Ethical Algorithm Design
CIS 4230/5230
Midterm Examination
Prof Michael Kearns
March 16, 2023

This exam is closed book and closed notes, with no calculators or phones. The only thing that should be on your desk is the exam and a writing implement.

If you need more space for a problem, feel free to use the back side of the sheets, but clearly label which problem you are continuing.

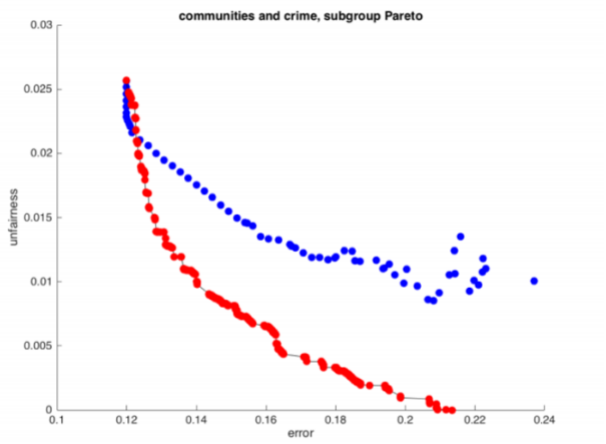
Problem 1:	**____/10**
Problem 2:	**____/10**
Problem 3:	**____/10**
Problem 4:	**____/10**
Problem 5:	**____/15**
Problem 6:	**____/10**
Problem 7:	**____/15**
Problem 8:	**____/10**
Problem 9:	**____/10**
Total:	**____/100**

1. **(10 points)** For each lettered item on the left, write the number of the item on the right that matches best.

(a) Northpointe ___	1. Error/unfairness trade-off
(b) One-time pad ___	2. Reconstruction attack
(c) Convergence of train and test errors ___	3. Impossibility theorem
(d) Randomized Response ___	4. Post-processing
(e) Bias bounties ___	5. Group fairness definition
(f) U.S. Census data ___	6. Plausible deniability
(g) Pareto frontier ___	7. (g,h) pairs
(h) Achieving many fairness notions at once ___	8. Cryptography
(i) Equalization of error rates ___	9. COMPAS
(j) “Bolt-on” fairness ___	10. Fundamental Theorem of ML

2. **(10 points)** In lecture we discussed a number of ways in which training a model in the "usual" manner (i.e. simply finding the model in our class with the smallest error on the training data) can result in a model that is unfair to one or more demographic groups. Briefly review the ways this can happen, and for each one, suggest any mitigation techniques or steps.

3. **(10 points)** The following image was discussed both in lecture and in one of the readings, when we were discussing the topic of “fairness gerrymandering”.



(a) Briefly but clearly describe what is meant by fairness gerrymandering.

(b) Briefly but clearly describe what the image is illustrating, including what the x and y axes measure, what each red and blue dot represent, and the main points the image is making.

4. **(10 points)** In lecture, we discussed why anonymization techniques or aggregation don't provide meaningful privacy guarantees. Clearly describe what is meant by anonymization and aggregation, and what types of attacks each is vulnerable to.

5. **(15 points)** You are an ML product manager, and your engineering team has delivered a classifier h(x) whose error rate on group A is 0.1, and whose error rate on group B is 0.2. Groups A and B are disjoint, and group A constitutes 0.6 or 60% of the overall population, and group B constitutes 0.4 or 40% of the population. In the following you can assume the inputs x indicate whether an individual is in group A or group B.

(a) What is the error rate of h(x) on the overall population? Give a numerical answer and show your work.

(b) For regulatory reasons, you can only use a classifier whose error rates on A and B are equal. Describe a randomized classifier g(x) that uses h(x), equalizes the error rates on A and B, and keeps the overall error rate as small as possible. You do not have to use the scheme described in lecture, but your description of g(x) should be precise.

(c) What is the error rate of your g(x) on groups A and B? Give a numerical answer and show your work.

(d) What is the overall error rate of your g(x)? Give a numerical answer and show your work.

(e) You subsequently learn that your engineering team chose the original h(x) to minimize overall error within some model class H. You ask them to instead find model in H that minimizes the overall error in H, but subject to the condition that the error rates on A and B be equal. Why might you prefer this to what you did in part (b)?

6. **(10 points)** Clearly label each of the following assertions as “true” or “false”.

(a) The inputs to the COMPAS risk assessment tool are the records of criminal defendants in the Broward County, FL dataset.

(b) Given an algorithm for training models without any fairness considerations, it is sometimes possible to design an algorithm training models obeying fairness considerations.

(c) The inclusion of some features in x may result in models that disfavor a subpopulation.

(d) In the error-unfairness Pareto diagrams studied in lectures, the only models worth considering were those that had nothing to their northeast.

(e) The one-time pad requires all pairs of users who wish to communicate securely to share a unique key.

(f) The “no harm whatsoever” notion of privacy discussed in lecture successfully balances the desire for privacy and important societal uses of data.

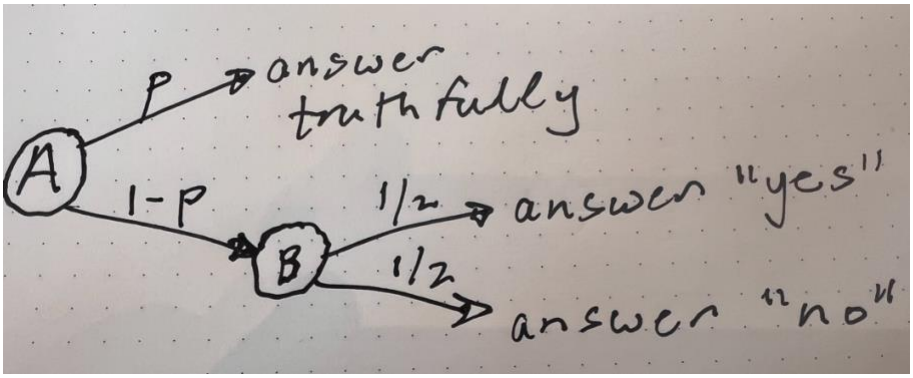
(g) As long as we train a model to be fair by race, and also fair by gender, it will automatically be fair by any race-gender pair (such as Black women).

