A brief introduction to fairness in supervised classification

Marco Saerens (UCLouvain)

Table of contents
Table of contents

- General introduction
- Sources of unfairness
- Fairness in supervised classification
  - Operational notions of fairness in predictive modeling
    - Group fairness
    - Individual fairness
    - Causality-based fairness
  - Mitigating bias in supervised classification
    - Pre-processing
    - Post-processing
    - In-training processing

Contents

These slides are largely inspired by the books, papers, and surveys cited at the end of the presentation
Fairness, interpretability and transparency in decisions are nowadays becoming a key challenge in society.

Many decisions are now supported by statistical, algorithmic, or artificial intelligence techniques;
  – Automated or partly automated decision-making based on data

Some automated decisions might significantly impact people’s life.
General introduction

- This includes
  - Selecting persons for a job
  - Deciding if a person will be granted a loan
  - Predicting future criminality
  - College admission
  - Selecting candidates for flat renting
  - Etc...

Indeed, softwares are currently used for taking decisions in
  - Healthcare
  - Banking
  - Criminal justice
  - Marketing
  - Education
  - Human resources
  - Finance
  - Etc…
General introduction

- Fairness has been widely studied for decades in almost all fields of social sciences: political sciences, economics, justice, philosophy, management, etc
  - The concept is rather general but it also depends on the context

- In short, fairness is about avoiding unjustified disparities of treatment among individuals
  - And, of course, “unjustified” depends on the situation and the overall goals within the society/institution
  - This is mainly studied in social sciences

General introduction

- In that context, general ethical questions that have to be answered are (Nielsen, 2020)
  - Rules of allocation: Who gets what and in which quantity?
  - Rules of decision: How do we decide who gets what?
  - Rules of political authority: Who decides who decides?

- Although essential, we will not discuss these fundamental societal/philosophical aspects
General introduction

- By now, decisions are also taken by computer programs
  - Which raises new questions
  - How can we increase or enforce fairness in statistical, ML, and AI systems relying on data?

- This is part of the sub-fields of “responsible AI”, “trustworthy AI”, “ethical AI”, “algorithmic fairness”, etc

Sources of unfairness
Sources of unfairness

- Two important problems arise (Castelnovo et al., 2022)
  - How can we measure and assess fairness of the outputs of predictive models, or more generally, algorithms?
  - How can we mitigate unfairness when present?

- Unfairness is usually related to some groups of persons whose treatment is differentiated (discrimination)
  - The group is then called a "protected group"
  - The variable identifying individuals from this protected group is called a "sensitive variable"

Sources of unfairness

- Some examples, depending on the context, are
  - Ethnicity
  - Gender
  - Age
  - Religious orientation
  - Political orientation
  - Sexual orientation
  - Income/social rank
  - Health-related attributes (e.g., handicap)
  - Etc…
Sources of unfairness

- Equality between groups can either be
  - Explicitly enforced *a priori* as a general principle through regulation rules (e.g., independence w.r.t. ethnicity)
- Or inequalities can be hidden in the model
  - There are systematic measurement/recording errors in the data
  - The training sample is not representative from the real population (automatic tap discriminating afro-americans)
  - There are historical/societal bias present in the data (woman discriminated for some job positions)
  - Some minority groups are not taken into account because they are not sufficiently represented (Bayes decision rule in supervised classification)
  - The classification model itself introduces or augments discrimination
  - Etc...

In that case, they could/should be identified and possibly mitigated

- Removing the sensitive variable from the model is sometimes not enough (e.g., using geographic location is often correlated with income, poverty or ethnicity)
- Indeed, other explanatory variables/features could be correlated with the sensitive variable = proxy
- Identifying *causal relationships* would be very interesting in this context
Fairness in supervised classification

Let us now consider \textit{supervised classification} problems

with only two classes for simplicity (binary)

