CS 594: 3- On Fairness and its Definitions

Abolfazl Asudeh
Fall 2020
9/10/2020

References:

1. Solon Barocas, Moritz Hardt, Arvind Narayanan. “Fairness in Machine Learning, Limitations and Opportunities” (book)
2. (Tutorial) Arvind Narayanan. "21 fairness definitions and their politics" In FAT*, 2018.
3. Ashudeep Singh and Thorsten Joachims. "Fairness of exposure in rankings" In KDD. 2018
4. Ke Yang and Julia Stoyanovich. "Measuring fairness in ranked outputs." In SSDBM. 2017
5. Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im) possibility of fairness." arXiv:1609.07236 (2016).
6. Cynthia Dwork, et al. "Fairness through awareness." In ITCS. 2012.
7. (Tutorial) A. Asudeh, H. V. Jagadish. Fairly Evaluating and Scoring Items in a Data Set. PVLDB, 13(12): 3445-3448, 2020

Fairness

Fairness is an important requirement for any automated decision system [popularly referred to as “AI system”, whether or not this actually uses AI techniques]..

Our focus in this lecture is score-based ranking and classificaiton.

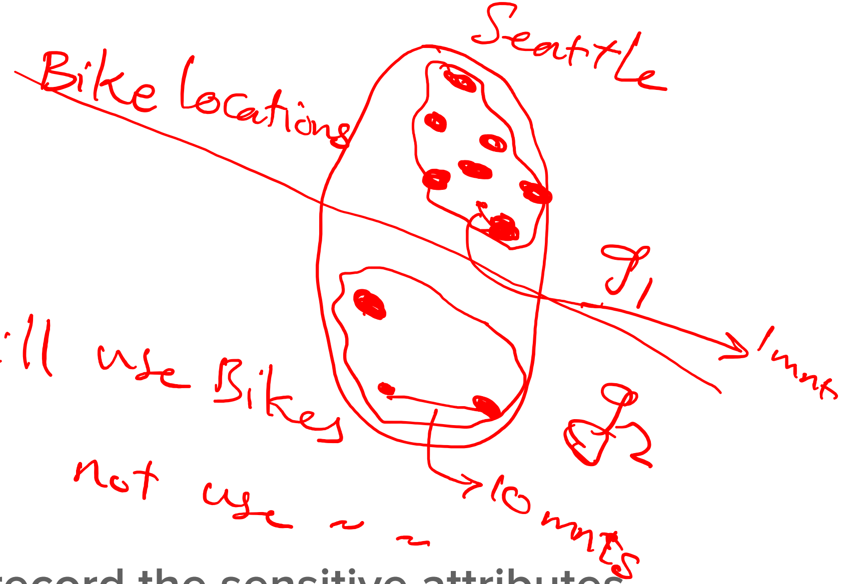
What is Fairness

We have already seen it is hard to define!

From whos prespective?

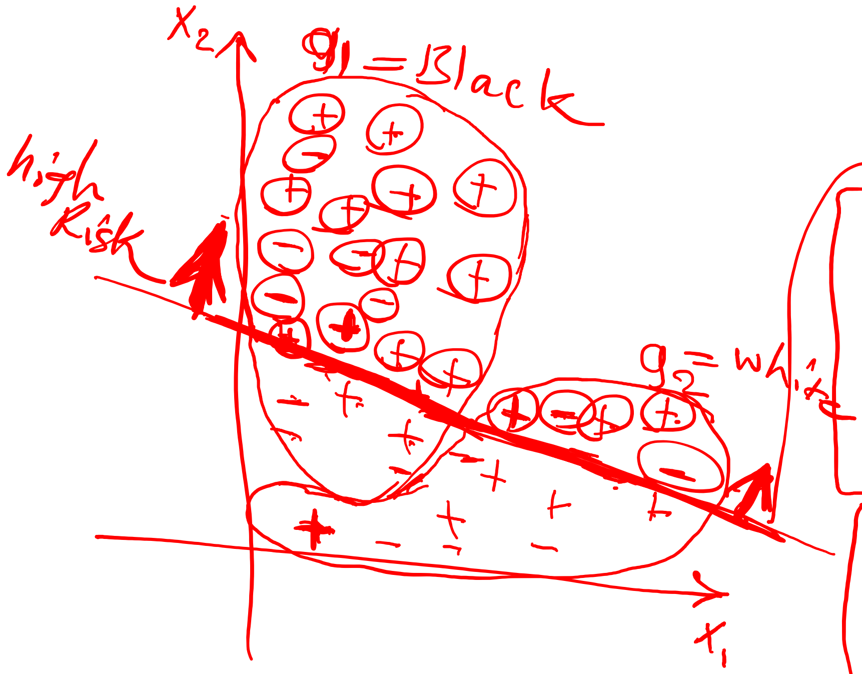
g1 : will use Bikes

g2 : " not use " "



