Example sheet 2

Bayesian inference

Foundations of Data Science—DJW—2019/2020

Question 1. I have a collection of numbers $x_1, \ldots, x_n$ which I take to be independent samples from the Normal(μ, σ_0^2) distribution. Here σ_0 is known, and μ is unknown. Using the prior distribution $M \sim \text{Normal}(\mu_0, \rho_0^2)$, show that the posterior density is

$$\Pr_M(\mu \mid x_1, \ldots, x_n) = \kappa e^{-(\mu-c)^2/2\tau^2}$$

and where κ is a normalizing constant, and where you should find formulae for c and τ in terms of x_i, σ_0, μ_0, and ρ_0, Hence deduce that the posterior distribution is Normal(c, τ^2). *[Note: 'M' is the upper-case form of the Greek letter 'μ', which is why I write* $\Pr_M(\mu)$*.]*

Question 2. In lecture notes exercise 1.9, page 14, we fitted a Gaussian mixture model for galaxy speeds. The model is a two-step random variable, and its fitted parameters are below.

```
def rx(p, μ, σ): # returns a random speed, in km/sec
    k = numpy.random.choice(range(len(p)), p=p)
    return numpy.random.normal(loc=μ[k], scale=σ[k])
p,μ,σ = [0.085, 0.278, 0.637], [9709, 19822, 22756], [422, 559, 3382]
```

For a galaxy with speed 22,000 km/sec, what is the probability it belongs to each one of the three clusters?

Question 3. Exercise 2.4 in lecture notes (page 33) describes the police stop-and-search dataset. Let the outcome for record i be $y_i \in \{0, 1\}$, where 1 denotes that the police found something and 0 denotes that they found nothing. Consider the probability model $Y_i \sim \text{Binom}(1, \beta_{\mathsf{eth}_i})$ where eth_i is the recorded ethnicity for the individual involved in record i, and where the parameters β_{As}, β_{Blk}, β_{Mix}, β_{Oth}, β_{Wh} are unknown. As a prior distribution, suppose that the five β parameters are all independent Beta$(1/2, 1/2)$ random variables.

(a) Write down the joint prior density for $(\beta_{\text{As}}, \beta_{\text{Blk}}, \beta_{\text{Mix}}, \beta_{\text{Oth}}, \beta_{\text{Wh}})$.

(b) Find the joint posterior distribution of $(\beta_{\text{As}}, \beta_{\text{Blk}}, \beta_{\text{Mix}}, \beta_{\text{Oth}}, \beta_{\text{Wh}})$ given the y data.

Question 4. In lecture notes section 2.2.5 we proposed a linear model for temperature increase:

$$\mathsf{temp} \approx \alpha + \beta_1 \sin(2\pi \mathsf{t}) + \beta_2 \cos(2\pi \mathsf{t}) + \gamma(\mathsf{t} - 2000).$$

Suggest a probability model for temp. Suggest Bayesian prior distributions for the unknown parameters α, β_1, β_2, and γ. Give pseudocode to find a 95% confidence interval for γ.

Please implement this, and post your confidence interval to Moodle, in the section labelled Practical 3. A survey of answers will be shared in lectures.

Question 5. I am prototyping a diagnostic test for a disease. In healthy patients, the test result is Normal$(0, 2.1^2)$. In sick patients it is Normal$(\mu, 3.2^2)$, but I have not yet established a firm value for μ.

In order to estimate μ, I trialled the test on 30 patients whom I know to be sick, and the mean test result was 10.3. I subsequently apply the test to a new patient, and get the answer 8.8. I wish to know whether this new patient is healthy or sick.

(a) Considering just the 30 trial patients, state the posterior distribution for μ, when the prior distribution is $M \sim \text{Normal}(5, 3^2)$.

(b) Let $h \in \{\text{healthy}, \text{sick}\}$ be the status of the new patient, and use the prior distribution $\Pr_H(\text{healthy}) = 0.99$, $\Pr_H(\text{sick}) = 0.01$. Write down the joint prior distribution for (M, H).

(c) Find the posterior density of (M, H), using readings from all 31 patients. Leave your answer as an unnormalized density function.

(d) Give pseudocode to compute the posterior distribution of H, i.e. compute $\mathbb{P}(H = \text{healthy}|\text{data})$ and $\mathbb{P}(H = \text{sick} \mid \text{data})$. Here 'data' refers to the readings from all 31 patients.

Hints and comments

Question 1. All Bayesian questions are answered in the same way: (a) write down the prior density for $\Pr_M(\mu)$, (b) write down the data density conditional on μ, (c) multiply them together, times a constant factor κ, to get the posterior density for $\Pr_M(\mu \mid x_1, \ldots, x_n)$.

