

CS561 - Artificial Intelligence

Fairness in Machine Learning



Overview

Overview



- **Personal Data, Sensitive Personal Data**
- **What is Processing Data? Exemptions**
- **Processing Sensitive Personal Data**
- **(Un)-Fairness three examples**
- **Different notions of (un)-fairness**
- **Models**
- **Tools**





Personal Data

Personal Data - 01



- The definition of personal information or personal data is the critical element which determines the zone of informational privacy guaranteed by a data protection legislation.
- The object of data protection regimes is to protect the autonomy of the individual by **protecting the identity of the individual**
- Information which is protected under the head of personal data must first and foremost be about such individual

Personal Data - 02



- For instance, a file maintained by a bank containing the KYC information of an individual is information about that individual.
- The relationship need not be as straightforward in all cases.
- For instance, information that a child is born with fetal alcohol syndrome is personal information both about the child and its mother
- All information about an individual is not personal data
- Protection of identity is central to informational privacy.
- So the information must be such that the individual is either identified or identifiable from such information.

Personal Data - 03



- **One important challenge to the definition of personal data arises from modern technologies which collect newer forms of data from newer sources.**
- **The current definition views personal data in terms of a binary, i.e. identifiable data and non-identifiable data.**
- **The workability of this definition has been called into question**
- **A well known example is of a data set of search queries released by AOL after having removed all identifiers which nonetheless resulted in the identification of an individual within days of release of the data set**

Personal Data - 04



- The AOL search data incident is a notable case that highlights the potential risks associated with the release of **anonymized data** and the importance of privacy protection in the digital age.
- In 2006, America Online (AOL), a major internet service provider at the time, released a large dataset containing search queries from around 650,000 of its users over a three-month period.
- The intent behind releasing this data was to facilitate academic research into search engine usage patterns.
- However, the dataset was not properly anonymized, meaning that while user names were replaced with unique identifiers, the search queries themselves were not sufficiently masked to prevent identification.
- As a result, researchers and journalists quickly realized that it was possible to trace many of the search queries back to individual users based on the specificity and personal nature of the queries.

Sensitive Personal Data

Sensitive Personal Data

- There are matters within this zone which are intimate matters in which there is a higher expectation of privacy
- Unauthorized use of such information of the individual may have severe consequences.
- The observations of the Supreme Court in Puttaswamy, on sexual orientation illustrate this aspect of sensitive information
- **"Sexual orientation is an essential attribute of privacy. Discrimination against an individual on the basis of sexual orientation is deeply offensive to the dignity and self-worth of the individual"**
- Such data, if revealed, may also be the basis of discriminatory action

What is Processing?

What is Processing?

- Having discussed the term personal data, it is important to demarcate actions performed on such data which would be the primary subject matter of the law.
- A term that is used to address any action involving data is the term —processing.
- Data protection laws across the globe have tried to develop definitions of data processing in such a manner that they cover all the associated activities that are performed on data.
- **European Union:** Any operation or set of operations which is performed on personal data collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction

What is Processing?

- **United Kingdom:** The means for obtaining, recording or holding the information or data or carrying out any operation or set of operations on the information or data, including organisation, adaptation, alteration, retrieval, consultation, disclosure by transmission, dissemination or otherwise making available, alignment, combination, blocking, erasure or destruction of the information or data
- Other countries has close meaning as discussed above.



Exemptions

Exemptions - 01



- There are some activities which cannot be brought under the purview of a data protection law.
- **Personal Use:** an individual who processes data for herself, or for household activities, such activity would be outside the scope of regulation. For instance, a personal diary maintained by an individual which may have references to friends and family, or an address book on a computer containing personal data of friends and acquaintances
- **Journalistic/Artistic/Literary Purposes:** For instance, newspapers routinely publish personal data of public figures or other individuals while reporting. Various data protection laws grant different levels of exemptions for processing of personal data for journalistic purposes

Exemptions - 02



- **Research/Historical and Statistical Purpose:** For instance, collection of personal data for Census
- **Investigation and detection of crime:** Several laws such as National Investigation Agency Act, 2008, the Prevention of Money Laundering Act, 2002 (PMLA) etc. empower law enforcement agencies and police officers to collect personal information for the purpose of investigation of a crime
- **National Security:** Processing of information in the interest of national security, or the security of the State and to prevent incitement to an offence is permissible as long as the law enforcement authority or the Government is able to demonstrate that processing of the information is necessary to achieve the purpose.



Processing Sensitive Information

Processing of sensitive information - 01



- It may be intuitively understood that an individual would consider it important to protect information relating to such core aspects of her being from being used or disclosed in a manner likely to cause harm to her
- In order to prevent harm, it may be necessary to categorise the types of information, which form an integral part of an individual's identity.
- The harms arise, of course, because information of the individual becomes available to others through a wide range of activities, collectively termed —data processing
- For instance, in some circumstances, disclosure of such information, is more likely to lead to discrimination, ridicule and reputational harm, especially where one's beliefs and choices form part of the minority view in society.

Processing of sensitive information - 02



- **Data protection law deals with the protection of personal data of an individual**
- **Personal data is understood as information relating to an identified or identifiable natural person.**
- **An identified person is one who can be identified directly or indirectly, with reference to one or more factors, which are specific to her physical, physiological, mental, economic, cultural or social identity**
- **Some of these identifying factors play an important role in forming an integral part of the individual's personality and being**

Processing of sensitive information - 03



- In order to guard against such harms, some jurisdictions recognise the necessity for certain pre-identified categories within the scope of personal data to grant individuals extra protection against misuse of these types of information, by prohibiting the collection, use and disclosure of this information without the explicit consent of the individual
- Such types of data are termed —sensitive and may include religious beliefs, physical or mental health, sexual orientation, biometric and genetic data, racial or ethnic origin and health information.

Processing of sensitive information - 04



- **Certain types of information have been identified as sensitive because there is a greater likelihood of harm caused to the individual if there is unauthorised collection, use and disclosure of this information.**
- **In order to understand the rationale behind identifying certain categories of information as sensitive, there may be a need to assess the harms, which are likely to arise**
- **For instance, there may be certain types of information, which are not classified under the law, but it could become sensitive because of its potential impact on individuals if this data is compromised in any manner.**
- **This could include unique identification numbers, passport numbers, and computer passwords**

(Un) Fairness – Three Examples

Example01 - 01



- Throughout the years, an employment bureau recorded various parameters of job candidates.
- Based on these parameters, the company wants to learn a model for partially automating the match making between a job and a job candidate.
- A match is labeled as successful if the company hires the applicant
- It turns out, however, that the historical data are biased
- For higher board functions, Caucasian males are systematically being favored

Example01 - 02



- A model learned directly on this data will learn this discriminatory behavior and apply it in future predictions.
- From an ethical and legal point of view it is of course unacceptable that a model discriminating in this way is deployed.

