LTCC

Applied Bayesian Methods

Aims of course

- To introduce the Bayesian approach to statistical inference
- To develop relevant methodology, theory and computational techniques for its implementation

Why Bayesian?

- Can incorporate prior information (ie info. before seeing data), rather than only data, into statistical modelling and inference
- Can build complex models
- "The philosophy of statistics", Dennis V. Lindley, 2000, JRSS-D

Course content & schedule

W1 §1 Introduction to Bayesian statistics

W2 §2 Bayesian inference

W3 §3 Prior distributions

W4 §4 Graphical models &
§5 Hierarchical models

W5 §6 Markov chain Monte Carlo (MCMC)

Texts

1. P.M. Lee, *Bayesian Statistics: An Introduction* (**Chapters 1–3**, 2004, 3rd Edition: Arnold).
2. J. Whittaker, *Graphical Models in Applied Multivariate Statistics* (**Chapters 1–3**, 1990, John Wiley & Sons).
3. C.M. Bishop, *Pattern Recognition and Machine Learning* (**Chapter 8** "Graphical models", 2006, Springer).
4. A. Gelman, J.B. Carlin, H.S. Stern & D.B. Rubin, *Bayesian Data Analysis* (**Chapter 5** "Hierarchical models", 2003, 2nd Edition: Chapman and Hall/CRC).
5. W.R. Gilks, S. Richardson & D.J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice* (**Chapters 1, 2 and 5**, 1996, Chapman & Hall/CRC).

Lecturer

- Prof Petros Dellaportas
  @room 131A, 1-19 Torrington Place
- p.dellaportas@ucl.ac.uk
  (*Your feedback on this course is welcome; the earlier the better!*)

Outline

1. Bayes' theorem

2. Interpretation of probability

3. Bayesian inference

4. Predictive distributions

1. Thomas Bayes (?-1761) and Bayes' theorem (1764)

Recall the *multiplication law of probability*,

$$P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A) \ .$$

Bayes' theorem:

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

In the case of probability distributions,

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{p(y)} \ .$$

Bayes' theorem:

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

Alternative forms of Bayes' theorem

Using the *law of total probability* that

$$P(B) = \sum_i P(B \cap A_i) \ ,$$

where $\{A_i : i = 1, 2, \ldots\}$ is a set of mutually exclusive and exhaustive events (ie $A_i \cap A_j = \emptyset \ , \forall i \neq j$ and $P(\bigcup_i A_i) = \sum_i P(A_i) = 1$), we obtain

$$\begin{aligned} P(A_1 \mid B) &= \frac{P(A_1)P(B \mid A_1)}{\sum_i P(B \cap A_i)} \\ &= \frac{P(A_1)P(B \mid A_1)}{\sum_i \{P(A_i)P(B \mid A_i)\}} \ . \end{aligned}$$

In the case of probability distributions,

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{p(y)} = \frac{p(\theta)\, p(y \mid \theta)}{\int p(\theta)\, p(y \mid \theta)\, d\theta} \ .$$

Example 1.1: diagnostic testing

A diagnostic test for HIV is claimed to have:

- 95% sensitivity ($\Rightarrow$ 95% of people who have HIV will test positive)
- 98% specificity ($\Rightarrow$ 98% of people who do not have HIV will test negative)

In a population with prior knowledge that an HIV prevalence is 1/1000, what is the chance that someone testing positive actually has HIV?

Solution

Let $Tr = 1$ be the event that the individual is truly HIV-positive, $Tr{=}0$ be truly HIV-negative.

Let $Te{=}1$ be the event that the individual tests positive, $Te{=}0$ be test-negative.

We have

- $P[Te{=}1 \mid Tr{=}1] = 0.95$
- $P[Te{=}0 \mid Tr{=}0] = 0.98$
- $P[Tr{=}1] = 0.001$

We want $P[Tr{=}1 \mid Te{=}1]$.

Based on Bayes' theorem, $P[Tr{=}1 \mid Te{=}1]$

$$
\begin{aligned}
&= \frac{P[Te{=}1 \mid Tr{=}1]\, P[Tr{=}1]}{P[Te{=}1]} \\
&= \frac{P[Te{=}1 \mid Tr{=}1]\, P[Tr{=}1]}{P[Te{=}1 \mid Tr{=}1]\, P[Tr{=}1] + P[Te{=}1 \mid Tr{=}0]\, P[Tr{=}0]} \\
&= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.02 \times 0.999} \\
&= 0.045
\end{aligned}
$$

So, over 95% of those testing positive will, in fact, NOT have HIV.

1. Importance of the prior.

   The disease prevalence can be regarded as a **'prior' probability** (0.001, ie unlikely). Although the test result points to disease, our calculation says unlikely.