How about a simple resolution?: Do not even record the sensitive attributes (senstive attributes are not part of the observation) diff. underlying distributions

Example 1: Demographic Disparity (Log. Regression)



$\min L_\theta(X)$

$L_\theta(g_1) = \frac{7}{23} \simeq 0.3$

$L_\theta(g_2) = \frac{7}{15} \simeq 0.5$

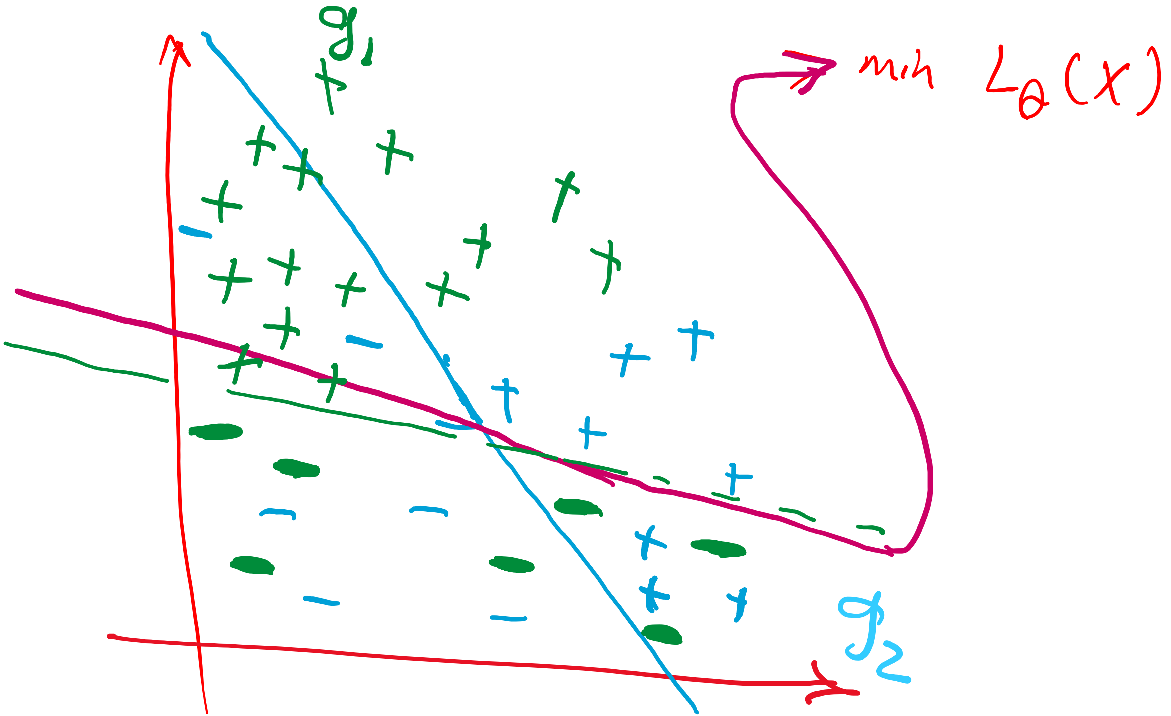
← misclassif

$DP(g_1) = \frac{17}{23} > 0.6$

$DP(g_2) = \frac{5}{15} \simeq 0.3$

different correlations

Example 2: Misclassification Disparity



A simple resolution?

How about a simple resolution?: Do not even record the sensitive attributes (senstive attributes are not part of the observation)

- No, it doesn’t work:
  - Different Deomgraphic groups may follow different distributions
  - Due to biases in data (we will discuss it later), the observations may be biased (e.g. correlated with sensitive attributes)

Simple Resolution 2

- How about building separate models for different groups?
- No!

1. We usually have few samples from minority groups → less accurate models for minorities
2. Observations across groups may help building more effective models
    - Not using all available training data → less (overall) performance
3. How about subgroups
4. How about individual fairness
5. Disparate Treatment

Disparate Treatment

Historically, and in law, we find two common “definitions” of fairness: Disparate Treatment and Disparate Outcome.

Individuals should not be treated differently on account of a sensitive attribute.

Do not explicitly use demographic information in decision making (as an observation):

- E.g.: do not have different rubrics for males and females in grading
- (still when designing the rubric you can be careful to implicitly take care of disparities)

Disparate Outcome

No disparate outcome is a group measure, and requires that the aggregate over the group of all individuals with a particular value of the sensitive attribute, the outcomes be similar.

- E.g.: fraction of women selected for a job corresponds to fraction of women who applied (or to fraction of women in the population).

At a high level