Example02 - 01



- A survey is being conducted by a team of researchers
- Each researcher visits a number of regionally co-located hospitals and enquires some patients.
- The survey contains ambiguous questions (e.g., “Is the patient anxious?”, “Is the patient suffering from delusions?”).
- Different enquirers will answer to these questions in different ways.
- Generalizing directly from the training set consisting of all surveys without taking into account these differences among the enquirers may easily result in misleading findings

Example02 - 02



- For example, if many surveys from hospitals in the Eindhoven area are supplied by an enquirer who more quickly than the other enquirers diagnoses anxiety
- Faulty conclusions such as “Patients in Eindhoven suffer from anxiety symptoms more often than other patients” may emerge.

Example03 - 01



- A bank that wants to use historical information on personal loans to learn models for predicting for new loan applicants the probability that they will default their loan.
- It could very well be that this data shows that members of certain ethnic groups are more likely to default their loan.
- From an ethical and legal point of view it is unacceptable to use the ethnicity of a person to deny the loan to him or her, as this would constitute an infringement of the discrimination laws.
- The ethnicity of a person is likely to be an information carrier rather than a distinguishing factor

Example03 - 02



- People from a certain ethnic group are more likely to default their loan because, e.g., the average level of education in this group is lower.
- In such a situation it is in general perfectly acceptable to use level of education for selecting loan candidates, even though this would lead to favoring one ethnic group over another.
- The bank could legally decide to split up the group of loan applicants according to their education level, and learn more fine-grained models for each of these groups separately.
- A prerequisite for this grouping or stratification approach is of course that the attribute education level is present in the dataset.

Example03 - 03



- The overall effect of stratification will be that one ethnic group may be favored over another.
- In each of the groups separately, the model should give equal probability to both classes.
- In the different strata, however, there may still be a strong dependency between ethnicity and loan defaulting.
- For example, it could be that the age distribution is different for the ethnic groups (e.g., one group has much more very young people), but the age of the loan applicants is not present in the dataset



Different Notions of Fairness

Different Notions of Fairness - 01

- **Three most popular notions of fairness used in machine learning**
- **Disparate Treatment**
- **Disparate Impact**
- **Disparate Mis-treatment**

Different Notions of Fairness - 02



Disparate Treatment

- A decision making system suffers from disparate treatment if it provides different outputs for groups of people with the same (or similar) values of non-sensitive features but different values of sensitive features
- That is basing the decision outcomes on the sensitive feature value amounts to disparate treatment

Different Notions of Fairness - 03



Disparate Treatment - Example

Gender	Clothing Bulge	Prox. Crime	Ground Truth
Male 1	1	1	1
Male 2	1	0	1
Male 3	0	1	0
Female 1	1	1	1
Female 2	1	0	0
Female 3	0	0	1

C1	C2	C3
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

Same values of clothing bulge, prox. Crime, prediction for male and female is different for C2

Different Notions of Fairness - 04



Disparate Treatment - Example

Gender	Clothing Bulge	Prox. Crime	Ground Truth
Male 1	1	1	1
Male 2	1	0	1
Male 3	0	1	0
Female 1	1	1	1
Female 2	1	0	0
Female 3	0	0	1

C1	C2	C3
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

Same values of clothing bulge, prox. Crime, prediction for male and female is different for C3

Different Notions of Fairness - 05



Disparate Impact

- A decision making system suffers from disparate impact if it provides outputs that benefit (hurt) a group of people sharing a value of a sensitive feature **more frequently** than other groups of people (also known as statistical parity)

Different Notions of Fairness - 06



Disparate Impact - Example

Gender	Clothing Bulge	Prox. Crime	Ground Truth
Male 1	1	1	1
Male 2	1	0	1
Male 3	0	1	0
Female 1	1	1	1
Female 2	1	0	0
Female 3	0	0	1

C1	C2	C3
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

For classifier C1,

Fraction of males: 1.00 (3 out of 3)

Fraction of females: 0.66 (2 out of 3)

Different Notions of Fairness - 07



Disparate Mis-treatment

- A decision making system suffers from disparate mistreatment if it achieves different classification accuracy (or conversely, error rate) for groups of people sharing different values of a sensitive feature
- This notion has been also referred to as equality of opportunity and predictive equality
- In addition to overall accuracy, a decision making system suffers from disparate mistreatment if individual **misclassification rates** (e.g., false positive rate, false negative rate) **are different for groups** of people sharing different values of a sensitive feature.

Different Notions of Fairness - 08



Disparate Mis-treatment - Example

Gender	Clothing Bulge	Prox. Crime	Ground Truth
Male 1	1	1	1
Male 2	1	0	1
Male 3	0	1	0
Female 1	1	1	1
Female 2	1	0	0
Female 3	0	0	1

C1	C2	C3
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

Classifiers C1 & C2 are unfair due to disparate mis-treatment

C1: FN_males = 0.0 and FN_females = 0.5

C2: FN_males = 0.0 and FN_females = 0.5

Different Notions of Fairness - 09



Differences

- **Direct unfairness:** Disparate treatment accounts for direct unfairness, i.e., a situation where a decision making process directly (or intentionally) uses the sensitive feature information to put a group of people sharing a value of a sensitive feature on relative disadvantage
- **Indirect unfairness:** disparate impact and disparate mistreatment account for indirect unfairness, i.e., a situation where the decision making process can indirectly or unintentionally leverage the correlation between sensitive features and class labels to put a sensitive feature group at relative disadvantage

Different Notions of Fairness – 10



Modeling

- **No disparate treatment:** A binary classifier does not suffer from disparate treatment if the probability that the classifier outputs a specific value of \hat{y} given a feature vector \mathbf{x} does not change after observing the sensitive feature z

$$P(\hat{y}|\mathbf{x}, z) = P(\hat{y}|\mathbf{x}).$$

Different Notions of Fairness – 11



Modeling

- **No disparate Impact:** A binary classifier does not suffer from disparate impact if the probability that a classifier assigns a user to the positive class, $\hat{y} = 1$, is the same for both values of the sensitive feature z

