DS-UA 202, Responsible Data Science, Spring 2023

Homework 1: Algorithmic Fairness
Due on Wednesday, March 10 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 **90 points**, or 10% of the course grade. Your grade for the programming portion (Problem 4) 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 questions 1,2, and 3 in a single PDF file. You may use LaTeX (If you are new to LaTeX, Overleaf is an easy way to get started), Microsoft Word, or Google Docs and export it as a PDF. Provide code in answer to Problem 4 in a Google Colaboratory notebook. Both the PDF and the notebook should be turned in as Homework 1 on Gradescope. Please clearly label each part of each question. Name the files in your submission *abc123*_hw1.pdf and *abc123*_hw1.ipynb (replace *abc123* with your UNI).

Problem 1 (10 points): Fairness from the point of view of different stakeholders

(a) (5 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 (ACC)
- **B:** Positive predictive value (PPV)
- **C:** False positive rate (FPR)
- **D:** False negative rate (FNR)
- **E:** Statistical parity (SP)

(b) (5 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 **A-C**:

- 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.

Problem 2 (20 points): Fairness impossibility results

Consider a binary classification problem where the population consists of two groups. The "Fair prediction with disparate impact" paper by Chouldechova 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 $p_A = 0.8$ be the probability that members of Group A have a positive outcome, and $p_B = 0.5$ be the probability that members of Group B have a positive outcome. Assume group A has 100 observations and group B has 80 observations.

(a) (5 points) Suppose that both groups A and B have equal false positive rates and equal false negative rates where FPR = 0.4 and FNR = 0.75. What are each of their respective values for TP, FN, FP, TN?

[*Hint: You may find it easier to fill in these values using a confusion matrix as shown below. You can refer to this Wikipedia article for confusion matrix definitions.]*

b) (5 points) What is the accuracy (ACC) for group A and group B? Which group has better accuracy?

Accuracy (ACC) = (TP + TN) / (P + N)

c) (5 points) What is the positive predictive value for group A and group B? Which group has better PPV?

Positive predictive value (PPV) = TP / PP

d) (5 points) How does this example show the fairness impossibility results described in the Chouldechova paper?

		Predicted Outcome		
		$\widehat{Y}$ = 1 Predicted Positive (PP)	$\widehat{Y}$ = 0 Predicted Negative (PN)	
Actual Outcome	Y=1 Positive (P)	True Positive (TP)	False Negative (FN)	False negative rate FNR = FN/P
	Y=0 Negative (N)	False Positive (FP)	True Negative (TN)	False positive rate FPR = FP/N

Problem 3 (15 points): Global perspectives on AI ethics

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)
- “Alexa vs Alice: Cultural Perspectives on the Impact of AI” (watch the lecture)
- “Trustworthy Cities: Ethical Urban Artificial Intelligence” (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-UA 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.

Problem 4 (45 points): Fairness-enhancing interventions

In this part of the assignment you will use Fairlearn 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 80% training and 20% test. Report all results on the withheld test dataset. We will be using the “Diabetes Hospital” dataset from Fairlearn (and [Strack et al., 2014]). **We select `gender` as the sensitive attribute** to analyze throughout this question.

You will evaluate performance using the following metrics, all of which are available through Fairlearn’s MetricFrame. You will report on the overall values for each, as well as group-specific (‘Male’ or ‘Female’) values for your models:

(i) Accuracy

(ii) Precision
(iii) Recall
(iv) FNR
(v) FPR

You will also evaluate performance using the following *fairness metrics* which *may* require conditioning on group membership (i.e. setting a `sensitive_features` hyperparameter).

(vi) false_negative_rate_difference
(vii) false_positive_rate_difference
(viii) demographic_parity_ratio
(ix) equalized_odds_ratio
(x) selection_rate_difference

(a) (5 points) Train a baseline **random forest** model (set its `n_estimators` as 1 and keep other arguments as the default value) to predict hospital readmittance. Report performance on the metrics listed above on the **test set**. Discuss your results in the report.

(b) (10 points) The random forest has two input parameters called `n_estimators` and `max_depth` which can really impact how well the random forest performs. So while we want to find good values for `n_estimators` and `max_depth`, this is hard to do without just guessing and checking a lot of different values. In general, `n_estimators` and `max_depth` are examples of **hyperparameters**, and the act of searching for good hyperparameters is called **hyperparameter tuning**.

