Machine Learning Exercise Sheet 14

Fairness

In-class Exercises

Formal Fairness Criteria

Problem 1: You are given data as shown on Table 1 where $X \in \mathbb{R}$ denotes the non-sensitive feature, $A \in \{a, b\}$ denotes the sensitive feature, and $Y \in \{0, 1\}$ denotes the ground-truth label.

Table 1: Fairness Data (each column is one data point)

ID	1	2	3	4	5	6	7
X	0.5	-1.0	-0.5	2.0	0.5	1.5	0.1
A	a	b	b	a	b	a	b
Y	1	1	0	0	0	0	1

a) Let the prediction $R = r(X)$ be some arbitrary function r that only depends on X. The sensitive attribute A is ignored. Can we conclude that the *Sufficiency* fairness criterion is satisfied for the data shown on Table 1? Justify your answer.

> No. We cannot conclude anything because "Fairness through Unawareness" does not work. There can be many highly correlated features that are proxies of the sensitive attribute.

b) Let the prediction $R \in \{0, 1\}$ be

$$R = \begin{cases} 0 & \text{if } 2 \cdot X > 2 \text{ and } A = a \\ 0 & \text{if } 4 \cdot X > 1 \text{ and } A = b \\ 1 & \text{otherwise} \end{cases}$$

Which ones of the following three fairness criteria *Independence, Separation, and Equality of Opportunity* are satisfied for the data shown on Table 1? Justify your answer.

> First we compute the predictions R to obtain:
>
> | ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
> |---|---|---|---|---|---|---|---|
> | X | 0.5 | -1.0 | -0.5 | 2.0 | 0.5 | 1.5 | 0.1 |
> | A | a | b | b | a | b | a | b |
> | R | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
> | Y | 1 | 1 | 0 | 0 | 0 | 0 | 1 |

Upload a single PDF file with your homework solution to Moodle by 09.02.2022, 11:59pm CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

We see that 1/3 instances in group a have $R = 1$ vs. 3/4 instances in group b. *Independence* is not satisfied.

For group a we have TP=1/1 and FP=0/2.

For group b we have TP=2/2 and FP=1/2.

Since only TP matches for both groups, *Equality of Opportunity* is satisfied and *Separation* is not satisfied.

c) Modify the *least* number of instances such that none of the above criteria are satisfied. You can only modify the non-sensitive features X. Write down the ID(s) of the modified instance(s) and their modified X value. Justify your answer!

We modify the instance with ID 1, changing the non-sensitive feature from $X = 0.5$ to $X = 1.5$. Thus, the prediction changes from $R = 1$ to $R = 0$.

Now we have 0/3 instances with $R = 1$ within its group, vs. 3/4 in the other group so *Independence* is still not satisfied. The TP rate has changed from 1/1 to 0/1 compared to 2/2 in the other group, which means that neither *Equality of Opportunity* nor *Separation* are satisfied.

Homework

Formal Fairness Criteria

Problem 2: As in the lecture, assume that we have non-sensitive features X, sensitive feature $A \in \{a, b\}$ and labels $Y \in \{0, 1\}$ following a joint data distribution $\mathcal{D}$, i.e. $((X, A), Y) \sim \mathcal{D}$, and that we have a binary clasifier f with $R = f(X, A)$.

Further assume that sensitive feature A and label Y have the joint distribution specified in Table 2,i.e. the sensitive feature A and the labels Y are entirely uncorrelated and both sub-populations are equally large.

	$A = a$	$A = b$
$Y = 0$	$\frac{1}{4}$	$\frac{1}{4}$
$Y = 1$	$\frac{1}{4}$	$\frac{1}{4}$

Table 2: $P(A, Y)$

a) We want to find the highest accuracy that can be achieved on data distributed as in Table 2 while ensuring the *independence* fairness criterion. Specify a joint distribution $P(A, Y, R)$ that

1. has marginal distribution $P(A, Y)$ specified in Table 2,
2. fulfills $R \perp\!\!\!\perp A$,

Upload a single PDF file with your homework solution to Moodle by 09.02.2022, 11:59pm CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

3. and maximizes the accuracy $\Pr[Y=0,R=0]+\Pr[Y=1,R=1]$,

by filling out Table 3. Justify your response. What is the maximum possible accuracy?

	$A=a$		$A=b$	
$Y=0$				
$Y=1$				
	$R=0$	$R=1$	$R=0$	$R=1$

Table 3: $P(A,Y,R)$.

