DCDS 500: Exploration in Computational and Data Sciences
Spring 2024

Project 1

Due 11:59 am, March 05, 2024 (on Canvas)

Table 1 tabulates the information for the college application example we reviewed in class. Recall that, in this problem, the population is *equally* divided into two groups g – an advantaged group A and a disadvantaged group D.

The features in this problem are two Boolean features x_1 and x_2. Applicants from group A, due to their advantaged background, have $x_i = 1$ with probability $\frac{2}{3}$ for both values of $i = \{1, 2\}$. On the other hand, applicants from group D, due to their disadvantaged background, have $x_i = 1$ with a smaller probability $\frac{1}{3}$ for both values of $i = \{1, 2\}$.

For simplicity, the potential of a student with feature vector x is defined by the function $f(x) = x_1 \wedge x_2$. Finally, w denotes the fraction of the population with the characteristics in each row. For example, $\frac{1}{18}$ of the population comes from a disadvantaged background D and has both $x_1 = x_2 = 1$.

The table is sorted in decreasing f-value. Thus, if the top $\frac{5}{18}$ of applicants with the highest f-value is admitted, then every admitted student will have an f-value of 1, $\frac{1}{5}$ of the admitted students will come from a disadvantaged background, and $\frac{4}{5}$ of the admitted students will come from an advantaged background.

x_1	x_2	g	f	w
1	1	D	1	$\frac{1}{18}$
1	1	A	1	$\frac{4}{18}$
1	0	D	0	$\frac{2}{18}$
1	0	A	0	$\frac{2}{18}$
0	1	D	0	$\frac{2}{18}$
0	1	A	0	$\frac{2}{18}$
0	0	D	0	$\frac{4}{18}$
0	0	A	0	$\frac{1}{18}$

Table 1: Example College Application Problem

In the class lecture, we discussed several scenarios in which the distribution of the groups of admitted students can differ if:

- Do not use x_2 and do not use group membership information.
- Do not use x_2, but use group membership information.
- Use x_2 for students in the disadvantaged group and do not use group membership information otherwise.

Part 1

You are to simulate this example college application problem in a programming language/environment of your choice and empirically verify if the efficiency (= average f-value of admitted students) and equity (= fraction of admitted students belonging to the disadvantaged group) are consistent with the theoretical efficiency and equity discussed in class.

Specifically, in your simulation, assume that there are 100 students in the entire population, where 50 students belong to the advantaged group, and the remaining 50 students belong to the disadvantaged group. Then, for each student, initialize their x_1 and x_2 values using the probabilities defined above and compute their f-values. Finally, for each of the four scenarios (the original scenario and the three variants described above), rank the students into groups appropriately and pick $\frac{5}{18}$ of the students to be admitted into college.

In your report, describe the efficiency and equity for each scenario. If they are inconsistent with the theoretical efficiency and equity, explain why.

Part 2

Now consider an evolutionary variant of this problem, where we refer to the initial population as Generation 1. At the end of the computation in a generation (i.e., after the determination of who is admitted into college and who is not), each person (i.e., parent) will "produce" a new person (i.e., child) to replace them in the next generation.

The characteristics of each child may depend on the characteristics of its parent and whether the parent was admitted into college. You are free to model how the characteristics of a child depend on the characteristics of the parent. However, provide a realistic rationale for your model.

As an example (you do not have to use this example!), among those admitted into college: (1) If they are from an advantaged group, then their child will be in an advantaged group; (2) If they are from a disadvantaged group, then their child will have a $p_1 = \frac{1}{2}$ of belonging to an advantaged group. Among those who were not admitted into college: (1) If they are from a disadvantaged group, then their child will be in a disadvantaged group; (2) If they are from an advantaged group, then their child will have a $p_2 = \frac{1}{3}$ of belonging to an advantaged group. The feature values x_1 and x_2 of the child will then be initialized based on their group membership, like in Generation 1.

Once the new generation is initialized, determine who is admitted into college and who is not in this new generation again, just like in Generation 1. Repeat this process for 100 generations.

Consider several questions of interest that are applicable in this evolutionary problem. For example, is there evidence of social mobility (i.e., the fraction

x_1	x_2	g	f	w	e
1	1	D	1	$\frac{1}{18}$	0.75
1	1	A	1	$\frac{4}{18}$	0.99
1	0	D	0	$\frac{2}{18}$	0.50
1	0	A	0	$\frac{2}{18}$	0.75
0	1	D	0	$\frac{2}{18}$	0.25
0	1	A	0	$\frac{2}{18}$	0.50
0	0	D	0	$\frac{4}{18}$	0.05
0	0	A	0	$\frac{1}{18}$	0.25

Table 2: Example College Application Problem with Outcome Probabilities

of the advantaged group grows over time)? If so, what is the difference in the mobility between the four scenarios? How are they dependent on the parameters in your model of how the characteristics of children depend on the characteristics of their parents? Are these results consistent with your observations in the real world?

Describe your evolutionary model and rationale for that model in your report. Also, plot graphs to illustrate your results and describe your observations in the report.

Part 3

Given insights from Part 1 and Part 2, consider ways an AI system could attempt to mitigate biases based on student background. Table 2 above adds the probabilities e that students secure gainful employment ten years after applying to college. The probabilities depend on x1, x2, background, and admission. Start by simulating the distribution of employment based on the model you produced for Part 1 and over time based on the evolutionary simulation for Part 2. Examine the results and select a bias mitigation approach based on readings and class discussion. Implement the algorithm and compare the results with no adjustment.

In your report, provide your rationale for selection and describe the results. Discuss key insights and implications for AI systems and algorithmic fairness.

Deliverables

You are to submit your report that includes your answers to Parts 1, 2, and 3 and a link to the source code for your implementation. Submit this report on Canvas before class on March 05, 2024.

You will also present your findings during class on February 27, 2024.