CS 189 Introduction to Machine Learning
Fall 2018 DIS10

1 Machine Bias

In 2016, ProPublica released an investigation of a criminal risk assessment tool called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions).[1] In the report, they claim that the tool, which is used for pretrial release and sentencing decisions by law enforcement agencies across the country, exhibited racial bias against blacks.

Examining data from the use of COMPAS in pretrial release decisions in Broward county, FL, the investigation revealed a racial disparity in the *error rates* of the tool. A higher rate of blacks than whites designated as "high risk" did not recividate,[2] while a higher rate of whites than blacks designated as "low risk" did.

Northpointe, the company that sells COMPAS, published a report in response,[3] arguing that their risk scores are equally accurate and predictive for whites and blacks. Could both of these arguments be true? To better understand, we will analyze the allegations and the response in the framework of formal non-discrimination criteria.

For simplicity, we will view this is a binary problem. (In reality, the COMPAS score is a value between 0 and 11.) We will use the random variable $\widehat{Y}$ to denote the designation of a defendant as "high risk" ($\widehat{Y} = 1$) or "low risk" ($\widehat{Y} = 0$) by their COMPAS score. We will use Y to denote whether or not an individual recidivated. Finally, we will use the random variable A to denote the race of the defendant. Recall the following definitions (specialized to binary decisions):

- **Independence** is satisfied if $P(\widehat{Y} = 1 \mid A = \text{black}) = P(\widehat{Y} = 1 \mid A = \text{white})$.
- **Separation** is satisfied if $P(\widehat{Y} = 1 \mid A = \text{black},\ Y = i) = P(\widehat{Y} = 1 \mid A = \text{white},\ Y = i)$ for $i = 0, 1$. This is an equality of *true and false positive rates*.
- **Sufficiency** is satisfied if $P(Y = 1 \mid A = \text{black},\ \widehat{Y} = i) = P(Y = 1 \mid A = \text{white},\ \widehat{Y} = i)$ for $i = 0, 1$. This implies an equality of *positive and negative predictive value*.

(a) In their report, ProPublica found that

> "Black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified. [...] White defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often."

In response, Northpointe pointed out that

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] as measured by arrest within two years

[3] http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

> "In comparison with whites, a slightly lower percentage of blacks were 'Labeled Higher Risk, But Didn't Re-Offend.' [...] A slightly higher percentage of blacks were 'Labeled Lower Risk, Yet Did Re-Offend.'"

How can we interpret these statements in terms of inequalities of probabilities and formal nondiscrimination criteria? (For simplicity, let's interpret 'slightly higher/lower' to be 'approximately equal.')

(b) Suppose we have that sufficiency is exactly satisfied. Then verify that the following relations are true

$$\mathrm{TPR}_a = \frac{\mathrm{PPV} \cdot p_a}{\mathrm{PPV} \cdot p_a + (1 - \mathrm{NPV})(1 - p_a)}, \ \mathrm{FPR}_a = \frac{(1 - \mathrm{PPV}) \cdot p_a}{(1 - \mathrm{PPV}) \cdot p_a + \mathrm{NPV} \cdot (1 - p_a)} \quad (1)$$

for $a \in \{\text{black,white}\}$, where we define p_a as the proportion of group a predicted to be high risk, i.e. $p_a = P(\widehat{Y} = 1 \mid A = a)$, true and false positive rates are defined for each group,

$$\mathrm{TPR}_a = P(\widehat{Y} = 1 \mid Y = 1,\ A = a), \ \ \mathrm{FPR}_a = P(\widehat{Y} = 1 \mid Y = 0,\ A = a),$$

and predictive values are group-independent (as a result of sufficiency),

$$\mathrm{PPV} = P(Y = 1 \mid \widehat{Y} = 1), \ \ \mathrm{NPV} = P(Y = 0 \mid \widehat{Y} = 0)\,.$$

(c) Show that if sufficiency is exactly satisfied and recidivism rates differ between groups (i.e. $P(Y = 1 \mid A = \text{black}) \neq P(Y = 1 \mid A = \text{white})$), then $p_{\text{black}} \neq p_{\text{white}}$.

