Fairness-Aware Data Mining

Toshihiro Kamishima
http://www.kamishima.net/
Last Update: 2018-03-23
Fairness-Aware Data Mining

Data analysis taking into account potential issues of fairness, discrimination, neutrality, or independence. It takes influence of the following sensitive information into account:

- maintaining social fairness (gender, race,...)
- restricted by law or contracts (insider or private information)
- any information whose influence data-analysts want to ignore

We here use the term ‘fairness-aware’ instead of an original term, ‘discrimination-aware’, because the term *discrimination* means classification in an ML context.
FADM was originally invented to eliminate socially discriminative outcomes when applying ML techniques to real-world problems.

More extensively, FADM methods would be helpful for correcting any type of biases, which are irrelevant to social discrimination, if what generates the biases is known.

Example:
Hotels’ occupancy rates are generally high, when room charges are high. Of course, the increase of occupancy rates are affected by factors besides room charges.

If such a factor is known to be a seasonal effect, FADM methods can be used for predicting a pure influence from room charges to occupancy rates.
Sources of Unfairness in Machine Learning
Sources of Unfairness in ML / DM

**Annotation Bias**
- Decisions whether to approve loan are unfair by reflecting on prejudice against a specific group in a historical record

**Sample Selection Bias**
- Records who have been able to pay off their loans are only available for those who have been approved the loans

**Inductive Bias**
- Records for minority individuals who have been able to pay off loans in a minority group can be ignored due to the assumption of ML algorithms
Annotation Bias

**Annotation Bias**: Target values or feature values in a training data are biased due to annotator’s cognitive bias or inappropriate observation schemes.

Biases in Labels or Targets
- Historical records of approvals for loan requests might be influenced by prejudice towards a specific group.
- Ratings are affected by predicted ratings displayed when users rate items.

[Cosley+ 03]

Biases in Features
- Uses of words in training corpus are affected by a gender bias.

[Bolukbasi+ 16]
- Admission to universities can be influenced by recommendation letters.
Online advertisements of sites providing arrest record information

Advertisements indicating arrest records were more frequently displayed for names that are more popular among individuals of African descent than those of European descent.

- African descent’s name
- European descent’s name

Arrested? negative ad-text

Located: neutral ad-text

[Sweeney 13]
Suspicious Placement Keyword-Matching Advertisement

Advertisement texts are chosen irrelevant to the actual existence of a prior arrest of the target name

African descent’s name
Actually, no prior arrest

European descent’s name
previously arrested

[Sweeney 13]
Selection of ad-texts was unintentional

Response from advertiser:

- Advertise texts are selected based on the last name, and no other information in exploited
- The selection scheme is adjusted so as to maximizing the crick-through rate based on the feedback records from users by displaying randomly chosen ad-texts

No sensitive information, e.g., race, is exploited in a selection model, but suspiciously discriminative ad-texts are generated

An annotation bias is caused due to the unfair feedbacks from users reflecting the users’ prejudice
Sample Selection Bias: Whether a datum is sampled depends on conditions or contents of the datum, and thus an observed dataset is not a representative of population.

*Strictly speaking, independence between the variables and the other variables needs to be considered.*

Labels are observed only for approved individuals, and counterfactual labels for declined individuals are unknown.

Simple prediction algorithms cannot learn appropriately from such a non-random dataset, depending on contents of the data.
Bandwagon Effect

Bandwagon Effects in ML
A bias in prediction by ML methods can produce a phenomenon, "richer gets richer"

Users’ cognitive bias
If others think that something is good, then I should, too

+ 

Algorithms’ inductive bias
popularity bias: A recommender system tends to select popular items

→

Incorrectly higher-rated items can be more popular, because a recommendation algorithm selects them

→

A undesirable feedback loop caused by undesired selection
Inductive Bias: a bias caused by an assumption adopted in an inductive machine learning algorithms

Inductive Machine Learning Algorithms:

- sample (training data)
- assumption (background knowledge)
- prediction function
- prediction rule

These assumptions are required for generalizing training data

The assumptions doesn’t always agree with a process of data generation in a real world

Inductive Bias
Example of Inductive Bias

- **Occam’s Razor**: Preference of ML algorithms to simpler hypothesis to improve generalization error
  - Missing exceptional minor patterns

- **Smoothness**: Smoother decision boundaries or curves to fit are preferred
  - Non-smooth changes cannot be represented

- **Sparseness**: Preference to hypothesis consisting of the smaller number of features
  - Abandoning less effective features

- **Model Bias**: A target hypothesis may not included in a model of candidate hypotheses
  - A learned hypothesis might not exactly match the target hypothesis
**US Census Data**: predict whether their income is high or low

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-Income</td>
<td>3,256</td>
<td>590</td>
</tr>
<tr>
<td>Low-income</td>
<td>7,604</td>
<td>4,831</td>
</tr>
</tbody>
</table>

Females are minority in the high-income class

In this original data set:
- The number of High-Male data is 5.5 times that of High-Female data
- While 30% of Male data are High income, only 11% of Females are
Inductive Bias: Example

**Odds ratio**: to evaluate the influence of a gender to an income ratio of the odds to be high-income for males to that for females

\[
\text{Odds ratio} = \frac{\Pr[\text{High, Male}] / \Pr[\text{Low, Male}]}{\Pr[\text{High, Female}] / \Pr[\text{Low, Female}]} 
\]

Directly derived from an observed sample

**odds ratio = 3.51**

Derived by a naive Bayes model w/o a gender feature

**odds ratio = 5.26**

The increase of the odds ratio implies that a gender has stronger impact on an income

**Due to an inductive bias, the minor information of high-income females is ignored**
Preliminaries
In fairness-aware data mining, we maintain the influence:

- Socially sensitive information
- Information restricted by law
- Information to be ignored

- University admission
- Credit scoring
- Click-through rate

Formal Fairness
The influence defined by a formal relation between sensitive feature, target variable, and other variables in a model.

