Bayesian probability theory

Brendon J. Brewer

Department of Statistics, The University of Auckland

https://www.stat.auckland.ac.nz/~brewer

@brendonbrewer

Propositions

Propositions are *statements that can be either true or false*.
Examples:

A: Hillary Clinton will be president of the USA on June 14th, 2017.
B: The current temperature in Auckland is greater than 14°C.
C: $\Omega_\Lambda \neq 0$.
D: David Hogg has had more than 500 mg of caffeine today.

Propositions

Propositions can be *combined* to make others, using the operators **and** ($\wedge$), **or** ($\vee$), and **not** ($\neg$):

$A \wedge B$
$C \vee A$
$\neg D$
$\neg(A \vee (C \wedge D))$

These things satisfy *Boolean Algebra*.

Notation

The operator **and** is sometimes denoted with a comma.

$$(A \wedge B) \equiv (A, B) \tag{1}$$

Quantifying propositions

Science is largely concerned with the *plausibility* of propositions. Plausibility depends on the information you have, so is a function of *two* propositions. Examples:

$P(A|B)$
$P(\neg D|(C \vee D))$
$P((C \wedge D)|\neg A)$

Read the "$|$" as "**given**" or "**conditional on**".

Constraints on plausibility

Some properties of plausibility:

$P(A \vee B) \geq P(A)$
$P(A \vee (B \vee C)) = P((A \vee B) \vee C)$

Where'd the RHS go? $||D$

Constraints on plausibility

Some properties of plausibility:

$$P(A \vee B) \geq P(A)$$
$$P(A \wedge B) \leq P(A)$$
$$P(A \vee (B \vee C)) = P((A \vee B) \vee C)$$

These imply the **sum rule**. Some other symmetries imply the **product rule**, and hence that **plausibilities are probabilities**.

Sum and product rule

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

$$P(A \wedge B) = P(A)P(B|A) = P(B)P(A|B)$$

Bayes' Rule $||I$

Bayes' Rule (consequence of product rule and commutativity of **and**)

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)} \tag{2}$$

where $P(D) = P(H)P(D|H) + P(\neg H)P(D|\neg H)$ (consequence of sum

rule)

Terminology $||I$

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)} \tag{3}$$

$$(\text{posterior probability}) = \frac{(\text{prior probability}) \times (\text{likelihood})}{(\text{marginal likelihood})} \tag{4}$$

The marginal likelihood is also known as the *evidence*. This is a physicist thing.

Another Bayes' Rule $||I$

For N mutually exclusive, exhaustive hypotheses $H_1, H_2, ..., H_N$, we have N posterior probabilities:

$$P(H_i|D) = \frac{P(H_i)P(D|H_i)}{P(D)} \tag{5}$$

where

$$P(D) = \sum_{i=1}^{N} P(H_i)P(D|H_i) \tag{6}$$

Another Bayes' Rule

This is the most important form of Bayes' rule.

$$P(H_i|D) = \frac{P(H_i)P(D|H_i)}{P(D)} \tag{7}$$

where

$$P(D) = \sum_{i=1}^{N} P(H_i)P(D|H_i) \tag{8}$$

Exercises!

Do questions 1 and 2 on the exercise sheet!

Parameter estimation

In most applications, we can use the "parameter estimation" story. E.g. Let θ be a quantity we want to know. Then the hypotheses might be:

$$
\begin{aligned}
H_1 &\equiv \theta = 5 \qquad (9)\\
H_2 &\equiv \theta = 6 \qquad (10)\\
H_3 &\equiv \theta = 7 \qquad (11)
\end{aligned}
$$

The data could also be a number. E.g. the conditioning proposition could be

$$D = 4 \qquad (12)$$

The previous version of Bayes' rule can be applied to each of the hypotheses.

Parameter estimation

In most applications, we can use the "parameter estimation" story. E.g. Let θ be a quantity we want to know. Then the hypotheses might be:

$$\begin{aligned} H_1 &\equiv \theta = 5 \quad (13) \\ H_2 &\equiv \theta = 6 \quad (14) \\ H_3 &\equiv \theta = 7 \quad (15) \end{aligned}$$

The data could also be a number. E.g. the conditioning proposition could be

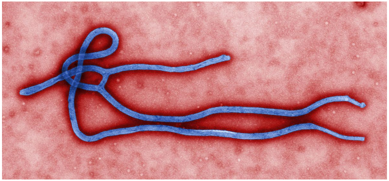
$$D = 4 \tag{16}$$

The previous version of Bayes' rule can be applied to each of the hypotheses.

Exercises

Do Question 3 on the exercise sheet!

