DS102 Fall 2019 - Midterm

First and Last Name: _____

Student ID: _____

- Please write your first and last name as well as your student ID at the top of the first sheet. Also write your student ID on the bottom of each page.
- You have 75 minutes: there are five questions on this exam, with each question being worth an equal amount of points.
- Make sure you have 14 pages. If you do not, let us know immediately.
- Question 1 (true/false) is required.
- For the remaining four questions (Questions 2-5), we will grade all of them and take the top three among these four. You may attempt all questions or skip one depending on time.
- Even if you are unsure about your answer it is better to write down as many details as possible so we can give you partial credit.
- You may, without proof, use theorems and facts that were given in the discussions, lectures or notes.
- We will only grade work on the front of each page unless you indicate otherwise. The exam is printed 1-sided so that you can use the back sides for scratch paper. If you do run out of space on the front, continue on the back side of the page and make a note at the bottom of this cover sheet to let us know.
- Make sure to write clearly. We can't give you credit if we can't read your solutions.

1. (10 points) For each of the following, answer true or false. **Circle T for true and F for false**. You don't need to justify your answer.

   (a) (1 point) (T / F) It is possible to control FDR over an infinite number of tests while always having nonzero probability of declaring a discovery on any given hypothesis.

   Solution: True, e.g LORD Algorithm

   (b) (1 point) (T / F) The following function is linear in x:

   $$f(x, \theta) = 1 + x + \sin(\theta x) + \theta^2 x.$$

   Solution: False, $sin(\theta x)$ is not a linear function of x.

   (c) (1 point) (T / F) Adding a regularization term to logistic regression prevents weights from diverging on linearly separable data.

   Solution: True, recall the plot in Lecture 7.

   (d) (1 point) (T / F) An equal tails credible interval and an interval of highest posterior density are equivalent if the prior distribution is symmetric.

   Solution: False, the intervals are the same if the *posterior* distribution is symmetric, which is not necessarily implied by the prior distribution being symmetric.

   (e) (1 point) (T / F) If we were to repeatedly compute a 90% credible interval from fresh samples, the fraction of calculated intervals that encompass the true parameter would be approximately 90%.

   Solution: False, this is the definition of a confidence interval. (a credible interval may have no relationship with the true parameter).

   (f) (1 point) (T / F) Rejection sampling is in general preferable to importance sampling, as the rejection sampling estimate has lower variance.

   Solution: False, it is the other way round.

   (g) (1 point) (T / F) Gibbs sampling updates one parameter at a time.

Solution: True, see algorithm from lecture 10.

(h) (1 point) (T / F) False discovery proportion can be thought of as a conditional probability that the reality is null (0), given that we made a discovery (1).

Solution: True, see Lecture 2 slides.

(i) (1 point) (T / F) The Hoeffding bound can be used for any random variable with finite variance.

Solution: False, the Hoeffding bound is only defined for bounded random variables.

(j) (1 point) (T / F) The Chebyshev bound allows us to construct confidence intervals for the mean of a bounded random variable.

Solution: True, this is the definition of the Chebyshev bound.

2. (10 points) Suppose you observe two independent Gaussian samples, $Z_1 \sim N(\mu_1, 1)$ and $Z_2 \sim N(\mu_2, 1)$. You want to test two hypotheses, one for each sample: the null hypotheses are $\mu_1 = 0$ and $\mu_2 = 0$, respectively. You compute two p-values as $P_i = \Phi(-Z_i)$, $i \in \{1, 2\}$, where Φ is the standard Gaussian CDF.

   Suppose that the ground truth is null, i.e. $Z_1, Z_2 \sim N(0, 1)$. We consider some $\alpha \in (0, 1)$.

   (a) (5 points) Suppose that you apply the simple decision rule of rejecting when $P_i \leq \alpha$, $i \in \{1, 2\}$. What is the false discovery rate (FDR) of this rule applied to P_1, P_2? Is it less than or equal to α?

   Solution: We can compute the FDR directly by definition:

   $$FDR = 0 \cdot (1 - \alpha)^2 + 1 \cdot (1 - (1 - \alpha)^2) = 2\alpha - \alpha^2.$$

   Since $\alpha^2 < \alpha$ for $\alpha \in (0, 1)$, this is greater than α.

(b) (5 points) Now suppose that you reject when $P_i \leq \frac{\alpha}{2}$, $i \in \{1, 2\}$. What is the false discovery rate (FDR) of this rule applied to P_1, P_2? Is it less than or equal to α?

Solution: By the same argument as in part (a):

$$FDR = 0 \cdot (1 - \alpha/2)^2 + 1 \cdot (1 - (1 - \alpha/2)^2) = \alpha - \frac{\alpha^2}{4} < \alpha.$$

3. (10 points) In this question we analyze the properties of the Benjamini-Hochberg (BH) procedure. Recall the steps of the procedure:

Algorithm 1 The Benjamini-Hochberg Procedure

input: FDR level α, set of n p-values $P_1, \ldots, P_n$

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

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

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

(a) (3 points) Suppose $P_1 = P_2 = \cdots = P_n = \alpha$, and we run BH under level α on these p-values. How many discoveries does BH make? Explain.