In this question, once you have the posterior density, rewrite it to look like a function of μ. This involves expanding quadratic terms and completing the square. Any terms that don't involve μ can be amalgamated with the constant factor κ.

When a question asks "find the posterior distribution", you should start by calculating the posterior density, leaving it unnormalized i.e. including a constant factor, call it κ. Then (a) if you recognize this as a standard density function, as in this case, just give its name; (b) if it's easy to find κ using "densities sum to one" then do so; (c) otherwise leave your answer as an unnormalized density function.

Question 2. Using the "densities sum to one" rule to normalize the posterior. Let (K, X) be a random choice of cluster and speed. The code tells us the prior distribution for K. The question is asking for the posterior density $\Pr_K(k \mid X = 22000)$, which you can find using Bayes's rule. Bayes's rule gives an answer with a normalizing constant κ, which you can calculate using the fact that the posterior density must sum to one, i.e. $\sum_{k \in \{0,1,2\}} \Pr_K(k \mid X = 22000) = 1$.

Question 3. A simple example of priors on multiple parameters. First, find the joint posterior density. Then give a 'user-friendly' description of this joint distribution using words or pseudocode.

Question 4. Confidence intervals. The prior is your choice. For the probability model, see section 2.4. For the prior distributions, it's your choice. For confidence intervals, see section 6.2.

Question 5. For Bayes update, use all the data + all the parameters. Dealing with nuisance parameters. Part (a) is just an application of question 1. The question doesn't tell you the individual readings of the 30 patients; let the readings be $x_1, \ldots, x_{30}$, and do the algebra, and it will turn out that it's sufficient to know the mean value which is 10.3.

For part (c): Let the test reading from a new patient be Y. The question tells us its density: it is

$$\Pr_Y(y \mid \mu, h) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2_{\text{healthy}}}} e^{-y^2/2\sigma^2_{\text{healthy}}} & \text{if } h = \text{healthy} \\ \frac{1}{\sqrt{2\pi\sigma^2_{\text{sick}}}} e^{-(y-\mu)^2/2\sigma^2_{\text{sick}}} & \text{if } h = \text{sick} \end{cases}$$

where $\sigma_{\text{healthy}} = 2.1$ and $\sigma_{\text{sick}} = 3.2$. Write this as an indicator function

$$\Pr_Y(y \mid \mu, h) = (\cdots)1_{h=\text{healthy}} + (\cdots)1_{h=\text{sick}}$$

and then you can more easily plug it into the Bayes update formula.

Part (d) is a question about marginals. See exercise 6.3 in lecture notes. Your pseudocode should make use of the data density, not the posterior density that you calculated in part (c).

Supplementary question sheet 2

Bayesian inference

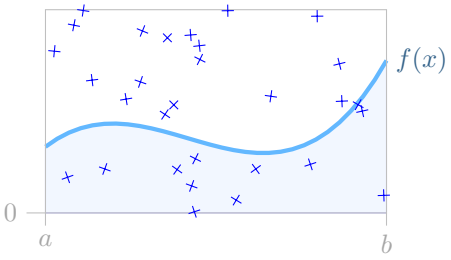
Foundations of Data Science—DJW—2019/2020

These questions are not intended for supervision (unless your supervisor directs you otherwise). Some of require careful maths, some are best answered with coding, some are philosophical.

Question 6. Suppose we're given a function $f(x) \geq 0$ and we want to evaluate

$$\int_{x=a}^{b} f(x)\, dx.$$

Here's an approximation method: (i) draw a box that contains $f(x)$ over the range $x \in [a, b]$, (ii) scatter points uniformly at random in this box, (iii) return $A \times p$ where A is the area of the box and p is the fraction of points that are under the curve. Explain why this is a special case of Monte Carlo integration.



Do NOT give a wishy-washy qualitative argument along the lines of "there are random points, and we're evaluating an integral, so it's a type of Monte Carlo". Monte Carlo has a precise meaning: $\mathbb{E}\, h(X) \approx n^{-1} \sum_i h(x_i)$. *In your answer you should (a) explain the random variable in question, (b) specify the* h *function, (c) give an explanation along the lines of page 38 of lecture notes.*

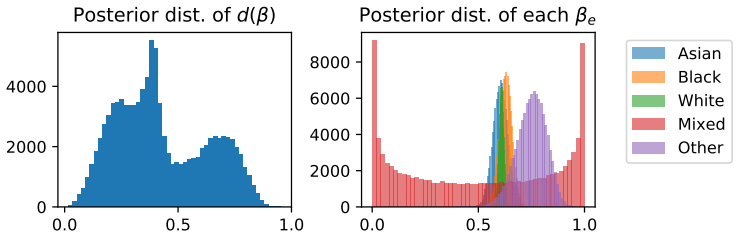
Question 7. In the setting of question 3, I wish to measure the amount of police bias. Given a 5-tuple of parameters $\beta = (\beta_{\text{As}}, \beta_{\text{Blk}}, \beta_{\text{Mix}}, \beta_{\text{Oth}}, \beta_{\text{Wh}})$, I define the overall bias score to be

$$d(\beta) = \max_{e,e'} \left| \beta_e - \beta_{e'} \right|.$$

If $d(\beta)$ is large, then there is *some* pair of ethnicities with very unequal treatment.