- How to relate these variables
- Which set of variables to be considered
- What states of sensitives or targets should be maintained
Notations of Variables

\( Y \)  
**target variable / object variable**

An objective of decision making, or what to predict
- \( Y \): true / population, \( \hat{Y} \): predicted, \( Y^\circ \): fairized

ex., loan approval, university admission, what to recommend

\( S \)  
**sensitive feature**

To ignore the influence to the sensitive feature from a target
- Specified by a user or an analyst depending on his/her purpose
- It may depend on a target or other features
- It can be multivariate

ex., socially sensitive information (gender, race), items’ brand

\( X \)  
**non-sensitive feature vector**

All features other than a sensitive feature
Other Notations

\[ D = \{ y_i, s_i, x_i \}_{i=1}^n \]  

**dataset**

Each datum is a triple of a target value, \( y_i \), a sensitive value, \( s_i \), and non-sensitive feature values, \( x_i \)

\[ D_s = \{ y_i, s_i, x_i \}_{i=1}^n \text{ s.t. } s_i = s \]  

**sensitive group**

a group consisting of the same sensitive value

If \( s_i = 0 \) indicates a minority individual to protect, \( D_0 \), is called a **protected group**, and the rest of dataset, \( D_1 \), is called a **non-protected group**

**E / \( \bar{E} \)**  

**explainable / unexplainable non-sensitive feature vector**

\[ \text{dom}(X) = \text{dom}(E) \times \text{dom}(\bar{E}) \]

Explainable variables are confounding variables with \( Y \) and \( S \), and their influence can be ignored because of legal or other reasons
**Independence**

**(unconditional) independence**

A pair sets of variables, Y and S, are not influenced from each other

\[ Y \perp S \]

**conditional independence**

Y and S are independent, if conditional variables, X, are fixed

\[ Y \perp S \mid X \]

Conditional independence doesn’t imply independence, and vice versa

**context-specific independence**

Y and S are independent, if X are fixed to specific values, x

\[ Y \perp S \mid X=x \]

* Notation with a symbol ‘\( \perp \)' (Unicode 2AEB) is called Dawid’s notation
Association Rule

**X ⇒ Y**  
**X**: antecedent, **Y**: consequent

If **X** is satisfied, **Y** is also satisfied with a high probability

**Ex.**

( milk ∈ Item ) ∧ ( bread ∈ Item ) ⇒ ( egg ∈ Item )

*Item*: a set of simultaneously bought items

A customer who buys milk (= X) and bread simultaneously will buy an egg (= Y) with high probability

**Support**

\[
support(X) = \frac{\text{# of data that satisfy } X}{\text{total # of data}} = Pr[X]
\]

**Confidence**

\[
conf(X, Y) = \frac{\text{# of data that satisfy both } X \text{ and } Y}{\text{# of data that satisfy } X} = Pr[Y | X]
\]
Taxonomy of Formal Fairness
Properties of Formal Fairness

Bias Type
- Which type of biases should be removed

Direct fairness — Indirect fairness
- Sensitive information influences targets directly, or indirectly

Disparate treatment — Disparate Impact
- Groups or individuals are intentionally treated differently, or
- Unintentional impact on distinct groups or individuals

Group Fairness — Individual Fairness
- Fairness for each group, OR fairness for each individual

Symmetric Property
- Protected and non-protected groups are treated symmetrically

Explainable Variables
- Excluding the effect of confounding factors between sensitives and targets
Three types of biases that undesirably corrupt outcomes

- **Annotation Bias**: unfair labeling by annotators; inappropriately observed feature values
- **Sample Selection Bias**: dataset that is not a representative of population to analyze
- **Inductive Bias**: propensity of ML algorithms caused by assumptions in the algorithms’ inductive process

Sources of undesired outcomes depends on problems

Formal fairness have to be selected by considering which type of biases tries to be removed
**Removing Annotation Bias**

**Annotation Bias**: Target values or feature values in a training data are biased due to annotator’s cognitive bias or inappropriate observation schemes.

Annotations are not reliable, and never accessible to a correct dataset.

Assumptions about the conditions that values or distributions of target variables and sensitive features should satisfy.

**Examples of assumptive conditions:**
- $Y \indep S \mid X=x$: $Y$ and $S$ are context-sensitive independent given $X=x$
- $Y \indep S \mid X$: $Y$ and $S$ are conditionally independent given $X$
- $Y \indep S$: $Y$ and $S$ are (unconditionally) independent
Removing annotation bias: $\hat{Y} \perp S$

Ratios between positives and negatives in prediction are matched

[Calders+ 10, Dwork+ 12]
Sample Selection Bias: Whether a datum is sampled depends on conditions or contents of the datum, and thus an observed dataset is not a representative of population.

Batch Learning: Training data violates a condition of random assignment in terms of sensitive information
- incorrectly annotated by an ML algorithm
  ➔ modify an inductive bias of the ML algorithm
- not sampled uniformly at random, as seen in a statistical survey
  ➔ modify data so as to satisfy a condition of random assignment

Online Learning: Selection of data to probe is biased in an ML tasks with a feedback loop, e.g., bandits, reinforcement learning, active learning
- biased selection of data to probe or investigate
  ➔ select randomly in terms of sensitive information
**Fair Bandit**

**Bandit problem**: maximize the cumulative rewards of selected arms

- **reward of the selected arm**
- **select one of arms**

If an arm that is selected initially returns a high-reward by chance, the other arms can be less frequently selected.

**original UCB**
- always select the arm whose upper confidence bound is the maximum
- deterministically select

**fair UCB**
- select arms whose confidence intervals overlap with equal prob.
- select with equal probability
Removing Inductive Bias

**Inductive Bias:** a bias caused by an assumption adopted in an inductive machine learning algorithms

Outcomes in a training dataset, $Y$, are assumed to be reliable, and are changed to predicted outcomes, $\hat{Y}$, by an induction process.