- Let $X$ be a $n \times p$ data matrix containing $n$ measurements of $p-1$ (random) explanatory variables or features $X_i$ and, as usual, one column of 1s (bias term, depending on the problem)
- For simplicity the \textit{sensitive variable} $Z$ is unique and discrete with two groups (binary variable). Group 1 is the \textit{protected group}
- The \textit{target}, dependent, discrete variable is $Y$ and is also \textit{binary}
- The output (score or rating) of the classification model is $\hat{Y}$ and is assumed to be \textit{calibrated}
  - This means that it estimates the \textit{a posteriori} probabilities of belonging to class 1 (the class of interest) – predicted probabilities
- The discretized binary output of the classification model is $\hat{Y}_d$ (after the \textit{decision}, usually $\geq 0.5$)
Fairness in supervised classification

- Let us first look at some operational notions/ measures of fairness used in predictive modeling
  - Group fairness
  - Individual fairness
  - Causality-based fairness

- Let’s also assume a situation where
  - Observing the binary target class \( Y = 1 \) means that the individual is “selected” – the “successful”, positive, outcome
  - The binary sensitive variable \( Z \) could be the gender
  - The binary decision of the classification model is \( Y_d \)

Operational measures: group fairness

- A first, rough, indicator that can be used to explore the difference of treatment between the two groups (protected and not protected) based on the data set is

- The average total variation (Zhang et al., 2018)
  \[
  \mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]
  \]

  - It is closely related to demographic parity (see later)
Operational measures: group fairness

- Another related measure computes the ratio instead,
\[
\frac{E[Y|Z = 1]}{E[Y|Z = 0]} = \frac{P(Y = 1|Z = 1)}{P(Y = 1|Z = 0)}
\]

- Intuitively, it is the ratio of the probability of being selected \((Y = 1)\) for the two groups of interest
  - If there is equality of treatment in average, the ratio is 1
  - It therefore provides an indication about the difference of treatment (here, “being selected”) within the two protected groups

Operational measures: group fairness

- Note that the difference of treatment could be due to a good reason – e.g. some legitimate variables that are correlated with the sensitive variable

- As an alternative, the odds ratio (popular in applied statistics) could also be measured,
\[
\left(\frac{P(Y = 1|Z = 1)}{P(Y = 0|Z = 1)}\right) / \left(\frac{P(Y = 1|Z = 0)}{P(Y = 0|Z = 0)}\right)
\]
Operational measures: group fairness

- Other measures involving the prediction model can usually be computed from the confusion matrix, based on \( n \) observed data.
  - Involving the actual observed binary class \( Y \) and the predicted binary class \( \hat{Y}_d \).

<table>
<thead>
<tr>
<th>Predicted class: ( \hat{Y}_d = 1 )</th>
<th>Predicted class: ( \hat{Y}_d = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual class: ( Y = 1 )</td>
<td>True Positives (TP)</td>
</tr>
<tr>
<td>Actual class: ( Y = 0 )</td>
<td>False Positives (FP)</td>
</tr>
</tbody>
</table>

- This quantity can be computed by

\[
\frac{TP(Z = 1) + FP(Z = 1)}{N(Z = 1)} = \frac{TP(Z = 0) + FP(Z = 0)}{N(Z = 0)}
\]

where \( N(Z = 0) \) means the number of data with \( Z = 0 \).

- This measure can therefore be used in order to assess the “fairness” of the predictions of a model.
Operational measures: group fairness

- The same measure can also be adapted to the predicted probability, or score/rating,
\[ \mathbb{E}[\hat{Y}|Z = 1] = \mathbb{E}[\hat{Y}|Z = 0] \]
  - Meaning that the expected a posteriori probability score of being selected is the same for male and female

- Moreover, the strict equality could be relaxed:
  - the difference between male and female should not exceed a given threshold
\[ |P(\hat{Y}_d = 1| Z = 1) - P(\hat{Y}_d = 1| Z = 0)| \leq \epsilon \]
\[ |\mathbb{E}[\hat{Y}|Z = 1] - \mathbb{E}[\hat{Y}|Z = 0]| \leq \epsilon \]

Operational measures: group fairness

- The previous measure can be extended to equal opportunity,
\[ P(\hat{Y}_d = 1| Y = 1, Z = 1) = P(\hat{Y}_d = 1| Y = 1, Z = 0) \]
  - Here, the chance of being selected by the model must be the same for male and female having the right profile (considering that \( Y \) reflects company’s (previous) judgment)

