COMS 4771 Machine Learning 2023 Spring
Homework 3

Danrui Wang - dw3031@columbia.edu

3/30/2023

1 Inconsistency of the fairness definitions

(i)

Think of the scenario that whether a person has bad records is the only criterion if the bank would approve the loan. If the person doesn't have any bad records, the bank would definitely approve the loan. If the person have any bad records, the bank would definitely refuse the loan.
X: whether the person has bad records ={1: has bad records, 0:doesn't has bad records} A: gender = {1:male, 0: female}
Y: whether the bank approve the loan = {1: approve the loan, 0: refuse the loan}

demographic parity Because whether the bank would approve the loan is only dependent on whether the person has bad records or not, the probability of whether the bank would approve the loan is the same for both gender.
If the person doesn't have any bad records (X = 0),

$$P_0[\hat{Y} = 1] = P_1[\hat{Y} = 1] = 1$$
$$P_0[\hat{Y} = 0] = P_1[\hat{Y} = 0] = 0$$

If the person has bad records (X = 1),

$$P_0[\hat{Y} = 1] = P_1[\hat{Y} = 1] = 0$$
$$P_0[\hat{Y} = 0] = P_1[\hat{Y} = 0] = 1$$

equalized odds Because Y is only dependent on whether the person has bad records or not, equalized odds would be the same for both gender.
If the person doesn't have any bad records (X = 0),

$$P_0[\hat{Y} = 1|Y = 1] = P_1[\hat{Y} = 1|Y = 1] = 1$$
$$P_0[\hat{Y} = 0|Y = 1] = P_1[\hat{Y} = 0|Y = 1] = 0$$

If the person has bad records (X = 1),

$$P_0[\hat{Y} = 0|Y = 0] = P_1[\hat{Y} = 0|Y = 0] = 1$$
$$P_0[\hat{Y} = 1|Y = 0] = P_1[\hat{Y} = 1|Y = 0] = 0$$

predictive parity Because there is no other criterion of Y apart from if the person has any bad records, the positive predictive and negative predictive value would be the same across the sensitive attribute. If the person doesn't have any bad records (X = 0),

$$P_0[Y = 1|\hat{Y} = 1] = P_1[Y = 1|\hat{Y} = 1] = 1$$
$$P_0[Y = 0|\hat{Y} = 1] = P_1[Y = 0|\hat{Y} = 1] = 0$$

If the person has bad records (X = 1),

$$P_0[Y = 0|\hat{Y} = 0] = P_1[Y = 0|\hat{Y} = 0] = 1$$
$$P_0[Y = 1|\hat{Y} = 0] = P_1[Y = 1|\hat{Y} = 0] = 0$$

In my example, Y is not dependent on our sensitive attribute A. All three definitions are satisfied simultaneously.

(ii)

If demographic parity holds true, $A \perp \hat{Y}$
If predictive parity holds true, $A \perp Y|\hat{Y}$
If both DP and PP hold true at the same time,

$$\begin{aligned} P(A \cap Y|\hat{Y}) &= P(A)P(Y|\hat{Y}) && \because A \perp Y|\hat{Y} \\ &= P(A|\hat{Y})P(Y|\hat{Y}) && \because A \perp \hat{Y} \end{aligned}$$

we get $A \perp Y$ ($\because P(A \cap Y|\hat{Y}) = P(A|\hat{Y})P(Y|\hat{Y})$)
Thus, if $A \not\perp Y$, DP and PP cannot hold true at the same time. (modus tollens)

(iii)

If demographic parity holds true, $A \perp \hat{Y}$
If Equalized Odds holds true, $A \perp \hat{Y}|Y$
Combining two conditions together:

$$\begin{aligned} P(A \cap \hat{Y}|Y) &= P(\hat{Y}|Y)P(A|Y) && \text{because } A \perp \hat{Y} \\ &= P(\hat{Y} \cap A|Y) && \text{indicate if } A \perp \hat{Y},\ A \perp \hat{Y}|Y \Longrightarrow \hat{Y} \perp A|Y \end{aligned}$$

