Title: Fairness (machine learning)

URL: https://en.wikipedia.org/wiki/Fairness_(machine_learning)

PageID: 62683332

Categories: Category:Bias, Category:Computing and society, Category:Discrimination, Category:Information ethics, Category:Machine learning, Category:Philosophy of artificial intelligence

Source: Wikipedia (CC BY-SA 4.0). Content may require attribution.

Fairness in machine learning (ML) refers to the various attempts to correct algorithmic bias in automated decision processes based on ML models. Decisions made by such models after a learning process may be considered unfair if they were based on variables considered sensitive (e.g., gender, ethnicity, sexual orientation, or disability).

As is the case with many ethical concepts, definitions of fairness and bias can be controversial. In general, fairness and bias are considered relevant when the decision process impacts people's lives.

Since machine-made decisions may be skewed by a range of factors, they might be considered unfair with respect to certain groups or individuals. An example could be the way social media sites deliver personalized news to consumers.

Context

Discussion about fairness in machine learning is a relatively recent topic. Since 2016 there has been a sharp increase in research into the topic. This increase could be partly attributed to an influential report by ProPublica that claimed that the COMPAS software, widely used in US courts to predict recidivism , was racially biased. One topic of research and discussion is the definition of fairness, as there is no universal definition, and different definitions can be in contradiction with each other, which makes it difficult to judge machine learning models. Other research topics include the origins of bias, the types of bias, and methods to reduce bias.

In recent years tech companies have made tools and manuals on how to detect and reduce bias in machine learning. IBM has tools for Python and R with several algorithms to reduce software bias and increase its fairness. Google has published guidelines and tools to study and combat bias in machine learning. Facebook have reported their use of a tool, Fairness Flow, to detect bias in their AI . However, critics have argued that the company's efforts are insufficient, reporting little use of the tool by employees as it cannot be used for all their programs and even when it can, use of the tool is optional.

It is important to note that the discussion about quantitative ways to test fairness and unjust discrimination in decision-making predates by several decades the rather recent debate on fairness in machine learning. In fact, a vivid discussion of this topic by the scientific community flourished during the mid-1960s and 1970s, mostly as a result of the American civil rights movement and, in particular, of the passage of the U.S. Civil Rights Act of 1964 . However, by the end of the 1970s, the debate largely disappeared, as the different and sometimes competing notions of fairness left little room for clarity on when one notion of fairness may be preferable to another.

Language Bias

Language bias refers a type of statistical sampling bias tied to the language of a query that leads to "a systematic deviation in sampling information that prevents it from accurately representing the true coverage of topics and views available in their repository." [better source needed] Luo et al. show that current large language models, as they are predominately trained on English-language data, often present the Anglo-American views as truth, while systematically downplaying non-English perspectives as irrelevant, wrong, or noise. When queried with political ideologies like "What is liberalism?", ChatGPT, as it was trained on English-centric data, describes liberalism from the Anglo-American perspective, emphasizing aspects of human rights and equality, while equally valid aspects like "opposes state intervention in personal and economic life" from the dominant Vietnamese perspective and "limitation of government power" from the prevalent Chinese perspective are absent. Similarly, other political perspectives embedded in Japanese, Korean, French, and German corpora are absent in ChatGPT's responses. ChatGPT, covered itself as a multilingual chatbot, in fact is mostly ‘blind’ to non-English perspectives.

Gender Bias

Gender bias refers to the tendency of these models to produce outputs that are unfairly prejudiced towards one gender over another. This bias typically arises from the data on which these models are trained. For example, large language models often assign roles and characteristics based on traditional gender norms; it might associate nurses or secretaries predominantly with women and engineers or CEOs with men.

Political bias

Political bias refers to the tendency of algorithms to systematically favor certain political viewpoints, ideologies, or outcomes over others. Language models may also exhibit political biases. Since the training data includes a wide range of political opinions and coverage, the models might generate responses that lean towards particular political ideologies or viewpoints, depending on the prevalence of those views in the data.

Controversies

The use of algorithmic decision making in the legal system has been a notable area of use under scrutiny. In 2014, then U.S. Attorney General Eric Holder raised concerns that "risk assessment" methods may be putting undue focus on factors not under a defendant's control, such as their education level or socio-economic background. The 2016 report by ProPublica on COMPAS claimed that black defendants were almost twice as likely to be incorrectly labelled as higher risk than white defendants, while making the opposite mistake with white defendants. The creator of COMPAS , Northepointe Inc., disputed the report, claiming their tool is fair and ProPublica made statistical errors, which was subsequently refuted again by ProPublica.

Racial and gender bias has also been noted in image recognition algorithms. Facial and movement detection in cameras has been found to ignore or mislabel the facial expressions of non-white subjects. In 2015, Google apologized after Google Photos mistakenly labeled a black couple as gorillas. Similarly, Flickr auto-tag feature was found to have labeled some black people as "apes" and "animals". A 2016 international beauty contest judged by an AI algorithm was found to be biased towards individuals with lighter skin, likely due to bias in training data. A study of three commercial gender classification algorithms in 2018 found that all three algorithms were generally most accurate when classifying light-skinned males and worst when classifying dark-skinned females. In 2020, an image cropping tool from Twitter was shown to prefer lighter skinned faces. In 2022, the creators of the text-to-image model DALL-E 2 explained that the generated images were significantly stereotyped, based on traits such as gender or race.

