Ethical Algorithm Design
CIS 4230/5230
Final Examination
Prof Michael Kearns
May 6, 2024

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.

Your Name:

Your Penn ID:

Problem 1: ____/10

Problem 2: ____/15

Problem 3: ____/15

Problem 4: ____/10

Problem 5: ____/10

Problem 6: ____/10

Problem 7: ____/10

Problem 8: ____/10

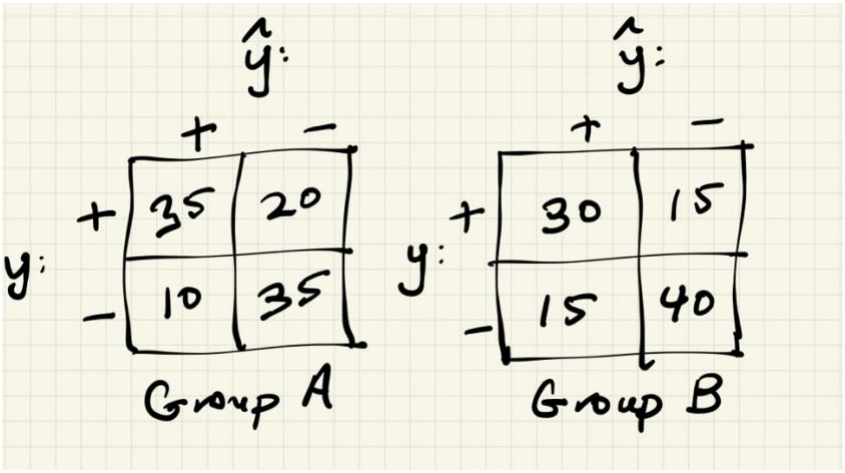
Problem 9: ____/10

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) Pareto frontier ___	1. Disgorgement type
(b) Change in function value ___	2. MIA
(c) Apple controversy ___	3. Privacy notion
(d) Reactive ___	4. Last mile of responsible AI
(e) No fairness/accuracy tradeoff ___	5. Sensitivity
(f) Privacy attack ___	6. Copyright concern
(g) Lack of downstream control ___	7. Convex
(h) No harm whatsoever ___	8. Privacy budget
(i) Regurgitation ___	9. Poison and cure
(j) DP/reconstruction attack ___	10. Bias bounties

2. **(15 points)** The tables below show the confusion matrices for a model on two disjoint subgroups of a dataset.



a) What are the overall error rates of this model on groups A and B? Would you consider this model to be fair with respect to overall error rates? Why or why not?

b) What are the false positive rates of this model on groups A and B? Would you consider this model to be fair with respect to false positive rates? Why or why not?

c) What are the false negative rates of this model on groups A and B? Would you consider this model to be fair with respect to false negative rates? Why or why not?

d) If you concluded that the model is fair in b) and c), simply write "N/A" below. If you concluded the model is unfair in b) and/or c), give specific application domains (e.g. recidivism prediction, consumer lending, etc.) in which one or the other group is "favored" or "disfavored" and explain why.

3. **(15 points)** Each of the functions f(x) below take a vector x of n real-valued inputs x_i, each of which is in the range [0,1]. For each function, write a precise expression for its sensitivity, describe the worst-case neighboring x and x' that prove this sensitivity, and then give an exact expression for the variance parameter b in the Laplace Mechanism in order to achieve ε-differential privacy in computing f(x).

a)

$$f(x) = \frac{1}{n} \sum_i \sqrt{x_i}$$

b)

$f(x)$ = value of first x_i such that $x_i \geq \max(x_1, x_2, \ldots, x_{i-1})$

c)

$$f(x) = \frac{1}{n} \sum_i 1/x_i$$

4. **(10 points)** The passage below is from one of the assigned readings. In as much detail as possible, discuss what the main topic of the article is, and what conclusion it reaches about the three proposed conditions. Then briefly discuss the broad implications of this finding for the ProPublica-COMPAS debate.