If both DP and EO holds true at the same time,
(1) $A \perp \hat{Y}|Y$

$$\begin{aligned} P(A \cap \hat{Y}|Y) &= P(A)P(\hat{Y}|Y) && \because A \perp \hat{Y}|Y \\ P(A \cap \hat{Y}|Y) &= P(A|Y)P(\hat{Y}|Y) && \because A \perp \hat{Y} \\ &\therefore P(A|Y) = P(A), \textbf{ indicating } A \perp Y \end{aligned}$$

(2) $\hat{Y} \perp A|Y$

$$\begin{aligned} P(\hat{Y} \cap A|Y) &= P(\hat{Y})P(A|Y) && \because \hat{Y} \perp A|Y \\ P(\hat{Y} \cap A|Y) &= P(\hat{Y}|Y)P(A|Y) && \because A \perp \hat{Y} \\ &\therefore P(\hat{Y}) = P(\hat{Y}|Y), \textbf{ indicating } \hat{Y} \perp Y \end{aligned}$$

Thus, if $A \not\perp Y$ and $\hat{Y} \not\perp Y$, Demographic Parity and Equalized Odds cannot hold at the same time.

(iv)

Because Equalized Odds needs to satisfy $P_0[\hat{Y} = \hat{y}|Y = y] = P_1[\hat{Y} = \hat{y}|Y = y]$, $\forall \hat{y}, y \in 0, 1$, meaning we need to prove the fairness of

$$FPR : P_a(\hat{Y} = 1|Y = 0) \qquad TNR : P_a(\hat{Y} = 0|Y = 0) = 1 - FPR$$
$$FNR : P_a(\hat{Y} = 0|Y = 1) \qquad TPR : P_a(\hat{Y} = 1|Y = 1) = 1 - FNR$$

We can only look at FPR and FNR.
Similarly , Predictive Parity needs to prove the fairness of

$$PPV : P_a(Y = 1|\hat{Y} = 1) \qquad FDR : P_a(Y = 0|\hat{Y} = 1) = 1 - PPV$$
$$NPV : P_a(Y = 0|\hat{Y} = 0) \qquad FOR : P_a(Y = 1|\hat{Y} = 0) = 1 - NPV$$

So, $P_0[Y = y|\hat{Y} = \hat{y}] = P_1[Y = y|\hat{Y} = \hat{y}]$, $\forall \hat{y}, y \in 0, 1$ is equivalent to $P_0[Y = 1|\hat{Y} = \hat{y}] = P_1[Y = 1|\hat{Y} = \hat{y}]$, $\forall \hat{y} \in 0, 1$.
Also, we can see that a necessary condition for PP is the equality of PPV, to prove the third statement, we prove a stronger statement: if A is dependent on Y , Equalized Odds and equality of Positive Predictive Value cannot hold at the same time, which is (FPR and FNR) and PPV .

As for FPR:
CN: condition negative, FP: false positive, FN: false negative, TP: true positive, TN, true negative
P = P(Y=1)

$$\begin{aligned}
FPR &= \frac{\sum FP}{\sum CN} \\
&= \frac{\sum FP}{\sum TN + \sum FP} \\
&= \frac{\sum TP + \sum FN}{\sum TN + \sum FP} \times \frac{\sum TP}{\sum TP + \sum FN} \times \frac{\sum FP}{\sum TP} \\
&= \frac{\frac{\sum TP + \sum FN}{\sum population}}{\frac{\sum FN + \sum FP}{\sum population}} \times (1 - \frac{\sum FN}{\sum FN + \sum TP}) \times \frac{\frac{\sum FP}{\sum TP + \sum FP}}{\frac{\sum TP}{\sum TP + \sum FP}} \\
&= \frac{P}{1 - P}(1 - FNR)\frac{1 - \frac{\sum TP}{\sum TP + \sum FP}}{\frac{\sum TP}{\sum TP + \sum FP}} \\
&= \frac{P}{1 - P}\frac{1 - PPV}{PPV}(1 - FNR)
\end{aligned}$$

