13 Fairness

Artificial intelligence and machine learning are being used more and more for automated decision making. This is also the case for sensitive issues such as whether a bank grants a loan, whether a citizen is enrolled in a social program, or whether a child is forcibly removed from their family.

The hope is that the use of AI will result in better decisions and outcomes because the algorithms can take into account more factors and learn from historical data as well as from its own mistakes. At the same time, there is a risk that learning from previous human decisions will reinforce existing biases and that reliance on quantitative data could introduce new forms of discrimination. As a result, there is a strong need to incorporate *fairness* into algorithms in order to protect disadvantaged groups. With the right approach, we might even be able to design AI algorithms that are optimal not only in terms of prediction and performance, but also in terms of fairness and equal opportunity.

A naïve approach to avoiding discrimination against a *protected attribute*, such as gender, race, or disability, would be to simply exclude that attribute from the model. However, this does not resolve the issue because the protected attribute may be correlated with other attributes that do enter the model: Even if the protected attribute is not directly used, it will influence the decision through its statistical dependence on other attributes. Therefore we require a better way to guarantee that a decision is not influenced by the protected attribute.

13.1 Fairness in binary decisions

Let us consider a setting where we want to make a decision $\hat{Y}$ based on attributes X while ensuring that the decision is fair with respect to a protected attribute G. To keep the discussion simple, we restrict ourselves to the binary setting, where $\hat{Y} \in \{0, 1\}$ is a binary decision, and the protected attribute $G \in \{a, b\}$ is a binary group membership variable, whereas X can be any attributes we include in making our decision $\hat{Y}$. Some fairness criteria also require us to know, at least in part, if the decision we made was correct: We will denote the correct decision by Y.[1]

[1] In some cases, it is not possible to ever know if the decision is correct. For example, in the case of a bank granting a loan, the correct decision could be defined as granting the loan to customers who repay but not to customers who default their payment. If the loan is granted, the bank will eventually learn if the decision was correct, but if the loan is not granted they have no direct way to find out if that was a correct decision.

We assume that we have access to a function

$$\hat{S} = f(X), \tag{13.1}$$

that outputs a score, $\hat{S}$, which we use to make our decision. This function could typically be based on a machine learning model, optimized on training data to predict some measure of success we would like to attain in our decision making.

While the protected group G does not enter directly into the model, it can be used either to evaluate the fairness of the decision procedure, or as a part of fitting the machine learning model to optimize fairness.

Once we have our function, we can use it to make decisions by selecting a threshold τ and deciding $\hat{Y} = 1$ when the score exceeds the threshold,

$$\hat{Y} = \begin{cases} 1 & \text{if } \hat{S} > \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{13.2}$$

In this setting we can summarize the decision procedure by a *decision matrix* for each of the protected groups,

	$G = a$		$G = b$	
	$\hat{Y} = 0$	$\hat{Y} = 1$	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	a_{00}	a_{01}	b_{00}	b_{01}
$Y = 1$	a_{10}	a_{11}	b_{10}	b_{11}

The matrix lists the number of times we made the decisions $\hat{Y} = 0$ and $\hat{Y} = 1$ when the correct decision would have been $Y = 0$ and $Y = 1$ for each of the two protected groups, $G = a$ and $G = b$.

To ensure that our decisions are fair, we need two things:

1. A method to adjust our classification procedure to optimize its fairness.

2. A clear, mathematical criterion that defines what it means to be *fair*, which we can either optimize or use to check if our decision procedure is fair.

13.2 Optimizing fairness

There are several methods that can be used to optimize a decision procedure to make it more fair. When we base our decision on a scoring function $f(X)$ learned from data, there are three fundamental approaches:

1. **Modify the training data to ensure that it is representative of a fair and unbiased decision process.** We need to make sure our data is representative for all protected groups: This might actually mean that we need an unbalanced data set, for example if a fair scoring function is more difficult to learn for a particular protected group.

2. **Modify the training method to ensure that the learned scoring function is fair and unbiased.** Several methods have been proposed which either directly optimize a mathematical fairness criterion, or ensure fairness in an implicit manner. An example of an implicit method is shown in Figure 13.1, where the idea is to train a classifier to predict the target Y based on features extracted from X in such a way that the protected group G cannot be predicted from the features. This ensures fairness in the sense that no information reflecting the protected group is used in the classification.

3. **Modify the way the scoring function is used to make decisions, so that they are fair and unbiased.** For example, we can modify or recalibrate the scores for the different protected groups to reduce bias. In the binary decision setting, as simple method is to choose different threshold values for each protected group.