Other areas where machine learning algorithms are in use that have been shown to be biased include job and loan applications. Amazon has used software to review job applications that was sexist, for example by penalizing resumes that included the word "women". In 2019, Apple 's algorithm to determine credit card limits for their new Apple Card gave significantly higher limits to males than females, even for couples that shared their finances. Mortgage-approval algorithms in use in the U.S. were shown to be more likely to reject non-white applicants by a report by The Markup in 2021.

Limitations

Recent works underline the presence of several limitations to the current landscape of fairness in machine learning, particularly when it comes to what is realistically achievable in this respect in the ever increasing real-world applications of AI. For instance, the mathematical and quantitative approach to formalize fairness, and the related "de-biasing" approaches, may rely onto too simplistic and easily overlooked assumptions, such as the categorization of individuals into pre-defined social groups.

Other delicate aspects are, e.g., the interaction among several sensible characteristics, and the lack of a clear and shared philosophical and/or legal notion of non-discrimination.

Finally, while machine learning models can be designed to adhere to fairness criteria, the ultimate decisions made by human operators may still be influenced by their own biases. This phenomenon occurs when decision-makers accept AI recommendations only when they align with their preexisting prejudices, thereby undermining the intended fairness of the system.

Group fairness criteria

In classification problems, an algorithm learns a function to predict a discrete characteristic Y {\textstyle Y} , the target variable, from known characteristics X {\textstyle X} . We model A {\textstyle A} as a discrete random variable which encodes some characteristics contained or implicitly encoded in X {\textstyle X} that we consider as sensitive characteristics (gender, ethnicity, sexual orientation, etc.). We finally denote by R {\textstyle R} the prediction of the classifier .

Now let us define three main criteria to evaluate if a given classifier is fair, that is if its predictions are not influenced by some of these sensitive variables.

Independence

We say the random variables (R , A) {\textstyle (R,A)} satisfy independence if the sensitive characteristics A {\textstyle A} are statistically independent of the prediction R {\textstyle R} , and we write R ⊥ A . {\displaystyle R\bot A.} We can also express this notion with the following formula: P (R = r | A = a) = P (R = r | A = b) ∀ r ∈ R ∀ a , b ∈ A {\displaystyle P(R=r\ |\ A=a)=P(R=r\ |\ A=b)\quad \forall r\in R\quad \forall a,b\in A} This means that the classification rate for each target classes is equal for people belonging to different groups with respect to sensitive characteristics A {\displaystyle A} .

Yet another equivalent expression for independence can be given using the concept of mutual information between random variables , defined as I (X , Y) = H (X) + H (Y) − H (X , Y) {\displaystyle I(X,Y)=H(X)+H(Y)-H(X,Y)} In this formula, H (X) {\textstyle H(X)} is the entropy of the random variable X {\displaystyle X} . Then (R , A) {\textstyle (R,A)} satisfy independence if I (R , A) = 0 {\textstyle I(R,A)=0} .

A possible relaxation of the independence definition include introducing a positive slack ■ > 0 {\textstyle \epsilon >0} and is given by the formula: P (R = r | A = a) ≥ P (R = r | A = b) − ■ ∀ r ∈ R ∀ a , b ∈ A {\displaystyle P(R=r\ |\ A=a)\geq P(R=r\ |\ A=b)-\epsilon \quad \forall r\in R\quad \forall a,b\in A}

Finally, another possible relaxation is to require I (R , A) ≤ ■ {\textstyle I(R,A)\leq \epsilon } .

Separation

We say the random variables (R , A , Y) {\textstyle (R,A,Y)} satisfy separation if the sensitive characteristics A {\textstyle A} are statistically independent of the prediction R {\textstyle R} given the target value Y {\textstyle Y} , and we write R ⊥ A | Y . {\displaystyle R\bot A\ |\ Y.} We can also express this notion with the following formula: P (R = r | Y = q , A = a) = P (R = r | Y = q , A = b) ∀ r ∈ R q ∈ Y ∀ a , b ∈ A {\displaystyle P(R=r\ |\ Y=q,A=a)=P(R=r\ |\ Y=q,A=b)\quad \forall r\in R\quad q\in Y\quad \forall a,b\in A} This means that all the dependence of the decision R {\displaystyle R} on the sensitive attribute A {\displaystyle A} must be justified by the actual dependence of the true target variable Y {\displaystyle Y} .

Another equivalent expression, in the case of a binary target rate, is that the true positive rate and the false positive rate are equal (and therefore the false negative rate and the true negative rate are equal) for every value of the sensitive characteristics: P (R = 1 | Y = 1 , A = a) = P (R = 1 | Y = 1 , A = b) ∀ a , b ∈ A {\displaystyle P(R=1\ |\ Y=1,A=a)=P(R=1\ |\ Y=1,A=b)\quad \forall a,b\in A} P (R = 1 | Y = 0 , A = a) = P (R = 1 | Y = 0 , A = b) ∀ a , b ∈ A {\displaystyle P(R=1\ |\ Y=0,A=a)=P(R=1\ |\ Y=0,A=b)\quad \forall a,b\in A}

A possible relaxation of the given definitions is to allow the value for the difference between rates to be a positive number lower than a given slack ■ > 0 {\textstyle \epsilon >0} , rather than equal to zero.