If both EO and PP holds at the same time, $FPR_0 = FPR_1$, $FNR_0 = FNR_1$, and $PPV_0 = PPV_1$:

$$\frac{P_0}{1-P_0}\frac{1-PPV_0}{PPV_0}(1-FNR_0) = \frac{P_1}{1-P_1}\frac{1-PPV_1}{PPV_1}(1-FNR_1)$$
$$\frac{P_0}{1-P_0} = \frac{P_1}{1-P_1}$$
$$P_0(Y=1) = P_1(Y=1)$$

Thus, if $A \not\perp Y$, EO and PP cannot hold true at the same time.

2 Combining multiple classifiers

(i)

$$
\begin{aligned}
D_{t+1}(i) &= \frac{D_t(i)\exp(-\alpha_t y_i f_t(x_i))}{Z_t} \\
D_{T+1}(i) &= D_1(i) \cdot \frac{\prod_t \exp(-\alpha_t y_i f_t(x_i))}{\prod_t Z_t} \\
&= \frac{\exp(-y_i \sum_t \alpha_t f_t(x_i))}{m \cdot \prod_t Z_t} \\
&= \frac{\exp(-y_i g(x_i))}{m \cdot \prod_t Z_t}
\end{aligned}
$$

(ii)

Because

$$
\begin{aligned}
y_i g(x_i) \leq 0 \rightarrow \exp(-y_i(g(x_i)) \geq 1 = 1[y_i \neq \text{sign}(g(x_i))] \\
y_i g(x_i) > 0 \rightarrow \exp(-y_i(g(x_i)) \geq 0 = 1[y_i \neq \text{sign}(g(x_i))]
\end{aligned}
$$

So

$$
\begin{aligned}
err(g) &= \frac{1}{m}\sum_i 1[y_i \neq \text{sign}(g(x_i))] \\
&= \frac{1}{m}\sum_i 1[y_i \cdot \text{sign}(g(x_i)) \leq 0] \\
&\leq \frac{1}{m}\sum_i \exp(-y_i g(x_i)) \\
&= \frac{1}{m}\sum_i m \cdot \prod_t Z_t D_{T+1}(i) \\
&= \prod_t Z_t \sum_i D_{T+1}(i) \\
&= \prod_t Z_t
\end{aligned}
$$

(iii)

$$\begin{aligned}
Z_t &= \sum_i D_t(i)\exp(-\alpha_t y_i f(t(x_i)) \\
&= \exp(\alpha_t) \sum_{i:y_i \neq f_t(x_i)} D_t(i) + \exp(-\alpha_t) \sum_{i:y_i = f_t(x_i)} D_t(i) \\
&= \exp(\alpha_t)\epsilon_t + \exp(-\alpha_t)(1-\epsilon_t) \\
&= \exp(\ln(\sqrt{\frac{1-\epsilon_t}{\epsilon_t}})\epsilon_t) + \exp(\ln(\sqrt{\frac{\epsilon_t}{1-\epsilon_t}})(1-\epsilon_t)) \\
&= \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\epsilon_t + \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}(1-\epsilon_t) \\
&= 2\sqrt{\epsilon_t(1-\epsilon_t)}
\end{aligned}$$

(iv)

$$\begin{aligned}
Z_t &= 2\sqrt{\epsilon_t(1-\epsilon_t)} \\
&= 2\sqrt{(0.5-\gamma_t)(1-(0.5-\gamma_t)} \\
&= 2\sqrt{0.25-\gamma_t^2} \\
&= \sqrt{1-4\gamma_t^2} \\
&\leq \exp(-2\gamma_t^2) \\
err(g) &= \prod_t Z_t \\
&\leq \prod_t \exp(-2\gamma_t^2) \\
&= \exp(-2\sum_t \gamma_t^2)
\end{aligned}$$

