MAS3301 Bayesian Statistics

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 2, 2008-9

1 Introduction to Bayesian Inference

1.1 What is "Bayesian inference"?

1.1.1 Inference

Some may have expected the word "statistics" rather than "inference." I have nothing against "statistics", of course, but it may give too narrow an impression. It might suggest that some things, e.g. decision making under uncertainty, manipulation of uncertainty when there may appear to be few, if any, "data" (e.g. in probabilistic expert systems), are excluded. "Statistics" might also have connotations for some people which I would prefer to avoid. It is necessary to get rid of preconceptions.

In fact, the wider topic of "Bayesian analysis" deals with problems which involve one or both of the following.

- Inference. That is learning from data, or perhaps just the propagation of uncertainty or information among a collection of "unknowns."
- Decision. Choosing an action. This falls within the scope of Bayesian analysis when the decision is made under uncertainty.

In this module we will be more concerned with inference than with decision but the two are closely linked.

1.1.2 "Bayesian"

Why do we use the word "Bayesian"? The truth is that "Bayesian" means slightly different things to different people. Even among people who call themselves "Bayesians" there are philosophical differences. (However these are generally small compared to the difference with non-Bayesians!) In this course we will take a fairly mainstream position, that is mainstream among Bayesians. We will adopt a subjectivist view of probability. (The word "Bayesian" is sometimes used to describe things which are not really Bayesian in the sense used here. Beware.)

So, what makes an analysis "Bayesian"?

- Full probability specification. The state of uncertainty over all of the unknown quantities (and statements) involved is described by a joint probability distribution.
- Probability represents "degree of belief."
- Probability is subjective.

(Note that, at one end of the "Bayesian" spectrum there is subjectivist work which uses expectation directly, rather than probability, and, at the other end, there are attempts to use "Bayesian methods" in a non-subjective context).

Where does Bayes come in? Bayes' theorem turns out to be crucial for many inference problems if we adopt this view. This is particularly true for the traditional "statistical" problems where Bayes' theorem is more or less always required. Thomas Bayes (1702-1761) was a Presbyterian minister in Tunbridge Wells, Kent. In 1763 a paper by Bayes was published posthumously by his friend Richard Price. This paper gave us Bayes' theorem. Bayes' theorem tells us how to turn round conditional probability so that the conditioning is in the opposite direction. We shall see why this is important.

1.2 Motivational example 1

The university where I once worked has two main campuses, St. Peter's and Chester Road. My office was at St. Peter's but, from time to time, I needed to visit Chester Road. Very often I wanted to do the journey in such a way that as little time as possible was spent on it. (Sometimes my requirements were slightly different. It might be that I had to arrive before a certain time, for example). The university's Campus Bus was scheduled to leave St. Peter's at 15 and 45 minutes past each hour and arrive at Chester Road seven minutes later. Of course, in practice, these times

varied a little. I guessed it took me about two minutes to leave my office and go to the bus stop. On the other hand it took around 15 minutes to do the whole journey on foot. At what time should I leave the office on a journey to Chester Road, bearing in mind all these uncertainties and others, such as the possibility that my watch is wrong? If I left too early I would waste time waiting for the bus. If I missed the bus I either had to walk or return to my office and try again later, having wasted some time. If I went to the bus stop, arriving close to the scheduled departure time and there was nobody around, should I wait? Should I walk? Should I return to my office?

In order to try to answer the questions in the preceding paragraph, we need some way to deal with uncertainty. The approach taken in this course is *subjective*. That is, uncertainty is regarded as a feature of a person's beliefs (in the example, my beliefs) rather than of the world around us. I need to make the decision and I try to do it in a way which is rational and consistent with my beliefs. I need to be able to compare what I expect to happen under each of my possible decisions. I need a way to use the evidence available to me when I arrive at the bus stop to inform me about whether it is likely that the bus has already left. We might attempt to analyse the problem by constructing a model, for example a model for the arrivals and departures of the bus. Such a model would involve probability because the behaviour of the bus varies in a way which is not entirely predictable to us. The model would involve parameters. We may be uncertain about the values of these parameters. We may also be uncertain about the validity of assumptions in the model. We can use probability to deal with all of these uncertainties provided we regard probability as degree of belief.