2. Importance of the data.

   The test result can change our belief that this person is HIV-positive: Before observing the test result, we think the probability of HIV-positive is 0.001; after observing a positive result, we modify the probability to 0.045. The latter is our **'posterior' probability** of HIV-positive.

2. Interpretation of probability

What does it actually mean to say that the probability of an event A is p, e.g. that the probability of heads in the toss of a fair coin is 0.5?

Frequentist interpretation

$$P[A] = \lim_{n\to\infty} \frac{m}{n} \ ,$$

where m is number of times the event A occurs in a sequence of n independent and identical 'experiments'.

- Involves hypothetical notion of a long sequence of repeatable experiments. OK for dice, maybe also OK for sampling from a real population, but what does it mean to think of e.g. 'nuclear war in next five years' as a repeatable experiment?

- Frequentist interpretation is objective in the sense that it is based on (potentially) observable events. However, this is also a limitation, since we often wish to consider the probabilities of unobservable quantities (e.g. drug efficacy)

Bayesian interpretation: subjective judgement

Probability of an event A is a measure of someone's degree of belief that A has/will occur.

- Depends on their own partial knowledge of the process of interest
- To emphasize this, we might denote the subjective probability of event A by $P(A \mid \text{personal knowledge})$

Example

Let A be the event that "FTSE 100 index rises above 6100 next week":

- I have no knowledge of the stock market, so say that

$$P(A \mid \text{personal knowledge}) = 0.5$$

- A stock market trader has considerable knowledge of how the FTSE index behaves, so might believe that

$$P(A \mid \text{personal knowledge}) = 0.95$$

Can use any number we choose to specify a subjective probability?

- No. Probabilities must *cohere*, ie obey the axioms of probability.
- If we want to be taken seriously, our probabilities must have some relationship with reality.

3. Bayesian inference

The fundamental principle of Bayesian statistics is: Our knowledge about *anything that is unknown* can be described by a probability or probability distribution.

Suppose there are observed data y and unknown parameters θ.

1. Classical inference:
   - Only the data y are regarded as random, while the parameters θ are treated as fixed but unknown.
   - Inference about θ is not just conditional on the observed data y, but also on what might have been observed under repeated sampling.

2. Bayesian inference:
   - Bayesians regard both y and θ as *random variables*.
   - Posterior inference about θ is conditional on the particular, actually observed, realisation of y.

We posit a model which specifies the *likelihood* $p(y \mid \theta)$.

From a Bayesian point of view:

- θ should have a *prior probability distribution* $p(\theta)$ reflecting our uncertainty about it, before seeing the data y;
- after seeing y, we should update our uncertainty about θ, by using a *posterior* distribution $p(\theta \mid y)$.

Bayes' theorem tells us how to calculate this:

$$
\begin{aligned}
p(\theta \mid y) &= \frac{p(\theta)\, p(y \mid \theta)}{p(y)} \\
&\propto p(\theta)\, p(y \mid \theta).
\end{aligned}
$$

posterior

prior

likelihood

Three components of Bayesian inference!

Example 1.2: Drug efficacy

The positive response rate of a drug is θ. Our prior knowledge about θ is its mean $m = 0.4$ with sd $\sqrt{v} = 0.1$.

1) Prior: How to translate our knowledge into a prior distribution $p(\theta)$?

Suppose $\theta \sim \text{Beta}(a, b)$, i.e.

$$\begin{aligned} p(\theta) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{a-1}(1-\theta)^{b-1} \end{aligned}$$

How to find a and b if $\theta \sim \text{Beta}(a, b)$?

For $\text{Beta}(a, b)$ distribution, we have

$$\begin{aligned} \text{mean } m &= a/(a+b) \\ \text{variance } v &= m(1-m)/(a+b+1) \\ \Rightarrow a &= m^2(1-m)/v - m \\ b &= m(1-m)^2/v - (1-m)\ . \end{aligned}$$

Solving gives $a = 9.2, b = 13.8$.

2) Likelihood: How to translate the observed data into likelihood?

Suppose we do an experiment and observe $y = 15$ positive responses in $n = 20$ trials (i.e. y successes in n independent trials): $Y \sim \text{Bin}(n, \theta)$

$$
\begin{aligned}
p(y \mid \theta) &= \binom{n}{y} \theta^y (1-\theta)^{n-y} \\
&\propto \theta^y (1-\theta)^{n-y}
\end{aligned}
$$

3) Posterior:

$$
\begin{aligned}
p(\theta \mid y) &\propto p(y \mid \theta) p(\theta) \\
&\propto \theta^y (1-\theta)^{n-y} \theta^{a-1} (1-\theta)^{b-1} \\
&= \theta^{y+a-1} (1-\theta)^{n-y+b-1}
\end{aligned}
$$

So, $\theta \mid y \sim \text{Beta}(y + a,\ n - y + b)$.

