Chapter 1

Bayesian Statistics

1.1 Introduction

Bayesian theory (named after the Rev. Thomas Bayes, an amateur 18th century English mathematician), provides an approach to statistical inference which is different in spirit from the familiar classical approach. We do not think of Bayesian statistics as a separate area within Statistics. Any statistical problem (Survival analysis; Multivariate analysis; Generalised linear models *etc.*) can be approached in a Bayesian way.

The basic philosophy underlying Bayesian inference is that **the only sensible measure of uncertainty is probability**. Data are still assumed to come from one of a parameterized family of distributions. However, whereas classical statistics considers the parameters to be *fixed but unknown*, the Bayesian approach treats them as random variables in their own right. Prior beliefs about θ are represented by the **prior**, $\pi(\theta)$, a probability density (or mass) function. The **posterior** density (mass function), $\pi(\theta|x_1, \dots, x_n)$ represents our *modified* belief about θ in the light of the observed data. We will do this in quite detail. Let us start from the basics.

1.2 Prior and Posterior Distributions

Theorem 1 (Bayes Theorem)

Let $B_1, B_2, \dots, B_k$ be a set of mutually exclusive and exhaustive events. For any new event A,

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}. \tag{1.1}$$

♡ **Example** 1.1. Suppose there are three production machines, I, II and III. Respectively 2%, 3% and 5% of the items produced by these machines are defective. Of total production machine I yields 30%, II yields 40% and III yields 30%. Suppose an item drawn randomly from the total production turns out to be defective. What is the chance that it was manufactured by machine I? □

Notation We are used to the notation that X is the random variable and x is its value. Now we will relax that little bit for the random variable θ only. We will use θ to denote the random variable and Θ to denote the parameter space. Often, however, Θ will be generically taken as $(-\infty, \infty)$.

Theorem 2 (Bayes Theorem for Random variables)

Suppose that two random variables X and θ are given with pdf's $f(x|\theta)$ and $\pi(\theta)$.

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} f(x|\theta)\pi(\theta)d\theta}. \tag{1.2}$$

Bayesian Inference Framework:

- X = Data or the notation $x_1, x_2, \cdots, x_n$.
- θ = Unknown parameters.
- $f(x_1, \cdots, x_n|\theta)$ = Likelihood of data given unknown parameters θ.
- $\pi(\theta)$ = Prior distribution or prior belief. A priori what you know for the unknown parameters.

By the above Bayes Theorem,

$$\pi(\theta|x_1, \cdots, x_n) = \frac{f(x_1, \cdots, x_n|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} f(x_1, \cdots, x_n \mid \theta)\pi(\theta)d\theta}.$$

This distribution is called the **posterior distribution**. Bayesian inference proceeds from this distribution. In practice, the denominator of the above equation needn't usually be calculated, and Bayes' rule is often just written,

$$\pi(\theta|x_1, \cdots, x_n) \propto f(x_1, \cdots, x_n \mid \theta)\pi(\theta).$$

Hence we always know the posterior distribution up-to a normalizing constant. Often we will be able to identify the posterior distribution of θ just by looking at the numerator. By Bayes Theorem we "update" $\pi(\theta)$ to $\pi(\theta|\mathbf{x})$.

Remark: Bayesian Learning:

$$\begin{aligned} \pi(\theta|x_1) &\propto f(x_1|\theta)\pi(\theta) \\ \pi(\theta|x_1, x_2) &\propto f(x_2|\theta)f(x_1|\theta)\pi(\theta) \\ &\propto f(x_2|\theta)\pi(\theta|x_1) \end{aligned}$$

Thus the Bayes theorem shows how the knowledge about the state of nature represented by θ is continually modified as new data becomes available.

♡ **Example** 1.2. Suppose $X \sim$ binomial(n, θ) where n is known and we assume Beta(α, β) prior for θ. Here the likelihood is

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Prior is

$$\pi(\theta) = \frac{1}{Beta(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Hence

$$\pi(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}.$$

Note that we have only written down the terms involving θ from the likelihood $\times$ the prior. We do not care care about the other terms which do not involve θ, like the $\binom{n}{x}$ or the constant $\frac{1}{Beta(\alpha,\beta)}$ because these cancel in the ratio. Now the posterior is recognised to be a Beta distribution with parameters $x + \alpha$ and $n - x + \beta$. Hence:

$$\pi(\theta|x) = \frac{1}{Beta(x+\alpha, n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}.$$

□

♡ **Example** 1.3. Suppose $X_1, \ldots, X_n$ is a random sample from the distribution with pdf $f(x|\theta) = \theta e^{-\theta x}$. Suppose the prior for θ is given by $\pi(\theta)$ and $\pi(\theta) = \mu e^{-\mu\theta}$ for some known $\mu > 0$.

