Homework 14: COMPAS and Algorithmic Fairness

36-462/662, Spring 2020

Due at 10 pm on Friday, 1 May 2020

> **Agenda**: Practice with the idea of algorithmic fairness; working with the black-boxed results of somebody else's data mining.
>
> **Reading**: Lecture 25 (Tuesday, 21 April), on fairness in prediction.

We're revisiting the COMPAS risk assessment data set from Homework 3 and Homework 4. You'll remember that previously we built models using this data set to predict violent redicivism, but we did not actually use the COMPAS score. This time, we'll use those scores, which are integers from 1 (lowest assessed risk) to 10 (highest).

Before using the data, filter it (as in HW 3) to remove all the arrestees who aren't either black or white.

Notation: In this problem set, Y is recidivism variable, 1 if the arrestee was re-arrested for violence within 2 years, and 0 otherwise. $\hat{Y}$ is the prediction of Y. The "positive" class will be recidivism, $Y = 1$, so a false positive means $Y = 0$ but $\hat{Y} = 1$, and a false negative means $Y = 1$ but $\hat{Y} = 0$.

1. *Features and race*
   a. (4) Using histograms or other suitable graphics, show the distribution of (i) age, (ii) number of priors and (iii) COMPAS scores for (α) everyone, (β) blacks and (γ) whites. (Ideally, you should have either a 3×3 array of plots, or 3 plots each with 3 curves.)
   b. (5) How easy would it be to infer whether an arrestee was white or black from their age? From their number of priors? From their COMPAS score? Explain in words, referring to the plots you draw in (a). (Calculations are not required but are fine.)
   c. (3) Is predicting redicivism from age just a disguised way of predicting recidivism from race? What about predicting recidivism from the number of priors? From the COMPAS score? Explain, by referring to parts (a) and (b).
2. *Accuracy and Error Rates of COMPAS* Suppose we predict recidivism for everyone whose COMPAS score reaches some threshold t, so $\hat{Y} = 1$ if $COMPAS \geq t$ and $\hat{Y} = 0$ otherwise. Since the scores are integers from 1 to 10, $t = 1$ would predict recidivism for everyone, and $t = 11$ would predict recidivism for no one.
   a. (3) *Accuracy* Plot the classification accuracy of the COMPAS score as a function of the threshold t. Include a horizontal line showing the baseline accuracy which we could achieve by predicting the same label for everyone. For what thresholds (if any) does COMPAS improve on this baseline?
   b. (2) *FNR* Plot the false negative rate of the COMPAS score as a function of the threshold t.
   c. (2) *FPR* Plot the false positive rate of the COMPAS score as a function of the threshold t.
   d. (3) *FNR vs. FPR* Plot the false negative rate against the false positive rate. (There should be 11 points on the plot, one for each value of t. [Or, if you make a line-type plot, the curve should have 11 corners.]) Describe the trade-off between the two types of error.
3. *Calibration of COMPAS*

a. (5) For each level (1–10) of the COMPAS score, find the actual frequency of recidivism, i.e., what fraction of arrestees with that score were, in fact, violent recidivists. Do this separately for (i) everyone, (ii) blacks and (iii) whites. Plot the results. (Ideally, you should have one plot with three curves.)

b. (3) Repeat you plot from (a), but now add suitable error bars to all your estimated proportions. *Hints*: (i) If n trials each have success probability p, successes are independent across trials, and we observe x total successes, we can estimate $\hat{p} = x/n$, with approximate standard error $\sqrt{\hat{p}(1-\hat{p})/n}$. (What's "success" here? What's n?) (ii) `segments()` may be helpful for drawing.

c. (5) Does the COMPAS score appear to be calibrated, or equally calibrated for both blacks and whites? Explain your answer by referring to the earlier parts of this problem.