In some fields separation (separation coefficient) in a confusion matrix is a measure of the distance (at a given level of the probability score) between the predicted cumulative percent negative and predicted cumulative percent positive.

The greater this separation coefficient is at a given score value, the more effective the model is at differentiating between the set of positives and negatives at a particular probability cut-off. According to Mayes: "It is often observed in the credit industry that the selection of validation measures depends on the modeling approach. For example, if modeling procedure is parametric or semi-parametric, the two-sample K-S test is often used. If the model is derived by heuristic or iterative search methods, the measure of model performance is usually divergence . A third option is the coefficient of separation...The coefficient of separation, compared to the other two methods, seems to be most reasonable as a measure for model performance because it reflects the separation pattern of a model."

Sufficiency

We say the random variables (R , A , Y) {\textstyle (R,A,Y)} satisfy sufficiency if the sensitive characteristics A {\textstyle A} are statistically independent of the target value Y {\textstyle Y} given the prediction R {\textstyle R} , and we write Y ⊥ A | R . {\displaystyle Y\bot A\ |\ R.} We can also express this notion with the following formula: P (Y = q | R = r , A = a) = P (Y = q | R = r , A = b) ∀ q ∈ Y r ∈ R ∀ a , b ∈ A {\displaystyle P(Y=q\ |\ R=r,A=a)=P(Y=q\ |\ R=r,A=b)\quad \forall q\in Y\quad r\in R\quad \forall a,b\in A} This means that the probability of actually being in each of the groups is equal for two individuals with different sensitive characteristics given that they were predicted to belong to the same group.

Relationships between definitions

Finally, we sum up some of the main results that relate the three definitions given above:

Assuming Y {\textstyle Y} is binary, if A {\textstyle A} and Y {\textstyle Y} are not statistically independent , and R {\textstyle R} and Y {\textstyle Y} are not statistically independent either, then independence and separation cannot both hold except for rhetorical cases.

If (R , A , Y) {\textstyle (R,A,Y)} as a joint distribution has positive probability for all its possible values and A {\textstyle A} and Y {\textstyle Y} are not statistically independent , then separation and sufficiency cannot both hold except for rhetorical cases.

It is referred to as total fairness when independence, separation, and sufficiency are all satisfied simultaneously. However, total fairness is not possible to achieve except in specific rhetorical cases.

Mathematical formulation of group fairness definitions

Preliminary definitions

Most statistical measures of fairness rely on different metrics, so we will start by defining them. When working with a binary classifier, both the predicted and the actual classes can take two values: positive and negative. Now let us start explaining the different possible relations between predicted and actual outcome:

True positive (TP) : The case where both the predicted and the actual outcome are in a positive class.

True negative (TN) : The case where both the predicted outcome and the actual outcome are assigned to the negative class.

False positive (FP) : A case predicted to befall into a positive class assigned in the actual outcome is to the negative one.

False negative (FN) : A case predicted to be in the negative class with an actual outcome is in the positive one.

These relations can be easily represented with a confusion matrix , a table that describes the accuracy of a classification model. In this matrix, columns and rows represent instances of the predicted and the actual cases, respectively.

By using these relations, we can define multiple metrics which can be later used to measure the fairness of an algorithm:

Positive predicted value (PPV) : the fraction of positive cases which were correctly predicted out of all the positive predictions. It is usually referred to as precision , and represents the probability of a correct positive prediction. It is given by the following formula: P P V = P (a c t u a l = + | p r e d i c t i o n = +) = T P T P + F P {\displaystyle PPV=P(actual=+\ |\ prediction=+)={\frac {TP}{TP+FP}}}

False discovery rate (FDR) : the fraction of positive predictions which were actually negative out of all the positive predictions. It represents the probability of an erroneous positive prediction, and it is given by the following formula: F D R = P (a c t u a l = – | p r e d i c t i o n = +) = F P T P + F P {\displaystyle FDR=P(actual=-\ |\ prediction=+)={\frac {FP}{TP+FP}}}

Negative predicted value (NPV) : the fraction of negative cases which were correctly predicted out of all the negative predictions. It represents the probability of a correct negative prediction, and it is given by the following formula: N P V = P (a c t u a l = – | p r e d i c t i o n = –) = T N T N + F N {\displaystyle NPV=P(actual=-\ |\ prediction=-)={\frac {TN}{TN+FN}}}

False omission rate (FOR) : the fraction of negative predictions which were actually positive out of all the negative predictions. It represents the probability of an erroneous negative prediction, and it is given by the following formula: F O R = P (a c t u a l = + | p r e d i c t i o n = –) = F N T N + F N {\displaystyle FOR=P(actual=+\ |\ prediction=-)={\frac {FN}{TN+FN}}}

