MS-A0503 First course in probability and statistics

5B More about Bayesian inference

Jukka Kohonen

Deparment of mathematics and systems analysis
Aalto SCI

Academic year 2020–2021
Period III

Contents

How to use the posterior distribution

Congratulations, you have the posterior distribution of the unknown parameter Θ. How can you use the distribution?

Like any distribution, you can use it in many ways, depending on

1. what question you want to answer
2. what is convenient to calculate.

Some typical uses:

- mode of the posterior distribution = where it is maximized
- mean of the posterior distribution = probability-weighted average
- median of the posterior distribution = 50% probability below
- credible interval, containing e.g. 95% of posterior probability
- report/visualize the full posterior distribution
- predictions of future data, based on posterior

Next, we will look closer into each alternative.

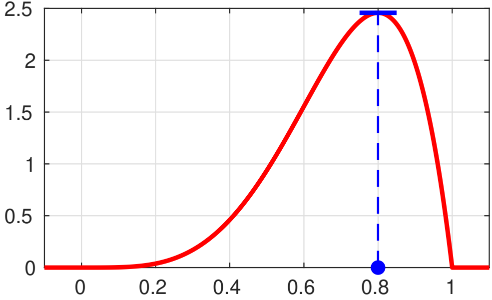
Posterior mode (MAP = **M**aximum **A** **P**osteriori estimate)

Unknown coin, uniform prior, observed 4 heads, 1 tails.
Posterior is Beta(5,2), with density (for $0 \leq \theta \leq 1$)

$$f(\theta \mid \vec{x}) = 30\theta^4(1-\theta).$$

To find the mode, inspect zeros of derivative, and ends of interval.

(The normalizing constant 30 plays no role in the maximization, so we could as well use the unnormalized posterior. Also compare to ML estimate.)



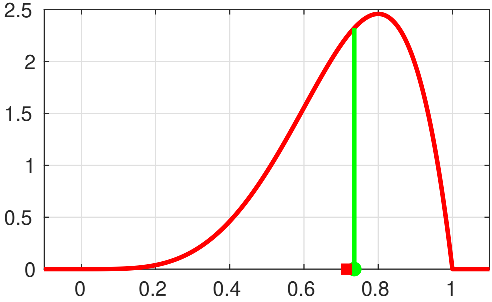
Mode = MAP estimate = 0.8

Posterior mean and median

Unknown coin, uniform prior, observed 4 heads, 1 tails.
Posterior is Beta(5,2), with density (for $0 \le \theta \le 1$)

$$f(\theta \mid \vec{x}) = 30\theta^4(1-\theta).$$

For mean, do the integral.
For median, solve where CDF=0.5. (R `qbeta` / Matlab `betainv`)



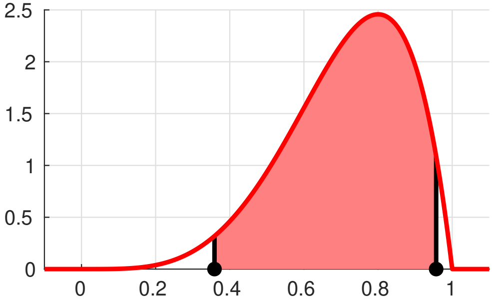
Mean = $5/7 \approx 0.7143$
Median ≈ 0.7356

Credible interval

Unknown coin, uniform prior, observed 4 heads, 1 tails.
Posterior is Beta(5,2), with density (for $0 \leq \theta \leq 1$)

$$f(\theta \mid \vec{x}) = 30\theta^4(1-\theta).$$

Find points where CDF is 0.025 and 0.975. qbeta / betainv
$\Rightarrow \Theta$ is between those points with 95% probability.



$$\mathbb{P}(\mathbf{0.3588} \leq \Theta \leq \mathbf{0.9567} \mid \vec{X} = \vec{x}) = 95\%$$

Prediction of future data

- The posterior distribution of Θ is our best knowledge of what the parameter value can be (combining prior and data).
- Usually the posterior distribution is **not** a single point. This openly shows our uncertainty; we do not pretend that we know the parameter value exactly.
- But the more data we obtain, the more precise the posterior becomes.

Question. After seeing five observations $\vec{x} = (1, 1, 1, 1, 0)$, we have the posterior $\Theta \sim \text{Beta}(5, 2)$.
What can we say about the next observation?

Answer. We form the (posterior) predictive distribution for it, applying law of total probability (= consider all possibilities and add up).

