Homework 2

Fall 23, CS 442: Trustworthy Machine Learning
Due Friday Oct. 27th 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 [30pts]

1.1 [10pts]

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 [20pts]

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 Basics in Information Theory

2.1 [20pts]

Let X be a categorical variable with k possible values, and P, Q be two probability distributions over X. Define a new random variable X' as follows:

$$X' = \begin{cases} X \sim P, & \text{if } B = 0, \\ X \sim Q, & \text{if } B = 1, \end{cases}$$

where $B \in \{0, 1\}$ is an independent and uniform distribution over $\{0, 1\}$.

2.1.1 [10pts]

Show that X' is distributed according to the mixture distribution $M := \frac{1}{2}(P + Q)$.

2.1.2 [10pts]

Show that $I(X'; B) = D_{\text{JS}}(P, Q)$, where $D_{\text{JS}}(P, Q)$ is the Jensen-Shannon divergence between P and Q, i.e., $D_{\text{JS}}(P, Q) = \frac{1}{2} D_{\text{KL}}(P \| M) + \frac{1}{2} D_{\text{KL}}(Q \| M)$.

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

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

$$\Delta_{\text{DP}}(g) := |\Pr_{A=0}(g(X) = 1) - \Pr_{A=1}(g(X) = 1)|,$$

where we use $\Pr_{A=a}(\cdot)$ to denote the conditional probability of an event conditioned on $A = a$. Clearly, if $\Delta_{\text{DP}}(g) = 0$, 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:

$$\varepsilon_{A=0}(h) + \varepsilon_{A=1}(h) \leq 1 - \Delta_{\text{DP}}(g),$$

where $\varepsilon_{A=a}(h) := \mathbb{E}_{A=a}[h(X) \neq a]$.

4 Fair Representations [40pts]

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 $\Delta_{\text{DP}}(\widehat{Y}) := |\Pr_{A=0}(\widehat{Y} = 1) - \Pr_{A=1}(\widehat{Y} = 1)|$.

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.

Dongheng Lin
October 13th, 2023

CS 442: Trustworthy Machine Learning
Homework 2

Solutions.

Q1

1.1

Let X and A be independent binary random variables satisfying $X \perp A$, taking values in 0, 1 with equal probabilities where:

$$\begin{aligned} Pr(X=1) &= Pr(X=0) = \frac{1}{2} \\ Pr(A=1) &= Pr(A=0) = \frac{1}{2} \end{aligned}$$

To satisfy the requirement that $X \not\perp A|Y$, we could have construct Y based on X and A:

$$Y = X \oplus A$$

Here, $\oplus$ represents the XOR operation. this means:

$$Y = \begin{cases} 1, & \text{if } X \neq A \\ 0, & \text{if } X = A \end{cases} \tag{1}$$

Next, let's prove that this setting satisfies $X \not\perp A|Y$.
Proof: Assume that $X \perp A|Y$, that means:

$$\begin{aligned} Pr(X,A|Y) &= Pr(X|Y)P(A|Y) \\ Pr(X=0,A=1|Y=1) &= Pr(X=0|Y=1)P(A=1|Y=1) \end{aligned}$$

- Under the construction above, since $X = 0$ given $Y = 1$ occurs only when $A = 1$ with a probability of 0.5 and $X \perp A$. We have:

$$\begin{aligned} Pr(X=0 \mid Y=1) &= Pr(X=0, A=1 \mid Y=1) + Pr(X=0, A=0 \mid Y=1) \\ &= \frac{Pr(X=0,A=1)}{Pr(X=0,A=1)+Pr(X=1,A=0)} = \frac{0.25}{0.25+0.25} = 0.5 \end{aligned}$$

- Similarly, we have $Pr(A=1|Y=1) = Pr(X=0) = 0.5$

Plug them in the equation gives:

$$\begin{aligned} Pr(X=0,A=1|Y=1) &= Pr(X=1,A=0|Y=1) = 0.5 \\ &\neq Pr(X=0|Y=1)Pr(A=1|Y=1) = 0.25 \end{aligned}$$

Hence, we have found a contradiction that proves this setting must satisfy $X \not\perp A|Y$. That means the constructed X is dependent on A given Y.