Then, explain why (1) implies that the separation and sufficiency criteria cannot be simultaneously met when rates of recidivism differ among different groups.

(d) What does all this mean for the use of COMPAS in the criminal justice system?

2 Simpson's Paradox

(For your convenience, we have reprinted the 2nd problem from last discussion on this worksheet.)

In 1973, overall admission rates to UC Berkeley graduate school displayed a significant gender imbalance (Figure 1), with male applicants being accepted more often than female applicants.

(a) Let Y be a random variable that denotes the admission decision (e.g. $Y = 1$ is the event of acceptance to graduate school). Let G be a random variable that denotes gender, which takes values in $\{\text{male}, \text{female}\}$. Use this notation to write the observation about overall acceptance rates as an inequality of probabilities.

(b) To investigate this problem, we look at the admissions practices of individual departments (Figure 2). Now it seems that the gender imbalance disappears or goes in the other direction!

Let D be a random variable that denotes the department, which takes values in $\{0, 1, 2, 3, 4, 5\}$. Use this notation to write the observation about acceptance rates by department as inequalities of probabilities.

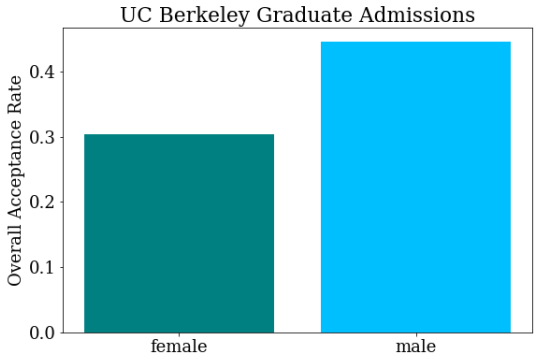

Figure 1: UC Berkeley Graduate Admissions by Gender

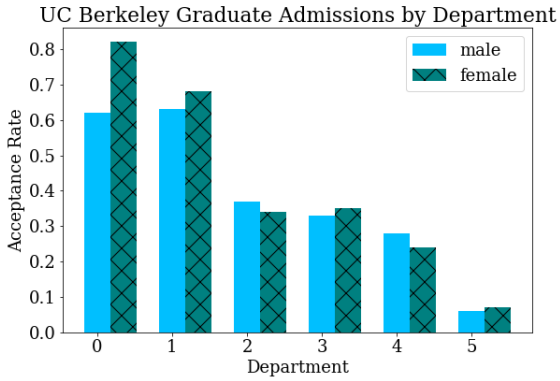

Figure 2: UC Berkeley Graduate Admissions by Department

(c) Write $P(Y = 1 \mid G = \text{female})$ in terms of $P(Y = 1 \mid G = \text{female}, D = i)$ for $i = 0, \ldots, 5$. Also write the expression for $P(Y = 1 \mid G = \text{male})$. Now, using the information in Figure 3, can you explain why the university-wide gender imbalance seems at odds with the pattern in individual departments?

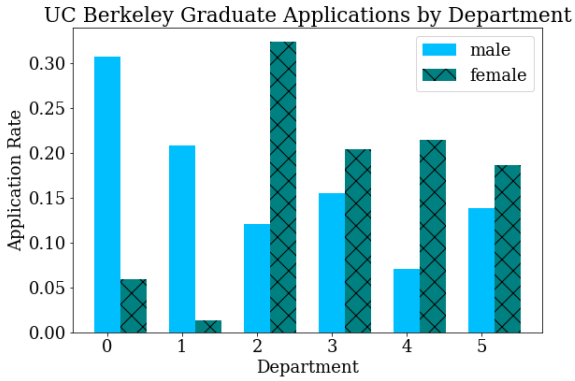

Figure 3: UC Berkeley Graduate Applications by Gender and Department

(d) This is an example of *Simpson's paradox*, which illustrates that drawing conclusions based on observational statistics may lead to incorrect conclusions. What does our statistical analysis suggest about the problem of gender imbalance overall?