The changes from $Y$ to $\hat{Y}$ are balanced between sensitive groups defined by $S$.

$\hat{Y} \perp S \mid Y$: $\hat{Y}$ and $S$ are conditionally independent given $Y$.

Empirical errors of $\hat{Y}$ over sample outcomes, $Y$, are equal for all groups consist of the same sensitive values.
Removing Inductive Bias

Removing inductive bias: \( \hat{Y} \perp S \mid Y \)

Ratios between true and predicted positives are matched

\[
\begin{array}{c|c|c|c}
S = 0 & \hat{Y} = 0 & \hat{Y} = 1 & Y = 0 \\
 & Y = 0 & Y = 1 & \\
S = 1 & \hat{Y} = 0 & \hat{Y} = 1 & Y = 0 \\
 & Y = 0 & Y = 1 &
\end{array}
\]
Direct Fairness / Indirect Fairness

- **Direct Fairness**: A target determination does not directly depend on a sensitive information.

- **Indirect Fairness**: A target determination does not only directly also indirectly depend on a sensitive feature.

**Direct Unfairness**: a target variable directly depends on a sensitive feature

A target ‘loan’ more frequently becomes ‘deny’ if a sensitive ‘race’ is ‘African’

\[ \Pr[\text{loan}=\text{deny} \mid \text{city}=\text{NYC}, \text{race}=\text{African}] \gg \Pr[\text{loan}=\text{deny} \mid \text{city}=\text{NYC}] \]
**Indirect Unfairness**: a target variable depends on a sensitive feature through a non-sensitive feature

A target ‘loan’ doesn’t directly depend on a sensitive ‘race’

\[ \Pr[ \text{loan}=\text{deny} \mid \text{city}=\text{NYC}, \text{ZIP}=10451 ] \gg \Pr[ \text{loan}=\text{deny} \mid \text{city}=\text{NYC} ] \]

‘loan=deny’ and ‘ZIP=10451’ are highly co-occurred

\[ \Pr[ \text{race}=\text{African} \mid \text{city}=\text{NYC}, \text{ZIP}=10451 ] \sim \text{high} \]
\[ \Pr[ \text{ZIP}=10451 \mid \text{city}=\text{NYC}, \text{race}=\text{African} ] \sim \text{high} \]

a target ‘loan’ in directly depends on a sensitive ‘race’

*Redescription*: the same set of objects are described by two different formulae or descriptions

Ex. A literal ‘city=NYC \land ZIP=10451’ is a redescription of ‘city=NYC \land race=African’
Disparate Treatment / Disparate Impact: Legal term about discrimination
While disparate treatment is intentional discrimination, disparate impact is unintentional one

From a formal viewpoint:

- **Disparate Treatment**: a value of target variable changes when its corresponding sensitive value changes
- **Disparate Impact**: distribution of target variables are different between sensitive groups
Group Fairness / Individual Fairness

- **Group Fairness**: individuals are fairly treated as a group
- **Individual Fairness**: each individual is fairly treated

* A protected group and a non-protected group consist of individuals such that $S=0$ and $S=1$, respectively

**Group Fairness**: Distributions of a target variable are equal for all possible sensitive groups

$$\Pr[Y | S] = \Pr[Y] \Rightarrow Y \perp S$$

- The effect of non-sensitive features are canceled by a random assignment due to the marginalization over each group
  - Achieving group fairness is much easier than individual fairness

- **Reverse Tokenism**: justify unfair treatment for members of a protected group by sacrificing a few superior members of a non-protected group
  - This cannot be prevented by achieving group fairness

[Dwork+ 12]
**Individual Fairness**: Distributions of a target variable are equal for all possible sensitive groups given a specific non-sensitive values

\[
\Pr[Y | S, X=x] = \Pr[Y | X=x] \Rightarrow Y \perp S | X=x
\]

- **The effect of non-sensitive features must be considered**
  - It is difficult to strictly satisfy a condition of group fairness

- **Situation Testing**: Legal notion of testing discrimination, comparing individuals having the same non-sensitive values except for a sensitive value

  [Luong+ 11]

**Difficulty**: hard to find individuals having the same non-sensitive values, but different sensitive values in a dataset \( \Rightarrow \) approximation

- accept the difference within a constant
  - [Dwork+ 12]

- use an aggregated property of neighbors
  - [Luong+ 11, Zhang+ 16]

- consider a upper bound of differences between sensitive groups, instead of their expectation
  - [Fukuchi+ 13]


Formal fairness or fairnes metric is **symmetric** if protected and non-protected groups are considered simultaneously.

If protected individuals are probed whether they are unfairly treated, but non-protected individuals are not, the formal fairness is **asymmetric**.

- **CV score**: $\text{Pr}[Y=1 \mid S=1] - \text{Pr}[Y=1 \mid S=0]$  
  This fairness metric is **asymmetric**.
  - This probes only whether protected individuals are unfairly treated.
  - Absolute of the CV score can be considered as symmetric metric.

- **Equilized Odds / Equal Opportunity**:  
  $\text{Pr}[\hat{Y}=1 \mid S=0, Y=y] = \text{Pr}[\hat{Y}=1 \mid S=0, Y=y]$  
  Equilized odds is **symmetric** formal fairness, because it is designed to consider both $y=0$ and $1$ cases.
  However, equal opportunity is **asymmetric** due to the ignorance of the $y=1$ case.
**Explainable Variable**: these variables influence both target and sensitive variables, and the influence is not semantically problematic.

In FADM, we are interested in the **pure effect** from a sensitive feature to a target **excluding the spurious effect of an explainable variable**.

**genuine occupational requirement**: the nature of the role makes it unsuitable for individuals with a particular sensitive value.

Ex. Fashion model for feminine clothes should be female.
In a context of causal inference, explainable variables can be considered as **confounding variables** that are semantically or legally explainable.