The Campus Bus example may seem to be of little importance but it has features which may be found in many practical problems which may involve very important consequences.

1.2.1 Motivational Example 2

A company wants to know what kinds of people buy its product so that advertising may be aimed at the right people. Suppose that the company can observe various attributes of a sequence of people, including whether or not they buy the product. This could be by a survey or, perhaps, by monitoring visits to the company's web site.

Here we wish to learn about the association of attributes with the particular attribute "buys the product." We want to do this so that we will be able to make predictions about whether or not other people will buy the product, based on their attributes. So, we need to express our prior beliefs about these things, including our uncertainties about the associations and our beliefs about how the people we observe are related to the more general population of people who might buy the product. Is there, for example, anything unusual about the people we observe or are all people, including those whom we observe, "exchangeable"?

1.3 Revision: Bayes' Theorem

See section 2.6.

2 Beliefs and uncertainty

2.1 Subjective probability

We are dealing with what has been called "the logic of uncertainty." Some may ask, "Why should we use probability?" However this is really the wrong question. It probably betrays a misunderstanding about what Bayesians mean by "probability." Bruno de Finetti, a hero of Bayesianism, wrote in his classic book, *The Theory of Probability,* that **"Probability does not exist."** It has no *objective* existence. It is not a feature of the world around us. It is a measure of degree of belief, your belief, my belief, someone else's belief, all of which could be different. So we do not ask, "Is probability a good model?" We start by thinking about how "degree of belief" should behave and derive the laws of probability from this. We say that, if we obey these rules, we are being *coherent.* (See below) Note then that this is a normative theory not a descriptive theory. We are talking about how a "rational" person "ought" to behave. People may behave differently in practice!

Students of MAS3301 will have come across expectation and probability earlier, of course, but here we will briefly present these concepts in Bayesian/subjectivisit terms.

2.2 Expectation

2.2.1 Definition

Consider a quantity, X, whose value is currently unknown to us but which will, at least in principle, be revealed to us eventually. E.g. the amount of money in my coat pocket, the height in cm., of the next person to walk along the corridor.

Rough definition: Suppose you could have either £X or £c where c is a fixed number. We choose c so that you can not choose between £X and £c. Then $c = \mathrm{E}(X)$, your *expectation* or *expected value* of X. Notice that this is subjective.

Tighter definition: The rough definition is not rigorous because of difficulties with the values people assign to gambles. (This is the subject of the theory of *utility*). Imagine instead that you will lose £$(c - X)^2$ and you can choose c. The value of c you choose is your expectation, $\mathrm{E}(X)$, of X.

There are some quantities, e.g. the mean weight of eggs, which we will never observe. However we can express expectations about them because of the equivalence of these expectations with expectations about observable quantities, e.g. the weight of an egg.

We can define expectations for functions of X, e.g. X^2. We can use this fact to indicate our uncertainty about the value of X.

2.2.2 Variance and Covariance

The *variance* of X is $\mathrm{E}\{[X - \mathrm{E}(X)]^2\} = \mathrm{var}(X)$. The *standard deviation* of X is $\sqrt{\mathrm{var}(X)}$. A large variance means that we are very uncertain about the value of X.

We call $\mathrm{E}(X)$ the *"price"* of X in the sense that we would have no preference between £$\mathrm{E}(X)$ and £X (but recall comment on utility theory). Clearly the "price" of $aX + bY$, where a and b are constants and X and Y are both unknowns is $a\mathrm{E}(X) + b\mathrm{E}(Y)$. It follows that

