))$, where $R(\theta, \delta(x)) = E_\theta[L(\theta, \delta(x))]$. For example in Figure 2, δ_2 would be chosen over δ_1 because its maximum worst-case risk (the grey dotted line) is lower.

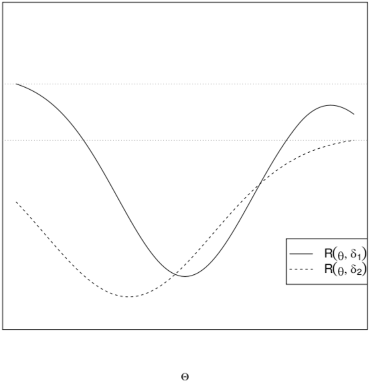

FIGURE 2.2: Minimax frequentist Risk

- A Bayesian answer is to introduce a weighting function $p(\theta)$ to tell which part of Θ is important and integrate with respect to $p(\theta)$.
- In some sense the frequentist approach is the opposite of the Bayesian approach. However, sometimes an equivalent Bayesian procedure can be derived using a certain prior.
- Before moving on, again note that $R(\theta, \delta(x)) = E_{\theta}[L(\theta, \delta(x))]$ is an expectation on X, assuming fixed θ.
- A Bayesian would only look at x, the data you observed, not all possible X.

Motivation for Bayes

The Bayesian approach can also be motivated by a set of principles. Some books and classes start with a long list of axioms and principles conceived in the 1950s and 1960s. However, we will focus on three main principles.

Conditionality Priciple : The idea here for a Bayesian is that we condition on the data x.

- Suppose we have an experiment concerning inference about θ that is chosen from a collection of possible experiments independently.
- Then any experiment not chosen is irrelevant to the inference (this is the opposite of what we do in frequentist inference).

For example, two different labs estimate the potency of drugs. Both have some error or noise in their measurements which can accurately estimate from past tests. Now we introduce a new drug. Then we test its potency at a randomly chosen lab. Suppose the sample sizes matter dramatically.

- Suppose the sample size of the first experiment (lab 1) is 1 and the sample size of the second experiment (lab 2) is 100.
- What happens if we're doing a frequentist experiment in terms of the variance ? Since this is a randomized experiment, we need to take into account all of the data. In essence, the variance will do some sort of averaging to take into account the sample sizes of each.
- However, taking a Bayesian approach, we just care about the data that we see. Thus, the variance calculation will only come from the actual data at the randomly chosen lab.

Thus, the question that we ask is should we use the noise level from the lab where it is tested or average over both ? Intuitively, we use the noise level from the lab where it was tested, but in some frequentist approaches, it is not always so straightforward.

For example, two different labs estimate the potency of drugs. Both have some error or noise in their measurements which can accurately estimate from past tests. Now we introduce a new drug. Then we test its potency at a randomly chosen lab. Suppose the sample sizes matter dramatically.

- Suppose the sample size of the first experiment (lab 1) is 1 and the sample size of the second experiment (lab 2) is 100.
- What happens if we're doing a frequentist experiment in terms of the variance ? Since this is a randomized experiment, we need to take into account all of the data. In essence, the variance will do some sort of averaging to take into account the sample sizes of each.
- However, taking a Bayesian approach, we just care about the data that we see. Thus, the variance calculation will only come from the actual data at the randomly chosen lab.

Thus, the question that we ask is should we use the noise level from the lab where it is tested or average over both ? Intuitively, we use the noise level from the lab where it was tested, but in some frequentist approaches, it is not always so straightforward.

For example, two different labs estimate the potency of drugs. Both have some error or noise in their measurements which can accurately estimate from past tests. Now we introduce a new drug. Then we test its potency at a randomly chosen lab. Suppose the sample sizes matter dramatically.

- Suppose the sample size of the first experiment (lab 1) is 1 and the sample size of the second experiment (lab 2) is 100.
- What happens if we're doing a frequentist experiment in terms of the variance ? Since this is a randomized experiment, we need to take into account all of the data. In essence, the variance will do some sort of averaging to take into account the sample sizes of each.
- However, taking a Bayesian approach, we just care about the data that we see. Thus, the variance calculation will only come from the actual data at the randomly chosen lab.