Updating Probabilities: Example



A patient goes to the doctor because he as a fever. Define

$$
\begin{aligned}
H \equiv &\quad \text{“The patient has Ebola”} \\
\neg H \equiv &\quad \text{“The patient does not have Ebola”.}
\end{aligned}
$$

Updating Probabilities: Example

Based on all of her knowledge, the doctor assigns probabilities to the two hypotheses.

$$\begin{aligned} P(H) &= 0.01 \\ P(\neg H) &= 0.99 \end{aligned}$$

But she wants to test the patient to make sure.

Updating Probabilities: Example

The patient is tested. Define

$$\begin{aligned} D &\equiv \quad \text{“The \textbf{test says} the patient has Ebola”} \\ \neg D &\equiv \quad \text{“The \textbf{test says} the patient does not have Ebola”.} \end{aligned}$$

If the test were perfect, we'd have $P(D|H) = 1$, $P(\neg D|H) = 0$, $P(D|\neg H) = 0$, and $P(\neg D|\neg H) = 1$.

Updating Probabilities: Example

The Ebola test isn't perfect. Suppose there's a 5% probability it simply gives the wrong answer. Then we have:

$$\begin{aligned} P(D|H) &= 0.95 \\ P(\neg D|H) &= 0.05 \\ P(D|\neg H) &= 0.05 \\ P(\neg D|\neg H) &= 0.95 \end{aligned}$$

Updating Probabilities: Example

Overall, there are four possibilities, considering whether the patient has Ebola or not, and what the test says.

$$\begin{gathered}(H, D)\\(\neg H, D)\\(H, \neg D)\\(\neg H, \neg D)\end{gathered}$$

Updating Probabilities: Example

The probabilities for these four possibilities can be found using the product rule.

$$\begin{aligned} P(H, D) &= 0.01 \times 0.95 \\ P(\neg H, D) &= 0.99 \times 0.05 \\ P(H, \neg D) &= 0.01 \times 0.05 \\ P(\neg H, \neg D) &= 0.99 \times 0.95 \end{aligned}$$

These four possibilities are **mutually exclusive** (only one of them is true) and exhaustive (it's not "something else"), so the probabilities add up to 1.

Updating Probabilities: Example

The test results come back and say that the patient has Ebola. That is, we've learned that D is true. So we can confidently rule out those possibilities where D is false:

$$\begin{aligned}
P(H, D) &= 0.01 \times 0.95 \\
P(\neg H, D) &= 0.99 \times 0.05 \\
P(H, \neg D) &= 0.01 \times 0.05 \\
P(\neg H, \neg D) &= 0.99 \times 0.95
\end{aligned}$$

Updating Probabilities: Example

We are left with these two possibilities.

$$\begin{aligned} P(H, D) &= 0.01 \times 0.95 \\ P(\neg H, D) &= 0.99 \times 0.05 \end{aligned}$$

It would be strange to modify these probabilities just because we deleted the other two. The only thing we have to do is renormalise them, by dividing by the total, so they sum to 1 again.

Updating Probabilities: Example

Normalising, we get

$$\begin{aligned} P(H|D) &= (0.01 \times 0.95)/(0.01 \times 0.95 + 0.99 \times 0.05) = 0.161 \\ P(\neg H|D) &= (0.99 \times 0.05)/(0.01 \times 0.95 + 0.99 \times 0.05) = 0.839 \end{aligned}$$

Moral

Bayesian updating is completely equivalent to this procedure:

1. Write a list of possible answers to your question.
2. Give a probability to each.
3. Delete the ones you discover are false.
4. Renormalise the remaining probabilities.

It just seems more complicated because we often apply it to more complex sets of hypotheses.

Parameter estimation

If you want to infer the value of a quantity θ (a "parameter") from the value of another quantity D (some "data"), define the prior distribution $p(\theta)$ and the sampling distribution or likelihood $p(\theta|D)$, then calculate the **posterior distribution** $p(\theta|D)$:

$$p(\theta|D) \propto p(\theta)p(D|\theta) \tag{17}$$

This notation hides **a lot**!

Ingredients I

Bayesian parameter estimation needs the following inputs:

- A **hypothesis space** describing the set of possible answers to our question ("parameter space" in fitting is the same concept).

- A **prior distribution** $p(\theta)$ describing how plausible each of the possible solutions is, not taking into account the data.

Ingredients II

Bayesian parameter estimation needs the following inputs:

- A **sampling distribution** $p(D|\theta)$ describing our knowledge about the connection between the parameters and the data.

When D is known, this is a function of θ called the **likelihood**.