(h) One of the properties of the “bias bounty” framework is the ability to reduce subgroup errors while also reducing the overall error.

(i) In our formal framework for discussing ML, we made strong assumptions on the distribution P over <x,y> pairs, but no assumptions on the models h(x) we fit to data.

(j) Security and privacy are distinct but complementary notions.

7. **(15 points)** The diagram below represents a generalization of the Randomized Response protocol discussed in lecture, for soliciting noisy answers to a potentially stigmatizing question such as "Were you privately rooting for the Kansas City Chiefs to win the Super Bowl?"



At step A, a participant is asked to flip a *biased* coin: with probability p (for some fixed p in [0,1]), the coin lands heads, in which case the participant follows the top arrow and is told to answer the question truthfully. With probability 1-p at step A, the coin lands tails, in which case the participant proceeds to step B, where they flip a fair/unbiased coin. If it lands heads (top branch from B), they are told to answer "yes" (regardless of what their true answer is), if it lands tails (bottom branch from B), they are told to answer "no" (regardless of what their true answer is). Note that standard Randomized Response uses the value p = ½ for the flip at A.

(a) In class we used the notation Pr["yes"|yes] (respectively, Pr["yes"|no]) for the probability over the coin flips that a participant whose true answer is yes (respectively, no) ends up answering "yes" in the protocol. Carefully compute these two conditional probabilities as a function of p, and simplify your answers.

(b) Based on your answers to part (a), compute the ratio Pr["yes"|yes]/Pr["yes"|no] and simplify your answer.

(c) The ratio in part (b) measures how many times more likely a “yes” response to the protocol is from someone whose truth is yes than whose truth is no, and thus can be viewed as an informal privacy measure. Based on your answer to part (b), what is the possible range of values for this ratio over all p in [0,1]?

(d) Based on your answers to parts (b) and (c), does a smaller or larger value of p provide more privacy? Briefly justify your answer.

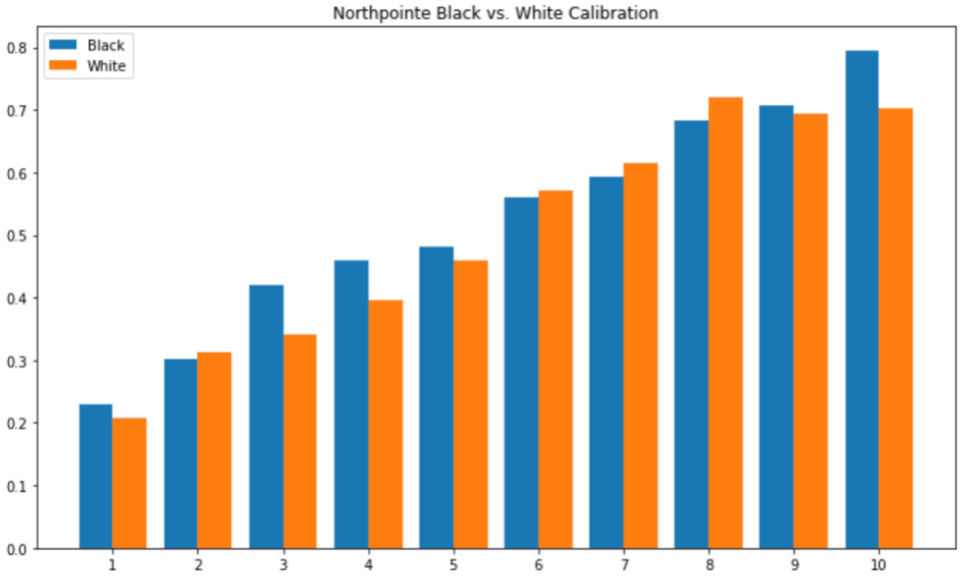
(e) Remembering that the goal of Randomized Response is to provide individual respondents “plausible deniability” while still allowing us to estimate the fraction of the population whose true answer is yes, and in light of your answer to part (d), what would be the disadvantage of choosing a value of p that provides an extremely high level of privacy? Justify your answer, ideally mathematically.

8. **(10 points)** Large language models such as ChatGPT take a natural language text prompt as input, and then generate rich extensions to the prompt by repeatedly predicting the probability distribution over the most likely next words, and then drawing from that distribution. The following is an example in which the first sentence is the prompt, and the text in green was generated by the language model:

Doctor Hansen looked at the chart with a slightly puzzled expression. He had been studying it for some time and was not sure what to make of it. He knew that the patient had a history of illness, but was not sure what was causing the current problems. He decided to order some additional tests to try to get to the bottom of it.

Write a brief essay (3-4 paragraphs) in which you discuss potential ethical or societal concerns with such models (a wide range of answers is possible/acceptable here, just demonstrate some careful thought). Be sure to include whether or not you think any of the technical approaches we have discussed in class are relevant to such models, and justify your answer.

9. **(10 points)** The image below was generated by a member of the class in response to a suggestion from Prof Kearns, and has its title preserved but the axis labels removed.



(a) Carefully and clearly describe what numerical quantity each axis represents, and what the difference between the blue and orange bars is.

(b) Concisely discuss what overarching message the figure is conveying. Which party's position is it supporting better, and why?