The effect of $E$ is removed by making $Y$ and $S$ **strongly ignorable**:

$$Y \perp S \mid E$$

**Techniques of causal inference are applicable**

**Ex.** randomization, propensity score
Propensity Score

**Propensity Score**: probability to be a protected group given an explainable values, $e(S) = \Pr[S=0 \mid E]$ 

Propensity score can be used for eliminating the effects of explainable variables due to its **balancing property**: $S \perp E \mid e(S)$

If $S$ is strongly ignorable given explainable variables, $S$ is strongly ignorable given a propensity score:

$$Y \perp S \mid E \implies Y \perp S \mid e(S)$$

The effect of explainable variables is removed by dividing a dataset into strata in which propensity scores are similar

<table>
<thead>
<tr>
<th>strata 1</th>
<th>$e(S) \in [0, 1/3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>strata 2</td>
<td>$e(S) \in [1/3, 2/3)$</td>
</tr>
<tr>
<td>strata 3</td>
<td>$e(S) \in [2/3, 1]$</td>
</tr>
</tbody>
</table>

[Calders+ 13]
Is the target determination fair in terms of a sensitive state

An example of university admission in [Žliobaitė+ 11]

**Sensitive feature:** $S$
- gender
  - male / female

**Target variable:** $Y$
- acceptance
  - accept / not accept

**Fair determination:** the gender does not influence the acceptance

**Statistical independence:** $Y \perp S$
Information-Theoretic Interpretation

Information theoretical view of a fair determination

Sensitive: $S$

$H(S | Y)$

$H(S)$

Target: $Y$

$H(Y)$

$I(S; Y)$

$H(Y | S)$

Statistical independence between $S$ and $Y$ implies zero mutual information: $I(S; Y) = 0$

The degree of influence $S$ to $Y$ can be measured by $I(S; Y)$
Causality with Explainable Features

An example of fair determination even if $S$ and $Y$ are not independent

**Sensitive feature:** $S$
- gender
  - male / female

**Target variable:** $Y$
- acceptance
  - accept / not accept

**Explainable feature:** $E$
- program
  - medicine / computer

Female $\rightarrow$ medicine $=$ high
Male $\rightarrow$ computer $=$ high

Removing the pure influence of $S$ to $Y$, excluding the effect of $E$

$Y \perp S \mid E$
Information Theoretic Interpretation

Sensitive: $S$

- $H(S \mid Y, E)$
- $I(S; E \mid Y)$
- $H(S)$

Explainable: $E$

- $I(S; Y; E)$
- $I(S; Y \mid E)$
- $I(Y; E \mid S)$
- $H(E \mid S, Y)$
- $H(E)$

Target: $Y$

- $H(Y)$
- $H(Y \mid S, E)$

the degree of conditional independence between $Y$ and $S$ given $E$

conditional mutual information: $I(S; Y \mid E)$

We can exploit additional information $I(S; Y; E)$ to obtain outcomes
Fairness Indexes
Calders-Verwer discrimination score (CV score)

$$\Pr[ Y=1 \mid S=1 ] - \Pr[ Y=1 \mid S=0 ]$$

The conditional probability of the advantageous decision for non-protected members subtracted by that for protected members

The condition of zero CV score implies unconditional independence between a target, $Y$, and a sensitive, $S$

$$\Pr[ Y \mid S ] = \Pr[ Y ]$$

$Y \perp S$

Sensitive information doesn’t influence outcomes
The condition elift = 1 means that no unfair treatments, and it implies

\[ \Pr[ Y=0 \mid S=0, X=x ] = \Pr[ Y=0 \mid X=x ] \]

when \( S \) and \( Y \) are additionally binary variables,

This condition is equivalent to the context-sensitive independence:

\[ Y \perp S \mid X=x \]

Useful for finding unfair effects from \( S \) to \( Y \) under the context of \( X=x \)
Formal Tasks of Fairness-Aware Data Mining
Formal Tasks of Fairness-Aware DM

Fairness-aware DM

Unfairness Prevention
predictor or transformation leading fair outcomes

Unfairness Discovery
finding unfair treatments

Taxonomy by Process
pre-process, in-process, post-process

Discovery from Datasets
finding unfair data or subgroups in a dataset

Taxonomy by Tasks
classification, regression, recommendation, etc…

Discovery from Models
finding unfair outcomes of a blackbox model

[Ruggieri+ 10]
Unfairness Discovery from Datasets: Find personal records or subgroups that are unfairly treated from a given dataset

Research Topics
- Definition of unfair records or subgroups in a dataset
- Efficiently searching patterns in the combinations of feature values
- How to deal with explainable variables
- Visualization of discovered records or subgroups
Unfairness Discovery from Models: When observing outcomes from a specific black-box model for personal records or subgroups, checking fairness of the outcomes.

Research Topics:
- Definition of unfair records or subgroups in a dataset
- Assumption on a set of black-box models
- How to generate records to test a black-box model
Unfairness Prevention: Pre-Process Approach

**Pre-Process:** potentially unfair data are transformed into fair data ①, and a standard classifier is applied ②.

- Any classifier can be used in this approach.
- The development of a mapping method might be difficult without making any assumption on a classifier.
Unfairness Prevention: In-Process Approach

**In-Process**: a fair model is learned directly from a potentially unfair dataset 🇨

- This approach can potentially achieve better trade-offs, because classifiers can be designed more freely
- It is technically difficult to formalize an objective function, or to optimize the objective function.
- A fair classifier must be developed for each distinct type of classifier
Unfairness Prevention: Post-Process Approach

Post-Process: a standard classifier is first learned $\odot$, and then the learned classifier is modified to satisfy a fairness constraint $\ominus$

- This approach adopts the rather restrictive assumption, obliviousness [Hardt+ 16], that fair class labels are determined based only on labels of a standard classifier and a sensitive value.
- This obliviousness assumption makes the development of a fairness-aware classifier easier.
Unfairness Prevention: Taxonomy by Tasks

