Decision Theory and Bayesian Analysis

Dr. Vilda Purutcuoglu

[1]

[1] Edited by Anil A. Aksu based on lecture notes of STAT 565 course by Dr. Vilda Purutcuoglu

Contents

LECTURE 1
Bayesian Paradigm

1.1. Bayes theorem for distributions

If A and B are two events,

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}. \tag{1.1}$$

This is just a direct consequence of the multiplication law of probabilities that says we can express $P(A \mid B)$ as either $P(A)P(B \mid A)$ or $P(B)P(A \mid B)$. For discrete distributions, if Z, Y are discrete random variables

$$P(Z = z \mid Y = y) = \frac{P(Z = z)P(Y = y \mid Z = z)}{P(Y = y)}. \tag{1.2}$$

- How many distributions do we deal with here?

We can express the denominator in terms of the distribution in the numerator[**1**].

$$P(Y = y) = \sum_z P(Y = y, Z = z) = \sum_z P(Z = z)P(Y = y \mid Z = z). \tag{1.3}$$

- This is sometimes called the law of total probability

In this context, it is just an expression of the fact that as z ranges over the possible values of Z , the probabilities on the left hand-side of equation 1.2 make up the distribution of Z given $Y = y$, and so they must add up to one. The extension to continuous distribution is easy. If Z, Y are continuous random variable,

$$f(Z \mid Y) = \frac{f(Z)f(Y \mid Z)}{f(Y)}. \tag{1.4}$$

where the denominator is now expressed as an integral:

$$f(Y) = \int f(Z)f(Y \mid Z)\mathrm{d}Z. \tag{1.5}$$

$$f = \begin{cases} continous \quad name? \\ discrete \quad name? \end{cases} \tag{1.6}$$

1.2. How Bayesian Statistics Uses Bayes Theorem

Theorem 1.7 (Bayes' theorem)**.**

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

$P(B)$=if we are interested in the event B, $P(B)$ is the initial or prior probability of the occurence of event B. Then we observe event A
$P(B \mid A)$ = How likely B is when A is known to have occurred is the posterior probability $P(B \mid A)$.

Bayes' theorem can be understood as a formula for updating from prior to posterior probability, the updating consists of multiplying by the ratio $P(B \mid A)/P(A)$. It describes how a probability changes as we learn new information. Observing the occurrence of A will increase the probability of B if $P(B \mid A) > P(A)$. From the law of total probability,

$$P(A) = P(A \mid B)P(B) + P(A \mid B^c) + P(A \mid B^c)P(B^c). \tag{1.8}$$

where $P(B^c) = 1 - P(B)$.

Lemma 1.9.

$$P(A \mid B) - P(A) = \frac{P(A) - P(A \mid B^c)P(B^c)}{1 - P(B^c)} - P(A)$$

Proof.

$$P(A \mid B) - P(A) = \frac{P(A) - P(A \mid B^c)P(B^c) - P(A) + P(A)P(B^c)}{P(B)}$$

$$P(A \mid B) - P(A) = \frac{P(B^c)(P(A) - P(A \mid B^c))}{P(B)}$$

$$P(A \mid B) - P(A) = P(B^c)(\frac{P(B)P(A \mid B) + P(B^c)P(A \mid B^c)}{P(B)} - \frac{P(A \mid B^c)}{P(B)})$$

$$P(A \mid B) - P(A) = P(B^c)(P(A \mid B) - \frac{P(A \mid B^c)(1 - P(B^c))}{P(B)})$$

$$P(A \mid B) - P(A) = P(B^c)(P(A \mid B) - P(A \mid B^c))$$

□

1.2.1. Generalization of the Bayes' Theorem

Let $B_1, ..., B_n$ be a set of mutually exclusive events. Then

$$P(B_r \mid A) = \frac{P(B_r)P(A \mid B_r)}{P(A)} = \frac{P(B_r)P(A \mid B_r)}{\sum_{i=1}^{n} P(B_r)P(A \mid B_r)}. \tag{1.10}$$

- Assuming that $P(B_r) > 0$,$P(A \mid B) > P(A)$ if and only if $P(A \mid B) > P(A \mid B^c)$.

