Fairness-Aware Machine Learning and Data Mining

Toshihiro Kamishima
http://www.kamishima.net/
Updated: 2020-03-18
Data analysis taking into account potential issues of fairness, discrimination, neutrality, or independence. It maintains the influence of these types of sensitive information:

- to enhance 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.
FAML was originally invented to eliminate socially discriminative outcomes when applying ML techniques to real-world problems.

More extensively, FAML 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, FAML methods can be used for predicting a pure influence from room charges to occupancy rates.
Growth of Fairness in ML

BRIEF HISTORY OF FAIRNESS IN ML

PAPERS


LOL FAIRNESS!!

OH, CRAP.
Sources of Unfairness in Machine Learning
Bias on the Web

[Figure 1: The vicious cycle of bias on the Web.]

Here, I explain each of the biases (in both the growth of the Web and its use.

[Figure 2: Shame effect (line with small trend direction) vs. minimal effort (notable trend direction) on number of links on U.K. webpages, with intersection between 12 and 13 links. Data at far right is probably due to pages having been written by people without Internet access.]

Consider the following survey of random selection of people. I found that the 0.05% of the most popular people attracted almost 50% of the followers. In 2011, a study by Wu et al. determined that 0.04% of the most popular Twitter users in the dataset were followed by only a few select celebrities. I was surprised to see that a month after publication of our results, my hunch was confirmed. In October 2015, Amazon began a corporate campaign against paid fake reviews that had been costing the company hundreds of millions of dollars a year. Saez-Trumper and I analyzed four datasets, and as I detail, the correlation between the average helpfulness of each review according to users and a proxy of text quality demonstrates the wisdom of the overall crowd.

Activity bias

Second-order bias

Self-selection bias

Interaction bias

Sampling bias

Algorithmic bias

Algorithm

Screen

Data bias

Sample selection bias

Inductive bias

Data at far right is probably due to pages having been written by people without Internet access. This is an illusion.

In light of such findings, a second-order of activity bias can be seen in most Web measures. The majority of Web users who only do a lot, possibly helping explain a big percentage of half the entries of English Wikipedia are the most biased, it represents the wisdom of the overall crowd.

Majority of Web users who only do a lot, possibly helping explain a big percentage of half the entries of English Wikipedia are the most biased, it represents the wisdom of the overall crowd.

On average, the first version of this article was researched and posted by 0.05% of all participants; the subsequent version was researched and posted by 0.04% of all participants; the result was only 2%. Finally, we learned that the first version was created through our interaction with major aspects of the research. For more detail, see the References and the author's activity bias.

Bias on the Web

In 2009 with almost 40,000 active users, its registered editors, or approximately 2,000 people, indicating only a small percentage of all users contribute to the start of English Wikipedia probably, which in itself is a form of self-selection bias. However, economic do a lot, possibly helping explain a big percentage of half the entries of English Wikipedia are the most biased, it represents the wisdom of the overall crowd.

Though the example of English Wikipedia is the most biased, it represents the wisdom of the overall crowd.

The 2,000 people at the start of English Wikipedia probably, which in itself is a form of self-selection bias. However, economic do a lot, possibly helping explain a big percentage of half the entries of English Wikipedia are the most biased, it represents the wisdom of the overall crowd.
Bias Sources in Data Mining

**Data / Annotation Bias:** bias of labels or features in data
- Decisions whether to approve loan are unfair by reflecting on prejudice against a specific group in a historical record

**Sample Selection Bias:** data are not a representative of population
- Records who have been able to pay off their loans are only available for those who have been approved the loans

**Inductive Bias:** a bias caused by a machine learning algorithm
- 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
Data / Annotation Bias

**Data 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 of Objects
- Use of word statistics of 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

Advertisment 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]
Suspicious Placement Keyword-Matching Advertisement

Selection of ad-texts was unintentional

Response from advertiser:

- Advertise texts are selected based on the last name, and no other information is exploited.
- The selection scheme is adjusted so as to maximizing the click-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.

A data 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.

---

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 to generalize training data.

The assumptions might not always agree with a process of data generation in a real world.

II
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
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

- Key decisions in the legal process have been historically 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**

[Angwin+ 16]
Defendants of African descents were often predicted to be more risky than they actually were, and vice versa.