Prediction of future data (coin example)

We have five observations $\vec{x}$, and wish to predict next observation Y. From the law of total probability, we have

$$f_{Y|\vec{X}}(y \mid \vec{x}) = \int f(y \mid \theta) f(\theta \mid \vec{x}) d\theta.$$

Different values of θ give different predictions for Y. These predictions are averaged, weighted by the posterior density of Θ.

This gives our best understanding of Y, considering what we now know about Θ.

- We do **not** choose just one value of θ, perhaps the "most probable" one, and use that as the probability of $Y = 1$. That might give quite erroneous predictions.
- We do not even reject the 5% tails; they are included in the calculation (and for good reason: they might actually affect the prediction, see following).

Prediction of future data (coin example)

5 observations $\vec{x} = (1,1,1,1,0)$, predict next observation Y.

- stochastic model as before: $\mathbb{P}(Y = 1 \mid \Theta = \theta) = \theta$.
- posterior for Θ is Beta(5,2).

So calculate:

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid \vec{X} = \vec{x}) &= f_{Y|\vec{X}}(1 \mid \vec{x}) \\
&= \int f_{Y|\Theta}(1 \mid \theta)\; f_{\Theta|\vec{X}}(\theta \mid \vec{x}) d\theta \\
&= \int_0^1 \theta \cdot 30\theta^4(1-\theta) d\theta \\
&= 30 \int_0^1 (\theta^5 - \theta^6) d\theta \\
&= 30 \cdot \left(\frac{1}{6} - \frac{1}{7}\right) \approx \mathbf{0.7143}.
\end{aligned}$$

For predicting one more data point, our probability is simply the posterior mean of Θ. But don't get too carried away ...

Prediction of future data (more predictions)

5 observations $\vec{x} = (1,1,1,1,0)$, calculate probability for $\vec{Y} = (1,1,1)$ (that next three are heads).

- stochastic model $\mathbb{P}(\vec{Y} = (1,1,1) \mid \Theta = \theta) = \theta^3$.
- posterior for Θ is Beta(5,2).

$$\begin{aligned}
\mathbb{P}(\vec{Y} = (1,1,1) \mid \vec{X} = \vec{x}) &= f_{\vec{Y}|\vec{X}}(1,1,1 \mid \vec{x}) \\
&= \int f_{\vec{Y}|\Theta}(1,1,1 \mid \theta)\, f_{\Theta|\vec{X}}(\theta \mid \vec{x}) d\theta \\
&= \int_0^1 \theta^3 \cdot 30\theta^4(1-\theta) d\theta \\
&= 30 \int_0^1 (\theta^7 - \theta^8) d\theta \\
&= 30 \cdot \left(\frac{1}{8} - \frac{1}{9}\right) \approx \mathbf{0.4167}.
\end{aligned}$$

It is not $0.7143^3 \approx 0.3645$, but bigger. Note where the cube goes.

Prediction of future data — drastic effect of uncertainty

Being honest about our uncertainty of Θ can have a big effect on predictive distributions.

We could do this with the continuous-θ coin, but let us do a simpler **discrete** example.

Consider the following two models:

- Model A: We have a fair coin, $\Theta = 0.5$ certainly.
- Model B: We have a coin that might be unfair: Θ is either 0, 0.5 or 1, with probabilities 0.01, 0.98, 0.01 respectively.

For predicting one result, the models are equivalent. Each says that the next result is heads with 50% probability.

For predicting next 100 results, the models disagree strongly.

Prediction of future data — drastic effect of uncertainty

What is the probability for the next 100 tosses to be all heads?

Model A: We know the coin is fair ($\Theta = 0.5$).

- Number of heads has Bin(100, 0.5) distribution, so
- 100 heads with probability $1/2^{100} \approx 8 \cdot 10^{-31}$

Model B: Value of Θ is 0, 0.5 or 1, with probabilities 0.01, 0.98, 0.01. (This could be our posterior from a small number of experiments.)

- By law of total probability, prob. of 100 heads is

$$(0.01 \cdot 0) + (0.98 \cdot 8 \cdot 10^{-31}) + (\mathbf{0.01} \cdot \mathbf{1}) \approx 0.01$$

Observe: If Model B is the best we know, then

- Using just the mode ($\theta = 0.5$) would go wrong
- Using just the mean ($\theta = 0.5$) would go wrong
- Rejecting "5% tails" would get rid of the two extreme possibilities, and would go wrong

Keep the uncertainty in your calculations and you get more truthful results!

Contents

Multiple categories

We worked with the binary model: data were 0-1-valued (or their counts), and we had a single probability parameter, discrete or continuous.

Next we consider sequences of categorical (nominal) data that have several categories (more than two).

