Introduction to Bayesian Statistics

Christiana Kartsonaki

February 11th, 2015

Introduction

Bayes' theorem is an elementary result in probability theory which relates the conditional probability P(A given B) to P(B given A), for two events A and B.

Bayes' theorem

$\mathbb{P}(A)$: probability of event A

$\mathbb{P}(A \mid B)$: conditional probability of event A given that event B has occurred

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

Bayesian inference

$$P(\mathit{conclusion} \mid \mathit{data}) \quad \propto \quad P(\mathit{data} \mid \mathit{conclusion}) \quad \times \quad P(\mathit{conclusion})$$

Bayesian inference

$$P(\textit{conclusion} \mid \textit{data}) \;\propto\; P(\textit{data} \mid \textit{conclusion}) \;\times\; P(\textit{conclusion})$$

posterior *likelihood* *prior*

Example

buses arrive in a random pattern (Poisson process), mean time interval unknown

you know something about the interval (e.g. likely not every 1 hour, but not 1 minute either)

the two sources of information are synthesized as a probability distribution

Example

buses arrive in a random pattern (Poisson process), mean time interval unknown data $\rightarrow$ **likelihood**

you know something about the interval (e.g. likely not every 1 hour, but not 1 minute either) **prior distribution**

the two sources of information are synthesized as a probability distribution **posterior distribution**

Example

buses arrive in a random pattern (Poisson process), mean time interval unknown   data $\rightarrow$ **likelihood**

you know something about the interval (e.g. likely not every 1 hour, but not 1 minute either)   **prior distribution**

the two sources of information are synthesized as a probability distribution **posterior distribution**

$\rightarrow$ merge information from data with 'external' information

Statistical inference

parameter θ

data $\mathbf{x}$

model $f(\mathbf{x}, \theta)$

Bayes' theorem

$$p(\theta \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \theta)\, \pi(\theta)}{p(\mathbf{x})}$$

$\pi(\theta)$: prior distribution

$p(\mathbf{x} \mid \theta)$: likelihood

$p(\theta \mid \mathbf{x})$: posterior distribution

$p(\mathbf{x})$: predictive probability of $\mathbf{x}$ (normalizing factor)

Bayes' theorem

$$p(\theta \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \theta)\, \pi(\theta)}{p(\mathbf{x})}$$

$\pi(\theta)$: prior distribution

$p(\mathbf{x} \mid \theta)$: likelihood

$p(\theta \mid \mathbf{x})$: posterior distribution

$p(\mathbf{x})$: predictive probability of $\mathbf{x}$ (normalizing factor)

posterior $\propto$ likelihood $\times$ prior

Bayesian inference

Any quantity that does not depend on θ cancels out from the denominator and numerator of Bayes' theorem.

So if we can recognise which density is proportional to the product of the likelihood and the prior, regarded solely as a function of θ, we know the posterior density of θ.

Frequentist and Bayesian statistics

Frequentist approaches $\rightarrow$ typically treat θ as an unknown constant

Bayesian approaches $\rightarrow$ treat it as a random variable

Likelihood

Likelihood $\rightarrow$ used in most approaches to formal statistical inference.

Describes the data generating process.

Prior and posterior distribution

prior distribution $\rightarrow$ represents information about the parameters other than that supplied by the data under analysis

posterior distribution $\rightarrow$ probability distribution as revised in the light of the data, determined by applying standard rules of probability theory

Prior and posterior distribution

prior distribution $\rightarrow$ represents information about the parameters other than that supplied by the data under analysis

posterior distribution $\rightarrow$ probability distribution as revised in the light of the data, determined by applying standard rules of probability theory

Sometimes the prior distribution is 'flat' over a particular scale, intended to represent the absence of initial information.

In complex problems with many nuisance parameters the use of flat prior distributions is suspect and, at the very least, needs careful study using sensitivity analyses.

Example

$X_1, \ldots, X_n$: random sample from an exponential distribution with density $f(x \mid \theta) = \theta e^{-\theta x}, \ x > 0$

prior $\pi(\theta) = \lambda e^{-\lambda\theta}, \ \theta > 0$, for some known value of λ

Example

$X_1, \ldots, X_n$: random sample from an exponential distribution with density $f(x \mid \theta) = \theta e^{-\theta x}, \ x > 0$

prior $\pi(\theta) = \lambda e^{-\lambda\theta}, \ \theta > 0$, for some known value of λ

Then the likelihood is

$$f(x_1, \ldots, x_n \mid \theta) = \theta^n \, e^{-\theta \sum_{i=1}^n x_i}.$$

Example

$X_1, \ldots, X_n$: random sample from an exponential distribution with density $f(x \mid \theta) = \theta e^{-\theta x},\ x > 0$

prior $\pi(\theta) = \lambda e^{-\lambda\theta},\ \theta > 0$, for some known value of λ