Overall Accuracy

<table>
<thead>
<tr>
<th></th>
<th>African</th>
<th>European</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Roughly the same</td>
<td>Not problematic</td>
</tr>
</tbody>
</table>

Recidivism Rates

<table>
<thead>
<tr>
<th></th>
<th>Actual</th>
<th>Predicted</th>
<th>False Positive Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>African</td>
<td><img src="actual.png" alt="Actual" /></td>
<td><img src="predicted.png" alt="Predicted" /></td>
<td><img src="FPR.png" alt="FPR" /></td>
</tr>
</tbody>
</table>

FPR (false positive ratio) = ratio of # of actually non-recidivated to # of people predicted to recidivate.
**US Census Data**: predict whether their income is high or low

**Females are minority in the high-income class**

<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>

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

\[\text{odds ratio} = 3.51\]

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

\[\text{odds ratio} = 5.26\]

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
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
Formal Fairness: Preliminary
In fairness-aware data mining, we maintain the influence:

- **Sensitive information**
  - socially sensitive information
  - information restricted by law
  - information to be ignored

- **Target / Objective**
  - university admission
  - credit scoring
  - crick-through rate

Formal Fairness

The desired condition defined by a formal relation between sensitive feature, target variable, and other variables in a model

- How to related 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
ex., loan approval, university admission, what to recommend

- $Y=1$ advantageous decision / $Y=0$ disadvantageous decision
- $Y$: observed / true, $\hat{Y}$: predicted, $Y^\circ$: fairized

**S** sensitive feature

To ignore the influence to the sensitive feature from a target
ex., socially sensitive information (gender, race), items’ brand

- $S=1$ non-protected group / $S=0$ disadvantageous decision
- Specified by a user or an analyst depending on his/her purpose
- It may depend on a target or other features

**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

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

Explainable variables are confounding variables with \( Y \) and \( S \), and their influence can be ignored because of legal or other reasons
Formal Fairness:
Based on Probabilities and Correlations
(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

[Boutilier+ 96]
Correlation Coefficient

\[ \rho = \frac{E[(X - E[X])(Y - E[Y])]}{\sqrt{E[(X - E[X])^2]} \sqrt{E[(Y - E[Y])^2]}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x - \bar{x})^2} \sqrt{\sum_i (y - \bar{y})^2}} \]

\* \( \bar{x} \) is a sample mean of \( x \)

Independence implies no-correlation, but no-correlation does not generally imply independence

**independence \( \Rightarrow \) no-correlation**

**Continuous Variable**
- If \( X \) and \( Y \) follows Gaussian, no-correlation implies independence

**Discrete Variable**
- If the rank of a frequency matrix for \( X \) and \( Y \) is 1, they are independent; If the matrix is singular, They are no-correlation
  - \( \Rightarrow \) If \( X \) and \( Y \) are binary, no-correlation implies independence
Bias Sources

Three sources of biases that undesirably corrupt outcomes

- **Data / 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 Data Bias

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

Data are not reliable, and never accessible to a fair dataset.

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

**Examples of assumptive conditions:**
- \( \hat{Y} \perp S \mid X=x \): \( Y \) and \( S \) are context-sensitive independent given \( X=x \)
- \( \hat{Y} \perp S \mid X \): \( Y \) and \( S \) are conditionally independent given \( X \)
- \( \hat{Y} \perp S \): \( Y \) and \( S \) are (unconditionally) independent
Independence / Statistical Parity

Remove data bias $\Rightarrow$ Independence / Statistical Parity: $\hat{Y} \perp S$

Ratios between positives and negatives in prediction are matched among all sensitive values

[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 test is biased in an ML tasks with a feedback loop, e.g., bandits, reinforcement learning, active learning
- biased selection of data to test or investigate
  - select randomly in terms of sensitive information
**Bandit problem**: maximize the cumulative rewards of selected arms

**Reward of the selected arm**

**Player** → **Bandits**

**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.
- **Fair UCB**: select arms whose confidence intervals overlap with equal probability.

**Deterministically select**

**Select with equal probability**
**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 the prediction, $\hat{Y}$, might be different from the observed, $Y$.

The changes from $Y$ to $\hat{Y}$ should be 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.
Remove inductive bias $\rightarrow$ Separation / Equalized Odds: $\hat{Y} \perp S \mid Y$

false positive ratio and true positive ratio must be matched between sensitive groups

[Hardt+ 16, Zafar+ 17]
Sufficiency

Remove inductive bias ➞ Sufficiency: \( Y \perp S \mid \hat{Y} \)

Precisions for positive and negative classes must be matched between sensitive groups.
Fairness in DM/ML has been discussed from 2010s

A statistics literature had discussed fairness criteria in 1960 — 70s after the US Civil Rights Act, 1964

**ML / DM**
- Independence
- Discovery & Prevention

**Statistics**
- Correlation
  - Discovery only

**Statistical Parity / Independence**
- Darlington (1971) criterion 4

**Equalized Odds / Separation**
- Cleary (1968), Darlington (1971) criterion (1), Linn (1973)

**Sufficiency**
- Darlington (1971) criterion (2)
Formal Fairness: Properties
Properties of Formal Fairness

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

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

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

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

Explainable Variable
- Exclusion of the explainable confounding effects between sensitives and targets
Disparate Treatment / Disparate Impact

**Disparate Treatment**
- **equality of opportunity**: tolerant to unequal outcome
- **procedural fairness**: eliminate sensitive information
- **intended**: direct or intentional reference of sensitive information

**Disparate Impact**
- **equality of outcome**: allow reverse discrimination
- **distributive justice**: fair allocation of goods
- **unintended**: indirect reference of sensitive information