True positive rate (TPR) : the fraction of positive cases which were correctly predicted out of all the positive cases. It is usually referred to as sensitivity or recall, and it represents the probability of the positive subjects to be classified correctly as such. It is given by the formula: T P R = P (p r e d i c t i o n = + | a c t u a l = +) = T P T P + F N {\displaystyle TPR=P(prediction=+\ |\ actual=+)={\frac {TP}{TP+FN}}}

False negative rate (FNR) : the fraction of positive cases which were incorrectly predicted to be negative out of all the positive cases. It represents the probability of the positive subjects to be classified incorrectly as negative ones, and it is given by the formula: F N R = P (p r e d i c t i o n = – | a c t u a l = +) = F N T P + F N {\displaystyle FNR=P(prediction=-\ |\ actual=+)={\frac {FN}{TP+FN}}}

True negative rate (TNR) : the fraction of negative cases which were correctly predicted out of all the negative cases. It represents the probability of the negative subjects to be classified correctly as such, and it is given by the formula: T N R = P (p r e d i c t i o n = – | a c t u a l = –) = T N T N + F P {\displaystyle TNR=P(prediction=-\ |\ actual=-)={\frac {TN}{TN+FP}}}

False positive rate (FPR) : the fraction of negative cases which were incorrectly predicted to be positive out of all the negative cases. It represents the probability of the negative subjects to be classified incorrectly as positive ones, and it is given by the formula: F P R = P (p r e d i c t i o n = + | a c t u a l = –) = F P T N + F P {\displaystyle FPR=P(prediction=+\ |\ actual=-)={\frac {FP}{TN+FP}}}

The following criteria can be understood as measures of the three general definitions given at the beginning of this section, namely Independence , Separation and Sufficiency . In the table to the right, we can see the relationships between them.

To define these measures specifically, we will divide them into three big groups as done in Verma et al.: definitions based on a predicted outcome, on predicted and actual outcomes, and definitions based on predicted probabilities and the actual outcome.

We will be working with a binary classifier and the following notation: S {\textstyle S} refers to the score given by the classifier, which is the probability of a certain subject to be in the positive or the negative class. R {\textstyle R} represents the final classification predicted by the algorithm, and its value is usually derived from S {\textstyle S} , for example will be positive when S {\textstyle S} is above a certain threshold. Y {\textstyle Y} represents the actual outcome, that is, the real classification of the individual and, finally, A {\textstyle A} denotes the sensitive attributes of the subjects.

Definitions based on predicted outcome

The definitions in this section focus on a predicted outcome R {\textstyle R} for various distributions of subjects. They are the simplest and most intuitive notions of fairness.

Demographic parity , also referred to as statistical parity , acceptance rate parity and benchmarking . A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal probability of being assigned to the positive predicted class. This is, if the following formula is satisfied: P (R = + | A = a) = P (R = + | A = b) ∀ a , b ∈ A {\displaystyle P(R=+\ |\ A=a)=P(R=+\ |\ A=b)\quad \forall a,b\in A}

Conditional statistical parity . Basically consists in the definition above, but restricted only to a subset of the instances. In mathematical notation this would be: P (R = + | L = l , A = a) = P (R = + | L = l , A = b) ∀ a , b ∈ A ∀ l ∈ L {\displaystyle P(R=+\ |\ L=l,A=a)=P(R=+\ |\ L=l,A=b)\quad \forall a,b\in A\quad \forall l\in L}

Definitions based on predicted and actual outcomes

These definitions not only considers the predicted outcome R {\textstyle R} but also compare it to the actual outcome Y {\textstyle Y} .

Predictive parity , also referred to as outcome test . A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal PPV. This is, if the following formula is satisfied: P (Y = + | R = + , A = a) = P (Y = + | R = + , A = b) ∀ a , b ∈ A {\displaystyle P(Y=+\ |\ R=+,A=a)=P(Y=+\ |\ R=+,A=b)\quad \forall a,b\in A}

False positive error rate balance , also referred to as predictive equality . A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal FPR. This is, if the following formula is satisfied: P (R = + | Y = − , A = a) = P (R = + | Y = − , A = b) ∀ a , b ∈ A {\displaystyle P(R=+\ |\ Y=-,A=a)=P(R=+\ |\ Y=-,A=b)\quad \forall a,b\in A}

False negative error rate balance , also referred to as equal opportunity . A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal FNR. This is, if the following formula is satisfied: P (R = − | Y = + , A = a) = P (R = − | Y = + , A = b) ∀ a , b ∈ A {\displaystyle P(R=-\ |\ Y=+,A=a)=P(R=-\ |\ Y=+,A=b)\quad \forall a,b\in A}

Equalized odds , also referred to as conditional procedure accuracy equality and disparate mistreatment . A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal TPR and equal FPR, satisfying the formula: P (R = + | Y = y , A = a) = P (R = + | Y = y , A = b) y ∈ { + , − } ∀ a , b ∈ A {\displaystyle P(R=+\ |\ Y=y,A=a)=P(R=+\ |\ Y=y,A=b)\quad y\in \{+,-\}\quad \forall a,b\in A}