Posterior is a Beta(15+9.2, 20-15+13.8) = Beta(24.2, 18.8) with mean and sd:
$m^* = (y + a)/(n + a + b) = 24.2/43 = 0.56$,
$\sqrt{m^*(1 - m^*)/(a + y + b + n - y + 1)} = 0.075$.

How the mean and sd are revised?

Frequentist inference vs. Bayesian inference

- Classical (frequentist) inference:

  What does the data y tell us about the unknown quantity θ?

- Bayesian inference:

  How should the data y change someone's opinion about θ?

(More on this in next weeks)

4. Predictive distributions

Aim: to use the observed data y to predict an unknown observable $\tilde{Y}$ from the same process (determined by unknown parameter θ).

How?

We shall use the predictive distribution, ie the distribution $p(\tilde{Y} \mid Y = y)$:

$$
\begin{aligned}
p(\tilde{Y} = \tilde{y} \mid Y = y) &= \int p(\tilde{y}, \theta \mid y) d\theta \\
&= \int p(\tilde{y} \mid \theta, y)\ p(\theta \mid y)\ d\theta \\
&= \int p(\tilde{y} \mid \theta)\ p(\theta \mid y)\ d\theta
\end{aligned}
$$

- 1st '=': law of total probability
- 2nd '=': multiplication law of prob.
- 3rd '=': $p(\tilde{y} \mid \theta, y) = p(\tilde{y} \mid \theta)$ ie $\tilde{Y}$ and Y are conditional independent, given θ

Example 1.3: Prediction

Recall Example 1.2. Suppose we wish to predict the outcome $\tilde{Y}$ of treating a new patient with the drug, given what we've observed

Recall that $\theta \mid y \sim \text{Beta}(24.2, 18.8)$

For continuous θ we have

$$
\begin{aligned}
p(\tilde{Y} = 1 \mid Y = y) &= \int p(\tilde{Y} = 1 \mid \theta) p(\theta \mid y) d\theta \\
&= \int \theta p(\theta \mid y) d\theta \\
&= E(\theta \mid y) \\
&= \frac{24.2}{24.2 + 18.8} \\
&= 0.56
\end{aligned}
$$

Sequential learning

Suppose we obtain data y_1 and form the posterior $p(\theta \mid y_1)$ and then we obtain further data y_2 (indep. of y_1 from the same process determined by θ, ie $p(y_2 \mid \theta, y_1) = p(y_2 \mid \theta)$).

A key aspect of Bayesian analysis is the ease with which sequential analysis can be performed. The posterior given y_1 and y_2 is:

$$p(\theta \mid y_2, y_1) \propto p(y_2 \mid \theta) \times p(\theta \mid y_1) \ .$$

'Today's posterior is tomorrow's prior'!

Why? Because:

$$p(\theta \mid y_2, y_1) \propto p(y_2, y_1 \mid \theta) \times p(\theta) \ ,$$

and y_1 and y_2 are conditional indep., we have

$$\begin{aligned} p(y_2, y_1 \mid \theta)p(\theta) &= p(y_2 \mid \theta)p(y_1 \mid \theta)p(\theta) \\ &\propto p(y_2 \mid \theta)p(\theta \mid y_1) \end{aligned}$$

Example 1.4: Sequential updating

Recall Example 1.2 (p18), in which we've already observed $y = 15$ positives in 20 trials. Suppose another trial is carried out in which we observe $z = 4$ positives in 12 trials.

1. Simultaneous analysis

$$
\begin{aligned}
\text{Prior: } \theta &\sim \text{Beta}(9.2, 13.8) \\
\text{Likelih: } p(y, z \mid \theta) &= \theta^{19}(1-\theta)^{32-19} \\
\text{Posterior: } \theta \mid y, z &\sim \text{Beta}(19 + 9.2, 13 + 13.8) \\
&= \text{Beta}(28.2, 26.8)
\end{aligned}
$$

2. Sequential analysis

$$
\begin{aligned}
\text{Prior: } \theta \mid y &\sim \text{Beta}(24.2, 18.8) \\
\text{Likelih: } p(z \mid \theta) &\propto \theta^{4}(1-\theta)^{12-4} \\
\text{Posterior: } \theta \mid y, z &\sim \text{Beta}(4 + 24.2, 8 + 18.8) \\
&= \text{Beta}(28.2, 26.8)
\end{aligned}
$$

Outline revisited

1. Bayes' theorem
2. Interpretation of probability
3. Bayesian inference
4. Predictive distributions

Next week: Bayesian Inference