Lecture - Bayesian Theory and Estimation Techniques

Dr. Arabin Kumar Dey

Assistant Professor
Department of Mathematics
Indian Institute of Technology Guwahati

August 18, 2014

Outline

1. **Bayesian Theory and other estimation techniques**

Outline

1. **Bayesian 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:

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

- $p(\theta|x)$ posterior
  $p(x|\theta)$ likelihood
  $p(\theta)$ prior

Basic Philosophical Difference

- Bayesians treat an unknown parameter θ as random and use probability to quantify their uncertainty about it.
- In contrast, frequentists treat θ as unknown but fixed, and they therefore believe that probability statements about θ are useless.

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.

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

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

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

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

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 ?

Credible Interval

- The Bayesian analogue of a $(1-\alpha)$ frequentist confidence interval is a $(1-\alpha)$ credible set, defined as the set C of values of θ whose posterior probability content is at least $1-\alpha$:

$$P(\theta \in C|y) = \int_C \pi(\theta|y)d\theta = (1-\alpha)$$

- Interpretation : The probability that θ lies in C given the observed data y is at least $(1-\alpha)$.

- For scalar θ, $C = (\theta_I, \theta_S)$ is equi-tailed if

$$P(\theta < \theta_I | y) = P(\theta > \theta_S | y) = \frac{\alpha}{2}$$

with (θ_I, θ_S) posterior $\frac{\alpha}{2}$ and $(1 - \frac{\alpha}{2})$ quantiles of θ.

Highest Posterior Region

- Let C chosen so that the posterior density for any θ in C is higher than that for any θ not in C.
  Then
  $$C(k_{1-\alpha}) = \{\theta \in \Theta : \pi(\theta|y) \geq k_{1-\alpha}\}$$
  is called a highest posterior density (HPD) credible set with posterior probability $(1-\alpha)$.
- Generally numerical techniques are used to compute a HPD credible set.