DS102 Spring 2020 - 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 80 minutes: there are five questions on this exam, with each question being worth an equal amount of points.
- Even if you are unsure about your answer it is better to write down partial solutions 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.

Honor Code

I will respect my classmates and the integrity of this exam by following this honor code.

I affirm:

- All of the work submitted here is my original work.
- I did not collaborate with anyone else on this exam.

Signature: ____

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) In the case of a simple null $H_0 : \theta = \theta_0$ and a simple alternative $H_1 : \theta = \theta_1$, the Neyman-Pearson lemma tells us the form of the test that has highest true positive rate, subject to a false positive rate being at most some value.

   Solution: True

   (b) (1 point) (T / F) When we apply Ordinary Least Squares for linear regression, adding more data points always strictly increases the sum of the squared errors with respect to the regression line.

   Solution: False. The added data points will give zero error if they are perfectly on the regression line!

   (c) (1 point) (T / F) Under the null hypothesis, the p-values are uniformly distributed.

   Solution: True

   (d) (1 point) (T / F) The Bonferroni correction controls the family-wise error rate only when all the p-values are independent of each other.

   Solution: False. The Bonferroni correction does not require any assumptions about dependence among the p-values or about how many of the null hypotheses are true.

   (e) (1 point) (T / F) The Gauss-Markov theorem assumes the errors of a linear regression model are identically distributed.

   Solution: False.

   (f) (1 point) (T / F) When running Metropolis-Hastings, the value of the next sample is dependent on values of samples we generated before it.

   Solution: False, it is only dependent on the one right before it.

   (g) (1 point) (T / F) Let $f_{\theta_0}(X)$ and $f_{\theta_1}(X)$ denote the likelihoods of the data X under the null and alternative distributions, respectively. Suppose $\theta_0 < \theta_1$ and we accept the null when $\frac{f_{\theta_0}(X)}{f_{\theta_1}(X)} > \eta$. If we want to create the test with the highest true positive rate, subject to the false positive rate being at most α, the threshold η must be the value such that $Pr(\frac{f_{\theta_0}(X)}{f_{\theta_1}(X)} \leq \eta \mid H_0) = \alpha$.

Solution: True. This is what the Neyman-Pearson Lemma states.

(h) (1 point) (T / F) Any point underneath a convex ROC curve (with false positive rate on the x axis and true positive rate on the y axis) is achievable by probabilistically choosing between two decision rules that lie on the ROC curve for each input sample.

Solution: False. Only convex combinations of points on an ROC curve are achievable this way, so for a convex ROC curve anything under the line $y = x$ is not achievable this way. (We went over this in Discussion 01.)

(i) (1 point) (T / F) In any binary classification problem, it is always possible to create a classifier that achieves a FPR of 0 by classifying every instance as *positive*.

Solution: False. The false positive rate is equal to $\frac{FP}{FP+TN}$. In order to guarantee a 0 FPR, we need FP to equal 0. We need to predict everything as negative to get no false positives.

(j) (1 point) (T / F) Assume:
$w^{(i)} \sim \text{Beta}(3, 5)$
$z^{(i)} \sim N(0, w^{(i)})$
$y^{(i)} = \langle \beta^*, x^{(i)} \rangle + \cos(x^{(i)}) \cdot z^{(i)}$
where the $z^{(i)}$ are independent of each other. The Gauss-Markov theorem allows us to conclude that $\hat{\beta}$, the OLS estimator, is an unbiased estimate of β^*.

Solution: True. All assumptions of the Gauss-Markov theorem needed to conclude unbiasedness of the OLS estimator are satisfied.

2. (10 points) In our discussion of multiple hypothesis testing and the reproducibility crisis in science, we saw that it is not uncommon for practitioners to misinterpret the results of their hypothesis tests. In this problem, we will calculate some metrics related to hypothesis testing which will highlight some limitations of p-values, and then explore the behavior of commonly-used quantities in multiple hypothesis testing problems.