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1).$$

Different Notions of Fairness – 12



Modeling

- **No disparate mis-treatment:** A binary classifier does not suffer from disparate mis-treatment if the misclassification rates for different groups of people having different values of the sensitive feature z are the same
- Overall misclassification rate

$$P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1),$$



Sensitive Attributes

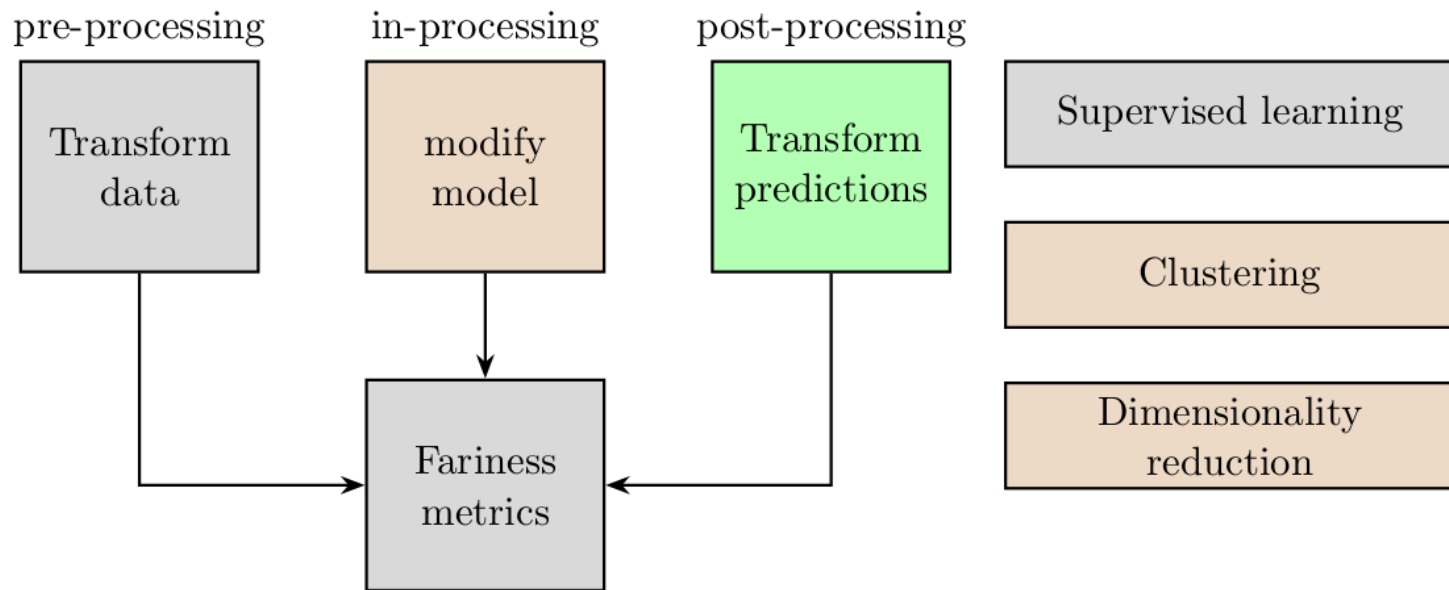
Sensitive Attributes



Sensitive Attribute	Example proxies
Gender	Education levels, Income, Occupation
Marital Status	Income, Education levels
Race	Keywords in user generated content
Disability	Personality test data



Discrimination Aware Classification - 01





In-processing methods



In-processing models - 01

- Given a training data set $D = \{ (\mathbf{x}_i, y_i) \}_{i=1}^N$, find a mapping function $f(\mathbf{x})$
- For decision boundary-based classifiers, finding $f(\mathbf{x})$ usually reduces to building a decision boundary in feature space that separates users in the training set according to their class labels
- One typically looks for a decision boundary, defined by a set of parameters θ^*
- The objective is to minimize loss function over training dataset $L(\theta)$

$$\theta^* = \operatorname{argmin}_{\theta} L(\theta).$$

In-processing models - 02

- Given an unseen feature vector \mathbf{x}_i from the test set, the classifier predicts the label as follows

$$\hat{y}_i = f_{\theta}(\mathbf{x}_i) = 1 \text{ if } d_{\theta^*}(\mathbf{x}_i) \geq 0$$

- Where $d_{\theta^*}(\mathbf{x}_i)$ denotes the signed distance from the feature vector \mathbf{x}_i to the decision boundary
- Fairness criteria as constraint during training is to incorporate explicitly fairness notions while performing training.
- The general optimization formulation is given as

In-processing models - 03

- The general optimization formulation is given as

$$\begin{array}{ll} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & P(.|z = 0) = P(.|z = 1) \end{array} \quad \left. \begin{array}{l} \\ \end{array} \right\} \begin{array}{l} \text{Classifier loss function} \\ \text{Fairness constraints,} \end{array}$$

- Where the probabilities in the constraint(s) can be replaced with the respective disparate impact and disparate mistreatment criteria

In-processing models - 04

- Covariance measure of decision boundary unfairness in the context of disparate impact
- The decision boundary (un)fairness due to disparate impact by means of the covariance between the users' sensitive attribute z and the signed distance from the users' feature vectors to the decision boundary $d_{\theta}(\mathbf{x})$
- The general objective function to minimize disparate impact is:

$$\begin{array}{ll}\text{minimize} & L(\theta) \\ \text{subject to} & \frac{1}{N} \sum_{(\mathbf{x}, z) \in \mathcal{D}} (z - \bar{z}) d_{\theta}(\mathbf{x}) \leq c, \\ & \frac{1}{N} \sum_{(\mathbf{x}, z) \in \mathcal{D}} (z - \bar{z}) d_{\theta}(\mathbf{x}) \geq -c,\end{array}$$

In-processing models - 05

- Covariance measure of decision boundary unfairness in the context of disparate impact
- The decision boundary (un)fairness due to disparate impact by means of the covariance between the users' sensitive attribute z and the signed distance from the users' feature vectors to the decision boundary $d_{\theta}(\mathbf{x})$
- The general objective function to minimize disparate impact is:

$$\begin{array}{ll}\text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{1}{N} \sum_{(\mathbf{x}, z) \in \mathcal{D}} (z - \bar{z}) d_{\boldsymbol{\theta}}(\mathbf{x}) \leq c, \\ & \frac{1}{N} \sum_{(\mathbf{x}, z) \in \mathcal{D}} (z - \bar{z}) d_{\boldsymbol{\theta}}(\mathbf{x}) \geq -c,\end{array}$$

