Assignment 2

Data, Privacy and Fairness

Version: 1.0

Due Date: March 9th, 2022, 11:59 PM
Submission: Assignment box on myCourses

Objective

The objective of this assignment is to examine your understanding of the fundamental concepts of fairness and privacy when it comes to applications of AI systems. As part of this assignments you will be asked to:

- Identify potential concerns with regards to privacy and fairness of the given machine learning application.
- Use some of the existing technical toolkits discussed in the tutorial to characterize and mitigate the concerns you have identified.

This assignment will be done individually.

Resources

We recommend that you use the concepts covered in the lectures and tutorials on data, privacy and fairness to complete the assignment. You can also refer to the following resources for completing parts 1 and 2 if you have any issues. We have covered IBM's libraries referenced below in the tutorials. You will be expected to complete the assignment using these libraries:

- IBM Differential Privacy library: https://github.com/IBM/differential-privacy-library
- IBM AI Fairness 360 toolkit: https://developer.ibm.com/open/projects/ai-fairness-360/

There are other resources available for you, such as these below, if you'd like to explore on your own. However, the grader will be running your code assuming the IBM libraries are used.

- Responsible AI toolkit by Microsoft
- Openmined blog on differential privacy
- Ted's Blog on differential privacy:
- Practical Fairness Book (Chapters 3 and 4)
- FairML book: https://fairmlbook.org/

Instructions

Case study

During the recent COVID-19 pandemic, a group of machine learning engineers saw a business opportunity in creating an application that allows nurses to triage patients within a resource constrained hospital environment. After a year of development and testing they have launched their company, *Prioritize*, and their first major pilot at a local hospital.

Prioritize executives have partnered up with a local hospital in Montreal and decided to initially deploy their automated triage algorithm, called TriageAssist within the emergency unit. Emergency units receive a lot of cases related to cardiovascular issues and it often takes a long time for nurses and doctors to assess the patient when they come into the ER. Any tool that would help speed up the process of triaging patients would be highly beneficial for the hospital.

TriageAssist is trained on a dataset containing medical and personal information of patients and whether they have a form of heart disease. Nurses and doctors could use TriageAssist to see whether a new patient at the ER is being classified as a patient with heart disease and use that to make decisions on their treatment of the patient. If a patient is classified as someone who has a heart disease then the ER staff can take more time with their diagnosis and treatment decision. Otherwise, they can speed up the process.

You have been hired by Prioritize to look into privacy and failure related issues with TriageAssist.

Dataset and classifier

For assignment 1, you used a version of the dataset used to train TriageAssist. For the purposes of this assignment, we have amended the dataset with one column of synthetic data on race. You can find the dataset for this assignment on myCourses. This dataset is originally from Kaggle and you can find a detailed description of how this dataset was created here. This information will be very useful for question 1 in part 1.

You can choose to use the classifier you created in assignment 1 for the following parts as a baseline comparison. If you'd like to experiment with other classifiers, you are free to do so. For example, this blogpost outlines classifiers for a version of this dataset. Note that your main focus should be on answering the questions in parts 1 and 2, rather than on creating the most accurate classifier.

Part 1: Data and privacy

1. Investigate the properties of this dataset and create a datasheet for this dataset for it based on your findings. You will not be able to fill out all of the questions in the

datasheet; however, please do your best to answer as many questions as possible. We recommend that you use the information on the Kaggle page of the original dataset and conduct basic statistical analysis of the assignment dataset to create a datasheet for the dataset.
Note: Please use the full template given on myCourses for creating this datasheet. You should be able to comfortably answer at least one question from each one of the datasheet sections (Motivation, Composition, etc). We are looking for 10 completed questions for a full mark for this question.

2. Considering contextual integrity framing of privacy, identify two forms of information flow in this scenario: one that is acceptable and another that is unacceptable?
   Tip: It may help to identify the stakeholders in this case study first.

