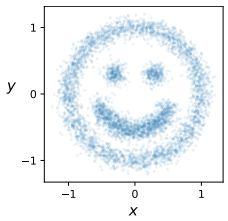
Example sheet 2

Bayesian inference

Data Science—DJW—2020/2021

Question 1. Define a function `rxy()` that produces a random pair of values (X, Y) which, when shown in a scatterplot, produces a smiley face like this. Also plot the marginal distributions of X and Y.



Question 2. I sample $x_1, \ldots, x_n$ from Uniform$[0, \theta]$. The parameter θ is unknown, and I shall use $\Theta \sim \text{Pareto}(b_0, \alpha_0)$ as my prior, where $b_0 > 0$ and $\alpha_0 > 1$ are known. This has the cumulative distribution function

$$\mathbb{P}(\Theta \leq \theta) = \begin{cases} 1 - \big(b_0/\theta\big)^{\alpha_0} & \text{if } \theta \geq b_0, \\ 0 & \text{if } \theta < b_0. \end{cases}$$

(a) Calculate the prior likelihood for Θ.

(b) Show that the posterior distribution of $(\Theta \mid x_1, \ldots, x_n)$ is Pareto, and find its parameters.

(c) Find a 95% posterior confidence interval for Θ.

(d) Find a different 95% posterior confidence interval. Which is better? Why?

Question 3. 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 $M \sim \text{Normal}(\mu_0, \rho_0^2)$ for μ, show that the posterior density is

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

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

Question 4. I have a collection of numbers

$$[4.3, 2.8, 3.9, 4.1, 9, 4.5, 3.3]$$

which look like they mostly come from a Gaussian distribution, but with the occasional outlier. Model the data as

$$X \text{ is } \begin{cases} \text{Normal}(\mu, 0.5^2) & \text{with probability 99\%} \\ \text{Cauchy} & \text{with probability 1\%.} \end{cases}$$

Use a Normal$(0, 5^2)$ prior distribution for μ. Give pseudocode to plot the posterior distribution. *[Note. The Cauchy random variable occasionally generates wildly huge values. The library function* `scipy.stats.cauchy.pdf(x)` *computes its pdf.]*

Question 5. In the lecture notes on linear modelling, we proposed a linear model for temperature increase:

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

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.

Question 6. In lecture notes section 2.6 we investigated a dataset of police stop-and-search actions. 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 $Y_i \sim \text{Binom}(1, \beta_{\texttt{eth}_i})$ where $\texttt{eth}_i$ 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 7. 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 be the status of the new patient, taking values in $\{\text{healthy}, \text{sick}\}$, and use the prior distribution

$$\Pr_H(h) = 0.991_{h=\text{healthy}} + 0.011_{h=\text{sick}}.$$

Write down the joint prior distribution for (M, H).

Clarification. The equation needs more symbols! It is

$$\Pr_H(h) = 0.99 \times 1_{h=\text{healthy}} + 0.01 \times 1_{h=\text{sick}}.$$

(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 $\mathbb{P}(H = h \mid \text{data})$ for both $h = \text{healthy}$ and $h = \text{sick}$. Here 'data' refers to the readings from all 31 patients.

Hints and comments

Question 1. Try extending the Gaussian mixture model from section 1. For plotting, here's some code. It assumes that you have stored your samples in a numpy array of shape $n \times 2$, one row per sample point, columns for x and y.

```
fig,((ax_x,dummy),(ax_xy,ax_y)) = plt.subplots(2,2, figsize=(4,4),
    sharex='col', sharey='row', gridspec_kw='height_ratios':[1,2], 'width_ratios':[2,1])
dummy.remove()
ax_xy.scatter(xy[:,0], xy[:,1], s=3, alpha=.1)
ax_x.hist(???, density=True, bins=60) # fill in the ???
ax_y.hist(???, density=True, bins=60, orientation='horizontal') # fill in the ???
plt.show()
```

Question 2. For part (a), just differentiate the cdf to get the pdf, i.e. the likelihood. Write it out using indicator function notation, $1_{\theta \geq b_0}$. This is often a good idea, when we're working with parameters that affect boundaries.

For the rest: **all Bayesian calculations start in exactly the same way.** First write out the likelihood of the observed data $\Pr(x_1, \ldots, x_n \mid \Theta = \theta)$, then (1) write down the prior likelihood $\Pr_\Theta(\theta)$, (2) apply Bayes's rule which says that the posterior likelihood is

$$\Pr_\Theta(\theta \mid x_1, \ldots, x_n) = \kappa \Pr_\Theta(\theta) \Pr(x_1, \ldots, x_n \mid \Theta = \theta).$$

In this question, write out the likelihood of the data using indicator notation, as in exercise sheet 1 question 4. Once you have the posterior density, gather together the θ terms, and you should end up with the density of another Pareto.