In Problem 4(b), we have already tuned `n_estimators=1000` and `max_depth=10` for you. We kept this choice of hyperparameters because they maximized accuracy. Use the hyperparameters provided in the notebook to train a 2nd **random forest** model. Report its performance on the metrics listed above on the test set.

In your report, discuss the impact of tuned hyperparameters compared to the baseline model on fairness and accuracy. Hypothesize about why certain hyperparameters led to more accurate/precise/fair models.

(c) (15 points) Consider the Adversarial Fairness Classifier from Fairlearn, an in-processing fairness-enhancing intervention by Zhang et. al ("Mitigating Unwanted Biases with Adversarial Learning" 2018). We will use a simple neural network as the classifier here instead of a random forest (this is already done in the code given to you). This algorithm also provides a parameter called **alpha** that controls the tradeoff between fairness and accuracy. In this question, you will measure the impact of **alpha** on fairness and accuracy.

Train four AdversarialFairnessClassifiers, each with one of the following 4 different **alpha** parameter values: **[0.0, 0.3, 0.7, 1.0]**.

Make sure to use the hyperparameters as stated in Problem 4(c) of the notebook, to avoid poor model behavior.

The Adversarial Fairness Classifier uses randomness in its initialization and training, which can impact performance. Retrain this across different 10 random seeds, and compute the metrics for each random seed and each alpha. This code will take a while to run, a bit over 1 hour. Plot the same metrics as above, but now using box-and-whiskers plots to show how the metrics vary across different choices of **alpha** (therefore, **alpha** is the x-axis; see here for example box-and-whisker code).

Discuss in your report how these results compare with the metrics from the tuned random forest model from **(b)**.

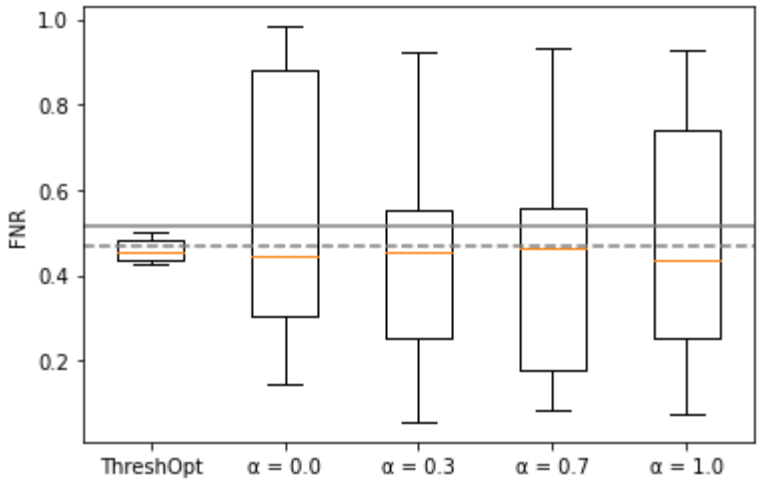
To help visualize this, optionally consider drawing an additional horizontal line (using `plt.axhline()`) on the boxplots to show how the hyperparameter-tuned random forest performed on each metric.

(d) (15 points) Mitigate the unfairness of the hyperparameter-tuned classifier from Problem 4(b) using ThresholdOptimizer, a post-processing algorithm by [Hardt et. al] ("Equality of Opportunity in Supervised Learning" 2016 https://arxiv.org/abs/1610.02413). We used this in Lab 3. Choose which **objective** (the `**constraints**` parameter in the API) best fits this fair classification scenario in the medical context, and **explain why**.

The way we split the training and test data is random, which can affect the quality of the final model. For this problem, **generate 10 different train/test splits**. For each split, use the training data to train both the random forest and ThresholdOptimizer, and use the test data to evaluate the same metrics as above.

Visualize the metrics by using a separate figure for each metric. Each figure should have 5 box-and-whisker plots on it, the first being the ThresholdOptimizer and the following four corresponding to each value of alpha from Problem 4(c). Lastly, each figure should have two gray horizontal lines on it, one dashed line showing how the un-tuned random forest from 4(a) performed on that metric, and a solid line showing how the tuned forest from 4(b) performed on that metric.

For instance, this is how our plot for FNR looks. Variations in the actual boxplots themselves will happen, and variations in the look of plot is fine, but this image should help show how to visualize one metric from all of these models in one figure:



Discuss in your report how these results compare with the metric from the models in Problems 4(b) and 4(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.