Thus, the question that we ask is should we use the noise level from the lab where it is tested or average over both ? Intuitively, we use the noise level from the lab where it was tested, but in some frequentist approaches, it is not always so straightforward.

For example, two different labs estimate the potency of drugs. Both have some error or noise in their measurements which can accurately estimate from past tests. Now we introduce a new drug. Then we test its potency at a randomly chosen lab. Suppose the sample sizes matter dramatically.

- Suppose the sample size of the first experiment (lab 1) is 1 and the sample size of the second experiment (lab 2) is 100.
- What happens if we're doing a frequentist experiment in terms of the variance ? Since this is a randomized experiment, we need to take into account all of the data. In essence, the variance will do some sort of averaging to take into account the sample sizes of each.
- However, taking a Bayesian approach, we just care about the data that we see. Thus, the variance calculation will only come from the actual data at the randomly chosen lab.

Thus, the question that we ask is should we use the noise level from the lab where it is tested or average over both ? Intuitively, we use the noise level from the lab where it was tested, but in some frequentist approaches, it is not always so straightforward.

Likelihood Principle: The relevant information in any inference about θ after observing x is contained entirely in the likelihood function.

- Remember the likelihood function $p(x|\theta)$ for fixed x is viewed as a function of θ, not x.
- For example in Bayesian approaches, $p(\theta|x) \propto p(x|\theta)p(\theta)$, so clearly inference about θ is based on the likelihood.
- Another approach based on the likelihood principle is Fisher's maximum likelihood estimation.
- In case this principle seems too indisputable, here is an example using hypothesis testing in coin tossing that shows how some reasonable procedures may not follow it.

Example

Let θ be the probability of a particular coin landing on heads and let

$$H_0 : \theta = \frac{1}{2}, H_1 : \theta > \frac{1}{2}$$

Suppose we observe the following sequence of flips :

$$H, H, T, H, T, H, H, H, H, H, H, T$$

(9 heads, 3 tails)
Then the likelihood is simply

$$p(x|\theta) \propto \theta^9(1-\theta)^3.$$

- Many non-Bayesian analyses would pick an experimental design that is reflected in $p(x|\theta)$, for example binomial (toss a coin until you get 3 tails).
- However the two lead to different probabilities over the sample space X. This results in different assumed tail probabilities and p-values.

Definition

Recall that for a data set $x = (x_1, \cdots, x_n)$, a sufficient statistic $T(x)$ is a function such that the likelihood $p(x|\theta) = p(x_1, \cdots, x_n|\theta)$ depends on $x_1, \cdots, x_n$ only through $T(x)$. Then the likelihood $p(x|\theta)$ may be written as $p(x|\theta) = g(\theta, T(x))h(x)$ for some functions g and h.

Definition

Exponential Families A family $\{P_\theta\}$ of distributions is said to form an s-dimensional exponential family if the distributions of P_θ have densities of the form

$$p_\theta(x) = exp[\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)]h(x)$$

Sufficiency Principle: The sufficiency principle states that if two different observations x and y have the same sufficient statistic $T(x) = T(y)$, then inference based on x and y should be the same. The sufficiency principle is the least controversial principle.

Theorem

The posterior distribution, $p(\theta|y)$ only depends on the data through the sufficient statistic, $T(y)$.

Proof. By the factorization theorem, if $T(y)$ is sufficient,

$$f(y|\theta) = g(\theta, T(y))h(y)$$

Then we know the posterior can be written

$$\begin{aligned} p(\theta|y) &= \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \\ &= \frac{g(\theta, T(y))h(y)\pi(\theta)}{\int g(\theta, T(y))h(y)\pi(\theta)d\theta} \\ &\propto g(\theta, T(y))p(\theta) \end{aligned}$$

which only depends on y through $T(y)$.