Then the likelihood is:

$$f(x_1, x_2, \ldots, x_n|\theta) = \prod \theta e^{-\theta x_i} = \theta^n e^{-\theta\sum_{i=1}^n x_i}.$$

Hence the posterior $\propto$ Likelihood $\times$ Prior is:

$$\pi(\theta|x_1,\ldots,x_n) \propto \theta^n e^{-\theta\sum_{i=1}^n x_i}\ \mu e^{-\mu\theta}.$$

Now collecting the terms involving θ only we see that:

$$\pi(\theta|x_1,\ldots,x_n) \propto \theta^n e^{-\theta(\mu+\sum_{i=1}^n x_i)}.$$

The above is the pdf of a Gamma random variable. □

♡ **Example** 1.4. Suppose $X_1,\ldots,X_n \sim N(\theta,\sigma^2)$, where σ^2 is known. Let $\pi(\theta) \sim N(\mu,\tau^2)$ for known μ and τ^2. The likelihood is:

$$f(x_1,x_2,\ldots,x_n|\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_i-\theta)^2}{\sigma^2}} = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2}\sum_{i=1}^n \frac{(x_i-\theta)^2}{\sigma^2}}.$$

The prior is:

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2}\frac{(\theta-\mu)^2}{\tau^2}}.$$

The posterior is proportional to the Likelihood $\times$ Prior. Hence we keep the terms involving θ only.

$$\pi(\theta|x_1,\ldots,x_n) \propto e^{-\frac{1}{2}\left[\sum_{i=1}^n \frac{(x_i-\theta)^2}{\sigma^2}+\frac{(\theta-\mu)^2}{\tau^2}\right]} = e^{-\frac{1}{2}M}.\ \ M \text{ is what's inside [and].}$$

Now we look at the exponent carefully. Notice that it is a quadratic in θ. OK. We should try to complete the square in θ and ultimately we may have that the posterior distribution of θ is a normal distribution. Now:

$$\begin{aligned} M &= \textstyle\sum_{i=1}^n \frac{(x_i-\theta)^2}{\sigma^2} + \frac{(\theta-\mu)^2}{\tau^2} \\ &= \frac{\sum x_i^2 - 2\theta\sum x_i + n\theta^2}{\sigma^2} + \frac{\theta^2-2\theta\mu+\mu^2}{\tau^2} \\ &= \theta^2(n/\sigma^2+1/\tau^2) - 2\theta(\textstyle\sum x_i/\sigma^2+\mu/\tau^2) + \textstyle\sum x_i^2/\sigma^2+\mu^2/\tau^2 \\ &= \theta^2\, a - 2\theta\, b + c \end{aligned}$$

where

$$a = n/\sigma^2+1/\tau^2, b = \sum x_i/\sigma^2+\mu/\tau^2, c = \sum x_i^2/\sigma^2+\mu^2/\tau^2.$$

Note that none of a, b and c involves θ. These are defined just for writing convenience. Now

$$\begin{aligned} M &= a(\theta^2 - 2\theta b/a) + c \\ &= a(\theta^2 - 2\theta b/a + b^2/a^2 - b^2/a^2) + c \\ &= a(\theta - b/a)^2 + b^2/a + c \end{aligned}$$

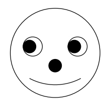
Note again that none of a, b and c involves θ, hence the first term only involves θ and the last two are rubbish!

$$\pi(\theta|x_1, \dots, x_n) \propto e^{-\frac{1}{2}a(\theta - b/a)^2}$$

which is easily recognised to be the pdf of a normal distribution with mean b/a and variance $1/a$. More explicitly

$$\pi(\theta|\mathbf{x}) = N\left(b/a = \frac{\sum x_i/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{n/\sigma^2 + 1/\tau^2}\right) = N\left(\frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{n/\sigma^2 + 1/\tau^2}\right)$$

