

# Data 102 Spring 2021

## Midterm 1

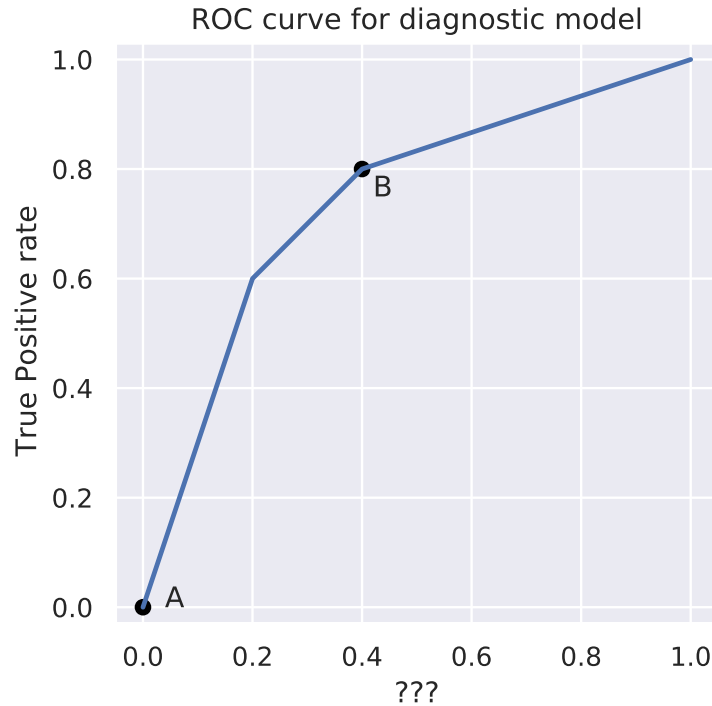
- Please write your solutions using either pen/pencil and paper, or a tablet. Each question should start on a new page. At the end of the exam period (or earlier), please upload your exam to the “Midterm 1” assignment on Gradescope. **It is your responsibility to make sure your work will be legible!**
- We will not answer any questions during the exam. If you think a question is unclear, state your assumptions and answer accordingly.
- You have 80 minutes to work on the exam: you must stop working at 11:00AM PT.
- This exam has 6 questions, for a total of 40 points. **You must complete all 6 questions to receive full credit.** There are multiple versions of this exam.
- Unless otherwise stated, you must show your work to receive full credit.
- You may, without proof, use theorems and facts that were given in the lectures, homework, lab, or discussions.
- **You must complete this honor pledge in order to receive credit on the exam:** We ask that you act in accordance with the honor code. Please copy the following statement by hand and sign your name, and include this in your submission.

<p><b>As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. These answers are my own work.</b></p>
---

0. **Make sure you complete the honor pledge on the previous page.**

1. (5 points) For each of the following, answer **True** or **False**. You don't need to justify your answer.
  - (a) (1 point) The value that minimizes the posterior risk with respect to zero-one loss is the posterior mean (the mean of the distribution  $p(\theta|X)$ ).
  - (b) (1 point) If, given  $n$   $p$ -values, we reject only those  $p$ -values at or below  $\alpha/n$ , then the false discovery rate will be less than or equal to  $\alpha$ .
  - (c) (1 point) Rejecting hypothesis  $t$  (as opposed to accepting it) in the LORD algorithm makes it more likely that hypothesis  $t + 1$  will be accepted.
  - (d) (1 point) When using Gibbs sampling, to update a particular hidden variable, we compute the most likely value of that variable conditioned on all the others.
  - (e) (1 point) When using Gibbs sampling, successive samples (for example,  $\theta^{(t)}$  and  $\theta^{(t+1)}$ ) are independent.
  