- The requirement can also be measured for persons who do not have the right profile,
\[ P(\hat{Y}_d = 1| Y = 0, Z = 1) = P(\hat{Y}_d = 1| Y = 0, Z = 0) \]
  - When both requirements are present, it is then called equalized odds
Operational measures: group fairness

- The counterpart for predicted scores/ratings is

\[ E[\hat{Y}|Y = 1, Z = 1] = E[\hat{Y}|Y = 1, Z = 0] \]

and

\[ E[\hat{Y}|Y = 0, Z = 1] = E[\hat{Y}|Y = 0, Z = 0] \]

- Notice that there are some incompatibility statements:
  - some criteria can be shown to be incompatible in some situations
  - See, e.g., Barocas et al., 2021

---

Operational measures: group fairness

- The notion of well-calibration fairness says that

\[ P(Y = 1|\hat{Y} = s, Z = 1) = P(Y = 1|\hat{Y} = s, Z = 0) = s \]

for all prediction levels \( s \) (Kleinberg et al., 2017).

- It means that all scores are well-calibrated and are the same for male and female
  - This is a strict notion of fairness

- Almost all these quantities can be estimated from the confusion matrix
  - Many additional measures were also introduced
Operational measures: individual fairness

- Most group fairness measures only require to satisfy conditions on average

- Individual fairness is based on a different principle, which is:
  - Similar individuals should receive similar treatments

- One example of such a condition is as follows
  - Assume that $x_i$ is the observed feature vector on individual $i$ (row $i$ of the data matrix viewed as a column vector)
  - The observed predicted probability and class provided by the model for $i$ are respectively $\hat{y}_i$ and $\hat{y}_d$
  - Same for individual $j$

Then the requirement (Dwork et al., 2012) could be formulated as

$$\text{dist}_Y(\hat{y}_i, \hat{y}_j) \leq \epsilon \times \text{dist}_X(x_i, x_j)$$

- Meaning that predictions are constrained to be more and more similar when the profiles of the individuals are more and more similar
- Individuals with similar profiles should receive similar treatments

- However, the practical choice of the distance and the scaling factor are not trivial
Operational measures: computational fairness

Still another proposition is to assess the impact of the explanatory variables on the protected variable (Feldman et al, 2015)

- We predict the binary protected variable $Z$ from the feature vector,
  $$\hat{z}_i = g(x_i)$$

- And then compute the average error rate per class (balanced error rate)
  $$\frac{P(\hat{Z}(X) = 1|Z = 0) + P(\hat{Z}(X) = 0|Z = 1)}{2}$$

- It indicates to which extend the set of explanatory variables is an important proxy of the sensitive variable
- It quantifies the amount of "information" about the sensitive variable present in the features
- And can be used in order to remove the variables that are highly associated with the sensitive variable, in a stepwise way

Operational measures: causality

Still another point of view is to work with causality

- The question then becomes (Khademi et al., 2019)

“Does the protected variable have a causal effect on the decision?”

- In general, answering to this question is not easy for observational data
- Moreover, it is in general assumed that all the variables having an impact on the dependent variable are known and are measured
- One attempt to estimate such causal effects is developed in (Khademi et al., 2019)
Operational measures: causality

- They introduce the “fair on average causal effect” (FACE)

- A decision is said to be $\epsilon$-fair on average on a population if (note that the measure is reinterpreted here because the notations were not completely understood)

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \left| \hat{Y}_i(z_i) - \hat{Y}_i(1-z_i) \right| \right] \leq \epsilon$$

- where $\hat{Y}_i(z_i)$ is the prediction score for individual $i$ having observed value $z_i$
- and $\hat{Y}_i(1-z_i)$ is the potential prediction score of $i$, had the value of its protected attribute been $(1-z_i)$ (that is, the opposite of what is observed)
- for an identical profile $x_i$

Operational measures: causality

- This quantity is estimated from the data by using statistical data matching techniques
  - Based on the profiles of the individuals ($x_i$)

- Additional work on fairness based on causality analysis is reviewed in (Makhlouf et al., 2021)
Mitigating bias in supervised classification

The idea here is to develop practical algorithms for improving the fairness of the outputs of predictive models. They can be classified in three different categories:

- Pre-processing techniques
- Post-processing techniques
- In-training techniques