	$A=a$		$A=b$	
$Y=0$	$\frac{1}{4}$	0	$\frac{1}{4}$	0
$Y=1$	0	$\frac{1}{4}$	0	$\frac{1}{4}$
	$R=0$	$R=1$	$R=0$	$R=1$

Condition 1. is fulfilled, because every combination of values for A and Y is still equally likely.

Condition 2. is fulfilled, because both groups have a 50% probability for $R=0$ and for $R=1$, i.e.

$$\Pr[R=r\mid A=a]=\frac{\frac{1}{4}}{\frac{1}{4}+\frac{1}{4}}=\frac{1}{2}=\frac{1}{4}+\frac{1}{4}=\Pr[R=r],$$

and

$$\Pr[R=r\mid A=b]=\frac{\frac{1}{4}}{\frac{1}{4}+\frac{1}{4}}=\frac{1}{2}=\frac{1}{4}+\frac{1}{4}=\Pr[R=r],$$

for $r\in\{0,1\}$.

Condition 3. is fulfilled, because the accuracy

$$\begin{aligned}
&\Pr[Y=0,R=0]+\Pr[Y=1,R=1] && (1)\\
=&\Pr[Y=0,A=a,R=0]+\Pr[Y=0,A=b,R=0]\\
&+\Pr[Y=1,A=a,R=1]+\Pr[Y=1,A=b,R=1] && (2)\\
=&\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4} && (3)\\
=&100\% && (4)
\end{aligned}$$

cannot possibly be higher than 100%.

Now, assume that the sensitive feature A and the label Y are heavily correlated, with joint distribution $P(A,Y)$ specified in Table 4.

	$A = a$	$A = b$
$Y = 0$	$\frac{3}{8}$	$\frac{1}{8}$
$Y = 1$	$\frac{1}{8}$	$\frac{3}{8}$

Table 4: $P(A, Y)$

b) Specify a joint distribution $P(A, Y, R)$ that

1. has marginal distribution $P(A, Y)$ specified in Table 4,
2. fulfills $R \perp\!\!\!\perp A$,
3. and maximizes the accuracy $\Pr[Y = 0, R = 0] + \Pr[Y = 1, R = 1]$,

by filling out Table 3. Justify your response. What is the maximum possible accuracy?

To maximize the accuracy, we need to maximize the sum of the "diagonal" elements

- $P(A = a, Y = 0, R = 0)$
- $P(A = a, Y = 1, R = 1)$
- $P(A = b, Y = 0, R = 0)$
- $P(A = a, Y = 1, R = 1)$

Due to the independence criterion (and because the two sub-populations are equally large), increasing the probability of $R = r$ with $r \in \{0, 1\}$ for one of the sub-populations ($A = a$ or $A = b$) forces us to also increase the probability of $R = r$ by an equal amount for the other sub-population.

The first 50% of the probability mass can be entirely allocated to correct predictions, without violating the fairness criterion.

	$A = a$		$A = b$	
$Y = 0$	$\frac{1}{8}$	0	$\frac{1}{8}$	0
$Y = 1$	0	$\frac{1}{8}$	0	$\frac{1}{8}$
	$R = 0$	$R = 1$	$R = 0$	$R = 1$

The entries $P(A = a, Y = 1, R = 1)$ and $P(A = b, Y = 0, R = 0)$ cannot be increased further, or the marginal probabilities $P(A = a, Y = 1)$ and $P(A = b, Y = 0)$ would exceed the values specified in Table 4.

Due to independence, increasing $P(A = a, Y = 0, R = 0)$ by ϵ forces us to also increase $P(A = b, Y = 1, R = 0)$ by ϵ. Increasing $P(A = b, Y = 1, R = 1)$ by ϵ forces us to also increase $P(A = a, Y = 0, R = 1)$ by ϵ. This leaves us with

	$A=a$		$A=b$	
$Y=0$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	0
$Y=1$	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$
	$R=0$	$R=1$	$R=0$	$R=1$

The overall accuracy is $\frac{2}{8}+\frac{1}{8}+\frac{1}{8}+\frac{2}{8}=75\%$.

c) Now, instead of enforcing independence, we want to determine the highest accuracy that can be achieved for our highly correlated distribution while enforcing the *separation* criterion. Specify a joint distribution $P(A,Y,R)$ that

1. has marginal distribution $P(A,Y)$ specified in Table 4,
2. fulfills $R \perp\!\!\!\perp A \mid Y$,
3. and maximizes the accuracy $\Pr[Y=0,R=0]+\Pr[Y=1,R=1]$,

by filling out Table 3. Justify your response. What is the maximum possible accuracy?