2. (4 points) For each question, **select all that apply**. If none of the answers are correct, write "None". You don't need to justify your answer.
  - (a) (1 point) Which sampling algorithm(s) involve an accept/reject step?
    - (A) Metropolis-Hastings
    - (B) Rejection sampling
  - (b) (1 point) Which sampling algorithms require us to know the distribution we're sampling from (including any normalization constants)?
    - (A) Metropolis-Hastings
    - (B) Rejection sampling
  - (c) (2 points) Which of the following statements about generalized linear models (GLMs) are true?
    - (A) The inverse link function in GLMs must always produce nonnegative outputs.
    - (B) Logistic regression is a kind of GLM.
    - (C) When estimating the parameters of a GLM, one must use Bayesian inference.
    - (D) Predictions from GLMs can only ever be accurate when interpolating, not when extrapolating.
    - (E) A Poisson likelihood model is only appropriate for cases where the variance of the data ( $y$ ) is much higher than the mean.

3. (9 points) We are trying to construct a spam filter for phone calls (an algorithm that classifies calls into spam vs not spam.) The model that we train has the following ROC curve.



- (a) (1 point) What is the  $x$ -axis label?
- (b) (2 points) Describe the decision rule that corresponds to the point  $A$ . Your answer should be one sentence or less.
- (c) (2 points) Let  $S$  be the binary random variable indicating whether a given call is spam ( $S = 1$ ) or not ( $S = 0$ ), and let  $D$  be the binary random variable indicating the algorithm's prediction for the call ( $D = 1$  for spam and  $D = 0$  otherwise). Write the false positive rate as a conditional probability in terms of  $S$  and  $D$ .
- (d) (2 points) We set our score threshold for the model so that the resulting test corresponds to point  $B$ . What fraction of calls does this test classify as being spam? Write your answer in terms of the true prevalence of spam calls  $\pi_1 = \mathbb{P}(S = 1)$ .
- (e) (2 points) The cost of misclassifying a real (non-spam) call as spam is 5 times more than that of misclassifying a spam call as non-spam. Write a loss function  $\ell(d, s)$  that expresses this belief.

4. (9 points) As part of a medical lab, you that would like to test five hypotheses using what you learned about multiple testing. Four lab tests have already been performed, resulting in the following  $P$ -values:  $P_1 = \frac{\alpha}{2}, P_2 = \frac{\alpha}{10}, P_3 = \frac{5\alpha}{6}, P_4 = \frac{\alpha}{4}$ , for some  $0 < \alpha < 1/2$ . We are currently waiting on the results for the fifth test, so the value for  $P_5$  is currently unknown. Unless otherwise stated, you should assume that  $P_5$  can take any value in  $[0, 1]$ .

Recall the steps of the Benjamini-Hochberg (BH) procedure:

---

**Algorithm 1** The Benjamini-Hochberg Procedure

---

**input:** FDR level  $\alpha$ , set of  $n$  p-values  $P_1, \dots, P_n$

Sort the p-values  $P_1, \dots, P_n$  in non-decreasing order  $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(n)}$

Find  $K = \max\{i \in \{1, \dots, n\} : P_{(i)} \leq \frac{\alpha}{n} i\}$

Reject the null hypotheses (declare discoveries) corresponding to  $P_{(1)}, \dots, P_{(K)}$

---

- (a) (2 points) If we use the Bonferroni method to control the family-wise error rate (FWER) at level  $2\alpha$  (**not**  $\alpha$ ) for all 5 tests, what is the minimum number of rejections made?
- (b) (3 points) Suppose  $P_5 \geq P_3$ . If we run the BH procedure to control the false discovery rate at level  $\alpha$  for all 5 tests, what are the possible sets of  $P$ -values that could be rejected? For example, a possible answer could be “ $\{P_1, P_2\}$  and  $\{P_1\}$ ”.
- (c) (2 points) What is the largest value of  $P_5$  so that BH makes the maximum number of rejections? You may write your answer in terms of  $\alpha$ .
- (d) (2 points) Suppose that for all five  $P$ -values, the null hypothesis is true. In this case, is it true that  $\text{FWER} = \text{FDR}$ ? Explain why or why not. *Hint: We use the convention that  $0/0 = 0$ .*