3 1-Norm Support Vector Machine

(i)

let $k_j = |w_j|$, $k_j \geq w_j, k_j \geq -w_j, j = 1, ..., n$
then:

$$\text{minimize } ||w||_1 = \min \sum_{j=1}^{n} |w_j| = \min \sum_{j=1}^{n} k_j$$
$$y_i(w \cdot x_i + w_0) \geq 1, i = 1, ..., m$$

for w_j and k_j, $j = 1, ..., n$, there are 2n+1 variables and m+2n constraints.

(ii)

Imaging shifting the $w \cdot x + w_0 = \pm 1$ until $w \cdot x + w_0 = -1$ passes through the origin, the Chebyshev distance between the two hyperplane becomes the Chebyshev distance between origin and $w \cdot x = 2$.
Let

$$W = (w_1, w_2, ..., w_n)$$
$$S = (sign(w_1), sign(w_2), ..., sign(w_n))$$
$$S \cdot W = sign(w_1) \cdot w_1 + ... + sign(w_n) \cdot w_n = \sum_{i=1}^{n} |w_i| = ||w||_1$$

Extend s to intersects $w \cdot x = 2$ and intersect at $x_i = (x_{i1}, x_{i2}, ..., x_{in}) = (s \cdot sign(w_1), ..., s \cdot sign(w_n))$.
Since

$$\begin{aligned} w \cdot w_i &= s \cdot sign(w_1) \cdot w_1 +, ... + s \cdot sign(w_n) \cdot w_n \\ &= s \cdot |w_1| + ... + s \cdot |w_n| = s \cdot ||w||_1 = 2 \\ s &= \frac{2}{||w||_1} \end{aligned}$$

The minimum Chebyshev distance from the origin to the point on $w \cdot x = 2$ is $\frac{2}{||w||_1}$.
Assume there is a point w_j on $w \cdot x = 2$ that has shorter Chebyshev distance than w_i.

$$\begin{aligned} \because w \cdot x_j &= w_1 x_{j1} + w_2 x_{j2} + ... + w_n x_{jn} \\ &\leq |w_1 x_{j1}| + |w_2 x_{j2}| + ... + |w_n x_{jn}| \\ &\leq |w_1| \frac{2}{||w||_1} + |w_2| \frac{2}{||w||_1} + .. + |w_n| \frac{2}{||w||_1} = 2 \end{aligned}$$

$\therefore x_j$ is not on the plane$w \cdot x + w_0 = 2$

$\therefore$ contradict with original assumption

$$\therefore x_j = x_i \longrightarrow \frac{2}{||w||_1} = max(|x_{j1}|, |x_{j2}, ..., |x_{jn}|)$$

$\therefore$ 1-norm SVM maximizes the Chebyshev distance between the hyper planes with $\frac{2}{||w||_1}$

(iii)

(iv)

1-norm SVM makes more sense. Because 1-norm SVM has it has many zero weights whereas 2-norm SVM has many non-zero small weights, if the output y depends only on a few input variables which the weigh vector is sparse, 1-norm can identify the most important input variables, disregard irrelvant features, improve the model interpretation, and reduce the model complexity.

4 Estimating Model Parameters for Regression

(i)

The optimization problem:

$$\min_{\beta}(-Q(\beta)) \qquad \text{-Q is objective function}$$
$$||\beta|| - 1 \leq 0 \qquad \text{inequality constraint}$$

First prove the objective function is convex.

