Homework 12: COMPAS Revisited

36-462/662, Spring 2022

Due at 6 pm on Thursday, 28 April 2022

Agenda: Practice with the idea of algorithmic fairness; working with the black-boxed results of somebody else's data mining.

Reading: Lecture 24 (Tuesday, 19 April), on fairness in prediction, and Kearns and Roth (2019), chapter 2.

We're revisiting the COMPAS risk assessment data set from Homework 6. 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 (except in the extra credit). This time, we'll use those scores, which are integers from 1 (lowest assessed risk) to 10 (highest).

Before using the data, filter it to remove all the arrestees who aren't either black or white[1]. When these questions refer to "everyone", it means "all blacks and all whites".

Notation: Y is the 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. **One last batch of reading questions** for O'Neil (2016) (links on the course homepage and on Canvas).

    a. (2) In chapter 6, O'Neil says that one of the issues with using personality tests to screen job applicants is a lack of "feedback". Describe, in your own words, what O'Neil means by "feedback", why it is absent, and why that makes the models worse.

    b. (2) In chapter 5, O'Neil asserts that recidivism-prediction models are "logically flawed". What, in your own words, is the logical flaw she identifies? (You do not have to say whether you think she is right, just describe her argument.)

    c. (2) In chapter 1, O'Neil lists some questions which are part of recidivism-prediction models which would not be allowed in (American) legal trials. Describe three of specific examples of such questions, in your own words.

    d. (2) In chapter 8, O'Neil writes that "E-scores ... analyze the individual through a veritable blizzard of proxies". Give two specific examples (according to O'Neil) of proxies used by e-scores, and two features which (she says) are *not* proxies.

    e. (2) In chapter 8, O'Neil complains about e-scores that there is "no feedback to set the system straight", *and* that e-scores "create a nasty feedback loop". Explain what she means in each case, and why she isn't contradicting herself.

2. **Features and race**

    a. (3) Using histograms or other suitable plots, 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 3 plots, each with 3 curves, but a 3×3 array of plots will get partial credit.)

[1] This is partly so that we don't have to worry about more than two-way comparisons, and partly because some of the other racial categories have only a small number of members in the data set, which would complicate your coding for some questions without teaching you much.

b. (5) How reliably can an arrestee's race be inferred from their age? From their priors? From their COMPAS score? Explain in words, referring to the plots you drew in Q2a. (Calculations are not required but are fine.)

c. (3) Consider the claim that "Predicting recidivism from age is just a disguised way of predicting recidivism from race". Give one reason in favor of this statement, and one against, based on Q2a–2b.

d. (3) Do the same for claim that "Predicting recidivism from the number of priors is just a disguised way of predicting recidivism from race".

e. (3) Do the same for the claim that "Predicting recidivism form COMPAS scores is just a disguised way of predicting recidivism from race".

3. **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. (2) **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? (There should be 11 points on this plot.)

   b. (3) **FNR vs. FPR** Plot the false negative rate (on the vertical axis) against the false positive rate (on the horizontal axis). (Again, there should be 11 points on the plot.) Include a diagonal line showing the performance of baseline randomized classifiers. Describe the trade-off between the two types of error.

4. **Calibration of COMPAS**

   a. (2) 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 for (i) blacks, (ii) whites and (iii) both together. Plot the results. (Ideally, you should have one plot with three curves, but three plots with one curve each will get partial credit.)

   b. (3) Repeat you plot from Q4a, 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? Justify your answer by referring to what you found in Q4a and Q4b.

5. **Disparity in COMPAS**

   a. (5) Predictions/decisions have **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 reach) demographic parity?

   b. (4) 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 from Q3a, 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. Make a plot of false positive rates against threshold, 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 **FPR disparity** as the ratio between the false positive rate for blacks and the false positive rate for whites. Make a plot showing the FPR disparity against the accuracy as the threshold varies. Describe the trade-off, if any, between parity and accuracy.

6. **Comparing COMPAS to Other Predictive Models** In these questions, randomly divide the data into an 80% training set and a 20% testing set; estimate all models on the training set, but evaluate their performance on the testing set.

   a. (3) In HW6, 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*: Solutions to HW6.

   b. (3) In HW6, 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:* Solutions to HW6.

   c. (5) Plot the FPR disparity against accuracy for the tree model, as in Q5d. Describe the trade-off, if any, between parity and accuracy for this model.

   d. (5) Plot the FPR disparity against accuracy for the logistic regression model, as in Q5d. Describe the trade-off, if any, between parity and accuracy for this model.

   e. (5) Make a plot which shows all the combinations of accuracy and FPR disparity that can be achieved using these three models. Highlight the points on the Pareto frontier. Are all of the frontier points from the same model, or do different models dominate in different parts of the frontier?

7. (7) **Advising Riverdale (Reprise)** Suppose that Riverdale County 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. Give an argument in favor of using COMPAS, an argument for using a different model instead of COMPAS, and an argument against using statistical model at all.

8. (1) **Timing** How long, roughly, did you spend on this assignment?

Presentation 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. 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 plots and tables are generated by code included in the R Markdown file. 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. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for. Text from the homework assignment, including this rubric, is included only when relevant, not blindly copied.

Extra Credit

A (10)

Berk and Elzarka (2020) suggest that when fairness is an issue, we might want to train a model using only data from the most privileged or advantaged group, and then apply this same model to everyone. The models then (presumably) treat everyone as though they were members of that most-privileged group. Create a training set consisting of 80% of the white arrestees, randomly selected. Use this training set to re-estimate your classification tree and logistic regression. Plot ROC curves for both models for (i) whites in the testing

(= not-training) set only, (ii) all black arrestees, and (iii) testing whites and blacks together. Why, in this approach, do we not need separate training and test sets for black arrestees? Similarly, create plots of accuracy versus FPR disparity for both models. Are these models fairer than the ones estimated from the whole data? Even if they are not *always* fairer, do they improve the fairness/accuracy frontier?

B (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 the parts of Q2, Q4, Q5 and Q6 which called for racial comparisons to look at the disparity between the sexes.

References

Berk, Richard A., and Ayya A. Elzarka. 2020. "Almost Politically Acceptable Criminal Justice Risk Assessment." *Criminology and Public Policy* 19:1231–1257. https://doi.org/10.1111/1745-9133.12500.

Kearns, Michael J., and Aaron Roth. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford: Oxford University Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.