- **Classification**
  - **Pre-Process** [Kamiran+ 12], [Dwork+ 12], [Hajian+ 13], [Zemel+ 13], [Mancuhan+ 14], [Feldman+ 15]
  - **In-Process** [Kamiran+ 10], [Kamishima+ 12], [Fukuchi+ 13], [Fukuchi+ 14], [Zafar+ 15], [Zafar+ 17]
  - **Post-Process** [Calders+ 10], [Kamiran+ 12], [Hardt+ 16]

- **Regression** [Fukuchi+ 13], [Calders+ 13]

- **Clustering** [Gondek+ 04], [Gondek+ 05]

- **Recommendation** [Kamishima+12, Kamishima+ 13] [Kamishima+ 16] [Kamishima+ 17]

- **Dimension Reduction** [Bolukbasi+ 16]
Unfairness Detection
**α-protection**

Association rules extracted from a data set

(a) **city=NYC ⇒ class=bad** (conf=0.25)
0.25 of NY residents are denied their credit application

(b) **city=NYC ∧ race=African ⇒ class=bad** (conf=0.75)
0.75 of NY residents whose race is African are denied their credit application

**extended lift (elift)**

\[
elift = \frac{\text{conf} ( A \land B \Rightarrow C )}{\text{conf} ( A \Rightarrow C )}
\]

the ratio of the confidence of a rule with additional condition to the confidence of a base rule

**α-protection**: considered as unfair if there exists association rules whose elift is larger than \( \alpha \)

**ex**: rule (b) isn’t \( \alpha \)-protected if \( a = 2 \), because \( \text{elift} = \frac{\text{conf}(b)}{\text{conf}(a)} = 3 \)
Situation Testing: When all the conditions are same other than a sensitive condition, people in a protected group are considered as unfairly treated if they received unfavorable decision.

- the statistics of decisions in $k$-nearest neighbors of data points in a protected group
- Condition of situation testing is $\Pr[ Y \mid \mathbf{E}, S=0 ] = \Pr[ Y \mid \mathbf{E}, S=1 ] \equiv Y \perp S \mid \mathbf{E} \ (\mathbf{E} : \text{explainable variables})$
Unfairness Prevention: Classification
Red-Lining Effect: Simple elimination of a sensitive features from training dataset fails to remove the influence of sensitive information to a target

- \( \Pr[ Y \mid X, S ] \): A model trained from a dataset with both sensitive and non-sensitive features
- \( \Pr[ Y \mid X ] \): A model that does not depend on S by eliminating a sensitive feature from a training dataset

\[
\Pr[ Y, X, S ] = \Pr[ Y \mid X, S ] \Pr[ S \mid X ] \Pr[ X ] \Rightarrow \Pr[ Y \mid X ] \Pr[ S \mid X ] \Pr[ X ]
\]

This is a condition \( Y \perp S \mid X \) (not \( Y \perp S \))

\( S \) still influences \( Y \) through \( X \)
Massaging: Pre-process type method

- A standard classifier is once applied, and class labels are modified so as to be balanced between sensitive groups
- Finally, a standard classifier is trained from the modified dataset

1. A standard classifier is applied, and training data are sorted according to the degree to be a positive class for each sensitive group

   non-protected $S=1$
   - - + + + + + + + + + +

   protected $S=0$
   - - - - - - + + + + +

2. Class labels are modified so that ratios of a positive class are balanced between sensitive groups

3. A final classifier is trained from the modified training dataset
Dwork’s Method

- **data owner**: loss function representing utilities for the vendor
- **vendor (data user)**: original data
- **archetype**: min loss function s.t. fairness constraint
- **fair decisions**: pre-process, individual fairness, removing annotation bias

**Fairness Constraint:**
- **Lipschitz condition**: similar original data are mapped to similar archetypes
- **Statistical Parity**: protected and non-protected groups are equally treated

[Dwork+ 12]
**Prejudice Remover Regularizer**

**Prejudice Remover**: a regularizer to impose a constraint of independence between a target and a sensitive feature, \( Y \perp S \)

The objective function is composed of classification loss and fairness constraint terms

\[
- \sum_{D} \ln \Pr[Y \mid X, S; \Theta] + \frac{\lambda}{2} \|\Theta\|_2^2 + \eta I(Y; S)
\]

- A class distribution, \( \Pr[Y \mid X, S; \Theta] \), is modeled by a set of logistic regression models, each of which corresponds to \( s \in \text{Dom}(S) \)
  \[
  \Pr[Y=1 \mid x, s] = \text{sig}(w^{(s)} \mid x)
  \]
- As a prejudice remover regularizer, we adopt a mutual information between a target and a sensitive feature, \( I(Y; S) \)
Discrimination-Aware Decision Tree

A decision tree maximizing the prediction performance under the fairness constraint

- in-process, annotation bias, group fairness

**Information gain**: at the internal nodes, a training dataset is divided by a feature maximizing information gain in terms of a sensitive feature, $S$, as well as a class variable, $Y$

- Every leave nodes contain data whose values of sensitive and class variables are the same

**Labels in leaves**

- **standard DT**: output labels of a majority class in a leave node
- **discrimination-aware DT**: class labels are inverted so as to improve fairness

- The evaluation measure for selecting leaves in which class labels are inverted

  \[ \text{[fairness index]} / \text{[decrease in accuracy]} \]
Unfair decisions are modeled by introducing the dependence of $\mathbf{X}$ on $S$ as well as on $Y$.

- $S$ and $\mathbf{X}$ are conditionally independent given $Y$.
- Non-sensitive features in $\mathbf{X}$ are conditionally independent given $Y$ and $S$.