Example

Example : Sufficiency Let $y = \sum_i y_i$. Consider

$$y_1, \cdots, y_n | \theta \sim Bin(1, \theta)$$

Then

$$p(y) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Let $p(\theta)$ represent a general prior. Then

$$p(\theta|y) \propto \theta^y (1-\theta)^{n-y} p(\theta)$$

which only depends on the data through the sufficient statistic y.

Bayesian Decision Theory

Earlier we discussed the frequentist approach to statistical decision theory. Now we discuss the Bayesian approach in which we condition on x and integrate over Θ (remember it was the other way around in the frequentist approach). The posterior risk is defined as

$$\rho(\pi, \delta(x)) = \int_{\Theta} L(\theta, \delta(x))\pi(\theta|x)dx$$

The Bayes action $\delta^*(x)$ for any fixed x is the decision $\delta(x)$ that minimizes the posterior risk. If the problem at hand is to estimate some unknown parameter θ, then we typically call this the Bayes estimator instead.

Theorem

Under squared error loss, the decision $\delta(x)$ that minimizes the posterior risk is the posterior mean.

Proof : Suppose that

$$L(\theta, \delta(x)) = (\theta - \delta(x))^2.$$

Now note that

$$\begin{aligned} \rho(\pi, \delta(x)) &= \int (\theta - \delta(x))^2 \pi(\theta|x) d\theta \\ &= \int \theta^2 \pi(\theta|x) d\theta + [\delta(x)]^2 \int \pi(\theta|x) d\theta - 2\delta(x) \int \theta\pi(\theta|x) d \end{aligned}$$

Then

$$\frac{\delta\rho(\pi, \delta(x))}{\delta\delta(x)} = 2\delta(x) - 2\int \theta\pi(\theta|x) d\theta = 0$$

$$\leftrightarrow \delta(x) = E(\theta|x)$$

and $\delta^2[\rho(\pi, \delta(x))]/\delta[\delta(x)]^2$, so $\delta(x) = E(\theta|x)$ is the minimizer.

Frequentist Interpretation: Risk

In frequentist usage, the parameter θ is fixed. Letting $R(\theta, \delta(x))$ denote the frequentist risk, recall that $R(\theta, \delta(x)) = E_\theta[L(\theta, \delta(x))]$. This expectation is taken over the data X, with the parameter θ held fixed.

Example : (Squared error loss). Let the loss function be squared error. In this case, the risk is

$$\begin{aligned} R(\theta, \delta(x)) &= E_\theta(\theta - \delta(x))^2 \\ &= E_\theta(\theta - E_\theta(\delta(x)) + E_\theta(\delta(x)) - \delta(x))^2 \\ &= \{\theta - E_\theta(\delta(x))\}^2 + E_\theta(\{\delta(x) - E_\theta(\delta(x))\}^2) \\ &= Bias^2 + Variance \end{aligned}$$

This result can be used to motivate frequentist ideas, e.g. minimum variance unbiased estimators (MVUEs).

Bayesian Parametric Models

For now we will consider parametric models, which means that the parameter θ is a fixed- dimensional vector of numbers. Let $x \in X$ be the observed data and $\theta \in \Theta$ be the parameter. Note that X may be called the sample space, while θ may be called the parameter space. Now we define some notation that we will reuse throughout the course:

$p(x\|\theta)$	likelihood
$\pi(\theta)$	prior
$p(x) = \int p(x\|\theta)\pi(\theta)d\theta$	marginal likelihood
$p(\theta\|x) = \frac{p(x\|\theta)\pi(\theta)}{p(x)}$	posterior probability
$p(x_{new}\|x) = \int p(x_{new}\|\theta)\pi(\theta\|x)d\theta$	predictive probability

Note that for the posterior distribution,

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \propto p(x|\theta)\pi(\theta)$$

and oftentimes it's best to not calculate the normalizing constant p(x) because you can recognize the form of $p(x|\theta)\pi(\theta)$ as a probability distribution you know. So don't normalize until the end! Two questions we still need to address are

- How do we choose priors ?
- How do we compute the aforementioned quantities, such as posterior distributions ?