They try to enforce the different notions of fairness based on the empirical data:

- This usually results in a trade-off between fairness and accuracy
- Let's investigate all of these
Mitigating bias: pre-processing

- Many different pre-processing techniques have been developed
- In general, they try to improve the “fairness” of the data matrix
- We will focus on a standard technique in applied statistics, valid for numerical features only
  - Partial linear regression
  - which is popular in causal linear regression modeling
- Usually, we first remove the sensitive variables from the data matrix – thus not used in the model

The main idea is to remove the linear dependence/correlation of each of the feature in the sensitive variable (Hayes, 2013; Pope, 2011)
- To do this, we first regress each feature $X_k$ in terms of the sensitive variables
- For instance, for individual/observation $i$,

$$\tilde{x}_{ik} = w_1 z_{i1} + w_2 z_{i2} + \cdots$$

and $e_{ik} = x_{ik} - \tilde{x}_{ik}$ is the residual computed on individual $i$ for feature $k$, so that

$$x_{ik} = w_1 z_{i1} + w_2 z_{i2} + \cdots + e_{ik}$$
Mitigating bias: pre-processing

- Then, the resulting residuals $e_{ik}$ are linearly uncorrelated with the sensitive variables
  - They are therefore a sort of “purified”, or “neutralized”, transformation of the original observations $x_{ik}$ of feature $X_i$
  - Which has been cleaned up from its linear association with the sensitive variables
  - This removes the part of linear variability of $X_i$ due to the sensitive variable

- The residuals are then used in the data matrix $X$, instead of the original variables
  - The features are thus cleaned up
  - However, the objective is often to guarantee independence w.r.t the sensitive variables
  - Which is much more difficult to achieve

Mitigating bias: post-processing

- **Post-processing**: here, we want to adjust the predicted scores in order to obtain fair transformed scores
  - For instance in terms of statistical parity
  - Or any other measure presented before

- This can be seen as a way to apply post-hoc “positive discrimination”
Mitigating bias: post-processing

- One of the main advantages of post-processing is the fact that it can easily be applied to a wide variety of classification models
  - Logistic regression, random forests, multilayer neural networks, etc
  - However, it is also sub-optimal because it is decoupled from the optimization problem used for fitting the model

- Post-processing could be done in the same way as in pre-processing
  - Assume a model has been trained, as for instance a logistic regression
  - and provides predicted probabilities $\hat{y}_i$
  - which are between 0 and 1

Mitigating bias: post-processing

- However, it is more tricky because the outputs of the model must be calibrated (a posteriori probabilities)
  - We could regress $\log \frac{\hat{y}_i}{1-\hat{y}_i}$ (instead of directly $\hat{y}_i$) in terms of the protected $Z_i$
  - and then compute the residuals as before
  - The linear part of the association (before the sigmoid) of the predicted values is then cleaned up
Mitigating bias: post-processing

- Here are other ways to enforce fairness by post-processing (Hardt et al., 2016, Kleinberg et al., 2018):
  - We only provide the intuition behind these techniques.
  - Assume we trained a model providing the predicted scores (a posteriori probabilities) for the individuals, \( \hat{y}_i = g(x_i) \).
  - Recall the definition of statistical parity,
    \[
    |P(\tilde{Y}_d = 1|Z = 1) - P(\tilde{Y}_d = 1|Z = 0)| \leq \epsilon
    \]
  - We want to select individuals based on \( \hat{Y} \) while guaranteeing statistical parity.
  - The selection decision will be based on a new variable \( \tilde{Y}_d \).
  - Which will be deduced from \( \hat{Y} \).

Mitigating bias: post-processing

- Kleinberg et al. showed that, under some reasonable assumptions, the optimal way of choosing the successful individuals is
  - by simply using two thresholds \( \theta_0 \) and \( \theta_1 \) on the \( \hat{y}_i \) for the two different groups \( Z = 0 \) and \( Z = 1 \).