$$\begin{aligned}
\mathrm{var}(X) &= \mathrm{E}\{[X - \mathrm{E}(X)]^2\} \\
&= \mathrm{E}\{X^2 - 2X\mathrm{E}(X) + [\mathrm{E}(X)]^2\} \\
&= \mathrm{E}(X^2) - 2[\mathrm{E}(X)]^2 + [\mathrm{E}(X)]^2 \\
&= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{var}(aX + bY) &= \mathrm{E}\{[aX + bY - a\mathrm{E}(X) - b\mathrm{E}(Y)]^2\} \\
&= \mathrm{E}\{a^2X^2 + b^2Y^2 + a^2[\mathrm{E}(X)]^2 + b^2[\mathrm{E}(Y)]^2 + 2abXY \\
&\quad -2a^2X\mathrm{E}(X) - 2abX\mathrm{E}(Y) - 2abY\mathrm{E}(X) - 2b^2Y\mathrm{E}(Y)\} \\
&\quad +2ab\mathrm{E}(X)\mathrm{E}(Y) \\
&= a^2\mathrm{var}(X) + b^2\mathrm{var}(Y) + 2ab\mathrm{covar}(X, Y).
\end{aligned}$$

where

$$\mathrm{covar}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}\{[X - \mathrm{E}(X)][Y - \mathrm{E}(Y)]\}$$

is called the *covariance* of X and Y. We also talk about the *correlation* of X, Y which is defined as $\mathrm{covar}(X, Y)/\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}$.

2.3 Probability

For non-quantitative unknowns, such as the occurrence of a particular event or the truth of a particular statement, we can introduce indicator variables. For example $I_R = 1$ if it rains tomorrow, $I_R = 0$ otherwise. We call $\mathrm{E}(I_R)$ the *probability* that it rains tomorrow, Pr(rain). For a discrete quantitative variable we can have, e.g., $I_3 = 1$ if $X = 3, I_3 = 0$ otherwise. For continuous quantitative variables we can have, e.g. , $I_x = 1$ if $X \leq x, I_x = 0$ otherwise so $\mathrm{E}(I_x) = \Pr(X \leq x) = F_X(x)$. Thus we can evaluate a probability distribution for an unknown quantity and $\Pr(X \leq x)$ becomes our degree of belief in the statement that $X \leq x$.

		Second animal		
		YES	NO	
First	YES	0.06	0.14	0.2
Animal	NO	0.14	0.66	0.8
		0.2	0.8	1.0

Table 1: Probabilities for possession of a particular gene in two animals.

2.4 Coherence

2.4.1 Concept

Can we assign our subjective probabilities to be whatever we like? No. We impose the conditions of *"coherence"*. These rule out beliefs which would appear to be irrational in the sense that they violate the "sure loser principle". This principle states that we should not hold "beliefs" which would force us to accept a series of bets which was bound to make us lose. The usual rules of probability can all be derived from the idea of coherence. We omit the details, except for the following simple example (2.4.2).

2.4.2 Example

Consider the statement S : "It will rain tomorrow."

Roughly: You pay me £$\Pr(S)$ and, in return, I give you £1 if S is true and £0 if it is not.

More carefully: You pay me £$k(1-p)^2$ if S true and £kp^2 if not. You get to choose p. Clearly $0 \leq p \leq 1$.

If $p < 0$ then $p = 0$ would always be better,

If $p > 1$ then $p = 1$ would always be better,

whether or not S true.

This is illustrated in figure 1. this shows the curves p^2 and $(1-p)^2$ plotted against p. That is, it shows the losses if S is false and if it is true plotted against our choice of p. We can see that, to the left of 0 and to the right of 1, both curves increase. So, it is always better to have a value of p such that $0 \leq p \leq 1$.

2.5 A simple example

Let us consider a very simple illustrative example. Suppose we are interested in the proportion of animals in some population which carry a particular gene and we have a means of determining whether any given animal does. (Very similar considerations will apply, of course, in many kinds of repeatable experiments with two possible outcomes). Suppose that we are going to test just two animals. There are only four possible outcomes and these are represented in table 1 which also assigns a probability to each of the possible outcomes. For now just regard these as the subjective probabilities of someone who has thought about what might happen.

The table also gives the *marginal* probabilities for each animal individually. We see that these are the same. The animals are said to be *exchangeable*.