4. *Fairness of COMPAS*

a. (5) Predictions (or decisions more generally) are said to show **demographic parity** when the fraction of positive predictions is the same across groups. For races, this would mean that $P(\hat{Y} = 1|\text{Race})$ is the same across races. Plot the fraction of arrestees with $\hat{Y} = 1$ as a function of threshold for (i) blacks alone, (ii) whites alone, and (iii) everyone. At what thresholds does COMPAS come closest to (or achieve) demographic parity?

b. (5) Predictions have **parity of predictive accuracy** when they are equally accurate for different groups in the population. Re-do your plot of accuracy against threshold t from (2a), showing separate curves for whites, for blacks, and for the whole population. At what thresholds does COMPAS come closest to (or achieve) parity of predictive accuracy?

c. (5) Predictions have **parity of error rates** when error rates are equal across different groups in the population. Re-do your plot of false positive rates against threshold t from (2c), showing separate curves for whites, for blacks, and for the whole population. At what thresholds does COMPAS come closest to (or achieve) parity of false positives?

d. (5) Define the **violation of FPR parity** as the absolute vale of the *difference* between the false positive rate for blacks and the false positive rate for whites. Make a plot showing the violation of FPR parity against the accuracy. (This should have 11 different points [or corners], one for each value of t.) Describe the trade-off, if any, between parity and accuracy.

5. *Comparing COMPAS to Other Predictive Models*

a. (5) In homework 3, we fit a classification tree with four leaves using just age and the number of priors as predictors. Re-fit that model to this data set. Create a plot of the false negative rate versus false positive rate as we vary the threshold for setting $\hat{Y} = 1$. *Hint*: See the solutions to homework 3.

b. (5) In homework 4, we fit a logistic regression using just age and the number of priors as predictors. Re-fit that model to this data set. Create a plot of the false negative rate versus false positive rate as we vary the threshold for setting $\hat{Y} = 1$. *Hint:* See the solutions to homework 4.

c. (5) Plot the violation of FPR parity against accuracy for the tree model, as in problem (5d). Describe the trade-off, if any, between parity and accuracy.

d. (5) Plot the violation of FPR parity against accuracy for the logistic regression model, as in problem (5d). Describe the trade-off, if any, between parity and accuracy.

e. (5) The COMPAS score is the output of proprietary, closed-source software. We don't know exactly how it's calculated, but we do know that it's supposed to be based on over 100 features. What advantages, if any, does it have over either the tree or the logistic regression model, in terms of accuracy, error rates, or fairness? Refer to your results in earlier parts of this homework in your answer.

6. (10) *Advising Riverdale* Suppose that Riverdale County, Florida, is considering adopting COMPAS, and that you have been hired by a member of the county council to advise them about this decision. (You can assume that Riverdale County, while fictional, is otherwise very similar to Broward County, where the data come from.) Summarize what you have learned from this analysis about the ways in which COMPAS is or is not accurate and fair. Based on this, how would you recommend that the county use COMPAS, if at all? Would you recommend an alternative tool? Would it make a difference to your recommendation whether the council member was black, white, or something else?

Rubric (10)

The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. All plots and tables are generated using code embedded in the document and automatically re-calculated from the data. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for.

Extra Credit

A (5)

We have assumed that if we use the COMPAS score, we need to apply the *same* threshold t to both whites and blacks. If we allowed there to be different thresholds for the two groups, could we achieve parity of false positive rates? If not, explain why not. If so, what would the common false positive rate be, what would the false negative rates be, and what would the accuracies be? Would you recommend doing this (assuming it's legal)?

B (5)

Using the methods from Homework 5, how much information does race give about recidivism, conditional on the COMPAS score? (That is, what is $I[\text{Recidivism}; \text{Race}|\text{COMPAS}]$?) How much information does sex give about recividism, conditional on the COMPAS score? How much information does the combination of race and sex give, conditional on the COMPAS score? (This is extra credit because you might need to modify some of the code from that homework.)

C (10)

The main problems have asked you to look at whether COMPAS is fair across races. We can also ask about whether it is fair across sexes. Re-do problems (3), (4) and (6c) and (6d) to look at the disparity between the sexes rather than the races. Would this modify your conclusions in (7)? Why or why not?