For the posterior confidence interval: the definition of a posterior confidence interval is in lecture notes section 7.2. You just have to solve the equations for `lo` and `hi`, using the cumulative distribution function for the Pareto.

Question 3. All Bayesian calculations start in exactly the same way. First write out the likelihood of the observed data $\Pr(x_1, \ldots, x_n \mid M = \mu)$, then (1) write down the prior likelihood $\Pr_M(\mu)$, (2) apply Bayes's rule which says that the posterior likelihood is

$$\Pr_M(\mu \mid x_1, \ldots, x_n) = \kappa \Pr_M(\mu) \Pr(x_1, \ldots, x_n \mid M = \mu).$$

Remember, this is a density function for a random variable M, and the argument is μ. Write your answer to gather together all the μ terms as much as you can. This involves expanding quadratic terms and completing the square. Any terms that don't involve μ can be amalgamated with the constant factor κ. What you end up with should look like a Normal density function, as a function of μ, and this lets you conclude that the posterior distribution is Normal.

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 4. All Bayesian computations start in exactly the same way. First write out the likelihood of the data, $\Pr(x_1, \ldots, x_n \mid M = \mu)$. The probability model here is very similar to a Gaussian mixture model, which we analysed in mock exam question 1. You'll need the cdf for the Cauchy, but you don't actually need to know a formula for it: just write $\text{cdf}_{\text{Cauchy}}(x)$ and $\text{pdf}_{\text{Cauchy}}(x)$. Then, (1) take a sample $\mu_1, \ldots, \mu_n$ from the prior distribution, (2) compute weights by evaluating the likelihood of the data at each one of these sampled μ-values, and rescaling so they sum to one.

For plotting the posterior distribution, see the examples in section 7.1.

Question 5. You should implement your proposed Bayesian model, and find a numerical value for the confidence interval. You can find a code skeleton at `https://github.com/damonjw/datasci/blob/master/ex2.ipynb`.

It's up to you to invent whatever probability distribution you like for `temp`; the simplest choice is to assume Gaussian errors as in section 2.4, and to pluck the noise parameter out of thin air. If you truly are uncertain about the noise parameter, then treat it as a random variable and invent a prior distribution for it.

It's up to you to invent whatever priors you like for the unknown parameters. It may seem totally arbitrary, but that's Bayesianism for you.

Question 6. This is a Bayesian question with multiple unknown parameters. You need to write down a joint prior density for all of them,

$$\Pr(\beta_{\mathrm{As}}, \beta_{\mathrm{Blk}}, \beta_{\mathrm{Mix}}, \beta_{\mathrm{Oth}}, \beta_{\mathrm{Wh}}).$$

See the mathematical solution to exercise 7.4 in lecture notes.

Bayes's rule, in its general form, says that

$$\Pr_\Theta(\theta \mid x) = \kappa \Pr_\Theta(\theta) \Pr_X(x|\Theta = \theta)$$

where θ denotes *all* the unknown parameters and x denotes *all* the dataset. Again, see exercise 7.4 in lecture notes. Leave your answer with κ.

After you've found the joint posterior density function, see if you can recognize it from the list of standard random variables.

Question 7. This is a question about multiple unknowns, using both the mathematical and the computational solutions. See exercise 7.4 from lecture notes.

Part (a) is just an application of question 3. 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), **all Bayesian calculations start the same.** First write out the likelihood function for the observed data. Here the observed data is $(x_1, \ldots, x_{30}, y)$, where $y = 0.88$ is the test reading from the new patient. You have to work out the joint density of this entire dataset, and your answer will depend on the unknown parameters (μ, h). Then, write out the joint prior likelihood for the unknown parameters. Then, multiply the two together. This gives you the posterior likelihood,

$$\Pr_{M,H}(\mu, h \mid x_1, \ldots, x_{30}, y).$$

Part (d) is a question about using marginalization to ignore nuisance parameters. See exercise 7.4 from lecture notes.