	$A=a$		$A=b$	
$Y=0$	$\frac{3}{8}$	0	$\frac{1}{8}$	0
$Y=1$	0	$\frac{1}{8}$	0	$\frac{3}{8}$
	$R=0$	$R=1$	$R=0$	$R=1$

Condition 1. is obviously fulfilled.

Condition 2. is also fulfilled: For $A=a$, we have:

$$\Pr[R=0 \mid A=a,Y=0]=\frac{\frac{3}{8}}{\frac{3}{8}+0}=1=\frac{\frac{3}{8}+\frac{1}{8}}{\frac{3}{8}+0+\frac{1}{8}+0}=\Pr[R=0|Y=0],$$

$$\Pr[R=0 \mid A=a,Y=1]=\frac{0}{0+\frac{1}{8}}=0=\frac{0+0}{0+\frac{1}{8}+0+\frac{3}{8}}=\Pr[R=0|Y=1],$$

$$\Pr[R=1 \mid A=a,Y=0]=\frac{0}{\frac{3}{8}+0}=0=\frac{0+0}{\frac{3}{8}+0+\frac{1}{8}+0}=\Pr[R=1|Y=0],$$

$$\Pr[R=1 \mid A=a,Y=1]=\frac{\frac{1}{8}}{0+\frac{1}{8}}=1=\frac{\frac{1}{8}+\frac{3}{8}}{0+\frac{1}{8}+0+\frac{3}{8}}=\Pr[R=1|Y=1].$$

The same holds, symmetrically for $A=b$.

Condition 3. is fulfilled, because the accuracy cannot possibly be higher than 100%.

Problem 3: Many fairness criteria are mutually exclusive. Prove that a joint distribution $P(A, Y, R)$ cannot simultaneously fulfill independence and sufficiency, if the label Y and the sensitive feature A are not independent.

More formally: Prove that, if $Y \not\perp\!\!\!\perp A$, then there is no $P(A, Y, R)$ with $R \perp\!\!\!\perp A$ and $Y \perp\!\!\!\perp A \mid R$.

Proof by contradiction. Assume that $R \perp\!\!\!\perp A$ and $Y \perp\!\!\!\perp A \mid R$, despite $Y \not\perp\!\!\!\perp A$.

We have

$$\begin{aligned} P(Y = y|A = a) &= \sum_{r=0}^{1} P(Y = y|R = r, A = a) \cdot P(R = r|A = a) \\ &= \sum_{r=0}^{1} P(Y = y|R = r) \cdot P(R = r|A = a) \\ &= \sum_{r=0}^{1} P(Y = y|R = r) \cdot P(R = r) \\ &= P(Y = y), \end{aligned}$$

where the first equality is the law of total probability, the second equality follows from sufficiency, the third equality follows from independence and the last equality is again the law of total probability.

This shows that, due to sufficiency and independence, we must have $Y \perp\!\!\!\perp A$. This contradicts our initial assumption that the label Y and the sensitive feature A are not independent.

Problem 4: In the lecture, we discussed how the classification threshold of a model can be adjusted to control its true-positive and false-positive rates and guarantee *separation*. However, we may only have block-box access to a classifier's binary outputs, and no access to its continuous score function.

As before, let $R = f(X, A)$ with $((X, A), Y) \sim \mathcal{D}$, where $Y, R \in \{0, 1\}$ and $A \in \{a, b\}$.
Furthermore, let

$$\begin{aligned} \text{TP} &= P(R = 1|Y = 1) \\ \text{FP} &= P(R = 1|Y = 0) \end{aligned}$$

be the true-positive and false-positive rates of classifier f.

a) Consider the random classifier $\hat{f}(x, a) = Z + (1 - Z) \cdot f(x, a)$ with $Z \sim \text{Bern}(p)$ and $\hat{R} = \hat{f}(X, A)$. It returns 1 with a probability of p and $f(x, a)$ with a probability of $1 - p$.

   Express the true-positive rate $P\left(\hat{R} = 1|Y = 1\right)$ and false-positive rate $P\left(\hat{R} = 1|Y = 0\right)$ of classifier $\hat{f}$ as functions of TP, FP and p.