In-processing models - 06



Fairness Constraints in Logistic Regression

- In logistic regression classifiers, one maps the feature vectors \mathbf{x}_i to the class labels y_i by means of a probability distribution

$$p(y = 1|\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}},$$

- Fairness constraints and objective function

$$\begin{array}{ll} \text{minimize} & -\sum_{(\mathbf{x}, y) \in \mathcal{D}} \log p(y|\mathbf{x}, \boldsymbol{\theta}) \\ \text{subject to} & \left. \begin{array}{l} \frac{1}{N} \sum_{(\mathbf{x}, z) \in \mathcal{D}} (z - \bar{z}) \boldsymbol{\theta}^T \mathbf{x} \leq c, \\ \frac{1}{N} \sum_{(\mathbf{x}, z) \in \mathcal{D}} (z - \bar{z}) \boldsymbol{\theta}^T \mathbf{x} \geq -c. \end{array} \right\} \end{array} \quad \begin{array}{l} \text{Logistic regression formulation} \\ \text{Disparate impact constraints} \end{array}$$

In-processing models - 07

Fairness Constraints in Non-linear SVM

- Incorporating the disparate impact constraints in non-linear SVMs result in the following optimization formulation

$$\begin{array}{ll}
 \text{minimize} & \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha} \\
 \text{subject to} & 0 \leq \boldsymbol{\alpha} \leq C, \\
 & \mathbf{y}^T \boldsymbol{\alpha} = 0, \\
 & \left. \begin{array}{l} \frac{1}{N} \sum_{(\mathbf{x}, z) \in \mathcal{D}} (z - \bar{z}) d_{\boldsymbol{\alpha}}(\mathbf{x}) \leq c, \\ \frac{1}{N} \sum_{(\mathbf{x}, z) \in \mathcal{D}} (z - \bar{z}) d_{\boldsymbol{\alpha}}(\mathbf{x}) \geq -c, \end{array} \right\} \begin{array}{l} \text{SVM formulation} \\ \text{Disparate impact constraints} \end{array}
 \end{array}$$



Models – Pre-processing model

Introduction - 01



- **Classifier construction is one of the most researched topics within the data mining and machine learning communities.**
- **Literally thousands of algorithms have been proposed.**
- **The quality of the learned models, however, depends critically on the quality of the training data.**
- **No matter which classifier inducer is applied, if the training data are incorrect, poor models will result.**

Introduction - 02



- The paper titled "**Data preprocessing techniques for classification without discrimination**" study cases in which the input data are discriminatory and learn a discrimination-free classifier for future classification.
- The input of the discrimination-aware classification problem is a labeled dataset and one or more sensitive attributes.
- The output is a classifier to predict the label that should not correlate with the sensitive attribute.
- The quality of the classifier is measured by its accuracy and discrimination; the more accurate, the better, and the less discriminatory, the better.

Introduction - 03



- In this lecture we restrict ourselves to one binary sensitive attribute S with domain $\{b, w\}$ and a binary classification problem with target attribute Class with domain $\{-, +\}$.
- “+” is the desirable class for the data subjects and the objects satisfying $S = b$ and $S = w$ represent, respectively, the deprived and the favored community.
- The discrimination of a classifier C is defined as

$$\text{disc}_{S=b} := P(C(X) = + \mid X(S) = w) - P(C(X) = + \mid X(S) = b) ,$$

- A discrimination larger than 0 reflects that a tuple for which S is w has a higher chance of being assigned the positive label by the classifier C than one where S is b .

Introduction - 04



- The following four methods for incorporating non-discrimination constraints into the classifier construction process
- **Suppression**
- **Massaging the dataset**
- **Reweighting.**
- **Sampling.**

Suppression

- Find the attributes that correlate most with the sensitive attribute S .
- To reduce the discrimination between the class labels and the attribute S , remove S and these most correlated attributes.
- This simple and straightforward approach will serve as the baseline



Introduction - 04

- The following four methods for incorporating non-discrimination constraints into the classifier construction process
- **Suppression**
- **Massaging the dataset**
- **Reweighting.**
- **Sampling.**

Massaging the dataset

- Change the labels of some objects in the dataset in order to remove the discrimination from the input data.
- A good selection of which labels to change is essential.
- To select the best candidates for relabeling, a ranker is used.
- Consider arbitrary combinations of ranker and learner.

Introduction - 04



- The following four methods for incorporating non-discrimination constraints into the classifier construction process

- **Suppression**

- **Massaging the dataset**

- **Reweighting.**

- **Sampling.**

Reweighting

- Instead of changing the labels, the tuples in the training dataset are assigned weights.
- By carefully choosing the weights, the training dataset can be made discrimination-free w.r.t. S without having to change any of the labels.
- The weights on the tuples can be used directly in any method based on frequency counts.

Introduction - 04



- The following four methods for incorporating non-discrimination constraints into the classifier construction process
- Suppression
- Massaging the dataset
- Reweighing.
- Sampling.

Sampling

- Uniform Sampling (US) apply uniform sampling with replacement.
- In this scheme, every object has a uniform probability to be duplicated to increase the size or to be skipped to decrease the size of a group.
- In Preferential Sampling (PS), borderline objects get high priority for being duplicated or being skipped.
- A ranker is used to decide which objects are at the border.



Discrimination Aware Classification

Discrimination Aware Classification - 01

- **Discrimination in Labelled Dataset:** Given a labeled dataset D , an attribute S and a value $b \in \text{dom}(S)$. The discrimination in D w.r.t. the group $S = b$, denoted $\text{disc}_{S=b}(D)$, is defined as

$$\text{disc}_{S=b}(D) := \frac{|\{X \in D \mid X(S) = w, X(\text{Class}) = +\}|}{|\{X \in D \mid X(S) = w\}|} - \frac{|\{X \in D \mid X(S) = b, X(\text{Class}) = +\}|}{|\{X \in D \mid X(S) = b\}|}.$$

- That is, the difference of the probability of being in the positive class between the tuples X in D having $X(S) = w$ in D and those having $X(S) = b$.