Solution: It makes n discoveries, because the highest p-value (equal to α) is less than or equal to $\frac{\alpha}{n} n = \alpha$.

(b) (3 points) Suppose $P_1 = P_2 = \cdots = P_{n-1} = \alpha, P_n = \alpha + 0.001\alpha$, and we run BH under level α on these p-values. How many discoveries does BH make? Explain.

Solution: It makes 0 discoveries, because no p-value is under the corresponding threshold $\frac{\alpha}{n}k$.

(c) (4 points) Suppose we run BH on $\{P_1, \ldots, P_n\}$, and we make $R < n$ discoveries. Now suppose we add an extra p-value equal to 0 to this set. Now we run BH on $\{P_1, \ldots, P_n, 0\}$ and get a new number of rejections R'. Which of the following are possible: $R' > R$, $R' = R$, $R' < R$? If multiple are possible, list all that are possible. Explain why.

Solution: The new p-value 0 is now the smallest one in the sequence. Therefore, if some given p-value was compared to $\frac{\alpha}{n}k$ before adding the extra p-value, now it's compared to $\frac{\alpha}{n+1}(k+1)$. Since $\frac{\alpha}{n+1}(k+1) > \frac{\alpha}{n}k$, it is now strictly easier to discover. Moreover, the new p-value 0 will definitely be discovered, so $R' > R$ is the only possiblity.

4. (10 points) Consider the Gaussian mixture model where your model asserts that the observed data is drawn from the following procedure:
For $i = 1, ..., n$

$$x_i \sim Bernoulli(\theta),$$
$$y_i \sim \mathcal{N}(\mu_{x_i}, \sigma^2_{x_i}),$$

where (x_i, y_i) pairs are observed, but the proportion θ, means μ_0, μ_1 and standard deviations σ_0, σ_1 are fixed and unknown. Recall that

$$\mathcal{N}(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

(a) (3 points) Express the likelihood function $p(x, y; \theta, \mu_0, \mu_1, \sigma_0, \sigma_1)$ in terms of the parameters and the data $x = (x_1, ..., x_n), y = (y_1, ..., y_n)$.

Solution:

$$p(x, y|\theta, \mu_0, \mu_1, \sigma_0, \sigma_1) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)}\mathcal{N}(y_i|\mu_{x_i}, \sigma_{x_i})$$

(b) (2 points) Write an expression for the log likelihood, $\ell(\theta, \mu_0, \mu_1, \sigma_0, \sigma_1; x, y) = \log p(x, y; \theta, \mu_0, \mu_1, \sigma_0, \sigma_1)$.

Solution: The log likelihood is:

$$\sum_{i=1}^{n} \{x_i \log(\theta) + (1 - x_i) \log(1 - \theta) + \log(\mathcal{N}(y_i | \mu_{x_i}, \sigma_{x_i})\})$$
$$= \sum_{i=1}^{n} \left\{ x_i \log(\theta) + (1 - x_i) \log(1 - \theta) - \frac{1}{2} \log(2\pi\sigma_{x_i}^2) - \frac{1}{2\sigma_{x_i}^2} (y_i - \mu_{x_i})^2 \right\}$$

(c) (2 points) Derive the maximum likelihood estimates (MLE) of θ ($\hat{\theta}_{MLE}$) as a function of the observed data $x_1, ..., x_n$ and $y_1, ..., y_n$. *hint: you may want to frame your estimate for $\hat{\theta}_{MLE}$ in terms of the quantity $C = \sum_{i=1}^{n} x_i$.*

Solution: We begin, as always, by writing out the log likelihood:

$$\ell(\theta, \mu_0, \mu_1 | x, y; \sigma_0, \sigma_1) = \sum_{i=1}^{n} \{x_i \log(\theta) + (1 - x_i) \log(1 - \theta) + \mathcal{N}(y_i | \mu_{x_i}, \sigma_{x_i})\}$$

Now differentiate w.r.t. θ:

$$\frac{\partial l}{\partial \theta} \ell(\theta, \mu_0, \mu_1 | x, y; \sigma_0, \sigma_1) = \sum_{i=1}^{n} \left\{ \frac{x_i}{\theta} - \frac{1 - x_i}{1 - \theta} \right\} \qquad = \frac{C}{\theta} + \frac{n - C}{1 - \theta}$$

Setting $\frac{\partial l}{\partial \theta} \ell(\theta, \mu_0, \mu_1 | x, y; \sigma_0, \sigma_1) = 0$ and solving for θ reveals that

$$\hat{\theta}_{MLE} = \frac{C}{n}$$

(d) (3 points) Derive the estimates $\hat{\mu_0}_{MLE}$ and $\hat{\mu_1}_{MLE}$ as a function of the observed data $x_1, ..., x_n$ and $y_1, ..., y_n$.

