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.')

Solution: ProPublica is pointing out that

$$P(\widehat{Y} = 1 \mid A = \text{black}, Y = 0) > P(\widehat{Y} = 1 \mid A = \text{white}, Y = 0), \text{ and}$$
$$P(\widehat{Y} = 0 \mid A = \text{white}, Y = 1) > P(\widehat{Y} = 0 \mid A = \text{black}, Y = 1)\,.$$

Then seeing that $P(\widehat{Y} = 0 \mid A, Y) = 1 - P(\widehat{Y} = 1 \mid A, Y)$, second inequality is equivalent to

$$P(\widehat{Y} = 1 \mid A = \text{white}, Y = 1) \leq P(\widehat{Y} = 1 \mid A = \text{black}, Y = 1)\,.$$

Therefore, ProPublica is pointing out that COMPAS does not satisfy the separation criteria. On the other hand, Northpointe is saying that

$$P(Y = 0 \mid A = \text{black}, \widehat{Y} = 1) \approx P(Y = 0 \mid A = \text{white}, \widehat{Y} = 1), \text{ and}$$
$$P(Y = 1 \mid A = \text{white}, \widehat{Y} = 0) \approx P(Y = 1 \mid A = \text{black}, \widehat{Y} = 0)\,.$$

These statements are equivalent to

$$P(Y = 1 \mid A = \text{white}, \widehat{Y} = 1) \approx P(Y = 1 \mid A = \text{black}, \widehat{Y} = 1), \text{ and}$$
$$P(Y = 0 \mid A = \text{black}, \widehat{Y} = 0) \approx P(Y = 0 \mid A = \text{white}, \widehat{Y} = 0)\,.$$

Therefore, Northpointe is pointing out that they approximately satisfy the sufficiency criteria.

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

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

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,

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

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

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

Solution: Starting with the given expression,

$$\text{TPR}_a = \frac{\text{PPV} \cdot p_a}{\text{PPV} \cdot p_a + (1 - \text{NPV})(1 - p_a)}$$

$$= \frac{P(Y=1 \mid \widehat{Y}=1) \cdot P(\widehat{Y}=1 \mid A=a)}{P(Y=1 \mid \widehat{Y}=1) \cdot P(\widehat{Y}=1 \mid A=a) + P(Y=1 \mid \widehat{Y}=0) P(\widehat{Y}=0 \mid A=a)}$$
$$= \frac{P(Y=1,\ \widehat{Y}=1 \mid A=a)}{P(Y=1,\ \widehat{Y}=1 \mid A=a) + P(Y=1\ , \widehat{Y}=0 \mid A=a)}$$
$$= \frac{P(Y=1,\ \widehat{Y}=1 \mid A=a)}{P(Y=1 \mid A=a)} = P(\widehat{Y}=1 \mid Y=1,\ A=a)\ .$$

as desired. Similarly for the second expression,

$$\text{FPR}_a = \frac{(1-\text{PPV}) \cdot p_a}{(1-\text{PPV}) \cdot p_a + \text{NPV} \cdot (1-p_a)}$$
$$= \frac{P(Y=0 \mid \widehat{Y}=1) \cdot P(\widehat{Y}=1 \mid A=a)}{P(Y=0 \mid \widehat{Y}=1) \cdot P(\widehat{Y}=1 \mid A=a) + P(Y=0 \mid \widehat{Y}=0) \cdot P(\widehat{Y}=0 \mid A=a)}$$
$$= \frac{P(Y=0,\ \widehat{Y}=1 \mid A=a)}{P(Y=0\ , \widehat{Y}=1 \mid A=a) + P(Y=0\ , \widehat{Y}=0 \mid A=a)}$$
$$= \frac{P(Y=0,\ \widehat{Y}=1 \mid A=a)}{P(Y=0 \mid A=a)} = P(\widehat{Y}=1 \mid Y=0,\ A=a)$$

as desired.

(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.

Solution: We have that

$$P(Y=1 \mid A=a) = P(Y=1,\ \widehat{Y}=1 \mid A=a) + P(Y=1,\ \widehat{Y}=0 \mid A=a)$$
$$= P(Y=1 \mid \widehat{Y}=1,\ A=a) \cdot P(\widehat{Y}=1 \mid A=a) + P(Y=1 \mid \widehat{Y}=0,\ A=a) \cdot P(\widehat{Y}=0 \mid A=a)$$
$$= \text{PPV} \cdot p_a + (1-\text{NPV}) \cdot (1-p_a)$$

If p_a were not different for different values, then $P(Y=1 \mid A=a)$ also would not be different. This is a contradiction. Therefore we must have that p_a is different for different values of a.

Looking at (1), we see that since PPV and NPV are equal for both groups, the fact that $p_{\text{black}} \neq p_{\text{white}}$ implies that $\text{TPR}_{\text{black}} \neq \text{TPR}_{\text{white}}$ and $\text{FPR}_{\text{black}} \neq \text{FPR}_{\text{white}}$.

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

Solution: We have just seen that ProPublica's observations were *inevitable* given that the sufficiency condition was met and that the rates of recidivism were different between races. Does these mean that the criticisms were invalid? Not necessarily, given the high cost of false positives to individuals and communities. But it does mean that simple metrics will not fix injustices in the legal system, and that there are inherent tradeoffs that cannot be addressed without understanding the context.

It may be interesting to bring up the point that judges can be biased. For example, judges are more statistically more likely to give harsher sentences the week after unexpected football game losses for the prominent college football team in the state.[4] Furthermore, the effect is larger for black defendants than for white defendants. Given this fact, should we accept even biased statistical scores for sentencing as being more objective?

