Applied Bayesian Data Analysis — Chapter 5

Kim Albertsson

CERN and Luleå University of Technology

kim.albertsson@ltu.se

November 13, 2019

Chapter 5

Gaining intuition about Bayes' rule

Introduction

Bayes' rule is a *central* concept of Bayesian statistics

$$p(c|r) = \frac{p(r|c)p(c)}{p(r)} \tag{5.5}$$

One primary usecase: Estimate probability of model paramters given data.

Derivation of Bayes' Rule

$$p(c|r) = \frac{p(r,c)}{p(r)} \tag{5.1}$$

$$p(r,c) = p(c|r)p(r) \tag{5.2}$$

$$p(r,c) = p(r|c)p(c) \tag{5.3}$$

$$p(c|r)p(r) = p(r|c)p(c) \tag{5.4}$$

Bayes' Rule:

$$p(c|r) = \frac{p(r|c)p(c)}{p(r)} \tag{5.5}$$

Remember that:

$$p(R=r) = \int p(R=r, C=c)\, dc$$

$$p(C=c|R=r) = \frac{p(R=r|C=c)p(C=c)}{\int p(R=r, C=c')\, dc'}$$

A Note on Notation

Given a probability distribution $p(r, c)$, what does $p(0.5)$ signify?

Unclear!

More clear alternatives: $p_r(0.5)$ or $p(R = 0.5)$

Mathematically there is no problem with: $p(R = c, C = r)$.
Keep track of you variables and observables!

Bayes' rule intuited from a two-way discrete table

$$p(c|r) = \frac{p(r|c)p(c)}{p(r)} \tag{5.5}$$

$$p(\mathrm{E} = \mathrm{Blue}|\mathrm{H} = \mathrm{Red}\}) = \frac{p(\mathrm{H} = \mathrm{Red}\}|\mathrm{E} = \mathrm{Blue})p(\mathrm{E} = \mathrm{Blue})}{p(\mathrm{H} = \mathrm{Red})}$$

Joint probabilities

	Black	Brown	Red	Blond	$\sum$
Brown	0.11	0.20	0.04	0.01	0.37
Blue	0.03	0.14	0.03	0.16	0.36
Hazel	0.03	0.09	0.02	0.02	0.16
Green	0.01	0.05	0.02	0.03	0.11
$\sum$	0.18	0.48	0.12	0.21	1.0

Example: Test for Disease

Diagnosis of rare disease.

Suppose one in a thousand has it. $\implies p(\theta = \ddot\frown) = 0.001$ Suppose test with true postitive rate 0.99 $\implies p(T = +|\theta = \ddot\frown) = 0.99$ Suppose also false postitive rate 0.05 $\implies p(T = +|\theta = \ddot\smile) = 0.05$

Question: Given a postitive test, what is the probability that the subject is ill?

$$p(\theta = \ddot\frown | T = +) = \frac{p(T = +|\theta = \ddot\frown)p(\theta = \ddot\frown)}{p(T = +)}$$

$$\begin{aligned} p(\theta = \ddot\smile) &= 1 - p(\theta = \ddot\frown) = 0.999 \\ p(T = +) &= p(T = +|\theta = \ddot\frown)p(\theta = \ddot\frown) \\ &\quad + p(T = +|\theta = \ddot\smile)p(\theta = \ddot\smile) \end{aligned}$$

Applied to Parameters and Data

Central question: Given some data, how likely are our model parameters?

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

Close analogy to the disease example, imagine big table.

$$\begin{aligned} p(\theta|X) &: \text{posterior} \\ p(X|\theta) &: \text{likelihood} \\ p(\theta) &: \text{prior} \\ p(X) &: \text{evidence} \end{aligned}$$

Data Order Invariance

h denotes probability of heads. $t_0 t_1 \dots$ is a sequence of coin flips.

$$
\begin{aligned}
p(h|t_0 t_1 \dots) &= p(h|t_0, t_1, \dots) \\
&= \frac{p(t_0, t_1, \dots |h)}{p(t_0, t_1, \dots)} p(h) \\
&\text{assume independence} \\
&= \frac{p(t_0|h)p(t_1|h)p(\dots|h)}{p(t_0)p(t_1)p(\dots)} p(h) \\
&= \dots \cdot \frac{p(t_1|h)}{p(t_1)} \cdot \frac{p(t_0|h)}{p(t_0)} \cdot p(h)
\end{aligned}
$$

Since multiplication is commutative, the order of the data does not matter.