Let S_1 be the statement, or *proposition*, that the first animal has the gene and S_2 be the proposition that the second animal has the gene. We write $\Pr(A)$ for the probability of a proposition A. We see that $\Pr(S_1) = \Pr(S_2) = 0.2$. Some readers will have noticed, however, that S_1 and S_2 are not *independent* since the probability of S_1 *and* S_2, $\Pr(S_1 \wedge S_2) = 0.06 > \Pr(S_1)\Pr(S_2) = 0.2 \times 0.2 = 0.04$. This is quite deliberate and reflects the fact that we do not know the underlying population proportion of animals with the gene. This is what makes it possible to learn from observation of one animal about what we are likely to see in another.

The *conditional probability* that the second animal has the gene given that the first does is

$$\Pr(S_2|S_1) = \frac{\Pr(S_1 \wedge S_2)}{\Pr(S_1)} = \frac{0.06}{0.2} = 0.3$$

Now, what happens to our beliefs about the second animal if we actually observe that the first has the gene? Well, unless something else causes us to change our minds, our probability that

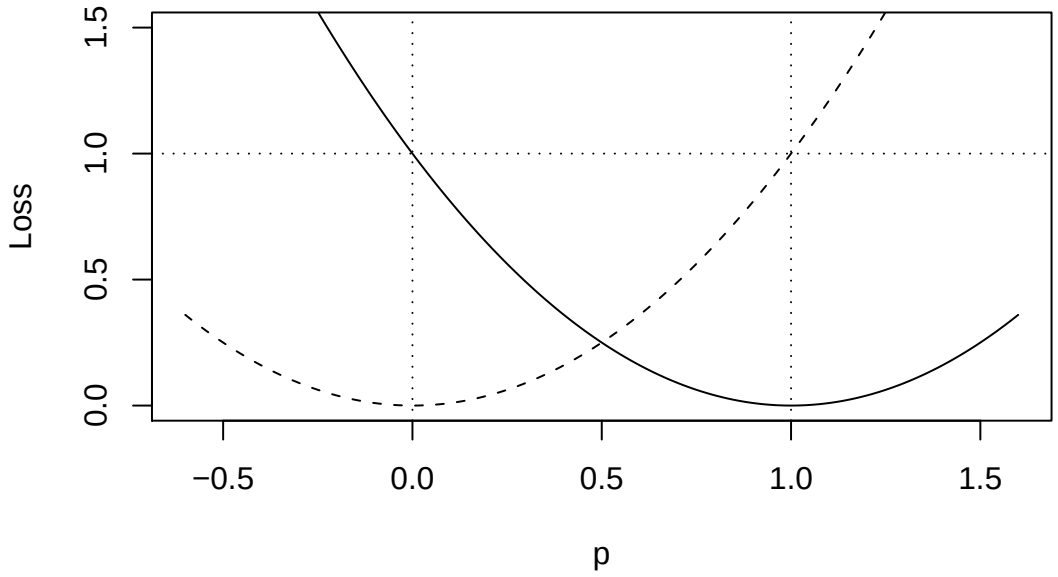

Figure 1: Losses if S is true (solid line) and if S is false (dashes) as a function of declared probability p.

the second animal has the gene should become the conditional probability that the second animal has the gene given that the first does. That is 0.3. There are deep philosophical arguments about whether you should *have* to adjust your beliefs in this way. You could have coherent beliefs before you see the first animal and a different set of coherent beliefs afterwards so that, at any point in time you are coherent. This Bayesian updating of beliefs requires beliefs before and after the observation to cohere with each other. Suffice it to say that, if you do not adjust beliefs in this way, you need to ask yourself why. It was the rational thing to do *before* you saw the data. If you programmed a computer to take observations and make decisions for you, and, of course, scientists are increasingly doing just this sort of thing, then, in order to be coherent at the moment you switch it on and leave it to run, you must program it to update beliefs in this way.

This adjustment of probabilities from *prior* probabilities to *posterior* probabilities by conditioning on observed data is fundamental to Bayesian inference.

