Homework 2

Spring 25, CS 442: Trustworthy Machine Learning
Due Friday Mar. 28th at 23:59 CT

Instructor: Han Zhao

Instructions for submission   All the homework submissions should be typeset in LaTeX. For all the questions, please clearly justify each step in your derivations or proofs.

1   Statistical Parity and Equalized Odds [10 pts]

1.1   [5 pts]

Construct three binary random variables X, A and Y such that X is independent of A, but X is dependent of A given Y.

1.2   [5 pts]

In the course we have seen the following incompatibility theorem between statistical parity and equalized odds for a binary classification problem:

> **Theorem 1: Incompatibility Theorem**
>
> Assume that Y and A are binary random variables, then for any binary classifier $\widehat{Y}$, statistical parity and equalized odds are mutually exclusive unless $A \perp Y$ or $\widehat{Y} \perp Y$.

Give an example of a classification problem where the target variable Y can take three distinct values, and such that statistical parity and equalized odds are simultaneously achievable.

2   Asymmetric Binary Distinguishing Game [20 pts]

Consider a binary distinguishing game between Alice (A) and Bob (B) as follows. Let P, Q be two distributions that are known to both A and B. Every round, A will flip a biased coin S, such that $\Pr(S = \text{Head}) = p$ and $\Pr(S = \text{Tail}) = 1 - p$ for some constant $0 < p < 1$. Then, upon observing the outcome of the coin flipping, A will draw sample X from a distribution D as follows:

$$D = \begin{cases} P & \text{If } S = \text{Head} \\ Q & \text{If } S = \text{Tail} \end{cases}$$

The outcome of the coin flipping will not be revealed to B. Instead, A will show the drawn sample $X \sim D$ to B, and the goal of B is to guess from which distribution $(P$ or $Q)$ the sample X is drawn from.

2.1 Error of a Randomized Strategy [10 pts]

Let $\eta : \mathcal{X} \to [0, 1]$ be the conditional probability of B guess that the given sample X is drawn from P, i.e.,
$$\eta(X) := \Pr(B \text{ guesses } X \sim P \mid X).$$
Derive the expected error of this strategy as a function of η.

2.2 Optimal Strategy and Optimal Error Rate [10 pts]

Based on your answer from the last question, derive the optimal strategy η^* and the corresponding optimal error rate of B.

3 Non-trivial Prediction of the Protected Attribute [10 pts]

Let the tuple (X, A, Y) be the random variables corresponding to input data, the protected attribute and the target variable, respectively. In many cases we can predict both Y and A from the same data X with reasonable accuracy. Suppose we have a classifier g to predict Y from X. Define the statistical disparity of g as
$$\Delta_{\mathrm{DP}}(g) := |\Pr_{A=0}(g(X) = 1) - \Pr_{A=1}(g(X) = 1)|,$$
where we use $\Pr_{A=a}(\cdot)$ to denote the conditional probability of an event conditioned on $A = a$. Clearly, if $\Delta_{\mathrm{DP}}(g) = 0$, then g satisfies the statistical parity condition. Show that there exists a classifier h to predict A from X such that the following error bound holds:
$$\varepsilon_{A=0}(h) + \varepsilon_{A=1}(h) \leq 1 - \Delta_{\mathrm{DP}}(g),$$
where $\varepsilon_{A=a}(h) := \Pr_{A=a}[h(X) \neq a]$.

4 Fair Representations [40 pts]

In this problem, we will show that fair representations whose distributions are conditionally aligned will not exacerbate the statistical disparity. Again, let the tuple (X, A, Y) be the random variables corresponding to input data, the protected attribute and the target variable, respectively. In this problem, we assume both A and Y to be binary variables.
Consider representations $Z = g(X)$ such that $Z \perp A \mid Y$. For a classifier $\widehat{Y} = h(Z)$ that acts on the representations Z, let $\Delta_{\mathrm{DP}}(\widehat{Y}) := |\Pr_{A=0}(\widehat{Y} = 1) - \Pr_{A=1}(\widehat{Y} = 1)|$.