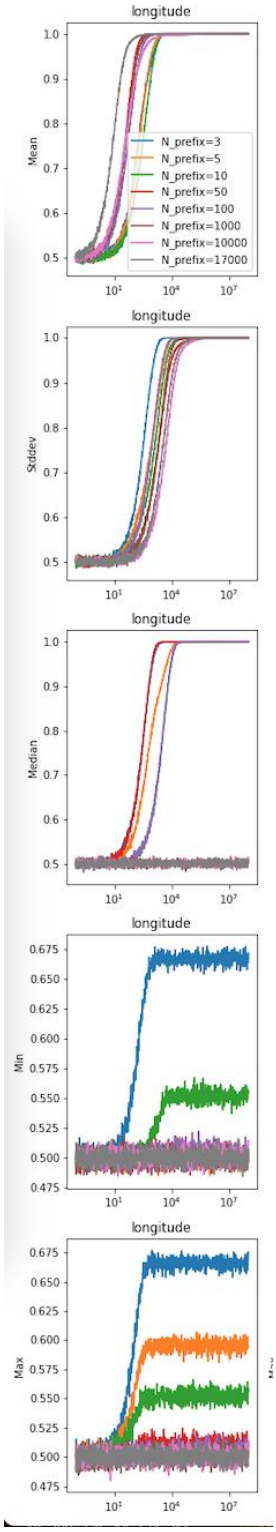
Fairness Properties for Risk Assignments. Within the model, we now express the three conditions discussed at the outset, each reflecting a potentially different notion of what it means for the risk assignment to be "fair."

(A) *Calibration within groups* requires that for each group t, and each bin b with associated score v_b, the expected number of people from group t in b who belong to the positive class should be a v_b fraction of the expected number of people from group t assigned to b.

(B) *Balance for the negative class* requires that the average score assigned to people of group 1 who belong to the negative class should be the same as the average score assigned to people of group 2 who belong to the negative class. In other words, the assignment of scores shouldn't be systematically more inaccurate for negative instances in one group than the other.

(C) *Balance for the positive class* symmetrically requires that the average score assigned to people of group 1 who belong to the positive class should be the same as the average score assigned to people of group 2 who belong to the positive class.

5. **(10 points)** The figure along the left below show sample plots from one of the homework assignments. Discuss this figure in as much detail as you can, including what the axes are measuring, what the different curves represent, why the different subplots look different, and how the figure relates to the theory discussed in class.



6. **(10 points)** In lectures, after giving the definition of differential privacy and a few different specific algorithms achieving DP, we then studied/proved a number of “nice” properties that DP algorithms enjoy in general. As precisely as you can, recall and describe what these nice properties are and why they are useful/important.

7. **(10 points)** Clearly label each of the following assertions as “true” or “false”.

(a) The definition of differential privacy ensures that no harm whatsoever will come to individuals as a result of DP computations using their data.

(b) Smaller values of ε in differential privacy correspond to adding less noise.

(c) Sometimes allowing race as an input to a model can allow the model to be more fair by race.

(d) Dataset emulation is a type of model disgorgement method.

(e) The composition of two ε-DP algorithms is also ε-DP.

(f) Tradeoffs between accuracy and fairness arise from expressing fair ML as a constrained optimization problem.

(g) If f(x) and g(x) are two functions that both obey some notion of fairness, then f(g(x)) will also obey that notion of fairness.

(h) The problem of finding the best linear separator for positively and negatively labeled points in high dimension has a fast algorithm.

(i) The problem of finding points on the Pareto frontier of the bolt-on fairness method has a fast algorithm.

(j) Modern optimization methods can be used both for DP synthetic data generation and for reconstruction attacks.

8. **(10 points)** Let f_1(x) be a function whose sensitivity is some expression s_1, and let f_2(x) be a function whose sensitivity is some expression s_2. For each of the functions f(x) below, write down an upper bound on its sensitivity.

    a) f(x) = f_1(x) + f_2(x)

    b) f(x) = f_1(x)*f_2(x)

    c) f(x) = max(f_1(x), f_2(x))

    d) f(x) = f_1(x)/f_2(x)

9. **(10 points)** For either our study of fairness in ML, or for our study of differential privacy, write a brief summary of what we learned. Include discussion of the problems/concerns, the scientific approaches we considered, their strengths and limitations, and the real-world case studies we examined.