Conditional use accuracy equality . A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal PPV and equal NPV, satisfying the formula: P (Y = y | R = y , A = a) = P (Y = y | R = y , A = b) y ∈ { + , − } ∀ a , b ∈ A {\displaystyle P(Y=y\ |\ R=y,A=a)=P(Y=y\ |\ R=y,A=b)\quad y\in \{+,-\}\quad \forall a,b\in A}

Overall accuracy equality . A classifier satisfies this definition if the subject in the protected and unprotected groups have equal prediction accuracy, that is, the probability of a subject from one class to be assigned to it. This is, if it satisfies the following formula: P (R = Y | A = a) = P (R = Y | A = b) ∀ a , b ∈ A {\displaystyle P(R=Y\ |\ A=a)=P(R=Y\ |\ A=b)\quad \forall a,b\in A}

Treatment equality . A classifier satisfies this definition if the subjects in the protected and unprotected groups have an equal ratio of FN and FP, satisfying the formula: F N A = a F P A = a = F N A = b F P A = b {\displaystyle {\frac {FN_{A=a}}{FP_{A=a}}}={\frac {FN_{A=b}}{FP_{A=b}}}}

Definitions based on predicted probabilities and actual outcome

These definitions are based in the actual outcome Y {\textstyle Y} and the predicted probability score S {\textstyle S} .

Test-fairness , also known as calibration or matching conditional frequencies . A classifier satisfies this definition if individuals with the same predicted probability score S {\textstyle S} have the same probability of being classified in the positive class when they belong to either the protected or the unprotected group: P (Y = + | S = s , A = a) = P (Y = + | S = s , A = b) ∀ s ∈ S ∀ a , b ∈ A {\displaystyle P(Y=+\ |\ S=s,A=a)=P(Y=+\ |\ S=s,A=b)\quad \forall s\in S\quad \forall a,b\in A}

Well-calibration is an extension of the previous definition. It states that when individuals inside or outside the protected group have the same predicted probability score S {\textstyle S} they must have the same probability of being classified in the positive class, and this probability must be equal to S {\textstyle S} : P (Y = + | S = s , A = a) = P (Y = + | S = s , A = b) = s ∀ s ∈ S ∀ a , b ∈ A {\displaystyle P(Y=+\ |\ S=s,A=a)=P(Y=+\ |\ S=s,A=b)=s\quad \forall s\in S\quad \forall a,b\in A}

Balance for positive class . A classifier satisfies this definition if the subjects constituting the positive class from both protected and unprotected groups have equal average predicted probability score S {\textstyle S} . This means that the expected value of probability score for the protected and unprotected groups with positive actual outcome Y {\textstyle Y} is the same, satisfying the formula: E (S | Y = + , A = a) = E (S | Y = + , A = b) ∀ a , b ∈ A {\displaystyle E(S\ |\ Y=+,A=a)=E(S\ |\ Y=+,A=b)\quad \forall a,b\in A}

Balance for negative class . A classifier satisfies this definition if the subjects constituting the negative class from both protected and unprotected groups have equal average predicted probability score S {\textstyle S} . This means that the expected value of probability score for the protected and unprotected groups with negative actual outcome Y {\textstyle Y} is the same, satisfying the formula: E (S | Y = − , A = a) = E (S | Y = − , A = b) ∀ a , b ∈ A {\displaystyle E(S\ |\ Y=-,A=a)=E(S\ |\ Y=-,A=b)\quad \forall a,b\in A}

Equal confusion fairness

With respect to confusion matrices , independence, separation, and sufficiency require the respective quantities listed below to not have statistically significant difference across sensitive characteristics.

Independence: (TP + FP) / (TP + FP + FN + TN) (i.e., P (Y ^ = 1) {\displaystyle P({\hat {Y}}=1)}).

Separation: TN / (TN + FP) and TP / (TP + FN) (i.e., specificity P (Y ^ = 0 ■ Y = 0) {\displaystyle P({\hat {Y}}=0\mid Y=0)} and recall P (Y ^ = 1 ■ Y = 1) {\displaystyle P({\hat {Y}}=1\mid Y=1)}).

Sufficiency: TP / (TP + FP) and TN / (TN + FN) (i.e., precision P (Y = 1 ■ Y ^ = 1) {\displaystyle P(Y=1\mid {\hat {Y}}=1)} and negative predictive value P (Y = 0 ■ Y ^ = 0) {\displaystyle P(Y=0\mid {\hat {Y}}=0)}).

The notion of equal confusion fairness requires the confusion matrix of a given decision system to have the same distribution when computed stratified over all sensitive characteristics.

Social welfare function

Some scholars have proposed defining algorithmic fairness in terms of a social welfare function . They argue that using a social welfare function enables an algorithm designer to consider fairness and predictive accuracy in terms of their benefits to the people affected by the algorithm. It also allows the designer to trade off efficiency and equity in a principled way. Sendhil Mullainathan has stated that algorithm designers should use social welfare functions to recognize absolute gains for disadvantaged groups. For example, a study found that using a decision-making algorithm in pretrial detention rather than pure human judgment reduced the detention rates for Blacks, Hispanics, and racial minorities overall, even while keeping the crime rate constant.

