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.

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.

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!

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 $R = f(X, A)$.

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.

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 $R \perp\!\!\!\perp A$,
   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: $P(A, Y, R)$.

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?

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.

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?

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 $Y \perp\!\!\!\perp A \mid R$.

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

$$\mathrm{TP} = P(R = 1 | Y = 1)$$
$$\mathrm{FP} = P(R = 1 | Y = 0)$$

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 $Z \sim \mathrm{Bern}(p)$ 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.

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.

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

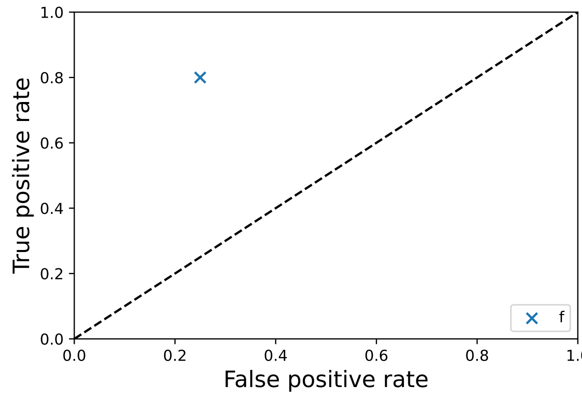

Figure 1: (FP, TP) achieved by classifier f

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 2, 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.

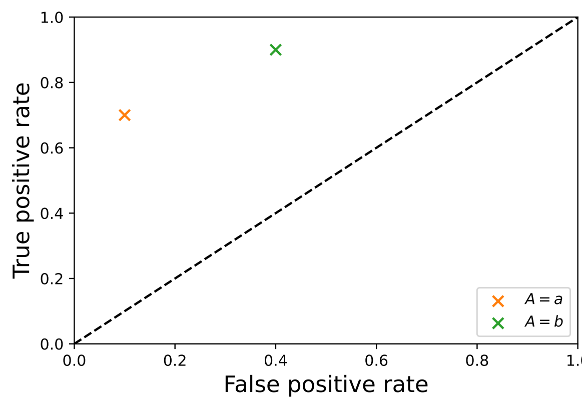

Figure 2: (FP, TP) achieved by classifier f