$\square$



1.3 Bayes Estimators

Given $\pi(\theta|x_1, \dots, x_n)$, we require a mechanism to choose a reasonable estimator $\hat{\theta}$. Suppose the true parameter is θ_0 which is unknown. Let a be our guess for it. In real life we may not have $a = \theta_0$. Then it is sensible to measure the penalty we have to pay for guessing incorrectly. The penalty may be measured by $(a - \theta_0)^2$ or $|a - \theta_0|$ or some other function. We should choose that value of a which minimizes the expected loss $E[L(a, \theta)]$, sometimes called the **risk**, where the expectation is taken with respect to the posterior distribution $\pi(\theta|x_1, \dots, x_n)$ of θ. Note that a should not be a function of θ, rather it should be a function of $x_1, \dots, x_n$, the random sample. The minimizer, $\hat{\theta}$ say, is called the Bayes estimator of θ.

1.3.1 Squared Error Loss Function

We consider the loss function:

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

See what happens to the expected loss = the risk, for squared error loss. Let

$$b = E_{\pi(\theta|x_1, x_2, \dots, x_n)}(\theta) = \int \theta \pi(\theta|x_1, x_2, \dots, x_n) d\theta.$$

$$\begin{aligned}
E[L(a,\theta)] &= \int L(a,\theta)\pi(\theta|x_1,\dots,x_n)d\theta \\
&= \int (a-b+b-\theta)^2\pi(\theta|x_1,\dots,x_n)d\theta \\
&= (a-b)^2 + \int (b-\theta)^2\pi(\theta|x_1,\dots,x_n)d\theta \\
&\geq \int (b-\theta)^2\pi(\theta|x_1,\dots,x_n)d\theta,
\end{aligned}$$

for any value of a. When will the above inequality be an equality? Ans: when $a = b$. Note that b is the posterior mean of θ. Hence we say that **the Bayes estimator under squared error loss is the posterior mean.**

1.3.2 Absolute Error Loss Function

What happens when we assume the absolute error loss, $L(a,\theta) = |a-\theta|$. Then by a theorem Degroot, page 210, $E[|a-\theta|]$ is minimized by taking a to be the median of the posterior distribution of θ. For this loss function the Bayes estimator is the posterior median. The median of a random variable Y with pdf $g(y)$ is defined as the value μ which solves:

$$\int_{-\infty}^{\mu} g(y)dy = \frac{1}{2}.$$

This is hard to find except for symmetric distributions. For example, for the normal example the Bayes estimator of θ under the absolute error loss is still the posterior mean because it is also the posterior median. For the other examples, we need a computer to find the posterior medians.

1.3.3 Step Function Loss

We consider the loss function:

$$\begin{aligned}
L(a,\theta) &= 0 \text{ if } |a-\theta| \leq \delta \\
&= 1 \text{ if } |a-\theta| > \delta
\end{aligned}$$

where δ is a given small positive number. Now let us find the expected loss, i.e. the risk. Note that expectation is to be taken under the posterior distribution.

$$\begin{aligned}
E[L(a,\theta)] &= \int_{\Theta} I(|a-\theta| > \delta)\pi(\theta|\mathbf{x})d\theta \\
&= \int_{\Theta} (1 - I(|a-\theta| \leq \delta))\,\pi(\theta|\mathbf{x})d\theta \\
&= 1 - \int_{a-\delta}^{a+\delta} \pi(\theta|\mathbf{x})d\theta \\
&\approx 1 - 2\delta\pi(a|\mathbf{x})
\end{aligned}$$

where $I(\cdot)$ is the indicator function. In order to minimise the risk we need to maximise $\pi(a|\mathbf{x})$ with respect to a and the Bayes estimator is the maximiser.

Therefore, the Bayes estimator is that value of θ which maximises the posterior, i.e. the modal value. This estimator is called the maximum a-posteriori (MAP) estimator.