As a Bayesian I view β as a random variable taking values in $[0, 1]^5$, therefore $d(\beta)$ is a random variable also. To investigate its distribution, I sample β from the posterior distribution that I found in question 3, I compute $d(\beta)$, and I plot a histogram. The output, shown on the left, is bizarre. To help me understand what's going on, I plot histograms of each of the individual β_e coefficients, shown on the right.

Explain the results. *[Note: The code for this analysis is available on Azure notebooks.]*



Question 8. I have a coin, which might be biased. I toss it n times and get x heads.

I am uncertain whether or not the coin is biased. Let $m \in \{\text{fair}, \text{biased}\}$ indicate which of the two cases is correct; and if it is biased let θ be the probability of heads. The probabilty of observing x heads is thus

$$\Pr(x \mid m, \theta) = \begin{cases} \binom{n}{x}\theta^x(1-\theta)^{n-x} & \text{if } m = \text{biased} \\ \binom{n}{x}(1/2)^x(1-1/2)^{n-x} & \text{if } m = \text{unbiased} \end{cases}$$

As a Bayesian I shall represent my uncertainty about m with a prior distribution, $\Pr_M(\text{fair}) = p$, $\Pr_M(\text{biased}) = 1 - p$. If it is biased, my prior belief is that the probability of heads is $\Theta \sim \text{Uniform}[0, 1]$.

(a) Write down the prior distribution for the pair (M, Θ), assuming independence as usual.

(b) Find the posterior distribution of (M, Θ) given x.

(c) Find $\mathbb{P}(M = \text{unbiased} \mid x)$, i.e. the posterior probability that the coin is unbiased.

This is a Bayesian question, and it's answered in the same way as any other Bayesian question: write down the prior density $\Pr_{M,\Theta}(m, \theta)$*, write down the data density* $\Pr(x \mid m, \theta)$*, and multiply them together (times a constant factor) to get the posterior* $\Pr_{M,\Theta}(m, \theta \mid x)$*. To keep track of all the cases, it may be helpful to use indicator functions, both for* $\Pr_M$ *and for* $\Pr(x \mid m, \theta)$*.*

Part (c) is about nuisance parameters, as in exercise 6.3 in notes (look at the mathematical solution of that exercise). Once we've found the posterior density, say $\Pr_{M,\Theta}(m, \theta) = \kappa f(m, \theta)$ *where* κ *is the normalizing constant, we have to integrate out* θ *to find the marginal distribution, as in section 3.4:*

$$\mathbb{P}(M = \mathit{fair} \mid x) = \int_\theta \kappa f(\mathit{fair}, \theta)\, d\theta \qquad \mathbb{P}(M = \mathit{biased} \mid x) = \int_\theta \kappa f(\mathit{biased}, \theta)\, d\theta.$$

Then solve for κ*, using the "densities sum to one" rule, as in question 2.*

This question is an illustration of Bayesian model selection, which you can read about in section 6.3 of lecture notes.

Question 9. Give an exact answer to question 5 part (d).

This is another question about nuisance parameters, like the previous question. The integrals require painstaking care, but they don't need any clever calculus tricks, just careful "completing the square" as in question 1 and then the "densities sum to one" rule.

Question 10. (a) Suppose we have a single observation x, drawn from $\text{Normal}(\mu + \nu, \sigma^2)$, where μ and ν are unknown parameters, and σ^2 is known. For μ use $\text{Normal}(\mu_0, \rho_0^2)$ as prior, and for ν use $\text{Normal}(\nu_0, \rho_0^2)$, where μ_0, ν_0, and ρ_0 are known. Find the posterior density of (μ, ν). Calculate the parameter values $(\hat{\mu}, \hat{\nu})$ where the posterior density is maximum. (These are called *maximum a posteriori estimates* or *MAP estimates*.)

(b) An engineer friend tells you "Bayesianism is the Apple of inference. You just work out the posterior, and everything Just Works™, and you don't need to worry about irritating things like confounded variables." What do you think?