   (a) (4 points) Suppose we are running a single binary hypothesis test. Denote by $H = 0$ that the null hypothesis is true, and by $H = 1$ that the alternative is true. Further, let X denote our observed data, let $p(X)$ be the corresponding p-value for our test, and let δ be our decision rule. We choose to reject the null hypothesis (which corresponds to $\delta(X) = 1$) if $p(X) \leq \alpha = 0.05$; otherwise, we fail to reject the null ($\delta(X) = 0$). Although we often do not know these things in practice, suppose that we also knew the following pieces of information:

      1. The prior probability of a true effect: $\mathbb{P}(H = 1) = 0.1$;
      2. The true positive rate of our testing procedure: $\mathbb{P}(\delta(X) = 1 \mid H = 1) = 0.5$.

   Calculate the posterior probability that the null hypothesis is true given that we reject the null hypothesis. Explain (in roughly 1 sentence) why this probability is not the same as the p-value.

Solution: Using Bayes' rule,

$$
\begin{aligned}
&\mathbb{P}(H = 0 \mid \delta(X) = 1) \\
&\quad= \frac{\mathbb{P}(\delta(X) = 1 \mid H = 0)\mathbb{P}(H = 0)}{\mathbb{P}(\delta(X) = 1)} \\
&\quad= \frac{\mathbb{P}(\delta(X) = 1 \mid H = 0)\mathbb{P}(H = 0)}{\mathbb{P}(\delta(X) = 1 \mid H = 0)\mathbb{P}(H = 0) + \mathbb{P}(\delta(X) = 1 \mid H = 1)\mathbb{P}(H = 1)} \\
&\quad= \frac{\mathbb{P}(p(X) \leq 0.05 \mid H = 0)(1 - \mathbb{P}(H = 1))}{\mathbb{P}(p(X) \leq 0.05 \mid H = 0)(1 - \mathbb{P}(H = 1)) + \mathbb{P}(\delta(X) = 1 \mid H = 1)\mathbb{P}(H = 1)} \\
&\quad= \frac{0.05 \cdot 0.9}{0.05 \cdot 0.9 + 0.5 \cdot 0.1} \\
&\quad\approx 0.47 \\
&\quad\gg 0.05.
\end{aligned}
$$

The p-value is often misinterpreted as the probability that the null hypothesis is true given the data. However, the p-value gives the probability under the null of seeing data as or more extreme that the data we observed. Thus, a small p-value can be thought of as providing evidence against the null hypothesis, but it does not actually quantify our belief that the null hypothesis is true based on the observed data. This is captured by the posterior probability that the null is true, which also takes into account the prior probability of nulls and the power of the test. This posterior probability precisely quantifies our belief that the null hypothesis is true given that the observed p-value is small.

(b) (2 points) Suppose we are running a single binary hypothesis test in the same setting as part (a). We would like to avoid being part of the reproducibility crisis, and thus are interested in understanding the chances that our discovery is real. Calculate the posterior probability that the alternative hypothesis is true given that we reject the null hypothesis. Explain (in roughly 1 sentence) why this is not the same as $1 - 0.05 = 0.95$.

Solution: We could solve this by either using

$$1 - \text{solution in part (a)}$$

or by using Bayes' rule,

$$\begin{aligned}
&\mathbb{P}(H = 1 \mid \delta(X) = 1)\\
&\quad= \frac{\mathbb{P}(\delta(X) = 1 \mid H = 1)\mathbb{P}(H = 1)}{\mathbb{P}(\delta(X) = 1)}\\
&\quad= \frac{\mathbb{P}(\delta(X) = 1 \mid H = 1)\mathbb{P}(H = 1)}{\mathbb{P}(\delta(X) = 1 \mid H = 0)\mathbb{P}(H = 0) + \mathbb{P}(\delta(X) = 1 \mid H = 1)\mathbb{P}(H = 1)}\\
&\quad= \frac{\mathbb{P}(\delta(X) = 1 \mid H = 1)\mathbb{P}(H = 1)}{\mathbb{P}(p(X) \leq 0.05 \mid H = 0)(1 - \mathbb{P}(H = 1)) + \mathbb{P}(\delta(X) = 1 \mid H = 1)\mathbb{P}(H = 1)}\\
&\quad= \frac{0.5 \cdot 0.1}{0.05 \cdot 0.9 + 0.5 \cdot 0.1}\\
&\quad\approx 0.53\\
&\quad\ll 0.95.
\end{aligned}$$