Examples:

- Rolls of a loaded die (3,6,6,2,6,1,3,4,6,6)
- Party stances in a sample (A,B,A,A,C,B,A,A,C,C)
- DNA sequence with four bases chosen randomly GTCTACCAG...
- Text, as a sequence of words, each word chosen randomly with some probabilities (the, quick, brown, fox, jumped, over, the, lazy, dog)

You can view the data either as a sequence of categorical variables, or as a vector of counts of the different values.

Multinomial model

- n independent observations $(X_1, X_2, \dots, X_n)$.
- Each X_i from the same discrete distribution over k possibilies
- The distribution has k probability parameters $\vec{p} = (p_1, p_2, \dots, p_k)$
- We can treat the probabilities as unknown, a random vector $\vec{P} = (P_1, P_2, \dots, P_k)$

We can use the familiar methods:

- Assume a prior distribution $f_{\vec{P}}(\vec{p})$
- Assume a stochastic model $f(x \mid \vec{p})$ (likelihood)
- After observations, work out posterior $f(\vec{p} \mid x)$

Stochastic model — Three-category example

A large population contains supporters of three parties A, B, C with proportions $\vec{p} = (p, q, r) = (0.5, 0.3, 0.2)$.

A random sample of $n = 10$ people is taken. Each person sampled has the probabilities $\vec{p}$ for the three parties.

Two questions:

- What kinds of (ordered) sequences are we likely to observe?
  example: AAAAAAAAAA or AAAABBBBCC
- What kinds of count vectors are we likely to observe?
  example: $(10, 0, 0)$ or $(4, 4, 2)$

For example,

- $\mathbb{P}(\text{AAAAAAAAAA}) = p^{10} \approx 0.000977$ Small
- $\mathbb{P}(\text{AAAABBBBCC}) = p^4 q^4 r^2 \approx 0.000020$ Smaller!?

Stochastic model — Three-category example

From elementary combinatorics, we know there are $3^{10} = 59049$ different 10-person strings from three letters. Let us list them, grouped by the counts of A,B,C. Recall $(p, q, r) = (0.5, 0.3, 0.2)$.

sequence	letter counts	$\mathbb{P}$(sequence)	
AAAAAAAAAA	$(10, 0, 0)$	$p^{10} = 0.000977$	} 1 sequence
...			
AAAABBBBCC	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	
BBCAABBAAC	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	
AABCCAABBB	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	} 3150 seq.
...			
CCBBBBAAAA	$(4, 4, 2)$	$p^4 q^4 r^2 = 0.000020$	
...			
CCCCCCCCCC	$(0, 0, 10)$	$r^{10} = 0.0000001$	} 1 sequence

$\mathbb{P}(\text{counts are } 10,0,0) = 1 \times 0.000977 \approx \mathbf{0.1\%}$

$\mathbb{P}(\text{counts are } 4,4,2) = 3150 \times 0.000020 = 0.0638 \approx \mathbf{6.4\%}$

Interlude — multinomial coefficients

Where did we get 3150 on the previous slide?
It is a multinomial coefficient, which tells: in how many ways can you order 4 A's, 4 B's and 2 C's, into a sequence of 10 letters.

Count the ways by the combinatorial product rule (Ross's "basic principle of counting") and the binomial coefficient:

- From the 10 places, choose 4 for the A: $\binom{10}{4} = 210$ ways
- From the remaining 6 places, choose 4 for B: $\binom{6}{4} = 15$ ways
- (From the remaining 2 places, choose 2 for C: $\binom{2}{2} = 1$ ways)

Product rule: $210 \cdot 15 \cdot 1 = 3150$ ways of placing the letters.

This can be written as the multinomial coefficient

$$\binom{10}{4, 4, 2} = \frac{10!}{4!\ 4!\ 2!} = 3150.$$

All possible count vectors, and their probabilities

$3^{10} = 59049$ different **sequences**, but 66 different **count vectors**.

