Overview

Submit your writeup including any code as a PDF via gradescope.[1] We recommend reading through the entire homework beforehand and carefully using functions for testing procedures, plotting, and running experiments. Taking the time to reuse code will help in the long run! Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you discuss the homework with your peers, include their names on your submission. Please make sure any handwritten answers are legible, as we may deduct points otherwise.

1. Bayesian fidget spinners

Nat's company manufactures fidget spinners. The company uses two factories, which we'll call factory 0 and factory 1. Each fidget spinner from factory k is defective with probability q_k ($k \in \{0, 1\}$). Nat knows that factory 0 produces fewer defective fidget spinners than factory 1 (in other words, $q_0 < q_1$).

She receives n boxes full of fidget spinners, but the boxes aren't labeled (in other words, she doesn't know which box is from which factory). For each box, she starts randomly pulling out fidget spinners until she finds a defective one, and records how many fidget spinners she pulled out (including the defective one). She calls this number x_i for box i, for $i = 1, \ldots, n$.

She wants to estimate the following pieces of information:

- Which boxes came from factory 0, and which came from factory 1? She defines a binary random variable for each box z_i with the factory label (i.e., $z_i = 0$ if box i is from factory 0, and $z_i = 1$ if box i is from factory 1).
- How reliable is each factory? In other words, what are q_0 and q_1?

Inspired by what she learned about Gaussian mixture models, she sets up the following probability model:

$$\begin{aligned} z_i &\sim \text{Bernoulli}(\pi) \\ q_k &\sim \text{Beta}(\alpha_k, \beta_k) \\ x_i | z_i, q_0, q_1 &\sim \text{Geometric}(q_{z_i}) \end{aligned}$$

(a) Draw a graphical model for the probability model described above if $n = 2$ (i.e., there are only two boxes of fidget spinners).

Nat decides to implement the model above setting the following hyperparameters:

$$\pi = 0.3, \quad q_0 \sim \text{Beta}(1, 5), \quad q_1 \sim \text{Beta}(5, 1)$$

[1] In Jupyter, you can download as PDF or print to save as PDF

(b) Which one of the following explains why Nat chose this value of π:

(i) Factory 0 produces more boxes than factory 1

(ii) Factory 0 produces fewer boxes than factory 1

(iii) Factory 0 is better (i.e., it is less likely to produce defective fidget spinners)

(iv) Factory 0 is worse (i.e., it is more likely to produce defective fidget spinners)

(c) Which one of the following explains why Nat chose these values of α and β?

(i) Factory 0 produces more boxes than factory 1

(ii) Factory 0 produces fewer boxes than factory 1

(iii) Factory 0 is better (i.e., it is less likely to produce defective fidget spinners)

(iv) Factory 0 is worse (i.e., it is more likely to produce defective fidget spinners)

(d) Use `data.py` to generate the data that Nat observes, then, using PyMC3, fit the model outlined above, setting the hyperparameters to the values that Nat chose. Obtain 1000 samples from the posterior distribution $p(q_0, q_1|x_1, \ldots, x_n)$, and generate a scatterplot (one point per sample).

(i) Based on the graphs, do q_0 and q_1 appear dependent or independent under the posterior distribution? What characteristics of the graph allow you to conclude this?

(ii) Approximately, how likely is it that there were more boxes from factory 0 than from factory 1?

(iii) What is your median estimate of factory 0's defect rate, based on the samples from the posterior?

Hint: when writing down the model in PyMC3, you should use fancy indexing. To refresh fancy indexing, here is a simple example.

```
my_binary_array = np.array([0, 0, 1, 1, 0, 1])
my_real_array = np.array([0.27, 0.34])
print(my_real_array[my_binary_array])
```

(e) Nat's friend Yaro suggests using Gibbs sampling. What is the Gibbs sampling update for q_i? Your answer should be in the form of a well-known distribution, along with values for the parameter(s) of that distribution. Justify your answer.
Hint: you can derive the update analytically, or you can use the fact that the Beta distribution is a conjugate prior for a Geometric likelihood.

2. Rejection Sampling

Consider the function

$$g(x) = \cos^2(12x) \times |x^3 + 6x - 2| \times \mathbb{1}_{x \in (-1, -.25) \cup (0,1)}.$$

In this problem, we use rejection sampling to generate random variables with pdf $f(x) = cg(x)$.

(a) Plot g over its domain. What is a uniform proposal distribution q that covers the support of f? What is the largest possible constant M such that the scaled target distribution $p(x) = Mg(x)$ satisfies $p(x) \leq q(x)$ for all x?

(b) Suppose you run rejection sampling with target p and proposal q from part (a) until you generate n samples and your sampler runs a total of $N \geq n$ times, including n acceptances and $N - n$ rejections. Explain how you can use n, N and M to estimate c.
Hint: the ratio of acceptances n to total runs N is an approximation of the ratio between the area under the curve $p(x)$ and the area under $q(x)$.
Hint: remember what happens if you integrate a pdf over its entire support.

(c) Use rejection sampling to generate a sample of size 10^3 from $p(x)$. Since $f(x)$ is a pdf and it's proportional to $p(x)$, we can display its estimate easily: plot a normalized histogram of your sample, and overlay a smooth kernel density estimate, that will provide more information on the shape of the estimated distribution.
Repeat the previous steps increasing the number of samples to 10^6.

3. Graphical models

Graphical models are often useful for modeling phenomena involving multiple variables. Consider the scenario in which you have parked your car, and you are thinking about the following fixed probabilities π_b and π_i, and binary random variables Z_b, Z_i, and X:

$$\pi_b = \mathbb{P}(\text{a burglar breaks into your car})$$

$$\pi_i = \mathbb{P}(\text{an innocent passerby touches your car})$$

$$Z_b = \begin{cases} 1 & \text{if there is a burglar} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_i = \begin{cases} 1 & \text{if there is an innocent passerby} \\ 0 & \text{otherwise} \end{cases}$$

$$X = \begin{cases} 1 & \text{if your car alarm goes off} \\ 0 & \text{otherwise} \end{cases}$$

You know that Z_b and Z_i are independent of each other, while the probability that your alarm goes off depends on both Z_b and Z_i (in other words, your car could go off because of a burglar or because of an innocent passerby). Draw a graphical model that represents the relationships between $\pi_b, \pi_i, Z_b, Z_i,$ and X?