Testing at level 0.05 guarantees that whenever reality is truly null our test has a high probability of not rejecting (e.g. $\mathbb{P}(\delta(X) = 0 | H = 0) \geq 0.95$). This is not the same as a guarantee that whenever our test rejects, the alternative hypothesis is true with a high probability (e.g. $\mathbb{P}(H = 1 \mid \delta(X) = 1)$).

(c) (4 points) Suppose now that, instead of testing a single hypothesis, we are running $m > 1$ independent binary hypothesis tests, H_i. Suppose as well that the prior probability of a true effect is now $\mathbb{P}(H_i = 1) = 0$; that is, all of the hypotheses are truly null. Show that in this setting, the family-wise error rate (probability of making at least one false discovery) and the false discovery rate (expected proportion of false discoveries) are equal.

Note: By convention, the proportion of false discoveries $= \frac{0}{0} = 0$ *when the total number of discoveries is* 0.

Solution: Let F be the number of false discoveries and T be the total number of discoveries. Since all the hypotheses are truly null, any discovery is a false discovery. This means that $F = T$ always holds. Therefore,

$$\begin{aligned}\mathrm{FDR} &= 0 * \mathbb{P}\left(\frac{F}{T} = 0\right) + 1 * \mathbb{P}\left(\frac{F}{T} = 1\right) \\ &= \mathbb{P}(F \geq 1) \\ &= \mathrm{FWER}.\end{aligned}$$

An alternative way to derive the relationship is to use an indicator random variable:

$$\begin{aligned}\mathrm{FDR} &= \mathbb{E}\left[\frac{F}{T}\right] \\ &= \mathbb{E}\left[\mathbb{1}\{F \geq 1\}\right] \\ &= \mathbb{P}(F \geq 1) \\ &= \mathrm{FWER}.\end{aligned}$$

3. (10 points) Graphical models are often useful for modeling phenomena involving multiple variables. In this problem, you'll formulate a graphical model, then demonstrate how to sample from the posterior using Gibbs sampling.

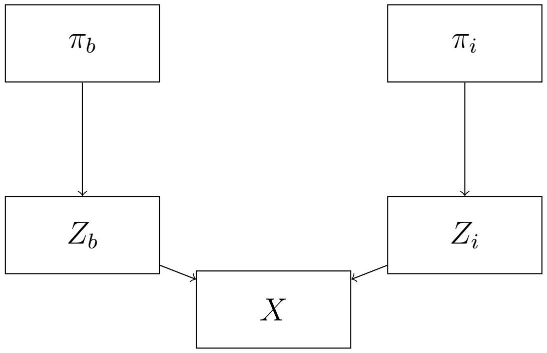
(a) (2 points) Consider the following scenario: suppose the probability that a burglar breaks into your car is π_b, and the probability that an innocent passerby accidentally touches your car is π_i. Let Z_b be a binary random variable that is 1 if there is a burglar, and 0 otherwise. Likewise, let Z_i be a binary random variable that is 1 if there is an innocent passerby, and 0 otherwise. Suppose Z_b and Z_i are independent of each other.

Let X be a binary random variable that is 1 if your car alarm goes off. The probability your car alarm goes off depends on Z_b and Z_i, and is known to be:

$\mathbb{P}(X = 1 \mid Z_b, Z_i)$	Z_b	Z_i
0	0	0
0.1	0	1
0.95	1	0
0.99	1	1

Draw the graphical model that describes the direct relationships between π_b, π_i, Z_b, Z_i, and X.

Solution:



(b) (5 points) Suppose you know the parameters π_b and π_i, as well as $\mathbb{P}(X = 1 \mid Z_b, Z_i)$ as specified in Part (a). X is the observed variable, and Z_i and Z_b are the latent (unobserved) variables. We want to sample from $\mathbb{P}(Z_i, Z_b \mid X, \pi_b, \pi_i)$, the posterior over the latent variables conditioned on everything else. We'll use Gibbs sampling to do this.