The Posterior Distribution

The data helps us by changing our prior distribution to the **posterior distribution**, given by

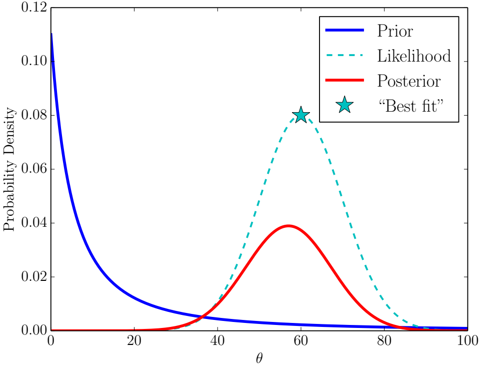
$$p(\theta|D) \quad = \quad \frac{p(\theta)p(D|\theta)}{p(D)}$$

where the denominator is the normalisation constant, usually called either the **marginal likelihood** or the **evidence**.

$$p(D) \quad = \quad \int p(\theta)p(D|\theta)\,d\theta.$$

Posterior Distribution vs. Maximum Likelihood

The practical difference between these two concepts is greater in higher dimensional problems.



Question Sheet

Do Question 4!

Parameter estimation

For a single unknown parameter, it's quite easy to do things numerically on a grid.

```
x = 5 # The data
mu = np.linspace(0.001, 30., 1001) # The hypothesis space
prior = 1./mu
likelihood = mu**5*np.exp(-mu)/scipy.misc.factorial(5)
posterior = prior*likelihood/np.trapz(prior*likelihood, x=mu)
plt.plot(mu, posterior)
plt.show()
```

In practice we often use the logarithms of the prior, likelihood, and posterior to avoid numerical problems.

Choosing priors

There are three broad approaches to assigning priors:

1. Use symmetries or otherwise try to define "prior ignorance". Often called "**objective Bayes**"
2. Carefully interview an expert in the field and try to capture their degrees of confidence accurately. Often called "**subjective Bayes**"
3. Just choose uniform, log-uniform, and gaussian distributions based on **tradition**.

Check your priors for **unintended consequences**.

Unintended consequences: example

Let $\theta =$ the mass of a galaxy, in solar masses.
"Prior ignorance" might motivate this prior:

$$\theta \sim U(0, 10^{15}).$$

Why use the log-uniform prior?

“Prior ignorance” might motivate this prior:

$$\theta \sim U(0, 10^{15}).$$

But this implies:

$$\begin{aligned} P(\theta \geq 10^{14}) &= 0.9 \\ P(\theta \geq 10^{12}) &= 0.999. \end{aligned}$$

i.e. we are not ignorant at all, with respect to some questions!

Why use the log-uniform prior?

$$\log_{10}(\theta) \sim U(5, 15).$$

implies:

$$\begin{aligned} P(\theta \geq 10^{14}) &= 0.1 \\ P(\theta \geq 10^{12}) &= 0.3 \end{aligned}$$

or

$$P(\theta \in [10^{10}, 10^{11}]) = P(\theta \in [10^{11}, 10^{12}]) = P(\theta \in [10^{12}, 10^{13}])...$$

The log-uniform prior

The log-uniform prior has density $p(\theta) \propto 1/\theta$.

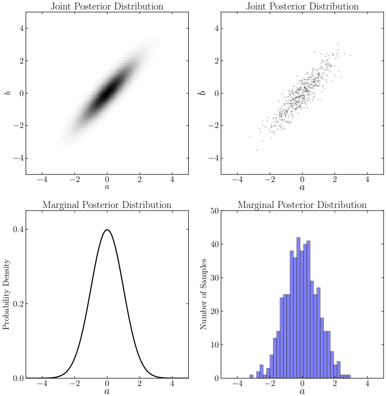
Common prior assumptions

Independent and identically distributed likelihood:

$$p(\mathbf{D}|\theta) = \prod_{i=1}^{N} f(D_i|\theta) \tag{18}$$

Marginalisation

This is why MCMC (next session) is useful when there's more than one unknown parameter, but it's also related to **prediction** in a Bayesian framework.



Prediction

$$\begin{aligned} p(D_{\mathrm{new}}|D) &= \int p(D_{\mathrm{new}}, \theta|D)\, d\theta && (19)\\ &= \int p(\theta|D) p(D_{\mathrm{new}}|\theta, D)\, d\theta && (20)\\ &= \int p(\theta|D) p(D_{\mathrm{new}}|\theta)\, d\theta && (21)\\ & && (22) \end{aligned}$$

(usually)
(posterior expectation of the conditional distribution for the prediction)

Question 5

There's a prediction question!