The uncertainties involved in this example can be thought of as falling into two categories. Some uncertainty arises because we do not know the underlying population proportion of animals with the gene. There is only one correct answer to this but we do not know it. Such uncertainty is described as *epistemic*. There is also uncertainty to do with the selection of animals to test. We might happen to choose one with the gene or one without. Uncertainty caused by such "randomness" is described as *aleatory*. Similarly, in identifying a protein from a mass spectrum there is epistemic uncertainty in that we do not know the true identity of the protein and aleatory uncertainty in the various kinds of "noise" which affect the observed spectrum.

2.6 Bayes' Theorem

Let $S_1, \ldots, S_n$ be events which form a *partition*. That is $\Pr(S_1 \vee \ldots \vee S_n) = 1$ and $\Pr(S_i \wedge S_j) = 0$ for any $i \neq j$. In other words one and only one of the events $S_1, \ldots, S_n$ must occur or one and only one of the statements $S_1, \ldots, S_n$ must be true. Let D be some other event (or statement). Provided $\Pr(S_i) \neq 0$, the conditional probability of D given S_i is

$$\Pr(D \mid S_i) = \frac{\Pr(S_i \wedge D)}{\Pr(S_i)}.$$

Hence the joint probability of S_i and D can be written as

$$\Pr(S_i \wedge D) = \Pr(S_i) \Pr(D \mid S_i).$$

Then the *law of total probability* says that

$$\begin{aligned} \Pr(D) &= \sum_{i=1}^{n} \Pr(S_i \wedge D) \\ &= \sum_{i=1}^{n} \Pr(S_i) \Pr(D \mid S_i) \end{aligned}$$

The conditional probability of S_k given D is

$$\begin{aligned} \Pr(S_k \mid D) &= \frac{\Pr(S_k \wedge D)}{\Pr(D)} \\ &= \frac{\Pr(S_k) \Pr(D \mid S_k)}{\sum_{i=1}^{n} \Pr(S_i) \Pr(D \mid S_i)}. \end{aligned}$$

2.7 Problems 1

1. Let E_1, E_2, E_3 be events. Let I_1, I_2, I_3 be the corresponding indicators so that $I_1 = 1$ if E_1 occurs and $I_1 = 0$ otherwise.

   (a) Let $I_A = 1 - (1 - I_1)(1 - I_2)$. Verify that I_A is the indicator for the event A where $A = (E_1 \vee E_2)$ (that is "E_1 or E_2") and show that

   $$\Pr(A) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \wedge E_2)$$

where $(E_1 \wedge E_2)$ is "E_1 and E_2".

(b) Find a formula, in terms of I_1, I_2, I_3 for I_B, the indicator for the event B where $B = (E_1 \vee E_2 \vee E_3)$ and derive a formula for $\Pr(B)$ in terms of $\Pr(E_1), \Pr(E_2), \Pr(E_3), \Pr(E_1 \wedge E_2), \Pr(E_1 \wedge E_3), \Pr(E_2 \wedge E_3), \Pr(E_1 \wedge E_2 \wedge E_3)$.

2. In a certain place it rains on one third of the days. The local evening newspaper attempts to predict whether or not it will rain the following day. Three quarters of rainy days and three fifths of dry days are correctly predicted by the previous evening's paper. Given that this evening's paper predicts rain, what is the probability that it will actually rain tomorrow?

3. A machine is built to make mass-produced items. Each item made by the machine has a probability p of being defective. Given the value of p, the items are independent of each other. Because of the way in which the machines are made, p could take one of several values. In fact $p = X/100$ where X has a discrete uniform distribution on the interval $[0, 5]$. The machine is tested by counting the number of items made before a defective is produced. Find the conditional probability distribution of X given that the first defective item is the thirteenth to be made.

4. There are five machines in a factory. Of these machines, three are working properly and two are defective. Machines which are working properly produce articles each of which has independently a probability of 0.1 of being imperfect. For the defective machines this probability is 0.2.

   A machine is chosen at random and five articles produced by the machine are examined. What is the probability that the machine chosen is defective given that, of the five articles examined, two are imperfect and three are perfect?