(i) Suppose we are running Gibbs sampling, and on each iteration we sample Z_b first then Z_i. We observed $X = 0$, and the values of Z_b and Z_i from iteration t are $Z_b^{(t)} = 0$ and $Z_i^{(t)} = 1$.

Derive the distribution used for the Gibbs sampling update of $Z_b^{(t+1)}$. Your solution should be in terms of π_b, π_i, and constants.

Solution: For the Gibbs sampling update of $Z_b^{(t+1)}$, we have

$$\begin{aligned}
\mathbb{P}(Z_b \mid Z_i, X, \pi_b, \pi_i) &= \frac{\mathbb{P}(Z_b, Z_i, X \mid \pi_b, \pi_i)}{\mathbb{P}(Z_i, X \mid \pi_b, \pi_i)} \\
&= \frac{\mathbb{P}(X \mid Z_b, Z_i, \pi_b, \pi_i)\mathbb{P}(Z_b, Z_i \mid \pi_b, \pi_i)}{\mathbb{P}(Z_i, X \mid \pi_b, \pi_i)} \\
&= \frac{\mathbb{P}(X \mid Z_b, Z_i)\mathbb{P}(Z_b \mid \pi_b)\mathbb{P}(Z_i \mid \pi_i)}{\mathbb{P}(Z_i, X \mid \pi_b, \pi_i)} \\
&= \frac{\mathbb{P}(X \mid Z_b, Z_i)\mathbb{P}(Z_b \mid \pi_b)\mathbb{P}(Z_i \mid \pi_i)}{\sum_{z \in \{0,1\}} \mathbb{P}(Z_b = z, Z_i, X \mid \pi_b, \pi_i)} \\
&= \frac{\mathbb{P}(X \mid Z_b, Z_i)\mathbb{P}(Z_b \mid \pi_b)\mathbb{P}(Z_i \mid \pi_i)}{\sum_{z \in \{0,1\}} \mathbb{P}(X \mid Z_b = z, Z_i)\mathbb{P}(Z_b = z \mid \pi_b)\mathbb{P}(Z_i \mid \pi_i)}.
\end{aligned}$$

Now we plug in the specific values given in the problem, including $X = 0$ and $Z_i = Z_i^{(t)} = 1$. Since Z_b is a binary random variable, to find its distribution we can just find $\mathbb{P}(Z_b = 1 \mid Z_i = 1, X = 0, \pi_b, \pi_i)$. Plugging in the values of $\mathbb{P}(X \mid Z_b = 1, Z_i = 1)$ given in Part (a), and $\mathbb{P}(Z_b = 1) = \pi_b$ and $\mathbb{P}(Z_i = 1) = \pi_i$, we have

$$\begin{aligned}
&\mathbb{P}(Z_b = 1 \mid Z_i = 1, X = 0, \pi_b, \pi_i) \\
&= \frac{\mathbb{P}(X = 0 \mid Z_b = 1, Z_i = 1)\mathbb{P}(Z_b = 1 \mid \pi_b)\mathbb{P}(Z_i = 1 \mid \pi_i)}{\sum_{z \in \{0,1\}} \mathbb{P}(X = 0 \mid Z_b = z, Z_i = 1)\mathbb{P}(Z_b = z \mid \pi_b)\mathbb{P}(Z_i = 1 \mid \pi_i)} \\
&= \frac{0.01 \cdot \pi_b \pi_i}{0.9 \cdot (1 - \pi_b) \cdot \pi_i + 0.01 \cdot \pi_b \pi_i}.
\end{aligned}$$

That is, $Z_b^{(t+1)}$ is a Bernoulli random variable with probability of one equal to

$$\frac{0.01 \cdot \pi_b \pi_i}{0.9 \cdot (1 - \pi_b) \cdot \pi_i + 0.01 \cdot \pi_b \pi_i}.$$

(ii) Now, suppose we draw $Z_b^{(t+1)} = 1$ from the distribution derived in Part (b.i). Derive the distribution used for the Gibbs sampling update of $Z_i^{(t+1)}$. Your solution should be in terms of π_b, π_i, and constants.