Then the likelihood is

$$f(x_1, \ldots, x_n \mid \theta) = \theta^n \, e^{-\theta \sum_{i=1}^n x_i}.$$

So the posterior distribution is

Example

$X_1, \ldots, X_n$: random sample from an exponential distribution with density $f(x \mid \theta) = \theta e^{-\theta x},\ x > 0$

prior $\pi(\theta) = \lambda e^{-\lambda\theta},\ \theta > 0$, for some known value of λ

Then the likelihood is

$$f(x_1, \ldots, x_n \mid \theta) = \theta^n \, e^{-\theta \sum_{i=1}^n x_i}.$$

So the posterior distribution is

$$p(\theta \mid \mathbf{x}) \propto \pi(\theta)\, f(\mathbf{x} \mid \theta)$$

Example

$X_1, \ldots, X_n$: random sample from an exponential distribution with density $f(x \mid \theta) = \theta e^{-\theta x}, \; x > 0$

prior $\pi(\theta) = \lambda e^{-\lambda\theta}, \; \theta > 0$, for some known value of λ

Then the likelihood is

$$f(x_1, \ldots, x_n \mid \theta) = \theta^n \, e^{-\theta \sum_{i=1}^n x_i}.$$

So the posterior distribution is

$$p(\theta \mid \mathbf{x}) \; \propto \; \lambda e^{-\lambda\theta} \, \theta^n e^{-\theta \sum x_i}$$

Example

$X_1, \ldots, X_n$: random sample from an exponential distribution with density $f(x \mid \theta) = \theta e^{-\theta x}, \; x > 0$

prior $\pi(\theta) = \lambda e^{-\lambda\theta}, \; \theta > 0$, for some known value of λ

Then the likelihood is

$$f(x_1, \ldots, x_n \mid \theta) = \theta^n \, e^{-\theta \sum_{i=1}^{n} x_i}.$$

So the posterior distribution is

$$\begin{aligned} p(\theta \mid \mathbf{x}) &\propto \lambda e^{-\lambda\theta} \, \theta^n e^{-\theta \sum x_i} \\ &\propto \theta^n e^{-\theta(\lambda + \sum x_i)} \end{aligned}$$

Example

$X_1, \ldots, X_n$: random sample from an exponential distribution with density $f(x \mid \theta) = \theta e^{-\theta x}, \; x > 0$

prior $\pi(\theta) = \lambda e^{-\lambda\theta}, \; \theta > 0$, for some known value of λ

Then the likelihood is

$$f(x_1, \ldots, x_n \mid \theta) = \theta^n \, e^{-\theta \sum_{i=1}^n x_i}.$$

So the posterior distribution is

$$p(\theta \mid \mathbf{x}) \; \propto \; \theta^n e^{-\theta(\lambda + \sum x_i)},$$

Example

$X_1, \ldots, X_n$: random sample from an exponential distribution with density $f(x \mid \theta) = \theta e^{-\theta x},\ x > 0$

prior $\pi(\theta) = \lambda e^{-\lambda\theta},\ \theta > 0$, for some known value of λ

Then the likelihood is

$$f(x_1, \ldots, x_n \mid \theta) = \theta^n \, e^{-\theta \sum_{i=1}^n x_i}.$$

So the posterior distribution is

$$p(\theta \mid \mathbf{x}) \ \propto \ \theta^n e^{-\theta(\lambda + \sum x_i)},$$

which is the Gamma$(n+1,\ \lambda + \sum x_i)$ density.

History

'inverse probability'

prior distribution intended to represent initial ignorance $\rightarrow$ used systematically in statistical analysis by Gauss and especially Laplace (circa 1800)

the approach was criticized during the 19th century

in the middle of the 20th century attention shifted to a personalistic view of probability – individual belief as expressed by individual choice in a decision-making context

Prior distribution

Controversial aspects concern prior. What does it mean?

- What is the prior probability that treatment and control have identical effect?
- What is the prior probability that the difference between two groups is between 5 and 15 units?

prior distribution must be specified explicitly, i.e. in effect numerically

Prior distribution

Broadly three approaches:

1. Summary of other data.
   Empirical Bayes.

2. Prior measures personalistic opinion of investigator about conclusions.
   Not useful for 'public' transmission of knowledge.

3. Objective degree of uncertainty.
   1. Agreed measure of uncertainty.
   2. Ignorance, reference, flat prior for interval estimate. Laplace's principle of indifference.