Using the law of total probability, we have

$$\begin{aligned}
&P\left(\hat{R}=1|Y=1\right)\\
=&P\left(\hat{R}=1 \mid R=1,Y=1\right)\cdot P\left(R=1 \mid Y=1\right)\\
&+P\left(\hat{R}=1 \mid R=0,Y=1\right)\cdot P\left(R=0 \mid Y=1\right)\\
=&P\left(\hat{R}=1 \mid R=1,Y=1\right)\cdot \mathrm{TP}+P\left(\hat{R}=1 \mid R=0,Y=1\right)\cdot(1-\mathrm{TP})\\
=&1\cdot \mathrm{TP}+p\cdot(1-\mathrm{TP})\\
=&p\cdot 1+(1-p)\cdot \mathrm{TP}.
\end{aligned}$$

The second equality results from inserting the definition of the true-positive rate. The third equality follows from the definition of $\hat{f}$.

Similarly, we can show that

$$\begin{aligned}
&P\left(\hat{R}=1|Y=0\right)\\
=&P\left(\hat{R}=1 \mid R=1,Y=0\right)\cdot P\left(R=1 \mid Y=0\right)\\
&+P\left(\hat{R}=1 \mid R=0,Y=0\right)\cdot P\left(R=0 \mid Y=0\right)\\
=&P\left(\hat{R}=1 \mid R=1,Y=0\right)\cdot \mathrm{FP}+P\left(\hat{R}=1 \mid R=0,Y=0\right)\cdot(1-\mathrm{FP})\\
=&1\cdot \mathrm{FP}+p\cdot(1-\mathrm{FP})\\
=&p\cdot 1+(1-p)\cdot \mathrm{FP}.
\end{aligned}$$

Evidently, random classifier f yields true-positive and false-positive rates that are linearly interpolated between $(\mathrm{TP},\mathrm{FP})$ and $(1,1)$.

b) Now consider the random classifier $\check{f}(x,a)=Z\cdot f(x,a)$ with $Z\sim\mathrm{Bern}(q)$ and $\check{R}=\check{f}(X,A)$. It returns 0 with a probability of $1-q$ and $f(x,a)$ with a probability of q.

Express the true-positive rate $P\left(\check{R}=1|Y=1\right)$ and false-positive rate $P\left(\check{R}=1|Y=0\right)$ of classifier $\check{f}$ as functions of TP, FP and q.

Similar to subtask a), we have

$$\begin{aligned}
&P\left(\check{R}=1|Y=1\right)\\
=&P\left(\check{R}=1 \mid R=1,Y=1\right)\cdot \mathrm{TP}+P\left(\check{R}=1 \mid R=0,Y=1\right)\cdot(1-\mathrm{TP})\\
=&q\cdot \mathrm{TP}+0\cdot(1-\mathrm{TP})\\
=&q\cdot \mathrm{TP}+(1-q)\cdot 0.
\end{aligned}$$

The second equality results from inserting the definition of the true-positive rate. The third equality follows from the definition of $\hat{f}$.

Similarly, we can show that

$$
\begin{aligned}
& P\left(\hat{R}=1|Y=0\right) \\
=& P\left(\hat{R}=1 \mid R=1, Y=0\right) \cdot P\left(R=1 \mid Y=0\right) \\
& + P\left(\hat{R}=1 \mid R=0, Y=0\right) \cdot P\left(R=0 \mid Y=0\right) \\
=& P\left(\hat{R}=1 \mid R=1, Y=0\right) \cdot \mathrm{FP} + P\left(\hat{R}=1 \mid R=0, Y=9\right) \cdot (1-\mathrm{FP}) \\
=& q \cdot \mathrm{FP} + 0 \cdot (1-\mathrm{FP}) \\
=& q\mathrm{FP} + (1-q) \cdot 0.
\end{aligned}
$$

Evidently, random classifier $\check{f}$ yields true-positive and false-positive rates that are linearly interpolated between $(\mathrm{TP}, \mathrm{FP})$ and $(0, 0)$.

c) Assume that classifier f achieves $(\mathrm{FP}, \mathrm{TP}) = (0.25, 0.8)$, as shown in Figure 1 below.

Copy the figure and indicate all $(\mathrm{FP}, \mathrm{TP})$-pairs that can be achieved by the random classifiers $\hat{f}$ and $\check{f}$ for $p, q \in [0, 1]$.