Solution:

By the same reasoning as Part (b.i), for the Gibbs sampling update of $Z_i^{(t+1)}$ we can focus on finding $\mathbb{P}(Z_i = 1 \mid Z_b = 1, X = 0, \pi_b, \pi_i)$ (since we drew $Z_b^{(t+1)} = 1$):

$$\begin{aligned}
&\mathbb{P}(Z_i = 1 \mid Z_b = 1, X = 0, \pi_b, \pi_i) \\
&= \frac{\mathbb{P}(X = 0 \mid Z_b = 1, Z_i = 1)\mathbb{P}(Z_b = 1 \mid \pi_b)\mathbb{P}(Z_i = 1 \mid \pi_i)}{\sum_{z \in \{0,1\}} \mathbb{P}(X = 0 \mid Z_b = 1, Z_i = z)\mathbb{P}(Z_i = z \mid \pi_i)\mathbb{P}(Z_b = 1 \mid \pi_b)} \\
&= \frac{0.01 \cdot \pi_b \pi_i}{0.05 \cdot (1 - \pi_i)\pi_b + 0.01 \cdot \pi_b \pi_i}.
\end{aligned}$$

4. (10 points) In this question, we will be working with ROC curves and fairness.

(a) (2 points) First, we will plot a single point on an ROC curve. Let $Y \in \{0, 1\}$ be a binary target, and let $X \in \mathbb{R}$ be a single feature used to predict Y. Consider the function $f(X) = \frac{e^X}{e^X+1}$. Suppose "Dataset A" contains six samples of $X, Y, f(X)$, shown in Table 1 below.

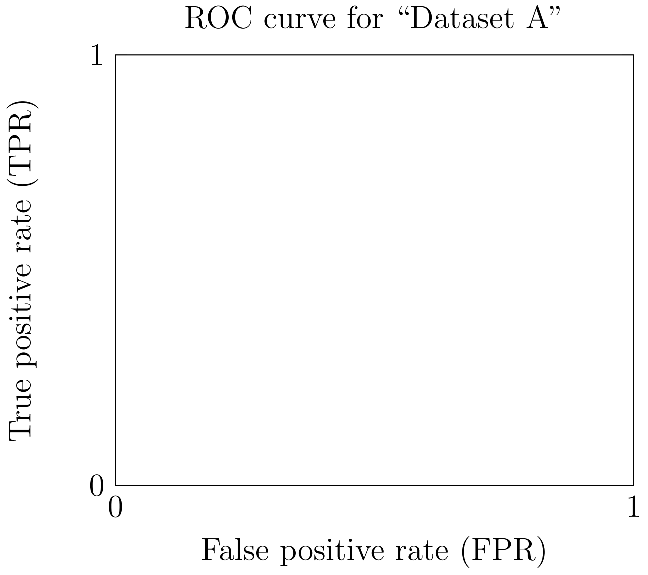
Table 1: Dataset A samples

Y	$f(X)$	X
0	0.1	-2.197
1	0.2	-1.386
0	0.4	-0.405
1	0.7	1.386
1	0.9	2.197
0	0.95	2.944

Suppose we have a decision rule,

$$\delta_t(X) = \begin{cases} 1 & \text{if } f(X) > t \\ 0 & \text{if } f(X) \leq t \end{cases}$$