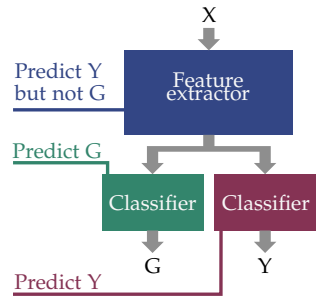

Figure 13.1: A neural network approach to create a fair classifier: The input layers (feature extractor) are optimized to learn feature that predict the target Y but *not* the protected attribute G. The two output blocks are trained to predict Y and G respectively.

13.3 Fairness criteria

13.3.1 Demographic parity

A simple fairness criterion is to require, that we decide $\hat{Y} = 1$ and $\hat{Y} = 0$ in the same fraction of cases in each protected group. Mathematically we may formulate this criterion as[2]

$$P(\hat{Y}=1|G=a) = P(\hat{Y}=1|G=b), \tag{13.3}$$

which we can read as: "The probability of deciding $\hat{Y} = 1$ for a case in protected group a is the same as the probability of deciding $\hat{Y} = 1$ for a case in protected group b." This does not take into account whether the decision was correct or not.

[2] This also implies that we decide $\hat{Y} = 0$ in the same fraction of cases in the two protected groups, which you will be asked to show as an exercise.

Based on the decision matrix, this criterion can be evaluated as

$$\frac{a_{01} + a_{11}}{n_a} = \frac{b_{01} + b_{11}}{n_b}, \tag{13.4}$$

where $n_a = a_{00} + a_{01} + a_{10} + a_{11}$ (and similarly n_b) is the total number of cases in the protected group.

While the demographic parity criterion is simple and intuitive, it has several potential drawbacks. If the target Y has a different distribution in the two protected groups, it might be too much to require that the percentages match in the decisions. Furthermore, if the scoring model is not equally good for the two groups, perhaps because of lack of training data for one of the groups, this will not be reflected in the decision.

> **Example 13.1 Demographic parity: University enrollment**
>
> A university uses a machine learning model to determine which applicants to accept as students. The system is trained on historical data to predict a student success score $\hat{S}$, and it depends on a number of student attributes X including high-school grades etc. This year, the university plans to enroll the 500 highest ranked applicants.
>
> Before making its final decision, the university decides to examine whether the decision process is fair for socially disadvantaged applicants. To do this, they divide the applicants into two groups, a and b, consisting of applicants from socially advantaged and disadvantaged groups respectively.
>
> According to the model, the university plans to accept 450 out of $n_a = 900$ applicants from group a and 50 out of

$n_b = 200$ applicants from group b.

Is this fair, according to the demographic parity criterion? No, because the fraction of students accepted from the two groups are not equal:

$$\text{Group } a: \quad \frac{450}{900} = 50\ \% \tag{13.5}$$

$$\text{Group } b: \quad \frac{50}{200} = 25\ \% \tag{13.6}$$

How could the university achieve demographic parity? Easy: Since the fraction of applicants in the two groups are known,

$$\text{Group } a: \quad \frac{900}{900+200} \approx 82\ \% \tag{13.7}$$

$$\text{Group } b: \quad \frac{200}{900+200} \approx 18\ \% \tag{13.8}$$

they could accept 82 % ($0.82 \cdot 500 = 410$ students) from group a and 18 % ($0.18 \cdot 500 = 90$ students) group b. This corresponds to choosing a different score threshold, τ, for each group.

While this would make the enrollment fair according to demographic parity with respect to the protected group, this necessarily also means that the university will accept some applicants from group b with a lower score than some rejected applicants from group a.

13.3.2 *Equalized odds*

A slightly more elaborate criterion, which avoids some of the downsides of demographic parity is the *equalized odds* criterion. This criterion requires that the probability of making a correct positive decision $\hat{Y} = 1$ among positive cases $Y = 1$ is the same in the two protected groups, and that the probability of making a correct negative decision $\hat{Y} = 0$ among negative cases $Y = 0$ is also the same in the two protected groups. Mathematically we can write this as the following two equations:

$$P(\hat{Y}=0|Y=0,G=a) = P(\hat{Y}=0|Y=0,G=b), \tag{13.9}$$

$$P(\hat{Y}=1|Y=1,G=a) = P(\hat{Y}=1|Y=1,G=b). \tag{13.10}$$

Based on the decision matrix, this can be computed as

$$\frac{a_{00}}{a_{00}+a_{01}} = \frac{b_{00}}{b_{00}+b_{01}}, \qquad \frac{a_{11}}{a_{10}+a_{11}} = \frac{b_{11}}{b_{10}+b_{11}}. \tag{13.11}$$

This criterion ensures that the decision has the same chance of being correct for both protected groups, or equivalently that the decision errors are equal. A potential drawback of this criterion is that because it requires the decision system to be tuned such that the error rates are equal across groups, it necessarily will perform worse for some group than it could.