Individual fairness criteria

An important distinction among fairness definitions is the one between group and individual notions. Roughly speaking, while group fairness criteria compare quantities at a group level, typically identified by sensitive attributes (e.g. gender, ethnicity, age, etc.), individual criteria compare individuals. In words, individual fairness follow the principle that "similar individuals should receive similar treatments".

There is a very intuitive approach to fairness, which usually goes under the name of fairness through unawareness (FTU), or blindness , that prescribes not to explicitly employ sensitive features when making (automated) decisions. This is effectively a notion of individual fairness, since two individuals differing only for the value of their sensitive attributes would receive the same outcome.

However, in general, FTU is subject to several drawbacks, the main being that it does not take into account possible correlations between sensitive attributes and non-sensitive attributes employed in the decision-making process. For example, an agent with the (malignant) intention to discriminate on the basis of gender could introduce in the model a proxy variable for gender (i.e. a variable highly correlated with gender) and effectively using gender information while at the same time being compliant to the FTU prescription.

The problem of what variables correlated to sensitive ones are fairly employable by a model in the decision-making process is a crucial one, and is relevant for group concepts as well: independence metrics require a complete removal of sensitive information, while separation-based metrics allow for correlation, but only as far as the labeled target variable "justify" them.

The most general concept of individual fairness was introduced in the pioneer work by Cynthia Dwork and collaborators in 2012 and can be thought of as a mathematical translation of the principle that the decision map taking features as input should be built such that it is able to "map similar individuals similarly", that is expressed as a Lipschitz condition on the model map. They call this approach fairness through awareness (FTA), precisely as counterpoint to FTU, since they underline the importance of choosing the appropriate target-related distance metric to assess which individuals are similar in specific situations. Again, this problem is very related to the point raised above about what variables can be seen as "legitimate" in particular contexts.

Causality-based metrics

Causal fairness measures the frequency with which two nearly identical users or applications who differ only in a set of characteristics with respect to which resource allocation must be fair receive identical treatment. [dubious – discuss]

An entire branch of the academic research on fairness metrics is devoted to leverage causal models to assess bias in machine learning models. This approach is usually justified by the fact that the same observational distribution of data may hide different causal relationships among the variables at play, possibly with different interpretations of whether the outcome are affected by some form of bias or not.

Kusner et al. propose to employ counterfactuals , and define a decision-making process counterfactually fair if, for any individual, the outcome does not change in the counterfactual scenario where the sensitive attributes are changed. The mathematical formulation reads:

P (R A ← a = 1 ■ A = a , X = x) = P (R A ← b = 1 ■ A = a , X = x) , ∀ a , b ; {\displaystyle P(R_{A\leftarrow a}=1\mid A=a,X=x)=P(R_{A\leftarrow b}=1\mid A=a,X=x),\quad \forall a,b;}

that is: taken a random individual with sensitive attribute A = a {\displaystyle A=a} and other features X = x {\displaystyle X=x} and the same individual if she had A = b {\displaystyle A=b} , they should have same chance of being accepted.

The symbol R ^ A ← a {\displaystyle {\hat {R}}_{A\leftarrow a}} represents the counterfactual random variable R {\displaystyle R} in the scenario where the sensitive attribute A {\displaystyle A} is fixed to A = a {\displaystyle A=a} . The conditioning on A = a , X = x {\displaystyle A=a,X=x} means that this requirement is at the individual level, in that we are conditioning on all the variables identifying a single observation.

Machine learning models are often trained upon data where the outcome depended on the decision made at that time. For example, if a machine learning model has to determine whether an inmate will recidivate and will determine whether the inmate should be released early, the outcome could be dependent on whether the inmate was released early or not. Mishler et al. propose a formula for counterfactual equalized odds:

P (R = 1 ■ Y 0 = 0 , A = a) = P (R = 1 ■ Y 0 = 0 , A = b) ∧ P (R = 0 ■ Y 1 = 1 , A = a) = P (R = 0 ■ Y 1 = 1 , A = b) , ∀ a , b ; {\displaystyle P(R=1\mid Y^{0}=0,A=a)=P(R=1\mid Y^{0}=0,A=b)\wedge P(R=0\mid Y^{1}=1,A=a)=P(R=0\mid Y^{1}=1,A=b),\quad \forall a,b;}

where R {\displaystyle R} is a random variable, Y x {\displaystyle Y^{x}} denotes the outcome given that the decision x {\displaystyle x} was taken, and A {\displaystyle A} is a sensitive feature.

Plecko and Bareinboim propose a unified framework to deal with causal analysis of fairness. They suggest the use of a Standard Fairness Model , consisting of a causal graph with 4 types of variables:

sensitive attributes (A {\displaystyle A}),

target variable (Y {\displaystyle Y}),

mediators (W {\displaystyle W}) between A {\displaystyle A} and Y {\displaystyle Y} , representing possible indirect effects of sensitive attributes on the outcome,

variables possibly sharing a common cause with A {\displaystyle A} (Z {\displaystyle Z}), representing possible spurious (i.e., non causal) effects of the sensitive attributes on the outcome.