Discrimination Aware Classification - 02

- **Discrimination in Classifier's Prediction:** Given an unlabeled dataset D , an attribute S and a value $b \in \text{dom}(S)$. The discrimination of the classifier C w.r.t. the group $S = b$ in dataset D , denoted $\text{disc}_{S=b}(D)$, is defined as:

$$\text{disc}_{S=b}(C, D) := \frac{|\{X \in D \mid X(S) = w, C(X) = +\}|}{|\{X \in D \mid X(S) = w\}|} - \frac{|\{X \in D \mid X(S) = b, C(X) = +\}|}{|\{X \in D \mid X(S) = b\}|}.$$

- That is, it is the difference in probability of being assigned the positive class by the classifier between the tuples of D having $X(S) = w$ and those having $X(S) = b$.

Discrimination Aware Classification - 03

Sex	Ethnicity	Highest Degree	Job Type	Class
M	Native	H. School	Board	+
M	Native	Univ.	Board	+
M	Native	H. School	Board	+
M	Non-nat.	H. School	Healthcare	+
M	Non-nat.	Univ	Healthcare	-
F	Non-nat	Univ.	Education	-
F	Native	H. School	Education	-
F	Native	None	Healthcare	+
F	Non-nat.	Univ	Education	-
F	Native	H. School	Board	+

In this dataset, the discrimination w.r.t. the attribute Sex and Class is

$$disc_{Sex=f}(D) = \frac{4}{5} - \frac{2}{5} = 40\%$$

It means that in the dataset, a female is, in absolute numbers, 40% less likely to be accepted for a job than a male

Discrimination Aware Classification - 04



- **The problem – Discrimination Aware Classification:** Given a labeled dataset D , an attribute S , and a value $b \in \text{dom}(S)$, learn a classifier C such that
- The accuracy of C for future predictions is high
- The discrimination of new examples classified by C is low.
- Clearly there will be a trade-off between the accuracy and the discrimination of the classifier.
- In general, lowering the discrimination will result in lowering the accuracy and vice versa.

Discrimination Aware Classification - 05



- Solutions to learn a non-discriminating classifier that uses the attribute S only during learning and not at prediction time.
- The solution is based on removing the discrimination from the training dataset
- A classifier is learned on this cleaned dataset
- As the classifier is trained on discrimination-free data, it is likely that its predictions will be (more) discrimination-free as well.

Discrimination Aware Classification - 06



Massaging the data

- In Massaging, we will change the labels of some objects X with $X(S) = b$ from $-$ to $+$,
- The same number of objects with $X(S) = w$ from $+$ to $-$.
- In this way the discrimination decreases, yet the overall class distribution is maintained
- The set **pr** of objects X with $X(S) = b$ and $X(\text{Class}) = -$ will be called the promotion candidates
- The set **dem** of objects X with $X(S) = w$ and $X(\text{Class}) = +$ will be called the demotion candidates.

Discrimination Aware Classification - 07



Massaging the data

- We will not randomly pick promotion and demotion candidates to relabel.
- On the training data, a ranker R for ranking the objects according to their positive class probability is learned.
- We assume that higher scores indicate a higher chance to be in the positive class.
- With this ranker, the promotion candidates are sorted according to **descending score** by R
- The demotion candidates according to **ascending score**.

Discrimination Aware Classification - 08



Massaging the data

- When selecting promotion and demotion candidates, first the top elements will be chosen.
- In this way, the objects closest to the decision border are selected first to be relabeled, leading to a minimal effect on the accuracy.
- The number M of pairs needed to be modified to make a dataset D discrimination-free can be calculated as follows.

$$M = \frac{disc(D) \times |D_b| \times |D_w|}{|D|}$$

Discrimination Aware Classification - 09



Massaging the data

- To reach zero discrimination, we hence have to make M modifications to dataset
- D_b and D_w denote objects in D with $S = b$ and $S = w$ respectively
- p_b and p_w are the number of positive objects with, respectively, $S = b$ and $S = w$



Massaging the data - Example

Discrimination Aware Classification - 10

Sex	Ethnicity	Highest Degree	Job Type	Class
M	Native	H. School	Board	+
M	Native	Univ.	Board	+
M	Native	H. School	Board	+
M	Non-nat.	H. School	Healthcare	+
M	Non-nat.	Univ	Healthcare	-
F	Non-nat	Univ.	Education	-
F	Native	H. School	Education	-
F	Native	None	Healthcare	+
F	Non-nat.	Univ	Education	-
F	Native	H. School	Board	+

Consider the data give to the left.

We want to learn a classifier to predict the class of objects for which the predictions are non-discriminatory toward Sex = F

In this example we rank the objects by their positive class probability given by a Naïve Bayes classification model

Discrimination Aware Classification - 11

Sex	Ethnicity	Highest Degree	Job Type	Class	Prob (%)
M	Native	H. School	Board	+	98
M	Native	Univ.	Board	+	89
M	Native	H. School	Board	+	98
M	Non-nat.	H. School	Healthcare	+	69
M	Non-nat.	Univ	Healthcare	-	30
F	Non-nat	Univ.	Education	-	2
F	Native	H. School	Education	-	40
F	Native	None	Healthcare	+	76
F	Non-nat.	Univ	Education	-	2
F	Native	H. School	Board	+	93

Consider the data give to the left.

We want to learn a classifier to predict the class of objects for which the predictions are non-discriminatory toward Sex = F

In this example we rank the objects by their positive class probability given by a Naïve Bayes classification model

The promotion candidates are sorted according to **descending score**

The demotion candidates according to **ascending score.**

Discrimination Aware Classification - 12

Sex	Ethnicity	Highest Degree	Job Type	Class	Prob (%)
M	Native	H. School	Board	+	98
M	Native	Univ.	Board	+	89
M	Native	H. School	Board	+	98
M	Non-nat.	H. School	Healthcare	+	69
M	Non-nat.	Univ	Healthcare	-	30
F	Non-nat	Univ.	Education	-	2
F	Native	H. School	Education	-	40
F	Native	None	Healthcare	+	76
F	Non-nat.	Univ	Education	-	2
F	Native	H. School	Board	+	93

Consider Female, -ve class label

Consider Male, +ve class label

The promotion candidates are sorted according to **descending score**

The demotion candidates according to **ascending score**.

