CS340 Assignment 2: Bias Assessment

Overview

In the lecture, we learned that AI systems could behave unfairly for a variety of reasons throughout the system pipeline. Sometimes, AI systems behave unfairly not because of societal biases, but because of characteristics of the data (e.g., too few data points about some group of people) or characteristics of the systems themselves.

As a first step to acknowledge the existence of bias in AI systems, you'll become a data scientist to identify, quantify, and analyze bias. The goal of this assessment is to answer the following question: **Which groups of people may be disproportionately negatively impacted by an AI system and in what ways and to what degree the model exhibits bias?** *To this end, you will work with a given dataset and your preferred model to measure a model bias from various angles.* The steps of the assessment are generalized as follows:

- Identify types of harms and the groups that might be harmed. (**2 points**)
- Train a model to obtain the predicted value of a test dataset. (**2 points**)
- Read the document <*Fairness Matrics.pdf*>, and learn to understand fairness metrics.
- Quantify the bias you discovered across the groups using the 4 given metrics. (**4×4=16 points**)

The details of these steps will be given later, and you will then present your interesting and valuable insights by writing a report. **The assignment requires a submission of the *source code* and the *report*, and takes up a total of 20 points in your final grading.**

Submission Guide

Please submit a zip file named <*StudentID-Name-Assignment2.zip*> to Blackboard, where the submitted zip includes two parts:
- The project folder (.zip), which includes all the source code and other relevant files necessary for running the project.
- A written report (.pdf), which is a complete and coherent description on how you evaluate bias and your insight on the results.

How to get help

If you encounter any problems, please do not hesitate to reach out to the TA team by either sending posts on the Blackboard discussion board or making in-person appointments via QQ. We are here to help. Our ultimate goal is for all of you to acquire knowledge through proper training instead of overwhelming you :)

Dataset

This project utilizes the dataset *<diabetic_preprocessed.csv>*, representing ten years of clinical care at 130 US hospitals. It includes 101,766 instances with 47 features. Each instance concerns hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days. Some variables that may be highly correlated with social inequity are shown below.

Columns	Description
race, gender, age	demographic features
medicare, medicaid	insurance information
admission_source_id	emergency, referral, or other
had_emergency, had_inpatient_days, had_outpatient_days	hospital visits in prior year
medical_specialty	admitting physician's specialty
time_in_hospital, num_lab_procedures, num_procedures, num_medications, primary_diagnosis, number_diagnoses, max_glu_serum, A1Cresult, insulinChange, diabetesMed	description of the hospital visit
discharge_disposition_id	discharched to home or not
readmitted, readmit_binary, readmit_30_days	readmission information

Table 1: Description of dataset

*https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

Now follow the steps

I. Identify potential bias (2 points)

First, we load the dataset and review the meaning of its columns.

[TODO] Understand the dataset *<diabetic_preprocessed.csv>* and then identify the data type of each columns, show whether column are binary, multi-categorical, or continuous data, etc.

In our case, **the sensitive features are <race>, <gender>, and <age>**, which are potentiality in harms of social bias. So, we have to figure out the composition of the dataset respecting these groups and the size of each group. Otherwise, the imbalance of the dataset may inject part of the bias into the AI system, and it is necessary to understand some basic profiles of the dataset about the sensitive attributes.

[TODO] Examine and summarize the sample sizes of the groups according to each sensitive feature. Describe whether the dataset respecting sensitive groups is balanced.

II. Train your model to obtain the predicted result (2 points)

We next **train a classification model using the *<diabetic_preprocessed_train.csv>* dataset to predict our target variable <readmit_30_days>**. There exists a wide variety of classification models to choose at your will, such as a deep neural network, a random forest model, or a statistical regression model. We may consider the choice of model for this project from the following aspects: interpretability, model expressiveness, training time, familiarity, etc. *Notice that the performance of the model does not affect the rating, as our focus is on the fairness assessment of the model.*

[TODO] Take training and testing datasets the <*diabetic_preprocessed_train.csv*> and <*diabetic_preprocessed_test.csv*>. Our target variable is **<readmit_30_days>**, and our sensitive feature for the purposes of fairness assessment is **<race>**. So you need to take **Y = <readmit_30_days>** and **X** the rest of the features **other than <race, discharge_disposition_id, readmitted, readmit_binary, readmit_30_days>.**

The training portion will be used to fit our model, and the test portion will be used to compute the fairness metrics.

[TODO] Now using the training dataset to train your preferable model (probably using *scikit-learn*), and **save its prediction results Y' = <readmit_30_days_pred> on a new column of the testing dataset** for analysis.

III. Quantify fairness with given metrics (16 points)

Compare quantified bias across the groups

Now we have a model (M), the model prediction values (Y') of the test data sample, and the actual values (Y) of the test data sample. And we have learnt the definition of each fairness measurement metric (File <Fairness Metrics>) . Next, we start quantifying the fairness of your model using these metrics: **1-demographic parity，2-equalized odds，3-equalized opportunity，4-conditional statistical parity**.

Step1: Y and Y' refer to the predicted and actual value of <readmit_30_days>, while A refers to the sensitive attribute <race>. Our sensitive feature race includes 5 categories (0-*'AfricanAmerican'*, 1-*'Caucasian'*, 2-*'Asian'*, 3-*'Hispanic'*, 4-*'Other'*), and therefore, 5 bias indicators respecting different race groups will be calculated for group-level metrics (a demographic parity example):

Pr(Y'=1 | A=0), Pr(Y'=1 | A=1), Pr(Y'=1 | A=2), Pr(Y'=1 | A=3), Pr(Y'=1 | A=4).

Step2: The fairness between two groups can be **synthesis to a ratio** aside in [0,1] as

$$0 < \Pr(Y'=1 \mid A=a1) / \Pr(Y'=1 \mid A=a2) < 1.$$

Step3: To summarize the disparities, we may report the fairness among multiple groups using **the smallest/ largest difference**, **the smallest/ maximum ratio**, etc. For example, we may summarize our findings with a table like the following:

| | Demographic Parity
| | AfricanAmerican | 0.43 | | Caucasian | 0.44 | | Asian | 0.52 | | Hispanic | 0.56 | | Unknown | 0.67 | |
| | Largest difference| 0.24 | | Smallest ratio | 0.64 | |maximum (worst-case) | 0.67 | |

Figure 1: The computation example for multi-categorical fairness metric.

[TODO] For group-level fairness (Metrics 1-4), we work the following as shown in Figure 1:

- **Step1:** Compute the indicator for each race group;
- **Step2:** Generate a synthesis result by pair-wisely computing the bias indicator, assessing the smallest difference, the largest difference, the smallest ratio, and the largest ratio;
- **Step3:** Write up how you evaluate and your analysis of the evaluation result in the report. The analysis may include a description of the fundamentals of fairness, a comparison of results under different measures, and the relationship between the fairness of the model to each sensitive group and the inequality of the training dataset.

* For legitimate factor L, take L= <age, gender>.