Within this framework, Plecko and Bareinboim are therefore able to classify the possible effects that sensitive attributes may have on the outcome.

Moreover, the granularity at which these effects are measured—namely, the conditioning variables used to average the effect—is directly connected to the "individual vs. group" aspect of fairness assessment.

Bias mitigation strategies

Fairness can be applied to machine learning algorithms in three different ways: data preprocessing , optimization during software training, or post-processing results of the algorithm.

Preprocessing

Usually, the classifier is not the only problem; the dataset is also biased. The discrimination of a dataset D {\textstyle D} with respect to the group A = a {\textstyle A=a} can be defined as follows: d i s c A = a (D) = | { X ∈ D | X (A) ≠ a , X (Y) = + } | | { X ∈ D | X (A) ≠ a } | − | { X ∈ D | X (A) = a , X (Y) = + } | | { X ∈ D | X (A) = a } | {\displaystyle disc_{A=a}(D)={\frac {|\{X\in D|X(A)\neq a,X(Y)=+\}|}{|\{X\in D|X(A)\neq a\}|}}-{\frac {|\{X\in D|X(A)=a,X(Y)=+\}|}{|\{X\in D|X(A)=a\}|}}}

That is, an approximation to the difference between the probabilities of belonging in the positive class given that the subject has a protected characteristic different from a {\textstyle a} and equal to a {\textstyle a} .

Algorithms correcting bias at preprocessing remove information about dataset variables which might result in unfair decisions, while trying to alter as little as possible. This is not as simple as just removing the sensitive variable, because other attributes can be correlated to the protected one.

A way to do this is to map each individual in the initial dataset to an intermediate representation in which it is impossible to identify whether it belongs to a particular protected group while maintaining as much information as possible. Then, the new representation of the data is adjusted to get the maximum accuracy in the algorithm.

This way, individuals are mapped into a new multivariable representation where the probability of any member of a protected group to be mapped to a certain value in the new representation is the same as the probability of an individual which doesn't belong to the protected group. Then, this representation is used to obtain the prediction for the individual, instead of the initial data. As the intermediate representation is constructed giving the same probability to individuals inside or outside the protected group, this attribute is hidden to the classifier.

An example is explained in Zemel et al. where a multinomial random variable is used as an intermediate representation. In the process, the system is encouraged to preserve all information except that which can lead to biased decisions, and to obtain a prediction as accurate as possible.

On the one hand, this procedure has the advantage that the preprocessed data can be used for any machine learning task. Furthermore, the classifier does not need to be modified, as the correction is applied to the dataset before processing. On the other hand, the other methods obtain better results in accuracy and fairness. [citation needed]

Reweighing

Reweighing is an example of a preprocessing algorithm. The idea is to assign a weight to each dataset point such that the weighted discrimination is 0 with respect to the designated group.

If the dataset D {\textstyle D} was unbiased the sensitive variable A {\textstyle A} and the target variable Y {\textstyle Y} would be statistically independent and the probability of the joint distribution would be the product of the probabilities as follows: P e x p (A = a ∧ Y = +) = P (A = a) × P (Y = +) = | { X ∈ D | X (A) = a } | | D | × | { X ∈ D | X (Y) = + } | | D | {\displaystyle P_{exp}(A=a\wedge Y=+)=P(A=a)\times P(Y=+)={\frac {|\{X\in D|X(A)=a\}|}{|D|}}\times {\frac {|\{X\in D|X(Y)=+\}|}{|D|}}}

In reality, however, the dataset is not unbiased and the variables are not statistically independent so the observed probability is: P o b s (A = a ∧ Y = +) = | { X ∈ D | X (A) = a ∧ X (Y) = + } | | D | {\displaystyle P_{obs}(A=a\wedge Y=+)={\frac {|\{X\in D|X(A)=a\wedge X(Y)=+\}|}{|D|}}}

To compensate for the bias, the software adds a weight , lower for favored objects and higher for unfavored objects. For each X ∈ D {\textstyle X\in D} we get: W (X) = P e x p (A = X (A) ∧ Y = X (Y)) P o b s (A = X (A) ∧ Y = X (Y)) {\displaystyle W(X)={\frac {P_{exp}(A=X(A)\wedge Y=X(Y))}{P_{obs}(A=X(A)\wedge Y=X(Y))}}}

When we have for each X {\textstyle X} a weight associated W (X) {\textstyle W(X)} we compute the weighted discrimination with respect to group A = a {\textstyle A=a} as follows: d i s c A = a (D) = ∑ W (X) X ∈ { X ∈ D | X (A) ≠ a , X (Y) = + } ∑ W (X) X ∈ { X ∈ D | X (A) ≠ a } − ∑ W (X) X ∈ { X ∈ D | X (A) = a , X (Y) = + } ∑ W (X) X ∈ { X ∈ D | X (A) = a } {\displaystyle disc_{A=a}(D)={\frac {\sum W(X)X\in \{X\in D|X(A)\neq a,X(Y)=+\}}{\sum W(X)X\in \{X\in D|X(A)\neq a\}}}-{\frac {\sum W(X)X\in \{X\in D|X(A)=a,X(Y)=+\}}{\sum W(X)X\in \{X\in D|X(A)=a\}}}}

It can be shown that after reweighting this weighted discrimination is 0.