(5,3,2)	**0.0851**	(2,5,3)	0.0122	(4,0,6)	0.000840
(6,2,2)	0.0709	(2,4,4)	0.0102	(2,8,0)	0.000738
(6,3,1)	0.0709	(4,6,0)	0.0096	(1,3,6)	0.000726
(4,4,2)	0.0638	(2,6,2)	0.0092	(1,8,1)	0.000590
(5,4,1)	0.0638	(3,2,5)	0.0091	(2,1,7)	0.000346
(5,2,3)	0.0567	(4,1,5)	0.0076	(0,6,4)	0.000245
(4,3,3)	0.0567	(7,0,3)	0.0075	(0,7,3)	0.000210
(7,2,1)	0.0506	(8,0,2)	0.0070	(1,2,7)	0.000207
(4,5,1)	0.0383	(9,1,0)	0.0059	(0,5,5)	0.000196
(3,4,3)	0.0340	(2,3,5)	0.0054	(3,0,7)	0.000192
(7,1,2)	0.0338	(6,0,4)	0.0053	(0,8,2)	0.000118
(6,1,3)	0.0315	(2,7,1)	0.0039	(0,4,6)	0.000109
(3,5,2)	0.0306	(9,0,1)	0.0039	(1,9,0)	0.000098
(4,2,4)	0.0284	(3,7,0)	0.0033	(0,3,7)	0.000041
(6,4,0)	0.0266	(5,0,5)	0.0025	(0,9,1)	0.000039
(7,3,0)	0.0253	(1,6,3)	0.0024	(1,1,8)	0.000035
(3,3,4)	0.0227	(1,5,4)	0.0024	(2,0,8)	0.000029
(8,1,1)	0.0211	(3,1,6)	0.0020	(0,2,8)	0.000010
(5,5,0)	0.0191	(2,2,6)	0.0018	(0,10,0)	0.000006
(5,1,4)	0.0189	(1,4,5)	0.0016	(1,0,9)	0.000003
(8,2,0)	0.0158	(1,7,2)	0.0016	(0,1,9)	0.000002
(3,6,1)	0.0153	(10,0,0)	0.000977	(0,0,10)	0.000000

Multinomial model — A discrete prior

For the probability parameter vector $\vec{P}$, in different situations we can have different kinds of priors.

Sometimes we just have a few possible values of the vector, perhaps just two, so the prior distribution is discrete.

E.g. we have just two kinds of dice in a bag: 9 fair and 1 loaded, and we know the loading. A randomly chosen die is

- with prob. 0.9 fair, with $\vec{p} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$
- with prob. 0.1 loaded, with $\vec{p} = (0.1, 0.1, 0.1, 0.1, 0.1, 0.5)$

Pick a random die, roll it four times with results $(3, 2, 6, 6)$. The likelihoods for the two possible parameter values are

- for a fair die: $(\frac{1}{6})^4 \approx 0.00077$
- for the loaded die: $0.1 \cdot 0.1 \cdot 0.5 \cdot 0.5 = 0.00250$

After this observation, we would have increased posterior probability for the die to be the loaded one. (But not certainty!)

Multinomial model — A continuous prior

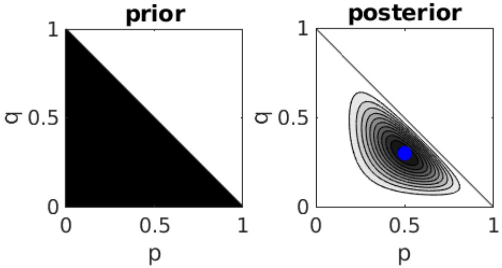
Or perhaps the values of the three probability parameters (p, q, r) are unknown real numbers. How do we handle this situation?

- Certainly all three are in the interval $[0, 1]$.
- They are not three freely chosen parameters, because we must have $p + q + r = 1$.
- We can consider a two-element parameter vector (p, q), and then $r = 1 - (p + q)$.
- We need $p \geq 0$ and $q \geq 0$ and $p + q \leq 1$. So (p, q) is constrained to be in a triangular area. (Picture!)
- Let the prior be the uniform density over the triangle,

$$f_{P,Q}(p, q) = 2 \qquad\qquad \text{if } p, q \geq 0 \text{ and } p + q \leq 1.$$

- We now have the likelihood and the prior, so we can proceed with Bayesian inference.

Multinomial model — Inference



After observing counts $(5, 3, 2)$, the posterior density of (P, Q) is

$$f(p, q \mid \vec{x}) = c \cdot p^5 q^3 (1 - p - q)^2$$

in the triangle, and c is again normalizing constant.
We can use the posterior density to compute posterior mode, posterior mean, 95% credible region, predictions etc. Posterior mode here shown as blue dot.

Contents

Choice of prior

Sometimes people are worried about the apparent subjectivity of Bayesian inference. If you want to report a certain posterior distribution you like, you could choose your prior so that you get the posterior you wanted?

- You should be honest in making your prior to be a fairly good representation of what is known about Θ before the data.
- Uniform priors often work out nice. Not always, in complicated models.
- Beware of assigning zero density to some parameter values that might actually be true. Zero prior leads to zero posterior, whatever your data are.
- With lots of data, the effect of the prior diminishes as the "data speaks for itself".
- When reporting your results, report the model and prior you used. Then your results are completely objective: anyone using that prior will get the same posterior.

Next week: Significance tests...