1.2

As an example, Let A be a protected attribute that can take values 0 or 1 with equal probability, i.e.,

$$P(A = 0) = P(A = 1) = 0.5$$

Let Y be the target variable that can take values 0, 1, or 2. The distribution of Y given A is:

$$\begin{aligned}
P(Y = 0|A = 0) &= \frac{1}{4} \\
P(Y = 1|A = 0) &= \frac{1}{2} \\
P(Y = 2|A = 0) &= \frac{1}{4} \\
P(Y = 0|A = 1) &= \frac{1}{3} \\
P(Y = 1|A = 1) &= \frac{1}{3} \\
P(Y = 2|A = 1) &= \frac{1}{3}
\end{aligned}$$

Here, we can see that A is not independent of Y since there is:

$$P(Y = 0|A = 0) \neq P(Y = 0|A = 1) \neq P(Y = 0)$$

Now, we define a classifier $\hat{Y} = h(\cdot)$ takes an input feature X where $X \perp A, X \not\perp Y$. Assume X is a feature ranges from {0, 1, 2} and the distribution of X given Y is:

$$\begin{aligned}
P(Y = 0|X = 0) &= 0.7 \\
P(Y = 1|X = 0) &= 0.2 \\
P(Y = 2|X = 0) &= 0.1 \\
P(Y = 0|X = 1) &= 0.1 \\
P(Y = 1|X = 1) &= 0.7 \\
P(Y = 2|X = 1) &= 0.2 \\
P(Y = 0|X = 2) &= 0.1 \\
P(Y = 1|X = 2) &= 0.2 \\
P(Y = 2|X = 2) &= 0.7
\end{aligned}$$

Now, we can build a classifier $\hat{Y} = h(X)$ which picks the value with highest probability given X as prediction. i.e. for such classifier, $h(X) = X$

Given this setup, the example classifier satisfies:

- **Statistical Parity:** Since the classifier $\hat{Y} = h(X)$ is based on X and $X \perp A$, the classifier's predictions will be independent of A, ensuring statistical parity. It's easy to verify that

$$P(\hat{Y} = y|A = 0) = P(\hat{Y} = y|A = 1), \forall y \in \{0, 1, 2\}$$

- **Equalized Odds**: We can observe that since A and $h(X), \hat{Y}$ is independent,

$$P(h(X) = \hat{Y}|A) = P(h(X) = \hat{Y})$$

That implies $\hat{Y} \perp A|Y$ ensuring Equalized Odds for multiple class classification [1].

A counter example like this should suffice for disproving the statement under $|Y| \geq 3$.

Q2

2.1.1

Since B is uniformly distributed over {0, 1}, we have:

$$Pr(B = 0) = Pr(B = 1) = \frac{1}{2}$$

Given a specific value x_i from the k possible values of X,

$$Pr(X' = x_i) = Pr_P(X' = x_i|B = 0)Pr(B = 0) + Pr_Q(X' = x_i|B = 1)Pr(B = 1)$$

Since B is independent of Y, and $X' = X \sim P$ when B = 0 and $X' = X\ Q$ when B = 1 we have:

$$\begin{aligned}
Pr(X' = x_i) &= Pr_P(X' = x_i|B = 0)Pr(B = 0) + Pr_Q(X' = x_i|B = 1)Pr(B = 1) \\
&= Pr_P(X' = x_i)Pr(B = 0) + Pr_Q(X' = x_i)Pr(B = 1) \\
&= \frac{1}{2}\left(Pr_P\left(X' = x_i\right) + Pr_Q\left(X' = x_i\right)\right)
\end{aligned}$$

i.e. $X' \sim M = \frac{1}{2}(P + Q)$ finished the proof.

2.1.2

Firstly, by definition we have:

$$I(X'|B) \quad = \quad H(X') - H(X'|B)$$

According to **2.1.1**, $X' \sim M = \frac{1}{2}(P+Q)$.

$$\begin{aligned} H(X') &= -\sum_{i=1}^{k} Pr_M(X'=x_i)\log Pr_M(X'=x_i) \\ &= -\sum_{i=1}^{k} M(x_i)\log M(x_i) \\ &= -\sum_{i=1}^{k} \frac{1}{2}\left(P(x_i)+Q(x_i)\right)\log\left(\frac{1}{2}\left(P(x_i)+Q(x_i)\right)\right) \end{aligned}$$