Complete Example: Estimating Bias in a Coin (I)

Likelihood for a coin flip:

$$p(t|\theta) = \theta^t(1-\theta)^{1-t}; t \in 0, 1 \text{ Bernoulli distribution}$$

For several flips:

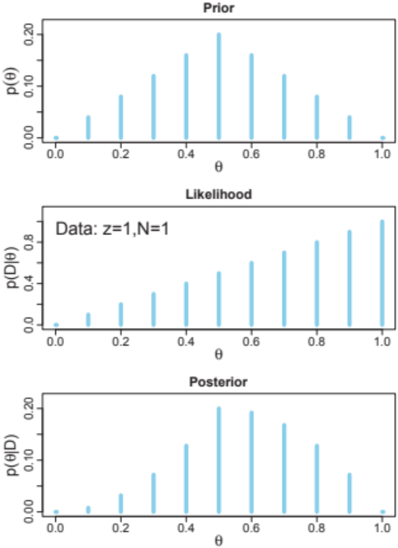
$$\begin{aligned} p(T|\theta) &= \prod_{t_i \in T} \theta^{t_i}(1-\theta)^{1-t_i} \\ &= \theta^z(1-\theta)^{N-z} \end{aligned}$$

where z is total number of heads and $N - z$ is total number of tails. That is, the probability of heads given a *propensity* for heads.

Complete Example: Estimating Bias in a Coin (II)

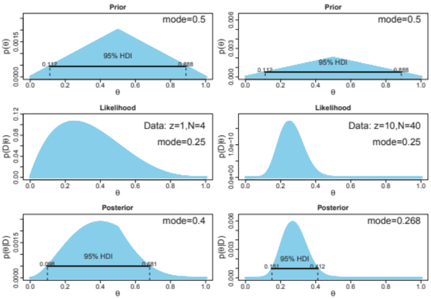
Qualitatively: If we assume a coin is most probably fair

If we measure a single flip and it is heads, this indicates we should shift our beliefs so heads is more likely.

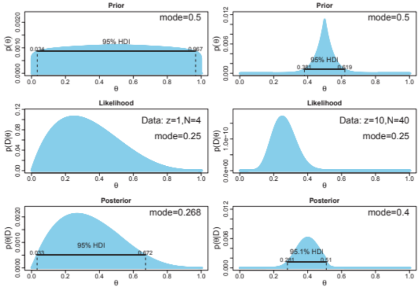


Influence of sample size on the posterior

A larger sample size makes us less reliant on the prior. Note that the prior is the same in both plots! This will be derived analytically for specific priors and likelihoods in the next chapter.



Compare to previous plot.
The prior can be thought of (through data order invariance) as incorporating previous data. Thus a flatter prior broades the posterior, while a sharper makes it more narrow (assuming it is reasonably correct).



Why Bayesian Inference Can Be Difficult

Bayes rule:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

Analytic solution: is not in general easy, or even possible. Exceptions include specific prior-likelihood combinations called conjugate priors. (A prior conjugate to the given likelihood.)
One can also approximate tricky integrals with easier to integrate functions. Called *variational approximation*.
Grid approximation: Discretisize, and numerically solve integrals. Fine for small parameter spaces, but quickly runs into the *curse of dimensionality*.
Sampling approximation: Randomly sample the posterior through *Markov-chain Monte-Carlo* methods. These methods skip the evaluation of the integrals. Lead to Bayesian methods gaining practical use.