COMP_SCI 397/497: Data Privacy (Winter 2022)

Homework 3 v1.0 (100 points)

Due date: March 10th, 2022, 11:59pm CST

Instructor: Sruti Bhagavatula

Your name: Hong Hong

Problem 1 Designing fair algorithms (25 points)

The local police department is hiring you to design a new machine learning model to predict which zip code to police at different times of the day (morning (9 am-12pm), afternoon (12 pm-5 pm), evening (5 pm-9 pm), night(9 pm-9 am)). They have provided you with historic arrest data. Each row in the dataset includes details from a separate police stop, including the date and time of the arrest, the arrested person's demographics (age, height, eye color, race, sex), the location of the arrest (zip code), and the charges (a categorical variable with the options: "arrest occurred" and "no arrest occurred").

Exercise 1. [10 pts] If we ignore fairness, what kind of model is the department looking for? I.e., what could be the relevant inputs and outputs (inputs and outputs can either be the raw data described above or may be some function/aggregate of any of them)? Can you describe an approach (in English) for training such a model based on the data you have been given?

Answer:
The department is looking for a model that can predict the possible number of arrested people in different zip code at different times of the day in order to decide which zip code to police.
Possible inputs: the zip code, date, time of the arrest, arrested person's demographics, and the charges
Possible outputs: predicted time of the arrest, the zip code, the charges
I may use the regression model that use these possible inputs to generate the possible outputs.

Exercise 2. [7 pts] What are some of the potential sources of bias that could arise from the process you described in the previous part? How would you measure this bias (i.e., which of the metrics we learned about in class would be most appropriate and why?)

Answer: Some attributes like "race" and "gender" may cause bias and force to make up a privileged group. I may use disparate impact to measure this bias for this metric.

Exercise 3. [8 pts] Describe a technique/procedure for fixing the bias you wrote about in Exercise 2 (beyond just removing sensitive attributes). This fix can either be on the historical data you were provided and/or the data/model/outputs of the model you are required to build. What are pros and cons of the technique you chose?

Answer: I plan to use prepossessing on the original data and use reweigh to deal with the unrepresentative data.
Pros: It can enforce fairness on the privileged and unprivileged groups because it will give the privileged group a lower positive outcome rate and the unprivileged group a higher positive outcome rate.
Cons: Reduced the model accuracy

Problem 2 Implementing fairness and privacy (55 points)

Data-trained predictive models are often used as black boxes that output a prediction that is used to aid decision making. Thus, it is hard to determine whether sensitive attributes such as race or gender are unduly influencing model predictions.

Several approaches have been proposed to detect and mitigate two types of bias: 1) bias in training data: bias is computed using only the training data 2) bias in classification: bias is computed using classification output and potentially training data

To check for and mitigate training and classification bias, IBM has developed a tool called `AI Fairness 360`[1].

2.1 Evaluating and implementing fairness (55 points)

In this exercise, you will use the IBM tool to study two the two main types of bias. You will be studying the following two bias metrics: 1) disparate impact[2] and 2) equal opportunity difference[3]. The first metric can be computed for dataset bias as well as classification bias (and has two separate implementations in the library) while the second metric is only used for computing classification bias. Two sample Jupyter Notebooks for working with AI Fairness 360 can be found here and here.

You must first install the tool following the setup instructions here: `https://github.com/IBM/AIF360`. Depending on which setup you follow, you may have to install several dependencies. You are expected to complete the following exercises by implementing a python program. You should refer to the tutorials and documentation of the Fairness 360 library[4][5].

In your code, you should:

- indicate through comments when a block of code corresponds to a TODO.
- write high-level comments throughout to describe what block of code does what (you will lose points if you have no helpful comments).

Exercise 4. [5 pts] For each of the two metrics above, indicate whether it can be used to evaluate training bias, classification bias or both. Justify why for each metric (if one metric can be used for both, describe how the metric would be modified to be used for both types of bias).

[1] `https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/`
[2] `https://arxiv.org/abs/1412.3756`
[3] `https://arxiv.org/abs/1610.02413`
[4] `https://aif360.readthedocs.io/en/latest/index.html`
[5] `https://aif360.mybluemix.net/`

Answer:
1)disparate impact:
this metric can be used to evaluate training bias and classification bias, as mentioned in the paper, a method linking disparate impact with BER can be used to evaluate the fairness metric on the input data.
2)equal opportunity difference: this metric can be used to evaluate classification bias instead of training bias because it only uses the post processing

Exercise 5. [14 pts] Compute the bias of the ProPublica dataset we've talked about in class (https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis) using the `disparate_impact` metric (`TODO #1`). Note that you can directly import this dataset from the AI Fairness 360 Library (look at "Datasets" section in the python docs). The loading of the Compas dataset takes two sensitive attributes by default, however, you must pick one and set the privileged and unprivileged groups accordingly. You may also want to drop protected features that are in the dataset by looking at the columns of the dataset and specifying what to drop when loading the dataset (see python doc). What is the value of disparate impact on the original training dataset? Is the training data biased? Why or why not?

Answer: In this section I picked the "sex" attribute and set the "Female" as protected groups and "Male" as unprotected groups. To avoid bias, I also dropped the feature about "race". The training data is not biased for the value of disparate feature is 0.8029 which is greater than 0.8.