The conditional entropy of $H(X'|B)$ can be calculated by following:

$$\begin{aligned} H\left(X' \mid B\right) &= \frac{1}{2}H(X \mid B=0) + \frac{1}{2}H(X \mid B=1) \\ &= \frac{1}{2}\left(-\sum_{i=1}^{k} P\left(x_i\right)\log P\left(x_i\right)\right) + \frac{1}{2}\left(-\sum_{i=1}^{k} Q\left(x_i\right)\log Q\left(x_i\right)\right) \end{aligned}$$

Combining the two terms we have:

$$\begin{aligned} I(X'|B) &= H(X') - H(X'|B) \\ &= -\sum_{i=1}^{k} M(x_i)\log M(x_i) + \frac{1}{2}\left(\sum_{i=1}^{k} P\left(x_i\right)\log P\left(x_i\right)\right) + \frac{1}{2}\left(\sum_{i=1}^{k} Q\left(x_i\right)\log Q\left(x_i\right)\right) \end{aligned}$$

Given that KL divergence is calculated by

$$D_{KL}(P||M) = \sum_{i=1}^{k} P\left(x_i\right)\log\frac{P\left(x_i\right)}{M\left(x_i\right)}$$

$$D_{KL}(Q||M) = \sum_{i=1}^{k} Q\left(x_i\right)\log\frac{Q\left(x_i\right)}{M\left(x_i\right)}$$

Since $D_{JS}(P,Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)$, we have:

$$\begin{aligned} D_{JS}(P,Q) &= \frac{1}{2}\sum_{i=1}^{k} P\left(x_i\right)\log\frac{P\left(x_i\right)}{M\left(x_i\right)} + \frac{1}{2}\sum_{i=1}^{k} Q\left(x_i\right)\log\frac{Q\left(x_i\right)}{M\left(x_i\right)} \\ &= \frac{1}{2}\sum_{i=1}^{k} P\left(x_i\right)\left(\log P\left(x_i\right) - \log M\left(x_i\right)\right) + \frac{1}{2}\sum_{i=1}^{k} Q\left(x_i\right)\left(\log Q\left(x_i\right) - \log M\left(x_i\right)\right) \\ &= -\sum_{i=1}^{k} M(x_i)\log M(x_i) + \frac{1}{2}\left(\sum_{i=1}^{k} P\left(x_i\right)\log P\left(x_i\right)\right) + \frac{1}{2}\left(\sum_{i=1}^{k} Q\left(x_i\right)\log Q\left(x_i\right)\right) \\ &= I(X'|B) \end{aligned}$$

Hence, we have proved that $I(X'|B) = D_{JS}(P,Q)$

Q3

Assume we have a classifier $h(X)$ having the outcome related to g(X):

$$h(X) = 0, \text{ if } g(X) = 1$$
$$h(X) = 1, \text{ if } g(X) = 0$$

For statistical disparity, denote $p_0 = Pr_{A=0}(g(X) = 1)$ and $p_1 = Pr_{A=1}(g(X) = 1)$, we have:

$$\begin{aligned} \Delta_{DP} &= |Pr_{A=0}(g(X) = 1) - Pr_{A=1}(g(X) = 1)| \\ &= |p_0 - p_1| \end{aligned}$$

Since $p_0 \geq 0$ and $p_1 \geq 0$, assume $p_0 \geq p_1$ without the loss of generalizablity, we have $\Delta_{DP} = p_0 - p_1$, thus:

$$1 - \Delta_{DP} = 1 - p_0 + p_1$$

Therefore, we can express the error of $h(X)$ with p_0 and p_1:

$$\begin{aligned} \epsilon_{A=0}(h) &= Pr(h(X) = 1 | A = 0) = Pr_{A=0}(g(X) = 0) = 1 - Pr_{A=0}(g(X) = 1) = 1 - p_0 \\ \epsilon_{A=1}(h) &= Pr(h(X) = 0 | A = 1) = Pr_{A=1}(g(X) = 1) = p_1 \end{aligned}$$

Adding the two terms, we can see that when $p_0 \leq p_1$:

$$\begin{aligned} \epsilon_{A=0}(h) + \epsilon_{A=1}(h) &= 1 - p_0 + p_1 \\ &\leq 1 - \Delta_{DP} \end{aligned}$$

Similarly, we can also find another classifier $h(X)$ for $p_0 \leq p_1$ that have the property can be verified this way, we could have:

$$h(X) = 1, \text{ if } g(X) = 1$$
$$h(X) = 0, \text{ if } g(X) = 0$$
$$\Delta_{DP} = p_0 - p_1$$

