DS-GA 1017, Responsible Data Science, Spring 2022

Homework 1: Algorithmic Fairness
Due on Thursday, March 3 at 11:59pm EST

Objectives

This assignment consists of written problems and programming exercises on algorithmic fairness.

After completing this assignment, you will:

- Understand that different notions of fairness correspond to points of view of different stakeholders, and are often mutually incompatible.
- Gain hands-on experience with incorporating fairness-enhancing interventions into machine learning pipelines.
- Learn about the trade-offs between fairness and accuracy.
- Observe the effect of hyperparameter tuning on performance, in terms of both accuracy and fairness.

You must work on this assignment individually. If you have questions about this assignment, please post a private message to all instructors on Piazza.

Grading

The homework is worth 75 points, or 10% of the course grade. Your grade for the programming portion (Problem 2) will be significantly impacted by the quality of your written report for that portion. In your report, you should explain your observations carefully.

You are allotted 2 (two) late days over the term, which you may use on a single homework, or on two homeworks, or not at all. If an assignment is submitted at most 24 hours late -- one day is used in full; if it's submitted between 24 and 48 hours late -- two days are used in full.

Submission instructions

Provide written answers to Problems 1, 2, and 3 in a single PDF file created using LaTeX. (If you are new to LaTeX, Overleaf is an easy way to get started.) Provide code in answer to Problem 2 in a Google Colaboratory notebook. Both the PDF and the notebook should be turned in as Homework 1 on BrightSpace. Please clearly label each part of each question.

Problem 1 (20 points): Fairness from the point of view of different stakeholders

(a) (3 points) Consider the COMPAS investigation by ProPublica and Northpointe's response. (You may also wish to consult Northpointe's report.) For each metric A-E below, explain in 1-2 sentences which stakeholders would benefit from a model that optimizes that metric, and why. If you believe that it would not be reasonable to optimize that metric in this case, state so and explain why.

A. Accuracy
B. Positive predictive value
C. False positive rate
D. False negative rate
E. Statistical parity (demographic parity among the individuals receiving any prediction)

(b) (3 points) Consider a hypothetical scenario in which *TechCorp*, a large technology company, is hiring for data scientist roles. Alex, a recruiter at *TechCorp*, uses a resume screening tool called *Prophecy* to help identify promising candidates. *Prophecy* takes applicant resumes as input and returns them in ranked (sorted) order, with the more promising applicants (according to the tool) appearing closer to the top of the ranked list. Alex takes the output of the *Prophecy* tool under advisement when deciding whom to invite for a job interview.

In their 1996 paper "Bias in computer systems", Friedman & Nissenbaum discussed three types of bias: **A.** pre-existing, **B.** technical, and **C.** emergent. We also discussed these types of bias in class and in the "All about that Bias" comic.

For each type of bias:
- give an example of how this type of bias may arise in the scenario described above;
- name a stakeholder group that may be harmed by this type of bias; and
- propose an intervention that may help mitigate this type of bias.

(c) (6 points) Consider a hypothetical scenario in which an admissions officer at *Best University* is evaluating applicants based on 3 features: SAT score, high school GPA, and family income bracket (low, medium, high). We discussed several equality of opportunity (EO) doctrines in class and in the "Fairness and Friends" comic: formal, substantive / luck egalitarian, and substantive / Rawlsian.

A. In a selection procedure that is fair according to formal EO, which of these features would the admissions officer use? Briefly justify your answer.
B. Suppose that income-based differences are observed in applicants' SAT scores: the median score is lower for applicants from low-income families, as compared to those from medium- and high-income families. Which EO doctrine(s) is/are consistent with the goal of correcting such differences in the applicant pool? Briefly justify your answer.
C. Describe an applicant selection procedure that is fair according to luck-egalitarian EO.

(d) (8 points) Consider a binary classification problem where the population consists of two groups. The "Fair prediction with disparate impact" by Chouldechova paper showed that if the base rate for the outcome of interest is different across groups -- that is, if fraction of each group with a positive outcome is different -- then no classifier can simultaneously achieve (i) equal positive predictive value, (ii) equal false positive rates, and (iii) equal false negative rates across groups.

Suppose we have Group A and Group B, with different base rates for the outcome of interest. Let $\mathbf{p_A}$ be the probability that members of Group A have a positive outcome, and $\mathbf{p_B}$ be the probability that members of Group B have a positive outcome. Prove mathematically that if $\mathbf{p_A} \neq \mathbf{p_B}$, then no classifier can simultaneously achieve the following three criteria: (i) equal accuracy, (ii) equal false positive rates, and (iii) equal false negative rates across groups.

Hint: Express each of these metrics in terms of the elements of a confusion matrix and solve for a relationship between them that depends on $\mathbf{p_A}$ or $\mathbf{p_B}$. You can refer to the table in the sensitivity and specificity Wikipedia article for a definition of these and other criteria.