Supplementary question sheet 2

Bayesian inference

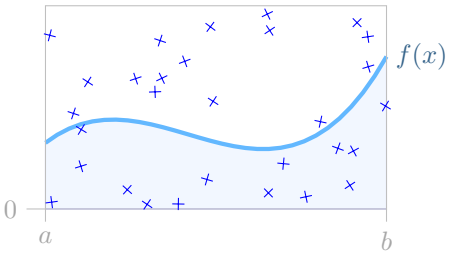
Data Science—DJW—2020/2021

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 8. 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 section 5.1 of lecture notes.*

Question 9. (a) For the random variables $X \sim \text{Uniform}[-1, 1]$ and $Y \sim \text{Normal}(X^2, 0.1^2)$, compute the conditional distribution of $(X \mid Y \in [0.5, 0.7])$. *[Hint. Let* $Z = 1_{Y \in [0.5, 0.7]}$ *and plot a histogram of* $(X \mid Z = 1)$*.]*

(b) Let $X \sim \text{Normal}(\mu, \sigma^2)$. Calculate the pdf and the cdf of $(X \mid X \geq 0)$. Leave your answer in terms of the pdf and cdf for the Normal distribution.

Question 10 (Iterated Bayes update). I have a coin, which might be biased. My data model is that each coin toss is $\text{Bin}(1, \Theta)$ where Θ is unknown. I plan to toss the coin 100 times. I get values $x_1, \ldots, x_{20}$, but then I get impatient and compute the posterior. I then toss it a further 80 times and get values $x_{21}, \ldots, x_{100}$.

(a) Using the prior $\Theta \sim \text{Uniform}[0, 1]$, find the posterior distribution $(\Theta \mid x_1, \ldots, x_{20})$.

(b) Using the distribution you found in (a) as the prior, find the posterior distribution conditional on $(x_{21}, \ldots, x_{100})$.

What do you notice?

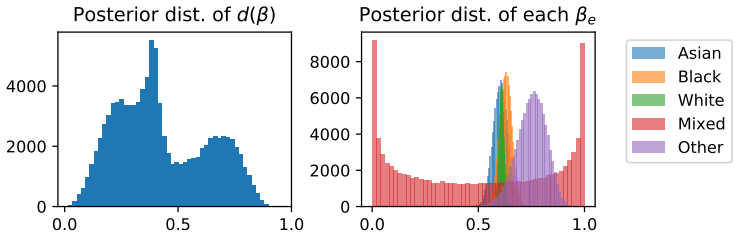
Question 11. In the setting of question 6, I wish to measure the amount of police bias. Given a 5-tuple of parameters $\beta = (\beta_{\text{As}}, \beta_{\text{Blk}}, \beta_{\text{Mix}}, \beta_{\text{Oth}}, \beta_{\text{Wh}})$, I define the overall bias score to be

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

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 6, 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. *[Hint. Explore the Beta distribution numerically. For what parameters does it have a bimodal distribution? What are the posterior distributions in this question?]*



Question 12. Consider the outlier model from question 4. How likely is it that the datapoint with value 9 is an outlier? *[Hint. Treat this as a two-parameter problem, like question 7.]*

Question 13. 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 7.4 in lecture 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 exercise 7.4:

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

Then solve for κ, using the "densities sum to one" rule, as in exercise 7.5 from lecture notes.

This question is an illustration of Bayesian model selection, which you can read about in section 7.4 of lecture notes.

Question 14. (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. Explain why the maximum likelihood estimates for μ and ν are non-identifiable.

(b) 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*.)

(c) 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 non-identifiability." What do you think?

Question 15. Here's my answer to question 1:

```
k = np.random.choice(4, p=[.6,.3,.05,.05], size=n)
t = np.random.uniform(size=n)
x = np.column_stack([np.sinπ(2**t), 0.55*np.sinπ(2**(0.4*t+0.3)), −0.3*np.ones(n), 0.3*np.ones(n)])
y = np.column_stack([np.cosπ(2**t), 0.55*np.cosπ(2**(0.4*t+0.3)), 0.3*np.ones(n), 0.3*np.ones(n)])
xy = np.column_stack([x[np.arange(n), k], y[np.arange(n), k]])
xy = np.random.normal(loc=xy, scale=.08)
```

Compute the distribution of $(X \mid Y = 0.3)$. Give your answer as a histogram.

You will need to derive your own method for sampling, along the lines of the derivation of computational Bayes in section 5.2. The difference here is that instead of using Bayes's rule

$$\Pr_X(x \mid Y = y) = \kappa \Pr_{X,Y}(x, y) = \kappa \Pr_X(x) \Pr_Y(y \mid X = x)$$

you will need to use a version more suited to the generation method used here,

$$Pr_{X,Y}(x, y) = \sum_k \int_t \Pr(x, y, k, t)\, dt = \sum_k \int_t \Pr_K(k) \Pr_T(t) \Pr_X(x \mid k, t) \Pr_Y(y \mid k, t)\, dt\,.$$

You should end up with a Monte Carlo integration that uses (K, T, X) samples.