Example 13.2 Equalized odds: Tiredness detection

A car manufacturer has created a new system that uses a camera to detect if the driver is tired, in order to take measures to avoid accidents. The system outputs a prediction $\hat{Y} = 1$ if it thinks the driver is tired. The system engineers are worried that the system might perform poorly for a protected group, so they examine the data they have gathered in their system test where they have also collected ground truth information, Y, about whether the driver was actually tired or not.

Their available data is summarized in the following table:

	$G = a$		$G = b$	
	$\hat{Y} = 0$	$\hat{Y} = 1$	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	70	15	30	10
$Y = 1$	20	130	5	30

First, let us examine the fraction of negative decisions among the negative cases, i.e., how large a fraction of non-tired drives were classified as non-tired:

$$\text{Group } a: \quad \frac{70}{90+15} \approx 74\ \% \tag{13.12}$$

$$\text{Group } b: \quad \frac{30}{30+10} = 75\ \% \tag{13.13}$$

These percentages seem fairly close (pun intended). Next, we look at the fraction of positive decisions among the positive cases, i.e., how large a fraction of tired drives were

classified as tired:

$$\text{Group } a: \quad \frac{130}{20+130} \approx 87\ \% \tag{13.14}$$

$$\text{Group } b: \quad \frac{30}{5+30} \approx 86\ \% \tag{13.15}$$

Again, the percentages match approximately. This indicates that the tiredness detector is fair according to the equalized odds criterion.[3]

[3] In this example we assess the fairness of the decision process by comparing sample proportions, but since we are actually interested in how fair the system is in the general population, we should take the uncertainty associated with the (small) sample size into account, for example by computing confidence intervals for the proportions.

13.3.3 *Equalized opportunity*

A third fairness criterion we will consider here is called the *equalized opportunity* criterion. According to this criterion, the probability that the decision is correct among cases where the decision is positive is required to be the same across protected groups. Mathematicall this can be written as

$$P(Y=1|\hat{Y}=1, G=a) = P(Y=1|\hat{Y}=1, G=b). \tag{13.16}$$

Based on the decision matrix, this criterion can be evaluated as

$$\frac{a_{11}}{a_{01}+a_{11}} = \frac{b_{11}}{b_{01}+b_{11}}. \tag{13.17}$$

The equalized opportunity criterion considers only whether there is a balanced chance of making the correct decision among cases for which the decision is positive. The cases where the decision is negative are not taken into account.

Example 13.3 Equalized opportunity: Granting a loan

A bank wants to examine if their procedure for granting or denying loans is fair according to the equalized opportunity criterion. Let $\hat{Y} = 1$ denote the decision to grant the loan, and let $Y = 1$ denote that the customer was able to repay the loan on time. Among the 10 000 most recent loans granted,

7500 were granted to customers from protected group a of which 7180 repaid on time, and

2500 were granted to customers from protected group b of which 2475 repaid.

The proportion of correct decisions (the loan was re-

paid) among positive decisions (the loan was granted) in the two groups are

$$\text{Group } a: \quad \frac{7180}{7500} \approx 96 \pm 0.5\ \% \tag{13.18}$$

$$\text{Group } b: \quad \frac{2470}{2500} \approx 99 \pm 0.4\ \% \tag{13.19}$$

Here we have included a confidence interval for the two proportions to make the statistical uncertainty clear.

Because the proportions are different for the two groups, the bank does not grant its loans in a fair manner according to the equalized opportunity criterion. To balance the opportunity, the bank could decide to grant more loans to group b, thereby increasing their risk, and/or grant fewer loans to group a (thereby reducing their risk).

Problems

1. Show that the demographic parity criterion in Eq. 13.4 implies $P(\hat{Y}=0|G=a) = P(\hat{Y}=0|G=b)$.
2. Is the tiredness detector in Example 13.2 fair according to the demographic parity criterion?

Solutions

1. We have $P(\hat{Y} = 1|G = a) = 1 - P(\hat{Y} = 0|G = a)$ since the probabilities must sum to one. Thus, we can write

$$\begin{aligned} P(\hat{Y}=1|G=a) &= P(\hat{Y}=1|G=b) && \Leftrightarrow \\ 1 - P(\hat{Y}=1|G=a) &= 1 - P(\hat{Y}=1|G=b) && \Leftrightarrow \\ P(\hat{Y}=0|G=a) &= P(\hat{Y}=0|G=b). \end{aligned}$$

2. Let us compute the fraction of drivers detected as tired in the two groups:

$$\text{Group } a: \quad \frac{130+15}{70+15+20+130} \approx 62\ \% \tag{13.20}$$

$$\text{Group } b: \quad \frac{10+30}{30+10+5+30} \approx 53\ \% \tag{13.21}$$

   This indicates that the tiredness detector is not fair according to demographic parity; although, due to the small sample size, the difference in percentages is not statistically significant. Anyway, demographic parity might not be desirable in this case, since drivers from group a appear to be more tired overall, which the decision system reflects.