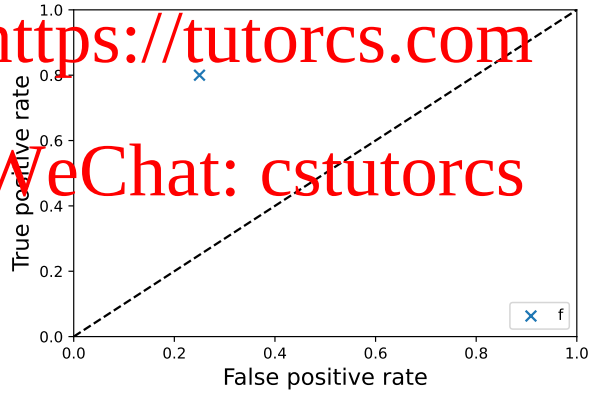

Figure 1: $(\mathrm{FP}, \mathrm{TP})$ achieved by classifier f

As mentioned before, $\hat{f}$ interpolates between $(\mathrm{TP}, \mathrm{FP})$ and $(1, 1)$, whereas $\check{f}$ interpolates between $(\mathrm{TP}, \mathrm{FP})$ and $(0, 0)$. This leads to to the following figure:

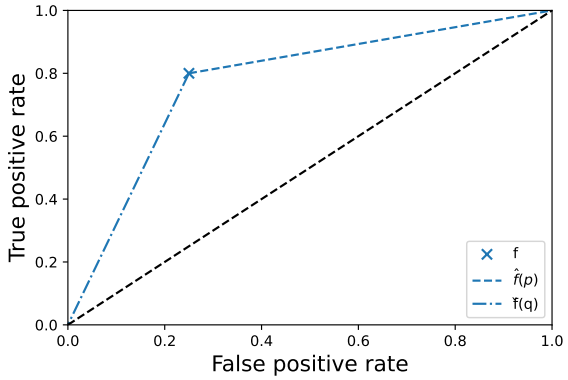

Figure 2: $(\mathrm{FP}, \mathrm{TP})$ achieved by f, $\hat{f}$ and $\check{f}$.

d) Now, we consider the true-positive and false-positive rates achieved by f on the two sub-populations $A = a$ and $A = b$ separately). As shown in Figure 3, classifier f has $(\mathrm{FP}, \mathrm{TP}) = (0.1, 0.7)$ for $A = a$ and $(\mathrm{FP}, \mathrm{TP}) = (0.4, 0.9)$ for $A = b$.

Describe how random classifiers can be used to guarantee that the *separation* criterion is fulfilled.

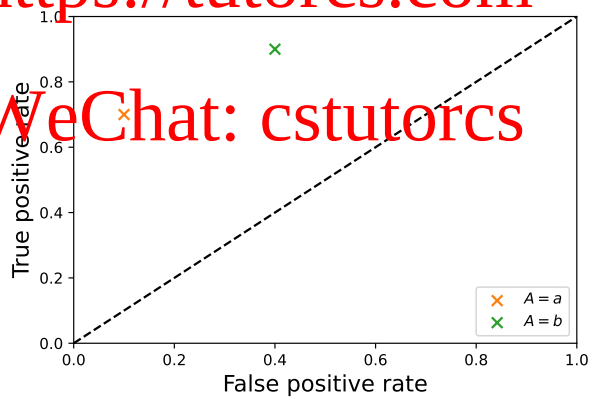

Figure 3: $(\mathrm{FP}, \mathrm{TP})$ achieved by classifier f

Separation requires that the false-positive and true-positive rates for both subpopulations are the same. To this end, we can construct an ensemble of two classifiers: A random classifier $\hat{f}_a$ for $A = a$ and a random classifier $\check{f}_b$ for $A = b$. With appropriately chosen p and q, we can cause both *ROC* curves to intersect, thus ensuring separation (see Figure 4).

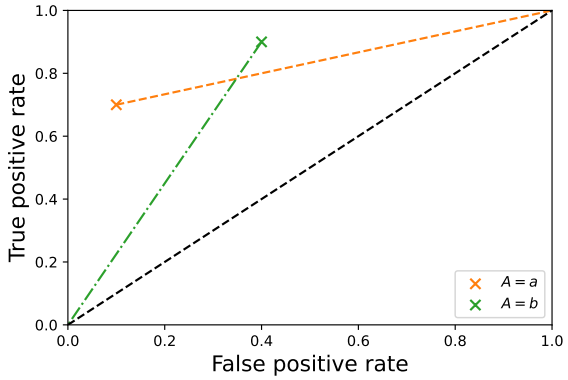

Figure 4: $(\mathrm{FP}, \mathrm{TP})$ achieved by $\hat{f}_a$ and $\check{f}_b$.