- In Bayesian inference we use Bayes' theorem in a particular way.
- Z is the parameter (vector) θ.
- Y is the data (vector) X.

So we have

$$f(\theta \mid X) = \frac{f(\theta) f(X \mid \theta)}{f(X)} \tag{1.11}$$

$$f(X) = \int f(\theta) f(X \mid \theta) \mathrm{d}\theta. \tag{1.12}$$

$$f(\theta) = \tag{1.13}$$

$$f(\theta \mid X) = \tag{1.14}$$

$$f(X \mid \theta) = \tag{1.15}$$

1.2.2. Interpreting our sense

How do we interpret the things we see, hear, feel, taste or smell?

Example 1.2.1. I hear a song on the radio I identify the singer as Robbie Williams. Why do I think it's Robbie Williams?. Because he sounds like that. Formally, $P($ What I hear Robbie Williams $) >> P($What I hear someone else $)$

Example 1.2.2. I look out of the window and see what appears to be a tree. It has a big, dark coloured part sticking up out of the ground that branches into thinner sticks and on the ends of these are small green things. Clearly, $P(view \mid tree)$ is high and $P(view \mid car)$ or $P(view \mid Robbie \quad Williams)$ are very small. But $P(view \mid carboard \quad cutout \quad cunningly \quad painted \quad to \quad look \quad like \quad a \quad tree)$ is also very high. Maybe even higher than $P(view \mid tree)$ in the sense that what I see looks almost like a tree.
Does this mean I should now believe that I am seeing a cardboard cutout cunningly painted to look like a tree? No because it is much less likely to begin with than a red tree.

In statistical terms, consider some data X and some unknown parameter θ. The first step in any statistical analysis is to build a model that links the data to unknown parameters and the main function of this model is to allow us to state the probability of observing any data given any specified values of the parameters. That is the model defines $f(x \mid \theta)$.

When we think of $f(x \mid \theta)$ as a function of θ for fixed observed data X, we call it likelihood function and it by $L(\theta, X)$.

- So how can we combine this with our example?

This perspective underlies the differences between the two main theories of statistical inference.

- Frequentist inference essentially uses only the likelihood, it does not recognize $f(\theta)$.
- Bayesian inference uses both likelihood and $f(\theta)$.

The principal distinguishing feature of Bayesian inference as opposed to frequentist inference is its use of $f(\theta)$.

1.3. Prior to Posterior

We refer to $f(\theta)$ as the prior distribution of θ. It represents knowledge about θ prior to observing the data X. We refer to $f(\theta \mid X)$ as the posterior distribution of θ and it represents knowledge about θ after observing X.

- So we have two sources of information about θ.
- Here $f(x)$ does not depend on θ. Thus $\int f(\theta \mid x)\mathrm{d}\theta = 1$. Since $f(x)$ is a constant within the integral, we can take it outside to get $1 = f^{-1}(x) \int f(\theta)f(x \mid \theta)\mathrm{d}\theta$.
- $f(\theta \mid x) =\propto f(\theta)f(x \mid \theta) \propto f(\theta)L(\theta; x)$ (the posterior is proportional to the prior times the likelihood).
- The constant that we require to scale the right hand side to integrate to 1 is usually called the normalizing constant. If we haven't dropped any constants form $f(\theta)$ or $f(x \mid \theta)$, then the normalising constant is just $f^{-1}(x)$, otherwise it also restores any dropped constants.

1.4. Triplot

If for any value of θ, we have either $f(\theta) = 0$ or $f(x \mid \theta) = 0$, then we will also have $f(\theta \mid x) = 0$. This is called the property of zero preservation. So if either:

- the prior information says that this θ value is impossible
- the data say that this value of θ is impossible because if it were the true value, then the observed data would have been impossible, then the posterior distribution confirms that this value of θ is impossible.