Exercise 6. [14 pts] Implement a logistic regression classifier using `scikit-learn` to predict if a defendant is likely to re-offend (`TODO #2`). You may use a classification threshold of 0.5 if you need to specify one. Compute the fairness of the classification algorithm according to the `disparate_impact` metric and the `equal_opportunity_difference` metrics (`TODO #3`). According to each metric, is the classifier fair and why?

Answer:
Disparate impact is 0.7940 which is smaller than 0.8, it is not fair in this metric.
Equal opportunity is 0.0, so it is fair in this metric.

Exercise 7. In this exercise, you will implement bias mitigation using pre-processing methods discussed in class which are also implemented in the Fairness 360 library.

(a) **(8 pts)** Implement bias mitigation by removing disparate impact on the original training data (`TODO #4`). Compute the `disparate_impact` metric once more. How much did the amount of disparate impact (bias) decrease or increase with pre-processing?

Answer: The amount of the disparate impact did not decrease or increase, it did not change with this preprocessing.

(b) **(8 pts)** Implement bias mitigation by reweighing the original training data (`TODO #5`). Compute the `disparate_impact` metric once more and report it here. How much did the amount of disparate impact (bias) decrease or increase with pre-processing?

Answer: The disparate impact increased to 1.0000.

Exercise 8. [6 pts] Using the same logistic regression classifier as above on the transformed data from either 7a or 7b (your choice), compute the fairness of the classifier using the classification `disparate_impact` and `equal_opportunity_difference` metrics (`TODO #6`). How has the fairness of the classifier changed (or not changed) after transforming the training data?

Answer: In this section, I chose the transformed data from 7b, the disparate impact increased to 0.9277 which is larger than 0.8, and the equal opportunity remained 0.0 which did not change. Hence it is fair for both metrics by using the transformed data.

Problem 3 Anonymous communication and Tor (20 points)

Although Tor was created to protect the privacy of internet users, there have been several ethical questions and issues surrounding its usage. Particularly due to its potential to be used as a tool for surveillance and its usage as a platform for cybercrime.

Answer the open-ended questions below after reading the following two articles. You will be graded based on the justification and thoughtfulness of your answer.

- https://www.theguardian.com/world/2013/oct/04/nsa-gchq-attack-tor-network-encryption
- https://www.wired.com/story/cia-sets-up-shop-on-tor/

Exercise 9. [6 pts] What are some positive and negative consequences of the creation of Tor? Do the positives outweigh the negatives? (You may do your own research on this and include your opinions; make sure to cite any sources.)

Answer:
Positive:
1.Protect the conversation from some of the journalists, activists and campaigners in different countries.
2.Users can bypass geographical restrictions when using the service
3.Users are protected from privacy theft or leak by Internet companies
Negative:
1.Being misused by people engaged in terrorism, the trade of child abuse images, and online drug dealing, cause the difficulty for government to fight against crime.
My opinion:
I do not think the positives outweigh the negatives, because the tor is a network that are not under government's control. I do not mean something under government's control is much more better, but that not under control has more possibility to be misused by some intend to commit crimes. Like organ trafficking, women and children trafficking, cross-border fraud. If the crimes are commit through Tor, that means we can hardly trace the criminals.

Exercise 10. [6 pts] Looking at the Guardian article, do you think it was ethical for government entities to unmask Tor users? Why or why not?

Answer: I think it is not ethical for government entities to directly unmask Tor users. Several attacks from government result in implanting malicious code on the computer of Tor users who visit particular websites. Even the agencies say they are targeting terrorists or criminals, it is not ethical the government classify users as the criminals when they visit some specific websites. And the unmask activity may also result in a wide range of Tor users in the specific area(geographic location), which may cause the the data leak.

Exercise 11. [8 pts] Looking at the Wired article, explain why the CIA created its own hidden service. What challenges arise from usage of this service (both for the maintainers of the service or its users)? What potential malicious usage could occur?

Answer: The hidden service created by CIA has the improved cryptographic algorithms and stronger authentication so that people around the world can browse the agency's website anonymously, which can create a safer platform for intelligence agencies and clandestine communications and monitor the criminals who try to steal the data.
Challenge:
For maintainers, there may be users who try to send false information to the site in large numbers, reducing the efficiency of the organization and diverting attention. And it is hard to trace the users' id.
For users, even the CIA claim that everyone can browse the agency's website anonymously, users do not actually know whether they are under third party's monitor. Some behavious may be profiled as possible criminals.
Malicious usage: Maintainers of the website may use techniques to track or profile users, or sell the data to the third parties. The attackers may also found some vulnerability to attack the website.

Exercise 12. **[Bonus 5 pts]** Is there an alternative platform besides an onion service that would better serve the CIA?s need?

Answer: Vivaldi(https://vivaldi.com/privacy/browser/): a browser claim that they do not track or profile users, don't do data collection and don't sell data to third parties. Not sure about the safety but this browser is the interesting one which tell user how the browser count without tracking.(https://vivaldi.com/zerotracking/) I would give it a try if I was the user.

Submission instructions

You will need to submit two files to Canvas under Homework 3 (you can submit multiple files by selecting "Add another file"):

1. A PDF file (named `<name>_<netID>_hw3.pdf`) with answers to problems 1, 2 (parts of it), and 3. You can add your answers to this PDF in the provided answer spaces (recommended), in the LaTex template provided (recommended), or a separate PDF with only answers.

2. A python file (named `<name>_<netID>_hw3.py`) containing your code for Problem 2.