♡ **Example** 1.5. **Binomial** As in Example 1.2. The Bayes estimator under squared error loss is

$$\hat{\theta} = \frac{x+\alpha}{x+\alpha+n-x+\beta} = \frac{x+\alpha}{n+\alpha+\beta}.$$

□

♡ **Example** 1.6. **Exponential** As in Example 1.3. The Bayes estimator is

$$\hat{\theta} = \frac{n+1}{\mu + \sum_{1=1}^{n} x_i}.$$

Note that μ is a given constant. □

♡ **Example** 1.7. **Normal** As in Example 1.4. The Bayes estimator under all three loss functions is

$$\hat{\theta} = \frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}$$

□

♡ **Example** 1.8. Let $X_1, \ldots, X_n \sim \textit{Poisson}(\theta)$. Also suppose the prior is $\pi(\theta) = e^{-\theta}, \theta > 0$. The likelihood is:

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \frac{1}{x_i!} e^{-\theta} \theta^{x_i} = e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i} \frac{1}{x_1! x_2! \cdots x_n!}$$

Now try to get the posterior distribution of θ as the Likelihood times the prior. We only need to collect the terms involving θ only.

$$\begin{aligned} \pi(\theta|x_1, \ldots, x_n) &\propto e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i} e^{-\theta} \\ &\propto e^{-(n+1)\theta} \theta^{\sum_{i=1}^{n} x_i} \end{aligned}$$

which is easily seen to be the pdf of Gamma$\left(1 + \sum_{i=1}^{n} x_i, \frac{1}{n+1}\right)$. Hence the Bayes estimator of θ under squared error loss is:

$$\hat{\theta} = \text{ Posterior mean } = \frac{1 + \sum_{i=1}^{n} x_i}{1+n}.$$

□

1.4 Credible Regions

Choose a set A such that $P(\theta \in A|\mathbf{x}) = 1 - \alpha$. Such a set A is called $100(1-\alpha)\%$ credible region for θ.

The set A is called a *Highest Posterior Density* (HPD) credible region if $\pi(\theta|\mathbf{x}) \geq \pi(\psi|\mathbf{x})$ for all $\theta \in A$ and $\psi \notin A$.

♡ **Example** 1.9. **Normal–Normal** Find a 95% HPD credible region for θ. It is simply given by

$$\frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} \pm 1.96\sqrt{\frac{1}{n/\sigma^2 + 1/\tau^2}}.$$

Put $\sigma^2 = 1, \mu = 0, n = 12, \bar{x} = 5, \tau^2 = 10$ to see a numerical example.

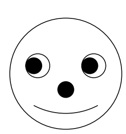
□

MA676: Bayesian Methods – Additional Exercises 1

1. A certain disease affects 0.1% of the population. A diagnostic test for the disease gives a positive response with probability 0.95 if the disease is present. It also gives a positive response with probability 0.02, however, if the disease is not present. If the test gives a positive response for a given person, what is the probability that the person has the disease?

2. The Japanese car company Maito make their Professor model in three countries, Japan, England and Germany, with one half of the cars being built in Japan, two tenths in England and three tenths in Germany. One percent of the cars built in Japan have to be returned to the dealer as faulty while the figures for England and Germany are four percent and two percent respectively. What proportion of Professors are faulty? If I buy a Professor and find it to be faulty, what is the chance that it was made in England?

3. Suppose that the number of defects on a roll of magnetic recording tape has a Poisson distribution for which the mean θ is unknown and that the prior distribution of θ is a gamma



distribution with parameters $\alpha = 3$ and $\beta = 1$. When five rolls of this tape are selected at random and inspected, the number of defects found on the rolls are 2, 2, 6, 0, and 3. If the squared error loss function is used what is the Bayes estimate of θ?

4. Suppose that $X_1, \dots, X_n$ is a random sample from the distribution with pdf

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose also that the value of the parameter θ is unknown ($\theta > 0$) and that the prior distribution of θ is a gamma distribution with parameters α and β ($\alpha > 0$ and $\beta > 0$). Determine the posterior distribution of θ and hence obtain the Bayes estimator of θ under a squared error loss function.

5. Suppose that $X_1, \dots, X_n$ is a random sample of size n from a normal distribution with unknown mean θ and variance 1. Suppose also that the prior distribution of θ is also a normal distribution with mean 0 and given variance τ^2. Derive the posterior distribution of θ given $X_1, \dots, X_n$. From the posterior distribution write the Bayes estimator under i) squared error loss and ii) absolute error loss, clearly stating the standard results you have used.