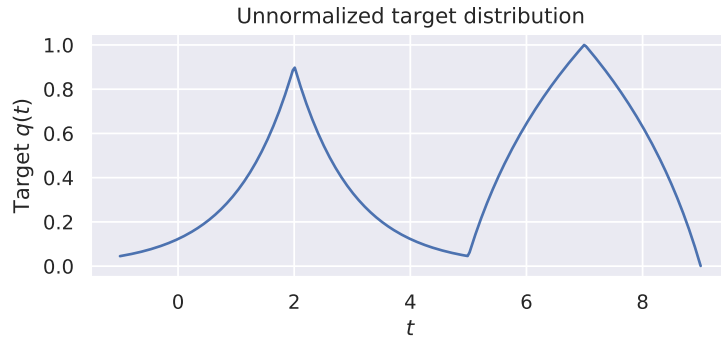
5. (8 points) **Grab Bag:** The parts of this question are all completely unrelated to each other.

(a) (2 points) Suppose we use a generalized linear model (GLM) to predict the number of Piazza followups for a Data 102 lecture thread. We'll let  $y$  be the number of followup posts, and we'll use the number of lecture videos, the length of the lecture videos, and the length of the discussion worksheet to construct  $x$ . Specify a **link function** and a **likelihood model** so that the predictions are always valid followup post counts.

(b) (4 points) Suppose that the prior over random variable  $\theta$  has the following form:  $p(\theta) \propto \theta^{a-1}e^{-b\theta}$ . This is a **Gamma distribution** with parameters  $a$  and  $b$  (called *shape* and *rate* respectively).

We observe two variables  $x$  and  $y$  that are conditionally independent given  $\theta$ . We know that  $x|\theta \sim \text{Exponential}(\theta)$ , and  $y|\theta \sim \text{Poisson}(\theta)$ . Show that the posterior distribution  $p(\theta|x, y)$  is also a Gamma distribution, and express its two parameters in terms of the prior parameters  $a$  and  $b$  and the observed values  $x$  and  $y$ .

(c) (2 points) Suppose we use rejection sampling to approximate the following unnormalized target distribution  $q(t)$  for random variable  $t$ , where  $t \in [-1, 9]$ :



Provide a sampling distribution  $p(t)$  that we can use with rejection sampling, and find a constant  $M$  such that  $Mq(t) \leq p(t)$ .

6. (5 points) **Bayesian fidget spinners**

Nat's company makes and sells fidget spinners. Suppose that when the factory manufactures them, each fidget spinner is defective with probability  $q$ : we'll call this the *defect rate*. She receives  $n$  boxes full of fidget spinners. For each box, she randomly pulls 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  $i$  ( $i = 1, \dots, n$ ). She defines the following Bayesian probability model:

$$\begin{aligned}q &\sim \text{Beta}(\alpha, \beta) \\x_i|q &\sim \text{Geometric}(q)\end{aligned}$$

(You may assume that the number of fidget spinners in each box is much larger than the number she pulls out, so her model is valid.)

- (a) (2 points) Nat is sure that the defect rate is very close to 0.1, and wants her prior distribution  $p(q)$  to reflect that certainty. Which of the following is the best choice for the parameters of her prior distribution (i.e.,  $\alpha$  and  $\beta$ )? **Select only one answer.**
- (A)  $\alpha = 1, \beta = 9$
  - (B)  $\alpha = 10, \beta = 90$
  - (C)  $\alpha = 9, \beta = 1$
  - (D)  $\alpha = 90, \beta = 10$
- (b) (3 points) Nat discovers that her boxes are actually from two different factories, A and B, which have *different* defect rates  $q_A$  and  $q_B$ . Unfortunately, the boxes aren't labeled with which factory they came from. Nat defines a new random variable  $z_i$ :

$$z_i = \begin{cases} 1 & \text{if box } i \text{ came from factory A} \\ 0 & \text{if box } i \text{ came from factory B} \end{cases}$$

She assumes that within any particular box, all the fidget spinners inside are from the same factory.

Suppose there are exactly three boxes. Draw a graphical model illustrating the relationship between  $q_A$ ,  $q_B$ ,  $z_i$ , and  $x_i$ , for  $i = 1, 2, 3$ .