- If the goal is to improve human biases and misjudgment, we should aim for a better tool that works in the context of sentencing from the perspective of legal professionals
- Do we really want predictive models of recidivism? (recall the prediction vs. freedom trade-off)
- Statistical noisiness between different judges (those who do not follow football, or were a fan of the other team) is fundamentally different from a single decision rule or algorithm that is deployed at scale

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.

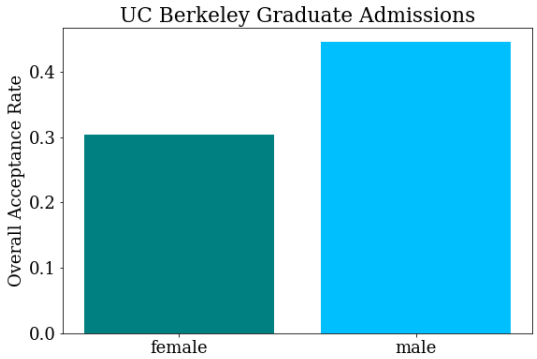

Figure 1: UC Berkeley Graduate Admissions by Gender

(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.

Solution:

$$P(Y = 1 \mid G = \text{male}) > P(Y = 1 \mid G = \text{female})$$

[4] https://www.nber.org/papers/w22611

(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!

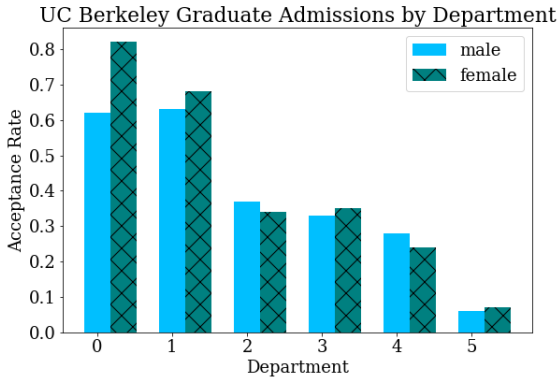

Figure 2: UC Berkeley Graduate Admissions by Department

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.

Solution:

$$
\begin{aligned}
&P(Y = 1 \mid G = \text{male},\ D = 0) < P(Y = 1 \mid G = \text{female},\ D = 0),\\
&P(Y = 1 \mid G = \text{male},\ D = 1) < P(Y = 1 \mid G = \text{female},\ D = 1),\\
&P(Y = 1 \mid G = \text{male},\ D = 2) \approx\ (>)\ P(Y = 1 \mid G = \text{female},\ D = 2),\\
&P(Y = 1 \mid G = \text{male},\ D = 3) \approx\ (<)\ P(Y = 1 \mid G = \text{female},\ D = 3),\\
&P(Y = 1 \mid G = \text{male},\ D = 4) \approx\ (>)\ P(Y = 1 \mid G = \text{female},\ D = 4),\\
&P(Y = 1 \mid G = \text{male},\ D = 5) \approx\ (<)\ P(Y = 1 \mid G = \text{female},\ D = 5)
\end{aligned}
$$

(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?

Solution:

$$
\begin{aligned}
P(Y = 1 \mid G = \text{female}) &= \sum_{i=0}^{5} P(Y = 1 \mid G = \text{female},\ D = i) P(D = i \mid G = \text{female}),\\
P(Y = 1 \mid G = \text{male}) &= \sum_{i=0}^{5} P(Y = 1 \mid G = \text{male},\ D = i) P(D = i \mid G = \text{male})
\end{aligned}
$$

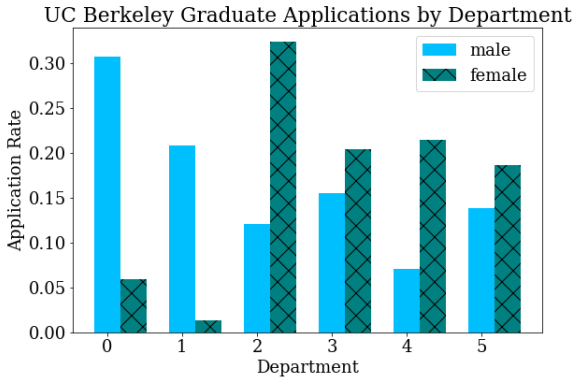

Figure 3: UC Berkeley Graduate Applications by Gender and Department

We can see that the overall acceptance rates are a weighted sum of the acceptance rate by department, where the weights correspond to application rates by gender. Looking at the chart, women apply to departments with low acceptance rate in greater proportion. In this case the confounding variable is the different admission practices between different departments.

(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?

Solution: Since admission decisions are made independently by departments, it is important to stratify the data in this way before drawing conclusions. If admissions were decided by a centralized body, then the original method for considering the data may have been appropriate.

The departments with higher admission rates were generally STEM departments which had relatively good funding and could therefore accept most qualified applications. The departments with lower admissions rates were generally humanities departments. Because of lower funding, admission was competitive even among highly qualified applicants.

The statistical analysis can only give limited information about the gender imbalance issue. For example:

- The different application patterns of male and female students could be considered a pipeline problem (e.g. female students are turned off of STEM areas during high school and therefore don't apply to graduate programs), and therefore not the responsibility of the university.
- On the other hand, the admission patterns could stem from toxic cultures within STEM departments, the reputations of which cause women to decide not to apply. This would point to a problem that the university should fix.