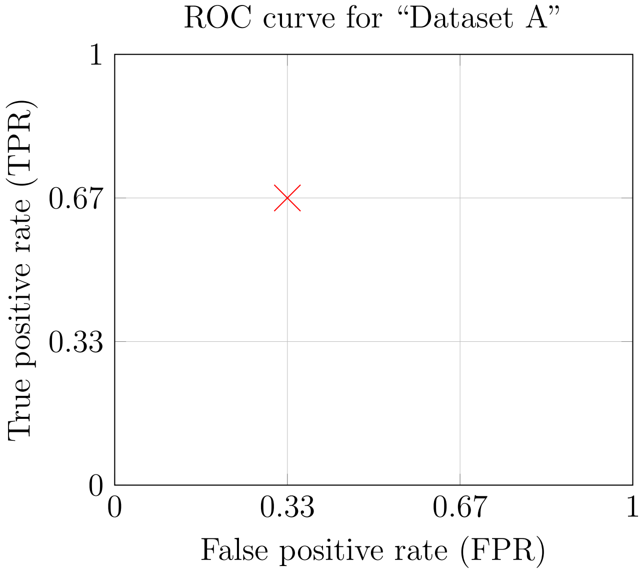
If we choose the decision threshold $t = 0.5$, what is the empirical true positive rate and the empirical false positive rate for the decision rule $\delta_t(X)$ on "Dataset A" above? In addition to calculating the true positive rate and false positive rate, plot the single point on the blank ROC curve below that corresponds to the decision rule $\delta_t(X)$ with decision threshold $t = 0.5$. You should both (i) write down numerical values for true positive rate and false positive rate, and (ii) draw a point in the plot. **Blank ROC curve: plot your answer here (you only need to plot a single point).** If writing answers on a separate sheet of paper, you may also copy this blank plot (with title and axes) onto your answer sheet, and plot your answer there.



Solution:

$$\text{TPR} = \frac{\#\text{ true positives}}{\#\text{ positives}} = \frac{2}{3}$$

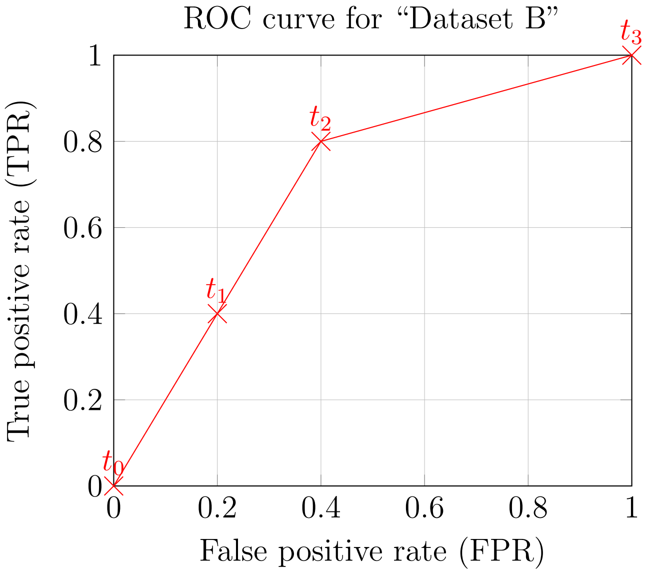
$$\text{FPR} = \frac{\#\text{ false positives}}{\#\text{ negatives}} = \frac{1}{3}$$



(b) (2 points) Now suppose that we have a different "Dataset B", where we again have samples of $X, Y, f(X)$ (similar to part (a)). Let $\delta_t(X)$ be defined as in part (a). This time, we want to choose a single decision threshold t that maximizes the *accuracy* of the decision rule $\delta_t(X)$, where

$$\begin{aligned} \text{accuracy} &= (\text{true positive rate}) * (\text{fraction of positives}) \\ &\quad + (\text{true negative rate}) * (\text{fraction of negatives}). \end{aligned}$$

Suppose "Dataset B" has the following ROC curve:



Each mark "×" represents the TPR and FPR for a single threshold t_0, t_1, t_2, or t_3, labeled on the plot.

Suppose that for Dataset B, (fraction of positives) = (fraction of negatives) = 0.5.

Which of the decision thresholds (t_0, t_1, t_2, or t_3) yields a decision rule with the highest accuracy? What is the accuracy for the decision threshold you chose?

Solution: t_2 has the highest accuracy, which is $0.5 * 0.8 + 0.5 * 0.6 = 0.7$.