- More precisely, if you decide to select \( n_1 \) individuals from group 1 and \( n_0 \) individuals from group 0 in order to satisfy statistical parity,
  - The best choice according to the assumptions is to select:
    - The \( n_1 \) best individuals in group 1 with the highest scores \( \hat{y}_i \).
    - The \( n_0 \) best individuals in group 0 with the highest scores \( \hat{y}_j \).
  - However, in practice, \( n_0 \) and \( n_1 \) are usually unknown a priori.
Mitigating bias: post-processing

- Still another technique on which we are currently working (De Schaetzen et al, 2021; Vancompernolle Vromman 2023), inspired by (Hardt, 2016; Zafar, 2017), is as follows
  
  - Again, we train a model providing the predicted scores (a posteriori probabilities), \( \hat{y}_i = g(x_i) \)
  
  - The idea is to compute new scores \( \tilde{y}_i \) satisfying the fairness constraints, for instance simple statistical parity,
    \[
    \left| \mathbb{E}[\tilde{Y} | Z = 1] - \mathbb{E}[\tilde{Y} | Z = 0] \right| \leq \epsilon
    \]
  
  - while remaining closest to the original predicted scores

This leads to the following optimization problem

\[
\begin{align*}
\text{minimize} & \quad \| \tilde{y} - \hat{y} \|^2 \\
\text{subject to} & \quad \frac{1}{n} |z^T \tilde{H} \tilde{y}| \leq \epsilon \\
& \quad \tilde{y} \geq 0 \\
& \quad \tilde{y} \leq e
\end{align*}
\]

- which is a simple least squares problem with linear inequality constraints
- The matrix \( \tilde{H} \) is the centering matrix and \( e \) is a column vector of 1s
- The result holds because \( z^T \tilde{H} \tilde{y} / n \) represents empirical statistical parity, up to a scaling factor
- It also represents the empirical covariance between \( z \) and \( \tilde{y} \) (Vancompernolle Vromman 2023)
Mitigating bias: post-processing

- Other techniques based on the swapping of classes are also very popular (Calders et al. 2010)
- We usually observe a trade-off between accuracy and the level of achieved fairness

Mitigating bias: in-training

- **In-training**: fairness components are added during the model fitting phase, within the optimization problem
  - As a regularisation term added to the objective function
  - As a constraint added to the optimization problem

- We will examine some examples of both
Mitigating bias: in-training

- An early proposition is to regularize the objective function by mutual information (Kamishima et al., 2012)
  - In the context of a logistic regression model
    \[
    \mathbb{E}_X \left[ \sum_{y,z \in \{0,1\}} P(\hat{Y}_d = y, Z = z | X) \log \left( \frac{P(\hat{Y}_d = y, Z = z | X)}{P(\hat{Y}_d = y | X)P(Z = z | X)} \right) \right]
    \]
    where the a posteriori probability is approximated by the model,
    \[P(\hat{Y}_d = 1 | X = x_i, Z = z_i) \approx g(x_i, z_i)\]
    - The smaller the regularisation term, the larger the independence between the prediction of the model and the sensitive variable

Mitigating bias: in-training

- Other work uses constraints in the optimization problem (Zafar et al., 2017)
  - Also in the context of a logistic regression model

- They constrain the covariance between (1) the sensitive variable and (2) the linear predicted score provided by the logistic regression to be low
  - Thus, below a certain threshold

- The linear predicted score is \(w^T x = x^T w\) and is proportional to the distance of \(x\) to the separating hyperplane
Mitigating bias: in-training

- The constrained optimization problem is

\[ \begin{align*}
\text{minimize} & \quad -\sum_{i=1}^{n} (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)) \\
\text{subject to} & \quad \frac{1}{n} |z^T H X w| \leq \epsilon
\end{align*} \]

- It minimizes minus the log-likelihood subject to the constraints
- Recall that the outputs (predicted scores) of the logistic regression model are
\[ \hat{y}_i = \sigma(w^T x_i) \] with \( \sigma(\cdot) \) being the sigmoid function

- A drawback of this method is that it does not impose fairness on the predicted scores
- But the model can be extended to SVMs

Mitigating bias: in-training

- Other examples in the same spirit is
  - Donini et al., 2018
  - Our proposal is to use the maximum entropy formulation (De Schaetzen et al., 2021; Vancompemolle Vromman 2023) of a logistic regression
References

- Vancompernolle Vromman F. et al. (2023) “Maximum entropy logistic regression for demographic parity in supervised classification. Submitted for publication.”
References