Empirical Bayes

'empirical' $\rightarrow$ frequency interpretation implied

e.g. an unknown parameter representing a mean of some measurement – likely to vary under different circumstances

can be represented by a widely dispersed distribution

leading to a posterior distribution with a frequency interpretation

Example: variances of gene expression for different probes on a microarray may be assumed to be a sample from a distribution with a common parameter

Personalistic prior

reflects the investigator's subjective beliefs

prior distribution is based on relatively informally recalled experience of a field, for example on data that have been seen only informally

Flat prior

a prior which aims to insert as little new information as possible

for relatively simple problems often limiting forms of the prior reproduce approximately or exactly posterior intervals equivalent to confidence intervals

Priors

- Is the prior distribution a positive insertion of evidence? If so, what is its basis and has the consistency of that evidence with the current data been checked?
- If flat/ignorance/reference priors have been used, how have they been chosen? Has there been a sensitivity analysis? If the number of parameters over which a prior distribution is defined is appreciable then the choice of a flat prior distribution could be misleading.
- Each of a substantial number of individuals may have been allocated a value of an unknown parameter, the values having a stable frequency distribution across individuals – empirical Bayes.

Posterior distribution

Conclusions can be summarized using for example

- posterior mean
- posterior variance
- credible intervals

Credible intervals

a region $C_\alpha(\mathbf{x})$ is a $100(1-\alpha)\%$ **credible region** for θ if

$$\int_{C_\alpha(\mathbf{x})} p(\theta \mid \mathbf{x})\, d\theta = 1 - \alpha$$

- there is posterior probability $1-\alpha$ that θ is in $C_\alpha(\mathbf{x})$

credible interval – special case of credible region

analogous to (frequentist) confidence intervals, different interpretation

Hypothesis testing

Frequentist approach to hypothesis testing $\rightarrow$ compares a null hypothesis H_0 with an alternative H_1 through a test statistic T that tends to be larger under H_1 than under H_0 and rejects H_0 for small p-values $p = \mathbb{P}_{H_0}(T \geq t_{obs})$, where t_{obs} is the value of T actually observed and the probability is computed as if H_0 were true

Bayesian approach $\rightarrow$ attaches prior probabilities to models corresponding to H_0 and H_1 and compares their posterior probabilities using the **Bayes factor**

$$B_{10} = \frac{\mathbb{P}(\mathbf{x} \mid H_1)}{\mathbb{P}(\mathbf{x} \mid H_0)}$$

Computation

conjugate prior $\rightarrow$ when the prior and the posterior are from the same family of distributions (for example normal prior and normal likelihood)

makes calculations easier

however, often unrealistic, so posterior distributions need to be evaluated numerically

Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo (MCMC): a stochastic simulation technique which is used for computing inferential quantities which cannot be obtained analytically

- MCMC simulates a discrete-time Markov chain
- it produces a dependent sequence of random variables $\{\theta^{(1)}, \ldots, \theta^{(M)}\}$ with approximate distribution the posterior distribution of interest
- MCMC is an iterative procedure, such that given the current state of the chain, $\theta(i)$, the algorithm makes a probabilistic update to $\theta(i+1)$
- Markov chains can automatically be constructed to match any posterior density

MCMC

Two of the most general procedures for MCMC simulation from a target distribution:

- Metropolis–Hastings algorithm
- Gibbs sampler

MCMC

Two of the most general procedures for MCMC simulation from a target distribution:

- Metropolis–Hastings algorithm
- Gibbs sampler

Software:

- WinBugs – a Windows version of BUGS (Bayesian analysis Using the Gibbs Sampler)
- CODA: a collection of convergence diagnostics and sample output analysis programs
- JAGS (Just Another Gibbs Sampler)

MCMC – priors

MCMC mostly uses flat priors.

- Flat for θ not same as flat for e.g. $\log(\theta)$.
- For models with fairly few parameters and reasonable data gives confidence level.
- For large number of parameters may give very bad answer. No general theory known.

Discussion

- Bayesian inference $\rightarrow$ based on Bayes' theorem
- differences in interpretation between Bayesian and frequentist inference
- choice of prior controversial
- computation usually done numerically; MCMC useful but to be used with caution

Further reading

- Cox, D. R. (2006). Frequentist and Bayesian Statistics: A Critique (Keynote Address). In *Statistical Problems in Particle Physics, Astrophysics and Cosmology* (Vol. 1, p. 3).
- Cox, D. R. and Donnelly, C. A. (2011). *Principles of Applied Statistics*. Cambridge University Press.
- Davison, A. C. (2003). *Statistical Models* (Vol. 11). Cambridge University Press.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman and Hall/CRC Texts in Statistical Science.
- WinBugs `http://www.mrc-bsu.cam.ac.uk/software/bugs/`