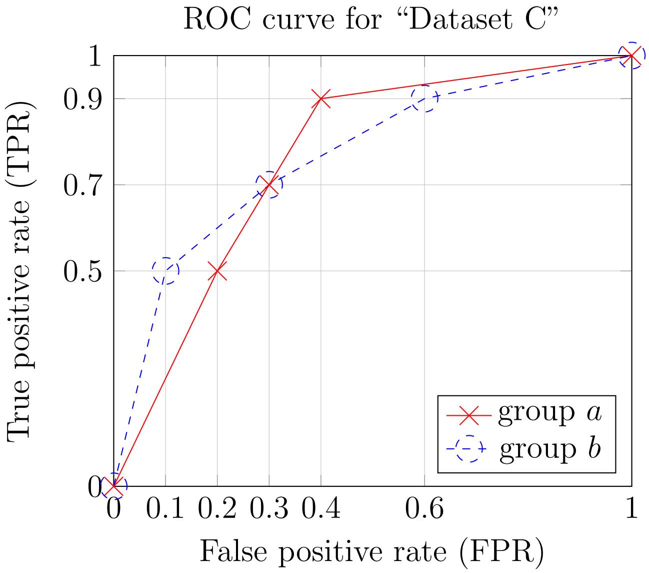
To find this, observe that accuracy $= 0.5 * \text{TPR} + 0.5 * (1 - \text{FPR})$. This implies that $\text{TPR} = \text{FPR} - 1 + 2 * \text{accuracy}$. This describes level sets of (TPR, FPR) pairs that have the same accuracies: for example, all points (TPR, FPR) on the line $\text{TPR} = \text{FPR} - 1$ have accuracy 0, and all points on the line $\text{TPR} = \text{FPR} + 1$ have accuracy 1.

Out of the possible decision thresholds, the decision threshold corresponding to the point on a line of slope 1 with the highest intercept is t_2, so t_2 must be the threshold that achieves the highest accuracy.

The students could also calculate the accuracies for each of the thresholds using brute force.

(c) (6 points) Now suppose we have a third dataset, “Dataset C”. Dataset C has samples of $X, Y, f(X)$ (similar to Datasets A and B), but also includes an additional variable $Z \in \{a, b\}$ that marks group membership. If $Z = a$, the data point belongs to group a, and if $Z = b$, the data point belongs to group b.

The following plot shows the ROC curves for each group:



Suppose we are choosing group dependent thresholds t_a and t_b (that is, we have decision rule $\delta_{t_a}(X)$ for group a, and decision rule $\delta_{t_b}(X)$ for group b, where $\delta_t(X)$ is defined as in part (a)).

(i) What is the maximum accuracy achievable in each group via group dependent thresholds t_a and t_b? Suppose as in part (b) that for each group, (fraction of positives) = (fraction of negatives) = 0.5. (Your solution should be two numbers. One accuracy number for each group.)

(ii) Suppose group a makes up 70% of the dataset and group b makes up 30% of the dataset. What is the overall accuracy of the classifier from part (i) on the whole dataset? (The answer is one accuracy number.)

(iii) Now determine the maximum accuracy on the whole dataset of a classifier that satisfies error rate parity. Recall, error rate parity requires the classifier to have the same TPR in both groups and the same FPR in both groups. (The answer is one number. The classifier may use group dependent thresholds.)

Solution:

(i) For group a, the maximum accuracy achievable is $0.6 * 0.5 + 0.9 * 0.5 = 0.75$. For group b, the maximum accuracy achievable is $0.5 * 0.9 + 0.5 * 0.5 = 0.7$.

(ii) The overall accuracy is $0.7 * 0.75 + 0.3 * 0.7 = 0.735$

(iii) As shown in Lecture 05 (link to scribe notes), any point that lies under all individual group ROC curves is a possible point where group dependent decision rules (possibly stochastic) can be chosen to achieve error rate parity. The point under both curves with the maximum accuracy is the point $(0.3, 0.7)$. The point $(0.3, 0.7)$ has accuracy $0.5 * 0.7 + 0.5 * 0.7 = 0.7$.