4.1 [10pts]

Show that for any classifier h that acts on the representations $Z = g(X)$, $\widehat{Y} = h(Z)$ satisfies equalized odds.

4.2 [20pts]

Define $\gamma_a := \Pr_{A=a}(Y = 0)$. Show that for any classifier h over Z, the following inequality holds:
$$\left| \Pr_{A=0}(\widehat{Y} = y) - \Pr_{A=1}(\widehat{Y} = y) \right| \leq |\gamma_0 - \gamma_1| \cdot \left(\Pr(\widehat{Y} = y \mid Y = 0) + \Pr(\widehat{Y} = y \mid Y = 1) \right), \forall y \in \{0, 1\}.$$

4.3 [10pts]

Prove that for any classifier $\widehat{Y} = h(Z)$, $\Delta_{\text{DP}}(h \circ g) \leq \Delta_{\text{BR}}$, where $\Delta_{\text{BR}} := |\gamma_0 - \gamma_1|$ is the difference of base rates. Note: this proposition states that if a classifier satisfies equalized odds, then it will not exacerbate the statistical disparity of the optimal classifier.

5 Coding [20 pts]

In this problem, you need to **1)** train a model to approximately achieve *statistical parity* using adversarial training (as introduced in Lecture 8), and **2)** train a model to approximately achieve *equalized odds* using conditional learning (as introduced in Lecture 10).

First, you need to download and unzip "cs442_hw2_code.zip". This folder consists of one subfolder and five files.

- The "data" subfolder includes two datasets—Adult and COMPAS.

- The "dataset.py" file processes these two datasets. Specifically, on the Adult dataset, "sex" is selected as the protected attribute while "income" is selected as the target attribute. On the COMPAS dataset, "African_American" is selected as the protected attribute while "Two_yr_Recidivism" is selected as the target attribute.

- The "models.py" file contains the model to be trained with adversarial training (called "FairNet") and the model to be trained with conditional learning (called "CFairNet").

- The "utils.py" file contains functions to compute errors.

- The "main_adult.py" file is the main file to run to get the performance on the Adult dataset. Once you implement the following functions, you can simply run:

  ```
  python main_adult.py -m fair
  python main_adult.py -m cfair-eo
  ```

  to get different errors. You can further specify "-u" to change the weight of adversarial classification loss. (Please specify as 0.1, 1, and 10, respectively.)

- The "main_compas.py" file is the main file to run to get the performance on the COMPAS dataset. Once you implement the following functions, you can simply run:

  ```
  python main_compas.py -m fair
  python main_compas.py -m cfair-eo
  ```

  to get different errors. You can further specify "-u" to change the weight of adversarial classification loss. (Please specify as 0.1, 1, and 10, respectively.)

For this problem, you will need to implement these functions following the corresponding docstrings.

  a. In "models.py" file, please implement the "forward" function and the "backward" function in the GradReverse Class.

   b. In "models.py" file, please implement the "forward" function in the FairNet Class.

   c. In "models.py" file, please implement the "forward" function in the CFairNet Class.

5.1 [10pts]

Please plot three figures for each dataset. In these figures, x-axis is the weight of adversarial classification loss (*i.e.*, 0.1, 1, and 10), while the y-axis is a) the overall predicted error ($\Pr(\hat{Y} \neq Y)$), b) the statistical parity gap ($|\Pr(\hat{Y} = 1 \mid A = 0) - \Pr(\hat{Y} = 1 \mid A = 1)|$), and c) the equalized odds gap ($0.5 * |\Pr(\hat{Y} = 1 \mid A = 0, Y = 0) - \Pr(\hat{Y} = 1 \mid A = 1, Y = 0)| + 0.5 * |\Pr(\hat{Y} = 1 \mid A = 0, Y = 1) - \Pr(\hat{Y} = 1 \mid A = 1, Y = 1)|$), respectively.

5.2 [5pts]

Perform an analysis on the results (*e.g.*, What are the relationships between the weight of adversarial classification loss and the above notions? What are the relationships between the overall predicted error and the other two fairness notions?)

5.3 [5pts]

Show your implemented codes.