Discrimination Aware Classification - 13

Sex	Ethnicity	Highest Degree	Job Type	Class	Prob (%)
M	Native	H. School	Board	+	98
M	Native	Univ.	Board	+	89
M	Native	H. School	Board	+	98
M	Non-nat.	H. School	Healthcare	+	69
F	Non-nat.	Univ.	Education	-	2
F	Native	H. School	Education	-	40
F	Non-nat.	Univ	Education	-	2

Consider Female, -ve class label

The promotion candidates are sorted according to **descending score**

Discrimination Aware Classification - 14

Sex	Ethnicity	Highest Degree	Job Type	Class	Prob (%)
M	Native	H. School	Board	+	98
M	Native	Univ.	Board	+	89
M	Native	H. School	Board	+	98
M	Non-nat.	H. School	Healthcare	+	69
F	Native	H. School	Education	-	40
F	Non-nat.	Univ.	Education	-	2
F	Non-nat.	Univ.	Education	-	2

Consider Female, -ve class label

The promotion candidates are sorted according to **descending score**

Discrimination Aware Classification - 15

Sex	Ethnicity	Highest Degree	Job Type	Class	Prob (%)
M	Non-nat.	H. School	Healthcare	+	69
M	Native	Univ.	Board	+	89
M	Native	H. School	Board	+	98
M	Native	H. School	Board	+	98
F	Native	H. School	Education	-	40
F	Non-nat.	Univ	Education	-	2
F	Non-nat	Univ.	Education	-	2

Consider demotion candidate Male, +ve class label

The demotion candidates according to **ascending score**.

Discrimination Aware Classification - 16

Sex	Ethnicity	Highest Degree	Job Type	Class	Prob (%)
M	Non-nat.	H. School	Healthcare	+	69
M	Native	Univ.	Board	+	89
M	Native	H. School	Board	+	98
M	Native	H. School	Board	+	98
F	Native	H. School	Education	-	40
F	Non-nat.	Univ	Education	-	2
F	Non-nat	Univ.	Education	-	2

For this dataset $\text{disc}(D) = 40\%$

$$|D_f| = |D_m| = 5;$$

$$|D| = 10$$

$$M = (0.40 * 5 * 5) / 10 = 1$$

Flip class label of one data point in D_f with least probability and flip class label

Flip class label of one data point in D_m with highest probability and flip class label

Discrimination Aware Classification - 17

Sex	Ethnicity	Highest Degree	Job Type	Class	Prob (%)
M	Non-nat.	H. School	Healthcare	-	69
M	Native	Univ.	Board	+	89
M	Native	H. School	Board	+	98
M	Native	H. School	Board	+	98
F	Native	H. School	Education	+	40
F	Non-nat.	Univ	Education	-	2
F	Non-nat	Univ.	Education	-	2

For this dataset $\text{disc}(D) = 40\%$

$$|D_f| = |D_m| = 5;$$

$$|D| = 10$$

$$M = (0.40 * 5 * 5)/10 = 1$$

Flip class label of one data point in D_f with least probability and flip class label

Flip class label of one data point in D_m with highest probability and flip class label

Discrimination Aware Classification - 18

Sex	Ethnicity	Highest Degree	Job Type	Class	Prob (%)
M	Native	H. School	Board	+	98
M	Native	Univ.	Board	+	89
M	Native	H. School	Board	+	98
M	Non-nat.	H. School	Healthcare	-	69
M	Non-nat.	Univ	Healthcare	-	30
F	Non-nat	Univ.	Education	-	2
F	Native	H. School	Education	+	40
F	Native	None	Healthcare	+	76
F	Non-nat.	Univ	Education	-	2
F	Native	H. School	Board	+	93

This dataset has **ZERO** discrimination as shown below

$$disc_{Sex=f}(D) = \frac{3}{5} - \frac{3}{5} = 0\%$$



Reweighing

Discrimination Aware Classification - 19



Reweighting the data

- The Massaging approach is rather intrusive as it changes the labels of the object
- This is a disadvantage and serious intrusion of labeled training dataset
- Reweighting does not have this disadvantage
- Instead of relabeling the data points, different weights will be attached to each data point
- For example, objects with $X(S) = F$ and $X(Class) = +$ will get higher weights than objects with $X(S) = F$ and $X(Class) = -$
- Objects with $X(S) = M$ and $X(Class) = +$ will get lower weights than objects with $X(S) = F$ and $X(Class) = -$.

Discrimination Aware Classification - 20



Reweighting the data

- Again, we assume that we want to reduce the discrimination to 0 while maintaining the overall positive class probability.
- If the dataset D is unbiased, i.e., **S and Class are statistically independent**, then the expected probability would be

$$P_{exp}(S = f \text{ and } class = +) = P(S = f) \times P(class = +)$$
$$\frac{|\{X \in D | X(S)=f\}|}{|D|} \times \frac{|\{X \in D | X(class)=+\}|}{|D|}$$

Discrimination Aware Classification - 21

Reweighting the data

- If the dataset D is unbiased, i.e., **S and Class are statistically independent**, then the observed probability would be

$$P_{obs}(S = f \text{ and } class = +) = P(S = f \text{ and } class = +) \\ \frac{|\{X \in D | X(S) = f \text{ and } class = +\}|}{|D|}$$

- To compensate for the bias, we will assign lower weights to objects that have been deprived or favored. Every object X will be assigned weight:

$$W(X) = \frac{P_{exp}(S=f \text{ and } class=+)}{P_{obs}(S=f \text{ and } class = +)}$$

Discrimination Aware Classification - 22



Reweighting the data

- In this way we assign a weight to **every data point** according to its **S** and **Class**-values.
- We will call the dataset **D** with the added weights, D_w .
- For the example dataset, the following are the characteristics
- 5 data points have female members
- 5 data points have male members
- 6 data points have + class
- $P_{\text{exp}}(\text{Sex} = F \ \& \ X(C) = +) = 0.5 * 0.6 = 30$
- $P_{\text{obs}}(\text{Sex} = F \ \& \ X(C) = +) = 0.2$
- $W(X) = 0.3/0.2 = 1.5$

$$W(X) := \begin{cases} 1.5 & \text{if } X(\text{Sex}) = f \text{ and } X(\text{Class}) = + \\ 0.67 & \text{if } X(\text{Sex}) = f \text{ and } X(\text{Class}) = - \\ 0.75 & \text{if } X(\text{Sex}) = m \text{ and } X(\text{Class}) = + \\ 2 & \text{if } X(\text{Sex}) = m \text{ and } X(\text{Class}) = - \end{cases}$$

Discrimination Aware Classification - 23

Sex	Ethnicity	Highest Degree	Job Type	Class	W(x)
M	Native	H. School	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. School	Board	+	0.75
M	Non-nat.	H. School	Healthcare	+	0.75
M	Non-nat.	Univ	Healthcare	-	2
F	Non-nat	Univ.	Education	-	0.67
F	Native	H. School	Education	-	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ	Education	-	0.67
F	Native	H. School	Board	+	1.5