Definition 1.16. Crowwell's Rule: If either information source completely rules out a specific θ, then the posterior must rule it out too.

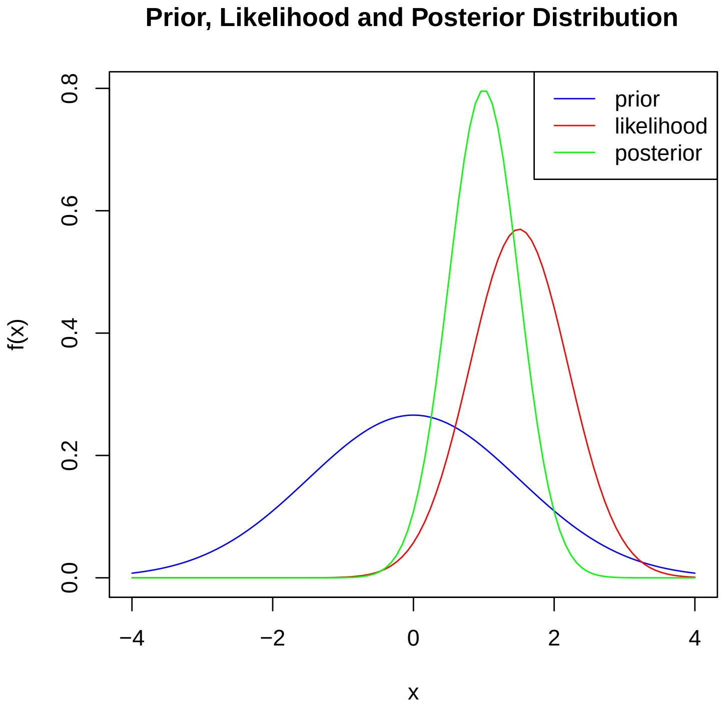

Figure 1. Triplot of prior, likelihood and posterior.

This means that we should be very careful about giving zero probability to something unless it is genuinely impossible. Once something has zero probability then no amount of further evidence can cause it to have a non-zero posterior probability.

- More generally, $f(\theta \mid x)$ will be low if either $f(\theta)$ is very small. We will tend to find that $f(x \mid \theta)$ is large when both $f(\theta)$ and $f(x \mid \theta)$ are relatively large, so that this θ value is given support by both information sources.

When θ is a scalar parameter, a useful diagram is the triplot, which shows the prior, likelihood and posterior on the same graph. An example is in Figure 1.[1]

A strong information source in the triplot is indicated by a crve that is narrow (and therefore, because it integrates to one, also has a high peak). A narrow curves concentrates on a small range of θ values, and thereby "rules out" all values of θ outside that range.

[1] All plots are generated in R, relevant codes are provided in Appendix R Codes

- Over the range $\theta < -1$, the likelihood:
- Over the range $\theta > 3$,the likelihood:
- Values of θ between -1 and 3, the likelihood:
- The maximum value of the posterior at:
- The MLE of θ is:

1.4.1. Normal Mean

For example, suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu,\, \sigma^2)$ and σ^2 is known. Then the likelihood is :

$$\begin{aligned} f(x \mid \mu) = \prod_{i=1}^{n} f(x_i \mid \mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right). \end{aligned} \tag{1.17}$$

As,

$$\begin{aligned} \sum(x_i-\bar{x}+\bar{x}-\mu)^2 &= \sum(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 + 2(\bar{x}-\mu)\sum(x_i-\bar{x}) \\ &= \sum(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2 \\ &\propto \exp\left(-\frac{1}{2\sigma^2}n(\bar{x}-\mu)^2\right). \end{aligned} \tag{1.18}$$