Solution:

$$\frac{\partial l}{\partial \mu_0}\ell(\theta, \mu_0, \mu_1|x, y; \sigma_0, \sigma_1) = \frac{\partial l}{\partial \mu_0}\left(\sum_{i=1}^{n} \mathbb{I}[x_i = 0]\mathcal{N}(y_i|\mu_0, \sigma_0)\right)$$

Apart from the indicator variable, this is very similar to the MLE for the mean of a sample with i.i.d. draws from a normal distribution with parameters μ_0, σ_0, where the only draws are those that have $x_i = 0$. Therefore, we know (or can derive) the MLE estimator to be the sample average of y_i's for which $x_i = 0$:

$$\hat{\theta}_{0MLE} = \frac{\sum_{i=1}^{n} \mathbb{I}[x_i = 0]y_i}{\sum_{i=1}^{n} \mathbb{I}[x_i = 0]}$$

Similarly, $\hat{\theta}_{1MLE} = \frac{\sum_{i=1}^{n} \mathbb{I}[x_i=1]y_i}{\sum_{i=1}^{n} \mathbb{I}[x_i=1]}$

5. (10 points) Here, we observe data points y_i as draws from the following procedure with *hidden variables* x_i, which we don't observe:

$$x_i = \begin{cases} 1 & \text{with probability } 1/2 \\ 0 & \text{with probability } 1/2 \end{cases}$$

If $x_i = 1$, then $y_i \sim \mathcal{N}(1, \sigma^2)$. Otherwise, $x_i = 0$, and $y_i \sim \mathcal{N}(-1, \sigma^2)$. Note these distributions share the same standard deviation σ.

We don't actually observe the variables x_i, but we'd like to infer them from our observations of y_i.

(a) (4 points) Show that for a given data point i,

$$p(x_i = 1|y_i) = \frac{1}{1 + e^{-f(y_i)}}$$

for some function $f(y_i)$. What is this function $f(y_i)$?

Solution: Using Bayes' rule:

$$\begin{aligned} p(x_i = 1|y_i) &= \frac{p(x_i = 1, y_i)}{p(y_i)} \\ &= \frac{p(y_i|x_i = 1) \cdot p(x_i = 1)}{p(y_i)} \\ &= \frac{p(y_i|x_i = 1)}{2p(y_i)} \end{aligned}$$

Similarly,

$$p(x_i = 0|y_i) = \frac{p(y_i|x_i = 0)}{2p(y_i)}$$

Since there are only two options for x_i,

$$\begin{aligned} p(x_i = 1|y_i) &= \frac{p(x_i = 1|y_i)}{p(x_i = 1|y_i) + p(x_i = 0|y_i)} \\ &= \frac{p(y_i|x_i = 1)}{p(y_i|x_i = 1) + p(y_i|x_i = 0)} \end{aligned}$$

Where that last line simplifies that $p(y_i)$'s and the 2's cancel out. Now we can plug in our Normal likelihood model:

$$\begin{aligned} p(x_i = 1|y_i) &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - 1)^2}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - 1)^2} + \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - (-1))^2}} \\ &= \frac{e^{-\frac{1}{2\sigma^2}(y_i - 1)^2}}{e^{-\frac{1}{2\sigma^2}(y_i - 1)^2} + e^{-\frac{1}{2\sigma^2}(y_i - (-1))^2}} \end{aligned}$$

Multiplying the numerator and denominator by $e^{\frac{1}{2\sigma^2}(y_i-1)^2}$ gives

$$p(x_i = 1|y_i) = \frac{1}{1 + e^{-\frac{1}{2\sigma^2}((y_i-(-1))^2+(y_i-1)^2)}}$$

$$\begin{aligned} f(y_i) &= \frac{1}{2\sigma^2}\left((y_i-(-1))^2 - (y_i-1)^2\right) \\ &= \frac{1}{2\sigma^2}\left((y_i+1)^2 - (y_i-1)^2\right) \\ &= \frac{2y_i}{\sigma^2} \end{aligned}$$

(b) (4 points) Suppose we want to declare a decision rule to classify point i as coming from either $\mathcal{N}(\mu_1 = 1, \sigma^2)$ or $\mathcal{N}(\mu_0 = -1, \sigma^2)$. We will do so by declaring:

$$\hat{x}_i = \arg \max_{x \in \{0,1\}} p(x_i = x | y_i)$$

State this decision rule in terms of the function $f(y)$ from part (a).

Solution: Since there are only two possibilities for x_i,

$$\begin{aligned}
\hat{x}_i &= \mathbb{I}[p(x_i = 1 | y_i; \theta = \frac{1}{2}, \mu_0 = -1, \mu_1 = 1, \sigma_0 = \sigma, \sigma_1 = \sigma) > 1/2] \\
&= \mathbb{I}[\frac{1}{1 + e^{-f(y_i)}} > 1/2] \\
&= \mathbb{I}[e^{-f(y_i)} < 1] \\
&= \mathbb{I}[f(y_i) > 0]
\end{aligned}$$

(c) (2 points) Given the form of $f(y)$ you solved for above, describe this decision rule in one sentence. (If you didn't get a solution to the previous part, you can take a guess given the geometry of the problem.)

Solution: We predict $\hat{x} = 1$ if the observed y is positive, or $\hat{x} = -1$ otherwise. (In general, this decision rule would predict $\hat{x}$ as the category with mean closest to the observed y_i.)