This dataset has **ZERO** discrimination

By multiply the frequency of every object by its weight, the discrimination would be 0

Discrimination Aware Classification - 24

Sex	Ethnicity	Highest Degree	Job Type	Class	W(x)
M	Native	H. School	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. School	Board	+	0.75
M	Non-nat.	H. School	Healthcare	+	0.75
M	Non-nat.	Univ	Healthcare	-	2
F	Non-nat	Univ.	Education	-	0.67
F	Native	H. School	Education	-	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ	Education	-	0.67
F	Native	H. School	Board	+	1.5

Algorithm

Train a classifier C on training set D_W , taking into account the weights

Discrimination Aware Classification - 25



- **Census Income**
- **Has 48,842 instances and contains demographic information of people**
- **The prediction task is to determine whether a person makes over 50K per year or not**
- **Each data object is described by 14 attributes, of which 8 are categorical and 6 are numerical attributes.**
- **The attributes in the dataset : age, type of work, education, years of education, marital status, occupation, type of relationship (husband, wife, not in family), sex, race, native country, capital gain, capital loss, and weekly working hours.**
- **Use Sex as discriminatory attribute.**
- **This dataset has discrimination of 19.45%**

Discrimination Aware Classification - 26



- **Communities and Crimes**
- **Has 1994 instances**
- **Give information about different communities and crimes within the United States.**
- **Each instance is described by 122 predictive attributes which are used to predict the total number of violent crimes per 100K population.**
- **Sensitive attribute Black**

Discrimination Aware Classification - 27



- Dutch census of the year 2001
- Has 189725
- The dataset is described by 13 attributes namely sex, age, household position, household size, place of previous residence, citizenship, country of birth, education level, economic status (economically active or inactive), current economic activity, marital status, weight, and occupation
- The attribute occupation as a class attribute
- Sex is the sensitive attribute.

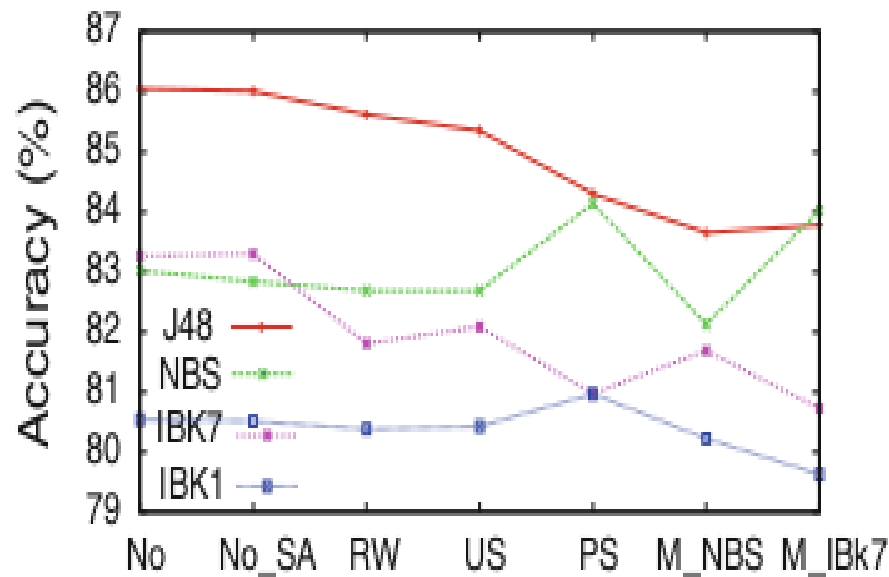
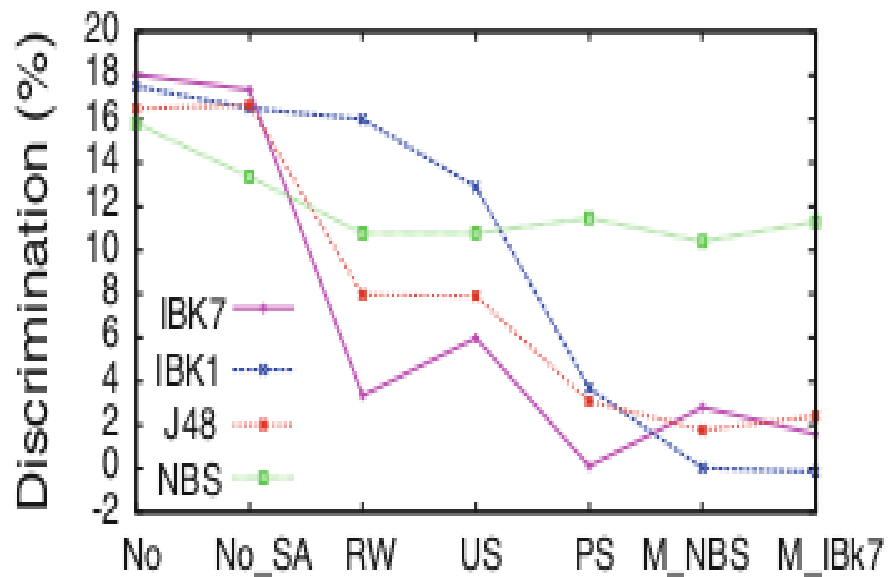
Discrimination Aware Classification - 28

- **Classifiers**
- **k-Nearest neighbor with $k = 1, 3, 7$**
- **Naïve Bayes Classifier**
- **Decision Tree learner**
- **Tool Box: Weka**