3. Identify any identifiers, quasi-identifiers and sensitive attributes in the given data. Calculate k-anonymity and l-diversity for the first 20 rows of this dataset. Provide a modified 5-anonymous dataset from these 20 rows.

Age	Sex	Race	ChestPain	RestingBP	Cholesterc	FastingBS	RestingEC	MaxHR	ExerciseAr	Oldpeak	ST_Slope	HeartDisea
40	M	A	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	O	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	O	ATA	130	283	0	ST	98	N	0	Up	0
48	F	W	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	O	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	W	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	H	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	B	ATA	110	208	0	Normal	142	N	0	Up	0
37	M	H	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
48	F	B	ATA	120	284	0	Normal	120	N	0	Up	0
37	F	B	NAP	130	211	0	Normal	142	N	0	Up	0
58	M	W	ATA	136	164	0	ST	99	Y	2	Flat	1
39	M	B	ATA	120	204	0	Normal	145	N	0	Up	0
49	M	W	ASY	140	234	0	Normal	140	Y	1	Flat	1
42	F	O	NAP	115	211	0	ST	137	N	0	Up	0
54	F	O	ATA	120	273	0	Normal	150	N	1.5	Flat	0
38	M	B	ASY	110	196	0	Normal	166	N	0	Flat	1
43	F	O	ATA	120	201	0	Normal	165	N	0	Up	0
60	M	H	ASY	100	248	0	Normal	125	N	1	Flat	1
36	M	H	ATA	120	267	0	Normal	160	N	3	Flat	1

Figure 1: A snapshot of the first 20 rows of data

4. Create a basic classifier for TriageAssist using logistic regression. Use the full dataset (not just the 20 rows in Q3).
5. In the tutorial, you were introduced to the IBM differential privacy package and created a basic differentially private classifier using this library. Create a differentially private classifier for this dataset using logistic regression.
6. Compare the differential private classifier's accuracy with the non-private classifier you created in step 4.
7. Calculate and illustrate (via a graph) how classification accuracy shifts with respect to different values of epsilon. You can find sample code for this in the tutorial demo on myCourses.

8. Discuss what value of epsilon might be appropriate for this scenario. What is the accuracy trade-off with the chosen value of epsilon?

Part 2: Fairness

1. Considering the stakeholders of this case study, answer the following questions:
    a. **Who gets access** to what resources/information in this case study?
    b. **Who decides** who gets access to these resources?
    c. **How do they decide** who gets what?
2. Based on your answers to question 1 and the dataset that you have, identify the privileged group(s) and the favored outcome.
3. Hypothesize and elaborate on some of the foreseeable fairness concerns for this case study before you do some further investigation in steps 4 - 6.
4. For the classifier that you created in step 4 (the non-DP classifier), use AIF 360 to set-up and calculate three different fairness metrics. Calculate these metrics for different groups based on the groups you identified as privileged and unprivileged in step 2.
5. Compare the fairness metrics and elaborate on what they mean for different groups in your dataset.
6. As part of your job at Prioritize, the team is asking you to recommend fairness metrics that should be used for this application. Which fairness metrics will best suit this application given the potential fairness concerns and cost of errors for this case study?
7. Prioritize has now asked you to do some pre-processing to minimize unfairness for TriageAssist. Using AIF 360, implement one pre-processing mitigation technique covered in class.
8. Report which pre-processing technique you used and how it affected the calculated fairness metrics you've chosen in step 6. Describe the benefits/shortcomings of the technique you chose and its appropriateness to the application. Discuss the trade-off between two pre-processing techniques: one that you've chosen and another you did not implement for this case study.

Deliverables

You should prepare a report (as a .PDF file) containing your response to the above mentioned parts. You should also prepare a .py or .ipynb file containing the code you've produced, and an accompanying readme file. You should submit all three files (report, code and readme file) as **one .zip file** in the same assignment box in myCourses.

Grading: Rubric for the assignment can be found in the assignment box on myCourses.