Problem 2 (40 points): Fairness-enhancing interventions in machine learning pipelines

In this part of the assignment you will use AIF360 to incorporate fairness-enhancing interventions into binary classification pipelines. You should use the provided Google Colaboratory notebook as the starting point for your implementation. Your grade will be based on the quality of your code and of your report: explain your findings clearly, and illustrate them with plots as appropriate.

In all experiments, split your data into 70% training, 10% validation, and 20% test. Use the validation dataset for hyperparameter tuning (see below). Report all results on the withheld test dataset.

Use the Folktables dataset, which was created as an update to the Adult dataset from the AIF360 toolkit. We have preloaded the ACSIncome dataset, which uses individuals' attributes to predict high vs low income. **We select sex as the sensitive attribute** to analyze throughout this question.

You will evaluate performance using the following metrics:

(i) Overall accuracy
(ii) Accuracy for the privileged group
(iii) Accuracy for the unprivileged group
(iv) Disparate Impact
(v) False positive rate difference

(a) (5 points) Train a baseline **random forest** model to predict income per year. Report performance on the metrics listed above on the **test set**. Discuss your results in the report.

(b) (5 points) Implement code to tune the hyperparameters of the baseline **random forest** model, tuning them for accuracy. Remember to use the **validation set** to select the best hyperparameters. Then, **for ten different train/validation/test splits**, report the same five metrics as above, calculated before and after hyperparameter tuning on the test set, for both models. Show performance in a plot; we suggest using a box-and-whiskers plot for this.

In your report, discuss the impact of hyperparameter tuning of both fairness and accuracy for both models, and hypothesize about any differences between the models.

(c) (10 points) Consider Disparate Impact Remover (DI-Remover), a pre-processing fairness-enhancing intervention by Feldman et al., 2015 (here) that is implemented in AIF360. This algorithm provides a parameter called the **repair level** that controls the trade-off between fairness and accuracy. In this question, you will measure the impact of repair level on fairness and accuracy.

Transform the original dataset using DI-Remover with **five different values of the repair level**. Train a **random forest** model on each transformed dataset again, using the same hyperparameters and train/validation/test split that you used in part **(b)**. Report the same five metrics again for each trained model.

Discuss in your report how these results compare with the metrics from the baseline random forest model from **(b)**, paying particular attention to the impact of repair level.

(d) (10 points) Train a model using the Prejudice Remover in-processing technique by Kamishima et al. 2012 (link) that is implemented in AIF360. This algorithm provides a parameter called **eta**, which controls the fairness regularization weight. Use the values [0.01, 0.1, 1] for the **eta** parameter. Plot both the accuracy and disparate impact as you adjust this parameter and discuss the results.

Discuss in your report how the effect of the eta parameter compares to what you observed for DI-Remover. (Remember: Prejudice Remover is not a pre-processing method that is combined with an existing Random Forest model. It's a different model altogether, which fits a Logistic Regression under the hood.)

(e) (10 points) Train a classifier using Reject Option Classification, a post-processing algorithm by Kamiran et al., 2012 (link) that is also implemented in AIF360. Do this for the same ten splits you used in part **(b)**. Report the same five metrics and compare those results to your results from parts **(b)** and **(c)**.

Discuss in your report how these results compare with the metrics from the baseline random forest model from **(b)**, paying particular attention to the impact of repair level, and how they compare with the results in **(c)**.

Conclude your report with any general observations about the trends and trade-offs you observed in the performance of the fairness enhancing interventions with respect to the accuracy and fairness metrics.

Problem 3 (15 points)

In the final part of the assignment, you will watch a lecture from the AI Ethics: Global Perspectives course and write a memo (500 words maximum) reflecting on issues of fairness raised in the lecture. You can watch either:

- "AI for whom?" (watch the lecture)
- "AI Powered Disability Discrimination: How Do You Lipread a Robot Recruiter" (watch the lecture)
- "Ethics in AI: A Challenging Task" (watch the lecture)
- "Alexa vs Alice: Cultural Perspectives on the Impact of AI" (watch the lecture)

Before watching the lecture, please register for the course at https://aiethicscourse.org/contact.html, specify "student" as your position/title, "New York University" as your organization, and enter DS-GA 1017 in the message box.

Your memo should include the following information:

- Identify and describe a data science application that is discussed in the lecture. What is the stated purpose of this data science application?
- Identify the stakeholders. In particular, which organization(s), industry, or population(s) could benefit from the data science application? Which population(s) or group(s) have been adversely affected, or are most likely to be adversely affected, by the data science application?
- **Option 1**: If applicable, identify examples of disparate treatment and/or disparate impact in the data science application and describe how these examples of disparate treatment or disparate impact relate to pre-existing bias, technical bias, and/or emergent bias.
- **Option 2**: If option 1 is inapplicable, give examples of harms that may be due to the use of the data science application, and explain or hypothesize about the data-related or other technical reasons that these harms may arise.

You may also discuss any other issue of fairness raised in the lecture.