Discrimination Aware Classification - 29



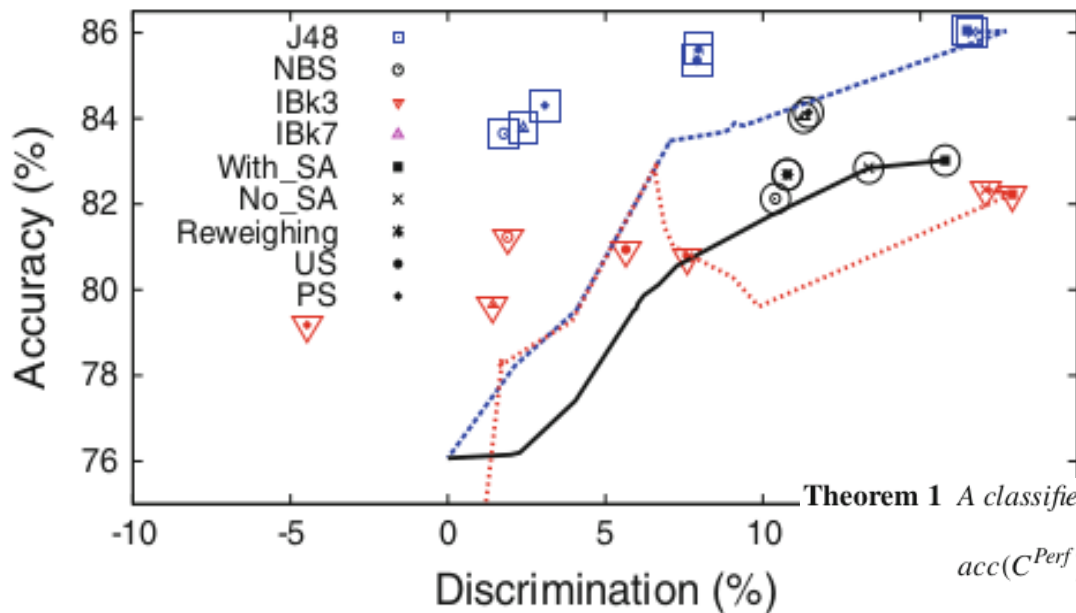
Census Income Dataset



Discrimination Aware Classification - 30



Census Income Dataset



An overview that allows to quickly assess which of the combinations are DA-optimal (discrimination-accuracy-optimal) among the classifiers

$C^{Perf}(X) = X(\text{Class})$ for all $X \in D$.

C^* all set of classifiers such that
 $P(C(X) = + \mid X \text{ in } D) =$
 $P(X(\text{class}) = + \mid X \text{ in } D)$

Theorem 1 A classifier C is DA-optimal in C_{all} iff

$$acc(C^{Perf}) - acc(C) = \frac{\min(d_b, d_w)}{d} (disc(C^{Perf}) - disc(C))$$

A classifier C is DA-optimal in C_{all}^* iff

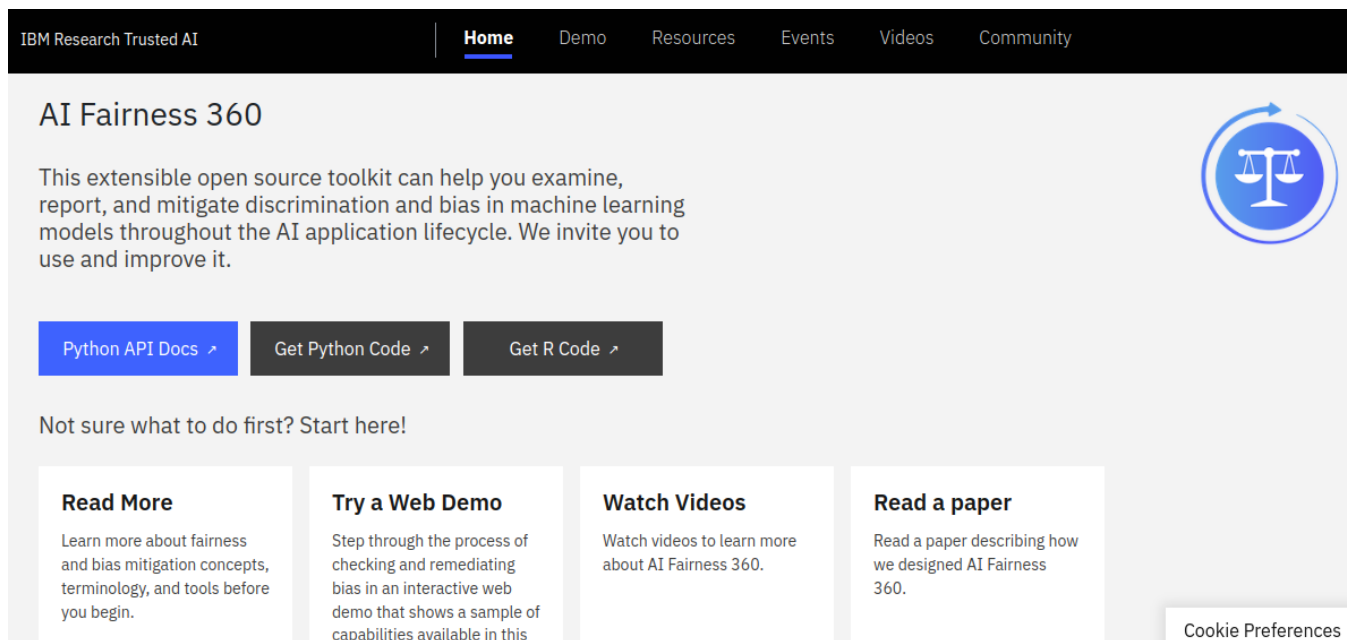
$$acc(C^{Perf}) - acc(C) = 2 \frac{d_b}{d} \frac{d_w}{d} (disc(C^{Perf}) - disc(C))$$





Tools

- The tool developed by IBM research for AI Fairness is accessible at: <https://aif360.res.ibm.com/>



The screenshot shows the homepage of the AI Fairness 360 toolkit. The header is black with white text for navigation: "IBM Research Trusted AI", "Home" (underlined), "Demo", "Resources", "Events", "Videos", and "Community". The main content area is light gray. It features a large heading "AI Fairness 360" and a paragraph describing the toolkit as an extensible open source toolkit for examining, reporting, and mitigating discrimination and bias. To the right is a circular logo with a scale of justice. Below the text are three buttons: "Python API Docs" (blue), "Get Python Code" (dark gray), and "Get R Code" (dark gray). A section titled "Not sure what to do first? Start here!" contains four cards: "Read More" (learn about fairness concepts), "Try a Web Demo" (step through the process of checking and remediating bias), "Watch Videos" (watch videos to learn more), and "Read a paper" (read a paper describing the design). A "Cookie Preferences" button is in the bottom right corner.

IBM Research Trusted AI | **Home** | Demo | Resources | Events | Videos | Community

AI Fairness 360

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

[Python API Docs](#) [Get Python Code](#) [Get R Code](#)

Not sure what to do first? Start here!

Read More

Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.

Try a Web Demo

Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this

Watch Videos

Watch videos to learn more about AI Fairness 360.

Read a paper

Read a paper describing how we designed AI Fairness 360.

[Cookie Preferences](#)



Thank You!