$$\begin{aligned}
f_\beta(x_i, y_i) &= P(Y = y_i, X = x_i) \\
&= P(Y = y_i|X = x_i) \cdot P(X = x_i) = P(Y = y_i|X = x_i) \cdot \prod_{j=1}^{d} P(X_J = x_{ij}) \\
&= \frac{1}{\sqrt{2\pi}||x_i||} \exp(-\frac{(y_i - x_i^T\beta)^2}{2||x_i||^2} \cdot \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x_{ij}^2}{2}) \\
\ln(f_\beta(x_i, y_i)) &= -\ln(\sqrt{2\pi}||x_i||) - \frac{(y_i - x_i^T\beta)^2}{2||x_i||^2} - \frac{d}{2}\ln(2\pi) - \sum_{j=1}^{d} \frac{x_{ij}^2}{2} \\
&= -\ln(\sqrt{2\pi}||x_i||) - \frac{(y_i - x_i^T\beta)^2}{2||x_i||^2} - \frac{d}{2}\ln(2\pi) - \frac{||x_i||^2}{2} \\
-Q(\beta) &= -\frac{1}{n}\sum_{i=1}^{n} \ln(f_\beta(x_i, y_i)) \\
\frac{\partial Q(\beta)}{\partial \beta_j} &= -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ln(f_\beta(x_i, y_i))}{\partial \beta_j} \\
&= -\frac{1}{n}\sum_{i=1}^{n} \frac{y_i - x_i^T\beta}{||x_i||^2} x_{ij}
\end{aligned}$$

For k $\in 1, ..., d,$ Hessian

$$\begin{aligned}
H_{jk} &= \frac{\partial^2 - Q(\beta)}{\partial\beta_j\partial\beta_k} = -\frac{1}{n}\sum_{i=1}^{n} x_{ij}\frac{1}{||x_i||^2}\frac{\partial y_i - x_i^T\beta}{\partial\beta_k} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_{ik}x_{ij}}{||x_i||^2} \\
\because \beta H\beta^T &= \sum_{j,k}\beta_j\beta_k H_{j,k} = \sum_{j,k}\beta_j\beta_k\frac{1}{n}\sum_{i=1}^{n}\frac{x_{ik}x_{ij}}{||x_i||^2} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{||x_i||^2}\sum_j x_{ij}\beta_j\sum_k x_{ik}\beta_k \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{||x_i||^2}(x_i^T\beta)^2 \geq 0
\end{aligned}$$

Thus, the Hessian is Positive semi-definite, and therefore the objective function is convex. Then, prove the inequality constraint functions are convex.
Let $K \in R^d$, subject to $||\beta|| - 1 \leq 0$,. For any $\lambda \in [0, 1]$ and β, $\beta' \in K$, we build point $\lambda\beta + (1 - \lambda)\beta'$
Since, $\lambda||\beta|| + (1 - \lambda)||\beta'|| \geq ||\lambda\beta + (1 - \lambda)\beta'||$ and $\lambda \leq 1$, $||\beta|| \leq 1$ and $||\beta'|| \leq 1$, we get $||\lambda\beta + (1 - \lambda)\beta'|| \leq \lambda||\beta|| + (1 - \lambda)||\beta'|| \leq 1$

Thus, any build point $\lambda\beta + (1 - \lambda)\beta'$ is in K and β is convex

(ii)

$$\begin{aligned}
\frac{\partial Q(\beta)}{\partial \beta_j} &= \frac{1}{n}\sum_{i=1}^{n} \frac{y_i - x_i^T\beta}{||x_i||^2} x_{ij} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{y_i x_{ij} - xijx_i^T\beta}{||x_i||^2} \\
\delta_\beta Q &= \frac{1}{n}\sum_{i=1}^{n} \frac{y_i - x_i^T\beta}{||x_i||^2} x_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{x_i y_i - xix_i^T\beta}{||x_i||^2} = 0 \\
&= \sum_{i=1}^{n} \frac{x_i y_i}{||x_i||^2} = \sum_{i=1}^{n} \frac{x_i x_i^T\beta}{||x_i||^2} \\
A &= \sum_{i=1}^{n} x_i x_i^T \\
b &= \sum_{i=1}^{n} x_i y_i
\end{aligned}$$