5. A crime has been committed. Assume that the crime was committed by exactly one person, that there are 1000 people who could have committed the crime and that, in the absence of any evidence, these people are all equally likely to be guilty of the crime.

   A piece of evidence is found. It is judged that this evidence would have a probability of 0.99 of being observed if the crime were committed by a particular individual, A, but a probability of only 0.0001 of being observed if the crime were committed by any other individual.

   Find the probability, given the evidence, that A committed the crime.

6. In an experiment on extra-sensory perception (ESP) a person, A, sits in a sealed room and points at one of four cards, each of which shows a different picture. In another sealed room a second person, B, attempts to select, from an identical set of four cards, the card at which A is pointing. This experiment is repeated ten times and the correct card is selected four times.

   Suppose that we consider three possible states of nature, as follows.

   State 1 : There is no ESP and, whichever card A chooses, B is equally likely to select any one of the four cards. That is, subject B has a probability of 0.25 of selecting the correct card.

   Before the experiment we give this state a probability of 0.7.

   State 2 : Subject B has a probability of 0.50 of selecting the correct card.

   Before the experiment we give this state a probability of 0.2.

   State 3 : Subject B has a probability of 0.75 of selecting the correct card.

   Before the experiment we give this state a probability of 0.1.

   Assume that, given the true state of nature, the ten trials can be considered to be independent.

   Find our probabilities after the experiment for the three possible states of nature.

   Can you think of a reason, apart from ESP, why the probability of selecting the correct card might be greater than 0.25?

7. In a certain small town there are n taxis which are clearly numbered $1, 2, \ldots, n$. Before we visit the town we do not know the value of n but our probabilities for the possible values of n are as follows.

n	0	1	2	3	4
Probability	0.00	0.11	0.12	0.13	0.14
n	5	6	7	8	≥ 9
Probability	0.14	0.13	0.12	0.11	0.00

On a visit to the town we take a taxi which we assume would be equally likely to be any of taxis $1, 2, \ldots, n$. It is taxi number 5. Find our new probabilities for the value of n.

8. A dishonest gambler has a box containing 10 dice which all look the same. However there are actually three types of dice.

- There are 6 dice of type A which are fair dice with $\Pr(6 \mid A) = 1/6$ (where $\Pr(6 \mid A)$ is the probability of getting a 6 in a throw of a type A die).
- There are 2 dice of type B which are biassed with $\Pr(6 \mid B) = 0.8$.
- There are 2 dice of type C which are biassed with $\Pr(6 \mid C) = 0.04$.

The gambler takes a die from the box at random and rolls it. Find the conditional probability that it is of type B given that it gives a 6.

9. In a forest area of Northern Europe there may be wild lynx. At a particular time the number X of lynx can be between 0 and 5 with

$$\Pr(X = x) = \binom{5}{x} 0.6^x 0.4^{5-x} \quad (x = 0, \ldots, 5).$$

A survey is made but the lynx is difficult to spot and, given that the number present is x, the number Y observed has a probability distribution with

$$\Pr(Y = y \mid X = x) = \begin{cases} \binom{x}{y} 0.3^y 0.7^{x-y} & (0 \leq y \leq x) \\ 0 & (x < y) \end{cases}.$$

Find the conditional probability distribution of X given that $Y = 2$.
(That is, find $\Pr(X = 0 \mid Y = 2), \ldots, \Pr(X = 5 \mid Y = 2)$).

10. A particular species of fish makes an annual migration up a river. On a particular day there is a probability of 0.4 that the migration will start. If it does then an observer will have to wait T minutes before seeing a fish, where T has an exponential distribution with mean 20 (i.e. an exponential(0.05) distribution). If the migration has not started then no fish will be seen.

(a) Find the conditional probability that the migration has not started given that no fish has been seen after one hour.

(b) How long does the observer have to wait without seeing a fish to be 90% sure that the migration has not started?

2.8 Homework 1

Solutions to Questions **8, 9, 10** *of Problems 1 are to be submitted in the Homework Letterbox no later than 4.00pm on Monday February 9th.*