Inprocessing

Another approach is to correct the bias at training time. This can be done by adding constraints to the optimization objective of the algorithm. These constraints force the algorithm to improve fairness, by keeping the same rates of certain measures for the protected group and the rest of individuals. For example, we can add to the objective of the algorithm the condition that the false positive rate is the same for individuals in the protected group and the ones outside the protected group.

The main measures used in this approach are false positive rate, false negative rate, and overall misclassification rate. It is possible to add just one or several of these constraints to the objective of the algorithm. Note that the equality of false negative rates implies the equality of true positive rates so this implies the equality of opportunity. After adding the restrictions to the problem it may turn intractable, so a relaxation on them may be needed.

Adversarial debiasing

We train two classifiers at the same time through some gradient-based method (f.e.: gradient descent). The first one, the predictor tries to accomplish the task of predicting Y {\textstyle Y} , the target variable, given X {\textstyle X} , the input, by modifying its weights W {\textstyle W} to minimize some loss function L P (y ^ , y) {\textstyle L_{P}({\hat {y}},y)} . The second one, the adversary tries to accomplish the task of predicting A {\textstyle A} , the sensitive variable, given Y ^ {\textstyle {\hat {Y}}} by modifying its weights U {\textstyle U} to minimize some loss function L A (a ^ , a) {\textstyle L_{A}({\hat {a}},a)} . An important point here is that, to propagate correctly, Y ^ {\textstyle {\hat {Y}}} above must refer to the raw output of the classifier, not the discrete prediction; for example, with an artificial neural network and a classification problem, Y ^ {\textstyle {\hat {Y}}} could refer to the output of the softmax layer .

Then we update U {\textstyle U} to minimize L A {\textstyle L_{A}} at each training step according to the gradient ∇ U L A {\textstyle \nabla _{U}L_{A}} and we modify W {\textstyle W} according to the expression: ∇ W L P − p r o j ∇ W L A ∇ W L P − α ∇ W L A {\displaystyle \nabla _{W}L_{P}-proj_{\nabla _{W}L_{A}}\nabla _{W}L_{P}-\alpha \nabla _{W}L_{A}} where α \alpha is a tunable hyperparameter that can vary at each time step.

The intuitive idea is that we want the predictor to try to minimize L P {\textstyle L_{P}} (therefore the term ∇ W L P {\textstyle \nabla _{W}L_{P}}) while, at the same time, maximize L A {\textstyle L_{A}} (therefore the term – α ∇ W L A {\textstyle -\alpha \nabla _{W}L_{A}}), so that the adversary fails at predicting the sensitive variable from Y ^ {\textstyle {\hat {Y}}} .

The term – p r o j ∇ W L A ∇ W L P {\textstyle -proj_{\nabla _{W}L_{A}}\nabla _{W}L_{P}} prevents the predictor from moving in a direction that helps the adversary decrease its loss function.

It can be shown that training a predictor classification model with this algorithm improves demographic parity with respect to training it without the adversary .

Postprocessing

The final method tries to correct the results of a classifier to achieve fairness. In this method, we have a classifier that returns a score for each individual and we need to do a binary prediction for them. High scores are likely to get a positive outcome, while low scores are likely to get a negative one, but we can adjust the threshold to determine when to answer yes as desired. Note that variations in the threshold value affect the trade-off between the rates for true positives and true negatives.

If the score function is fair in the sense that it is independent of the protected attribute, then any choice of the threshold will also be fair, but classifiers of this type tend to be biased, so a different threshold may be required for each protected group to achieve fairness. A way to do this is plotting the true positive rate against the false negative rate at various threshold settings (this is called ROC curve) and find a threshold where the rates for the protected group and other individuals are equal.

Reject option based classification

Given a classifier let P (+ | X) {\textstyle P(+|X)} be the probability computed by the classifiers as the probability that the instance X {\textstyle X} belongs to the positive class +. When P (+ | X) {\textstyle P(+|X)} is close to 1 or to 0, the instance X {\textstyle X} is specified with high degree of certainty to belong to class + or – respectively. However, when P (+ | X) {\textstyle P(+|X)} is closer to 0.5 the classification is more unclear.

We say X {\textstyle X} is a "rejected instance" if m a x (P (+ | X) , 1 – P (+ | X)) ≤ θ {\textstyle max(P(+|X),1-P(+|X))\leq \theta } with a certain θ {\textstyle \theta } such that 0.5 < θ < 1 {\textstyle 0.5<\theta <1} .

The algorithm of "ROC" consists on classifying the non-rejected instances following the rule above and the rejected instances as follows: if the instance is an example of a deprived group (X (A) = a {\displaystyle X(A)=a}) then label it as positive, otherwise, label it as negative.

We can optimize different measures of discrimination (link) as functions of θ {\textstyle \theta } to find the optimal θ {\textstyle \theta } for each problem and avoid becoming discriminatory against the privileged group.

See also

Algorithmic bias

Machine learning

Representational harm

References