* It is as if two naive Bayes classifiers are learned depending on each value of the sensitive feature; that is why this method was named by the 2-naive-Bayes.
Calders-Verwer’s 2-Naive-Bayes

\[ \hat{Pr}[Y, X, S] = \hat{Pr}[Y, S] \prod_i \hat{Pr}[X_i|Y, S] \]

\( \hat{Pr}[Y, S] \) is modified so as to improve the fairness

Estimated model \( \hat{Pr}[Y, S] \) \quad \text{fair} \quad \text{fair estimated model} \( \hat{Pr}^{\circ}[Y, S] \)

keep the updated marginal distribution close to the \( \hat{Pr}[Y] \)

while \( \Pr[Y=1 | S=1] - \Pr[Y=1 | S=0] > 0 \)

if \# of data classified as “1” < # of “1” samples in original data then
increase \( \Pr[Y=1, S=0] \), decrease \( \Pr[Y=0, S=0] \)
else
increase \( \Pr[Y=0, S=1] \), decrease \( \Pr[Y=1, S=1] \)
reclassify samples using updated model \( \Pr[Y, S] \)

update the joint distribution so that its fairness is enhanced
Even if $Y$ and $S$ are independent, actual class labels may not satisfy a fairness constraint

**Deterministic decision rule**: Class labels are generated not probabilistically, but deterministically by a decision rule

**Difference**: $\Pr[Y, S] - \Pr[Y] \Pr[S]$

**Always Independent**
Labels probabilistically generated according to $\Pr[Y] \Pr[S] \Pr[X \mid Y, S]$

**Not Independent in general**
Bayes optimal Labels are generated by a deterministic decision rule:
$y^* \leftarrow \arg \max_y \Pr[y \mid x, s]$

**Model bias**: Models doesn’t contain true distribution to learn in general
**Model-based Independence** : Class labels are assumed to be generated probabilistically

\[ \tilde{Y} \perp S, \text{ where } (\tilde{Y}, S) \sim \Pr^o[\tilde{Y}, S] \]

**Actual Independence** : Class labels are assumed to be deterministically generated by applying a decision rule

\[ \tilde{Y} \perp S, \text{ where } (\tilde{Y}, S) \sim \hat{\Pr}^o[\tilde{Y}, S] = \frac{1}{n} \sum_{x \in D_s} \hat{\Pr}[\tilde{Y}|x, s] \]

\[ \begin{cases} 
\hat{\Pr}[\tilde{y}|x, s] = 1, & \text{if } \tilde{y} = \arg \max_y \hat{\Pr}[y|x, s] \\
\hat{\Pr}[\tilde{y}|x, s] = 0, & \text{otherwise}
\end{cases} \]

satisfy actual independence instead of model-based independence

**Fairness in class labels can be greatly improved**
Unfairness Prevention: Recommendation
Independence Enhanced Recommendation

Predicting Ratings: a task to predict a rating value that a user would provide to an item

Dataset
\[ D = \{(x_i, y_i, r_i, s_i)\} \]

Prediction Function
\[ \hat{r}(x, y, s) \]

Fairness constraint: \( R \perp S \)

Finding Good Items: Another type of recommendation task to find at least one item would be preferred by a user
Probabilistic Matrix Factorization Model
predict a preference rating of an item \( y \) rated by a user \( x \)
well-performed and widely used

**Prediction Function**

\[
\hat{r}(x, y) = \mu + b_x + c_y + p_x q_y^T
\]

- global bias
- cross effect of users and items
- user-dependent bias
- item-dependent bias

**Objective Function**

\[
\sum_D (r_i - \hat{r}(x_i, y_i))^2 + \lambda \| \Theta \|^2
\]

- squared loss function
- L2 regularizer
- regularization parameter

For a given training dataset, model parameters are learned by minimizing the squared loss function with an L2 regularizer.
Independence Enhanced PMF

Prediction Function

A prediction function is selected according to a sensitive value

\[ \hat{r}(x, y, s) = \mu^{(s)} + b_x^{(s)} + c_y^{(s)} + p_x^{(s)} q_y^{(s)} \top \]

Objective Function

**Independence parameter**: control the balance between the independence and accuracy

\[ \sum_D (r_i - \hat{r}(x_i, y_i))^2 - \eta \text{indep}(R, S) + \lambda \| \Theta \|^2 \]

**Independence term**: a regularizer to constrain independence

- The larger value indicates that ratings and sensitive values are more independent
- Matching means of predicted ratings for two sensitive values

[Kamishima+ 12, Kamishima+ 13]
Latent Class Model: A probabilistic model for collaborative filtering

Latent topic variable

A basic topic model, pLSA

extended so as to be able to deal with ratings $r$ given by users $x$ to items $y$

Model parameters can be learned by an EM algorithm

Prediction:

$$\hat{r}(x, y) = E_{Pr[r|x, y]}[level(r)] = \sum_r Pr[r|x, y] level(r)$$

the $r$-th rating value

A rating value can be predicted by the expectation of ratings

[Hofmann 99]
Independence-Enhancement by a Model-based Approach

A sensitive variable is embedded into the original LCM

- A rating and a sensitive variable are mutually independent
- A user, an item, and a rating are conditionally independent given $Z$

Type 1 model

Type 2 model

Experimental results show that the performance of these two models are nearly equal
Relation to Other Research Topics
Privacy-Preserving Data Mining

Fairness in Data Mining
the independence between an objective $Y$ and a sensitive feature $S$

from an information theoretic perspective,
mutual information between $Y$ and $S$ is zero: $\text{I}(Y; S) = 0$

from the viewpoint of privacy-preservation,
protection of sensitive information if an objective is exposed

Different points from PPDM
• introducing randomness is occasionally inappropriate for severe decisions, such as job application
• disclosure of identity isn’t problematic in FADM, generally
Causal Inference

Causal inference: a general technique for dealing with probabilistic causal relations

FADM and causal inference can commonly cope with the influence of sensitive information / cause to a target / outcome

Causal Inference

- **Objective**: estimating the degree of the effect between variables
- **Approach**: based on a causal graph representing strict causal relationships

Fairness-aware Data Mining

- **Objective**: finding or removing pure effect from a sensitive to a target
- **Approach**: Light-weight approach ignoring information irrelevant to a sensitive feature
Other Relationships with Causal Inference