Which error rate can be expressed by:

$$\begin{aligned} \epsilon_{A=0}(h) &= Pr_{A=0}(h(X) = 1) = Pr_{A=0}(g(X) = 1) = p_0 \\ \epsilon_{A=1}(h) &= Pr_{A=1}(h(X) = 0) = 1 - Pr_{A=1}(g(X) = 1) = 1 - p_1 \end{aligned}$$

Adding the two terms, we can see that when $p_0 \leq p_1$:

$$\begin{aligned} \epsilon_{A=0}(h) + \epsilon_{A=1}(h) &= 1 - p_1 + p_0 \\ &\leq 1 - \Delta_{DP} \end{aligned}$$

Hence we proved, $\exists h(X), \epsilon_{A=0}(h) + \epsilon_{A=1}(h) \leq 1 - \Delta_{DP}$

Q4

4.1

To prove that equalized odds holds for $h(Z)$, we want to show the property where:

$$Pr(\hat{Y} = 1 \mid Y = y, A = 0) = Pr(\hat{Y} = 1 \mid Y = y, A = 1), \forall y \in \{0, 1\}$$

When y = 1, we can show that the equity between true positive rates on different groups of A holds:

$$Pr(\hat{Y} = 1 \mid Y = 1, A = 0) = Pr(\hat{Y} = 1 \mid Y = 1, A = 1)$$

Starting with the left-hand side, since $\hat{Y} = h(Z)$, it can be written as:

$$Pr(h(Z) = 1 \mid Y = 1, A = 0)$$

Now, given that $Z \perp A \mid Y$, which implies:

$$Pr(h(Z) = 1 \mid Y = 1, A = 0) = Pr(h(Z) = 1 \mid Y = 1)$$

Similarly, for the right-hand side:

$$Pr(\hat{Y} = 1 \mid Y = 1, A = 1) = Pr(h(Z) = 1 \mid Y = 1, A = 1)$$

Again, since $Z \perp A \mid Y$:

$$\begin{aligned} Pr(h(Z) = 1 \mid Y = 1, A = 1) &= Pr(h(Z) = 1 \mid Y = 1) \\ &= Pr(h(Z) = 1 \mid Y = 1, A = 0) \end{aligned}$$

Similarly, we can also prove that for false positive cases:

$$\begin{aligned} Pr(h(Z) = 1 \mid Y = 0, A = 1) &= Pr(h(Z) = 1 \mid Y = 0) \\ &= Pr(h(Z) = 0 \mid Y = 0, A = 0) \end{aligned}$$

That means, we have the condition $Pr(\hat{Y} = 1 \mid Y = y, A = 0) = Pr(\hat{Y} = 1 \mid Y = y, A = 1), \forall y \in \{0, 1\}$ hold which is equivalent to the statement where $h(X)$ satisfies Equalized Odds [2].

4.2

Using the law of total probability, we can express each term of the left hand side as:

$$\begin{aligned} Pr_{A=0}(\hat{Y}=y) &= Pr(\hat{Y}=y \mid Y=0, A=0) \times Pr(Y=0|A=0) \\ &\quad + Pr(\hat{Y}=y \mid Y=1, A=0) \times Pr(Y=1|A=0) \end{aligned}$$

Similarly for $A=1$.

$$\begin{aligned} Pr_{A=1}(\hat{Y}=y) &= Pr(\hat{Y}=y \mid Y=0, A=1) \times Pr(Y=0|A=1) \\ &\quad + Pr(\hat{Y}=y \mid Y=1, A=1) \times Pr(Y=1|A=1) \end{aligned}$$

Given $Z \perp A \mid Y$, we have:

$$Pr(\hat{Y}=y \mid Y=0, A=0) = Pr(\hat{Y}=y \mid Y=0, A=1) = Pr(\hat{Y}=y \mid Y=0)$$
$$Pr(\hat{Y}=y \mid Y=1, A=0) = Pr(\hat{Y}=y \mid Y=1, A=1) = Pr(\hat{Y}=y \mid Y=1)$$

Plugging the above results into the LHS expression, we get:

$$LHS = \mid Pr(\hat{Y}=y \mid Y=0) \times (Pr(Y=0|A=0) - Pr(Y=0|A=1)) +$$
$$Pr(\hat{Y}=y \mid Y=1) \times (Pr(Y=1|A=0) - Pr(Y=1|A=1)) \mid$$