Note that $2(\bar{x}-\mu)\sum(x_i-\bar{x}) = 0$ as $\sum(x_i-\bar{x}) = 0$. Suppose the prior distribution for μ is normal:

$$\mu \sim \mathcal{N}(m, v). \tag{1.19}$$

Then applying Bayes' theorem we have:

$$\begin{aligned} f(\mu \mid x) \propto \underbrace{\exp\left(-\frac{1}{2\sigma^2}n(\bar{x}-\mu)^2\right)}_{f(x\mid\mu)} & \underbrace{\exp\left(-\frac{1}{2\sigma^2}n(\mu-m)^2\right)}_{f(\mu)} \\ &= \exp\left(-\frac{\theta}{2}\right). \end{aligned} \tag{1.20}$$

Note that

$$\theta = n\sigma^{-2}(\bar{x}-\mu) + v^{-1}(\mu-m)^2 = (v^*)^{-1}(\mu-m^*)^2 + R \tag{1.21}$$

and

$$v^* = (n\sigma^{-2} + v^{-1})^{-1} \tag{1.22}$$

$$m^* = (n\sigma^{-2} + v^{-1})^{-1}(n\sigma^{-2}\bar{x} + v^{-1}m) = a\bar{x} + (1-a)m \tag{1.23}$$

where $a = n\sigma^{-}2/(n\sigma^{-2} + v^{-1})$

$$R = (n^{-1}\sigma^2 + v)(\bar{x} - m)^2 \tag{1.24}$$

Therefore,

$$f(\mu \mid x) \propto \exp\left(-\frac{1}{2\sigma^2} n(\mu - m)^2\right) \tag{1.25}$$

and we have shown that the posterior distribution is normal too: $\mu \mid x \sim \mathcal{N}(m^*, v^*)$

- m^* = weighted average of the mean m and the usual frequentist data-only estimate $\bar{x}$.
  The weights $\propto$:
- Bayes' theorem typically works in this way. We usually find that posterior estimates are compromises between prior estimates and data based estimates and tend to be closer whichever information source is stronger. And we usually find that the posterior variance is smaller than the prior variance.

1.4.2. Weak Prior Information

It is the case where the prior information is much weaker that the data. This will occur, for instance, if we do not have strong information about Q before seeing the data, and if there are lots of data. Then in triplot, the prior distribution will be much broader and flatter that the likelihood. So the posterior is approximately proportional to the likelihood.

Example 1.4.1. In the normal mean analysis, we get weak prior information by letting the prior precision of v^{-1} become small. Then $m^* \to \bar{x}$ and $v^* \to \sigma^2/n$ so that the posterior distribution of μ corresponds very closely with standard frequentist theory.

LECTURE 2
Some Common Probability Distributions

LECTURE 3
Inference

Basic Statistics

R Codes

Listing 1. Triplot Code in R

```
####################################################
#                                                #
#      A Sample Triplot by Anil Aksu               #
#   It is developed to show some basics of R       #
#                                    #
####################################################

## the range of sampling
x=seq(−4,4,length=101)
## this function gets numbers from console
prior=dnorm(x, mean = 0, sd = 1.5, log = FALSE)
likelihood=dnorm(x, mean = 1.5, sd = 0.7, log = FALSE)
posterior=dnorm(x, mean = 1, sd = 0.5, log = FALSE)


## let's plot them
plot(range(x), range(c(likelihood,prior,posterior)), ...
    type='n', xlab="x", ylab="f(x)")
lines(x, prior, type='l', col='blue')
lines(x, likelihood, type='l', col='red')
lines(x, posterior, type='l', col='green')

title("Prior, Likelihood and Posterior Distribution")
legend(
  "topright",
  lty=c(1,1,1),
  col=c("blue", "red", "green"),
  legend = c("prior", "likelihood","posterior")
)
```

BIBLIOGRAPHY

1. Allen B. Dawney. *Think Bayes: Bayesian Statistics in Python.* O'REILLY, 2013.