**bounds of probability of a counterfactual case** [Pearl 2009]

- An individual of $S=1$ actually receives $Y=1$
- If $S$ was changed to 0, how was the probability of receiving $Y=0$?
- Under the condition called exogenous, the range of this probability is

$$\max\{0, \Pr[Y=1|S=1] - \Pr[Y=1|S=0]\} \leq \text{PNS} \leq \min\{\Pr[Y=1,X=1], \Pr[Y=0,X=0]\}$$

This lower bound coincides with a CV score

- The use of a propensity score to remove the influence of an explainable variable [Calders+ 13]
- Using a similarity measure where the effect of changing the feature values is considered based on a causal graph [Zhang+ 16]
Cost-Sensitive Learning: learning classifiers so as to optimize classification costs, instead of maximizing prediction accuracies

FADM can be regarded as a kind of cost-sensitive learning that pays the costs for taking fairness into consideration

**Cost matrix** $C(i \mid j)$: cost if a true class $j$ is predicted as class $i$

Total cost to minimize is formally defined as (if class $Y = 1$ or $0$):

$$\mathcal{L}(x, i) = \sum_j \Pr[j \mid x] C(i \mid j)$$

An object $x$ is classified into the class $i$ whose cost is minimized
Theorem 1 in [Elkan 2001]
If negative examples in a data set is over-sampled by the factor of

\[
\frac{C(1|0)}{C(0|1)}
\]

and a classifier is learned from this samples, a classifier to optimize specified costs is obtained.

In a FADM case, an over-sampling technique is used for avoiding unfair treatments.

A corresponding cost matrix can be computed by this theorem, which connects a cost matrix and the class ratio in training data.

* Note that this over-sampling technique is simple and effective for avoiding unfair decisions, but its weak point that it completely ignores non-sensitive features.
Other Connected Techniques

Legitimacy / Leakage
- Data mining models can be deployed in the real world

Independent Component Analysis
- Transformation while maintaining the independence between features

Surrogate Data
- To perform statistical tests, specific information is removed from data sets

Dummy Query
- Dummy queries are inputted for protecting users’ demographics into search engines or recommender systems

Visual Anonymization
- To protect identities of persons in images, faces or other information is blurred
Applications
Recidivism Risk Score

- **COMPAS** (Correctional Offender Management Profiling for Alternative Sanctions) developed by Northpointe, used in many states
- Evaluate the re-offending risk by a ten-point-scale
- Judges are given the scores in the process of pretrial release, or even in sentencing to parole

**Merits and Concerns**

- For two centuries, key decisions in the legal process have been affected by personal biases
- Scores can be exploited not for the designed purposes
  - Judges in Wisconsin used the score for sentencing, though it is designed for the probation decision
- **Scores must accurately predict which defendants likely to re-offend**
Defendants of African descents were often predicted to be more risky than they actually were, and vice versa.  

* FPR (false positive ratio) = ratios of # of actually non-recidivated to # of people predicted to recidivate
Restrictions by Regulations or Laws

Quantitative restrictions by regulations or laws against discrimination:

**Anti-Discrimination Act (Australia, Queensland)**
- a person treats, or proposes to treat, a person with an attribute less favorably than another person without the attribute

**Racial Equality Directive (EU)**
- shall be taken to occur where one person is treated less favorably than another is in a comparable situation on grounds of racial or ethnic origin

**Uniform Guidelines on Employee Selection Procedure (US)**
- a selection rate for any race, sex, or ethnic group which is less than four-fifths (or eighty percent) of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact
Non-Redundant Clustering

non-redundant clustering: find clusters that are as independent from a given uninteresting partition as possible

clustering facial images

- A simple clustering method finds two clusters: one contains only faces, and the other contains faces with shoulders
- A data analyst considers this clustering is useless and uninteresting
- By ignoring this uninteresting information, more meaningful female- and male-like clusters could be obtained

The influence of uninteresting information can be ignored
System managers should fairly treat their content providers

Fair treatment in search engines

- The US FTC has investigated Google to determine whether the search engine ranks its own services higher than those of competitors

Fair trading in recommendation

- E-commerce platform sites should not abuse their position to recommend their own items more frequently than tenants' items
- In sites comparing hotel booking or insurance products, or in sites for recruiting employees, all of which will want to treat their information suppliers fairly
Filter Bubble: To fit for Pariser’s preference, conservative people are eliminated from his friend recommendation list in FaceBook.

If a political conviction of a friend candidate doesn’t influence who will be recommended, Information about whether a candidate is conservative or progressive can be ignored in a recommendation process.
RecSys 2011 Panel on Filter Bubble

- Are there “filter bubbles?”
- To what degree is personalized filtering a problem?
- What should we as a community do to address the filter bubble issue?


Intrinsic trade-off

- providing a diversity of topics
- focusing on users’ interests

To select something is not to select other things
Personalized filtering is a necessity

Personalized filtering is a very effective tool to find interesting things from the flood of information.

recipes for alleviating undesirable influence of personalized filtering

- capture the users’ long-term interests
- consider preference of item portfolio, not individual items
- follow the changes of users’ preference pattern
- give users to control perspective to see the world through other eyes
Fairness-Aware Data Mining

http://www.kamishima.net/fadm/
Auditing black-box models for indirect influence.
In Proc. of the 16th IEEE Int’l Conf. on Data Mining, pages 1–10, 2016.

R. Agrawal and R. Srikant.
Fast algorithms for mining association rules.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner.
Machine bias, 2016.

S. Athey.
Beyond prediction: Using big data for policy problems.

A. Barr.
Google mistakenly tags black people as ‘gorillas,’ showing limits of algorithms.
⟨http://on.wsj.com/1CaCNlb⟩.
B. Berendt and S. Preibusch.
Exploring discrimination: A user-centric evaluation of discrimination-aware data mining.
In Proc. of the IEEE Int’l Workshop on Discrimination and Privacy-Aware Data Mining,

B. Berendt and S. Preibusch.
Better decision support through exploratory discrimination-aware data mining:
Foundations and empirical evidence.

Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

C. Boutiller, N. Friedman, M. Goldszmidt, and D. Koller.
Context-specific independence in Bayesian networks.
In Uncertainty in Artificial Intelligence 12, pages 115–123, 1996.

Controlling attribute effect in linear regression.

T. Calders and S. Verwer.
Three naive Bayes approaches for discrimination-free classification.
Ò. Celma and P. Cano.
From hits to niches?: or how popular artists can bias music recommendation and discovery.

Is seeing believing? how recommender interfaces affect users’ opinions.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel.
Fairness through awareness.

C. Elkan.
The foundations of cost-sensitive learning.

Certifying and removing disparate impact.
In Proc. of the 21st ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining, pages 259–268, 2015.
D. Fleder and K. Hosanagar.
Recommender systems and their impact on sales diversity.

S. Forden.
Google said to face ultimatum from FTC in antitrust talks.
Bloomberg, Nov. 13 2012.
⟨http://bloom.bg/PPNEaS⟩.

K. Fukuchi and J. Sakuma.
Neutralized empirical risk minimization with generalization neutrality bound.

Prediction with model-based neutrality.
[LNCS 8189].

B. Gao and B. Berendt.
Visual data mining for higher-level patterns: Discrimination-aware data mining and beyond.
D. Gondek and T. Hofmann.
Non-redundant data clustering.
In *Proc. of the 4th IEEE Int’l Conf. on Data Mining*, pages 75–82, 2004.

D. Gondek and T. Hofmann.
Non-redundant clustering with conditional ensembles.
In *Proc. of the 11th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining*, pages 70–77, 2005.

S. Hajian and J. Domingo-Ferrer.
A methodology for direct and indirect discrimination prevention in data mining.

S. Hajian, J. Domingo-Ferrer, and O. Farràs.
Generalization-based privacy preservation and discrimination prevention in data publishing and mining.
*Data Mining and Knowledge Discovery*, 2014.

M. Hardt, E. Price, and N. Srebro.
Equality of opportunity in supervised learning.

T. Hofmann and J. Puzicha.
Latent class models for collaborative filtering.
M. Joseph, M. Kearns, J. Morgenstern, and A. Roth.
Fairness in learning: Classic and contextual bandits.

F. Kamiran and T. Calders.
Data preprocessing techniques for classification without discrimination.

F. Kamiran, T. Calders, and M. Pechenizkiy.
Discrimination aware decision tree learning.

Decision theory for discrimination-aware classification.
In *Proc. of the 12th IEEE Int’l Conf. on Data Mining*, pages 924–929, 2012.

F. Kamiran, I. Žliobaitė, and T. Calders.
Quantifying explainable discrimination and removing illegal discrimination in automated decision making.

T. Kamishima and S. Akaho.
Considerations on recommendation independence for a find-good-items task.
Enhancement of the neutrality in recommendation.
In *The 2nd Workshop on Human Decision Making in Recommender Systems*, 2012.

Fairness-aware classifier with prejudice remover regularizer.

Efficiency improvement of neutrality-enhanced recommendation.

Model-based and actual independence for fairness-aware classification.
*Data Mining and Knowledge Discovery*, 32:258–286, 2018.

Model-based approaches for independence-enhanced recommendation.
In *Proc. of the IEEE 16th Int’l Conf. on Data Mining Workshops*, pages 860–867, 2016.
T. Kamishima and K. Fukuchi.
Future directions of fairness-aware data mining: Recommendation, causality, and theoretical aspects.

Y. Koren.
Factorization meets the neighborhood: A multifaceted collaborative filtering model.
In *Proc. of the 14th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining*, pages 426–434, 2008.

B. T. Luong, S. Ruggieri, and F. Turini.
k-NN as an implementation of situation testing for discrimination discovery and prevention.
In *Proc. of the 17th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining*, pages 502–510, 2011.

K. Mancuhan and C. Clifton.
Combating discrimination using bayesian networks.

P. Miettinen and E. Galbrun.
An introduction to redescription mining.
E. Pariser.
*The Filter Bubble: What The Internet Is Hiding From You.*
Viking, 2011.

D. Pedreschi, S. Ruggieri, and F. Turini.
Discrimination-aware data mining.
In *Proc. of the 14th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining*, pages 560–568, 2008.

D. Pedreschi, S. Ruggieri, and F. Turini.
Measuring discrimination in socially-sensitive decision records.

P. Resnick, J. Konstan, and A. Jameson.
Panel on the filter bubble.
The 5th ACM Conf. on Recommender Systems, 2011.
⟨http://acmrecsys.wordpress.com/2011/10/25/panel-on-the-filter-bubble/⟩.

Anti-discrimination analysis using privacy attack strategies.
[LNCS 8725].
S. Ruggieri, D. Pedreschi, and F. Turini.
Data mining for discrimination discovery.  
*ACM Transactions on Knowledge Discovery from Data*, 4(2), 2010.

R. Salakhutdinov and A. Mnih.
Probabilistic matrix factorization.  

The bandwagon effect of collaborative filtering technology.  

L. Sweeney.
Discrimination in online ad delivery.  

B. Zadrozny.
Learning and evaluating classifiers under sample selection bias.  

M. B. Zafar, I. V. Martinez, M. G. Rodriguez, and K. Gummadi.
Fairness constraints: A mechanism for fair classification.  
Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork.
Learning fair representations.

L. Zhang, Y. Wu, and X. Wu.
Situation testing-based discrimination discovery: A causal inference approach.

I. Žliobaitė.
On the relation between accuracy and fairness in binary classification.

I. Žliobaitė, F. Kamiran, and T. Calders.
Handling conditional discrimination.
In Proc. of the 11th IEEE Int’l Conf. on Data Mining, 2011.