For brevity, let's denote:

$$a := Pr(\hat{Y}=y \mid Y=0) \times (Pr(Y=0|A=0) - Pr(Y=0|A=1))$$
$$b := Pr(\hat{Y}=y \mid Y=1) \times (Pr(Y=1|A=0) - Pr(Y=1|A=1))$$

Given $\gamma_a := Pr_{A=a}(Y=0)$, we have:

$$\gamma_0 = Pr(Y=0 \mid A=0)$$
$$\gamma_1 = Pr(Y=0 \mid A=1)$$

And that means:

$$Pr(Y=0 \mid A=0) = \gamma_0$$
$$Pr(Y=0 \mid A=1) = \gamma_1$$
$$(1 - Pr(Y=0 \mid A=0) = (1-\gamma_0)$$
$$(1 - Pr(Y=0 \mid A=1)) = (1-\gamma_1)$$

Substitute what we have above for $|a|, |b|$:

$$\begin{aligned} |a| &= Pr(\hat{Y}=y \mid Y=0) \times |\gamma_0 - \gamma_1| \\ &= Pr(\hat{Y}=y \mid Y=0) \times |\gamma_0 - \gamma_1| \end{aligned}$$

$$\begin{aligned} |b| &= Pr(\hat{Y}=y \mid Y=1) \times |(1-\gamma_0) - (1-\gamma_1)| \\ &= Pr(\hat{Y}=y \mid Y=1) \times |\gamma_0 - \gamma_1| \end{aligned}$$

By triangle inequality, we know that $|a+b| \leq |a| + |b|$, therefore we can know that:

$$\begin{aligned} LHS = |a+b| &\leq |a| + |b| \leq |\gamma_0 - \gamma_1| \times (Pr(\hat{Y} = y \mid Y = 0) + Pr(\hat{Y} = y \mid Y = 1)) \\ &\leq RHS \end{aligned}$$

The inequality above shows the upper bound of $|a+b|$, which is $|a| + |b|$ consistent with RHS, hence we proved the inequality.

4.3

From 4.2, we know that given $y \in \{0, 1\}$:

$$\begin{aligned}
\Delta_{DP}(h \circ g) = |Pr_{A=0}(\hat{Y} = 1) - Pr_{A=1}(\hat{Y} = 1)| &\leq |\gamma_0 - \gamma_1| \times (Pr(\hat{Y} = 1 \mid Y = 0) + Pr(\hat{Y} = 1 \mid Y = 1)) \\
&\leq \Delta_{BR} \times (Pr(\hat{Y} = 1 \mid Y = 0) + Pr(\hat{Y} = 1 \mid Y = 1)) \\
|1 - Pr_{A=0}(\hat{Y} = 1) - 1 + Pr_{A=1}(\hat{Y} = 1)| &\leq |\gamma_0 - \gamma_1| \times (Pr(\hat{Y} = 0 \mid Y = 0) + Pr(\hat{Y} = 0 \mid Y = 1)) \\
&\leq \Delta_{BR} \times (Pr(\hat{Y} = 0 \mid Y = 0) + Pr(\hat{Y} = 0 \mid Y = 1))
\end{aligned}$$

Since $\Delta_{DP}(h \circ g) = |Pr_{A=0}(\hat{Y} = 1) - Pr_{A=1}(\hat{Y} = 1)| = |Pr_{A=1}(\hat{Y} = 1) - Pr_{A=0}(\hat{Y} = 1)|$, adding the two inequalities above we have:

$$2\Delta_{DP}(h \circ g) = 2|Pr_{A=0}(\hat{Y} = 1) - Pr_{A=1}(\hat{Y} = 1)| \quad \leq \quad \Delta_{BR}(TPR + FPR + TNR + FNR)$$

$\because TPR + FNR = 1, FPR + TNR = 1$, hold under any classifier, that means the inequality above simplifies to:

$$\begin{aligned}
\therefore 2\Delta_{DP}(h \circ g) &\leq 2\Delta_{BR} \\
\Delta_{DP}(h \circ g) &\leq \Delta_{BR}
\end{aligned}$$

Hence, we proved that the upper bound of $\Delta_{DP}(h \circ g)$ is actually Δ_{BR}.

References

[1] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953. PMLR, 07–10 Jul 2017.

[2] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.