**legal notions about fairness**

[Barocas+ 17, Feldman+ 15]
Direct Fairness / Indirect Fairness

Technical notions about fairness

Direct Fairness
a model does not directly refer sensitive information

\[ Y \perp S \mid X \]

Indirect Fairness
remove indirect reference to sensitive information

\[ Y \perp S, \hat{Y} \perp S \mid Y, \ldots \]

These technical notions are often expressed by legal terms

**Disparate Treatment**
Strictly speaking, disparate treatment includes intended indirect reference to sensitive information

**Disparate Impact**
Strictly speaking, whether or not the reference is intended should be cared in a disparate impact case

[Pedreschi+ 08]
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

This is a condition $\hat{Y} \perp S \mid X$ (not $\hat{Y} \perp S$)

$s$ still influences $Y$ through $X$
Red-Lining Effect: Simple elimination of a sensitive feature from training dataset fails to remove the influence of sensitive information to a target.

Ex: People of the same race frequently resident in a specific region.

Even if their race are not explicitly referred, the information is included in that of their residential region.

Distributive justice cannot be satisfied, under procedural fairness.
Causality and Direct / Indirect Fairness

Causal Inference

**Direct Effect**
- $S$: from a specific value 0 to 1
- $Y$: how the outcome changes through the direct path

$S: 0 \rightarrow 1 \quad S \xrightarrow{} Y \quad \Pr[Y]$?

**Indirect Effect**
- changes through the indirect path

$S: 0 \rightarrow 1 \quad S \xrightarrow{} Y \quad \Pr[Y]$?

Fairness-aware Data Mining

**Direct Fairness**
- $S$: don’t care the value of $S$
- $Y$: not directly depend on $Y$

$S: ? \quad S \xrightarrow{} Y \quad \Pr[Y | X]$?

**Indirect Fairness**
- $S$ is totally independent from $Y$
- Equivalent to zero total effect

$S \perp Y$

**total effect** = \[ \text{direct effect} + \text{indirect effect} \]
**Group Fairness** / **Individual Fairness**

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

---

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

**Independence (statistical parity) case**

\[ \Pr[Y | S=s] = \Pr[Y], \forall s \in \text{Dom}(S) \Rightarrow Y \perp S \]

Equivalent to the indirect fairness condition

---

**Limitations of Group Fairness**

- **Individuals are differently treated in each group**
  - some protected individual may receive disadvantageous decision
- **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**: In addition to the group fairness condition, distributions of a target variable are equal for all possible sensitive groups given a specific non-sensitive values.

**Independence (statistical parity) case**
\[ \Pr[Y \mid S, X=x] = \Pr[Y \mid X=x], \forall x \in \text{Dom}(X) \rightarrow Y \perp\!\!\!\perp S \mid X \]

Equivalent to the direct fairness condition.

- In addition to the individual fairness condition, if $X$ is also independent of $S$, the group fairness condition is satisfied.

\[ Y \perp S \mid X \text{ and } S \perp X \Rightarrow Y \perp S \]

**Situation Testing** [Luong+ 11]

- Legal notion of testing discrimination, comparing individuals having the same non-sensitive values except for a sensitive value.
Approximation of Individual Fairness

Probability distributions must be estimated for all non-sensitive values
\[ \Pr[Y | S, X=x] = \Pr[Y | X=x], \forall x \in \text{Dom}(X) \Leftrightarrow Y \perp S | X \]

It is practically impossible to collect data
whose non-sensitive values are exactly same
need an approximation technique

allow the difference within a constant  
[Dwork+ 12, Zemel+ 13]

- Given a predictor, \( X' \rightarrow Y \), a map, \( f \), that projects similar points in an original \( X \) space to similar points in a fair \( X' \) space

  Lipschitz condition \( d(x_1, x_2) \leq L d(f(x_1), f(x_2)), x_1, x_2 \in \text{Dom}(X) \)

aggregate information of its neighbors  
[Luong+ 11, Zhang+ 16]

- A probability distribution, \( \Pr[Y | S, X=x] \), is estimated from a dataset composed of the k-nearest neighbor of the point, \( x \)
A formal fairness condition is **symmetric** if all the statuses of sensitive features and outcomes are considered equally; otherwise the fairness condition is **asymmetric**

**α-protection**

\[
\frac{\Pr[Y=1 \mid S=0, X=x]}{\Pr[Y=1 \mid X=x]} \leq \alpha
\]

- **α-protection** is an **asymmetric** fairness condition, because it takes only protected group into account

**Equalized Odds / Equal Opportunity**

\[
\Pr[ \hat{Y}=1 \mid S=0, Y=y ] = \Pr[ \hat{Y}=1 \mid S=0, Y=y ]
\]

- Equalized 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 \( Y=0 \) case
**Explainable Variable:** these variables influence both target and sensitive variables, and the influence is not semantically problematic.

In FAML, 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**

$$\hat{Y} \perp S \mid E$$

Techniques of causal inference are applicable

Ex. randomization, propensity score
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:** $Y \perp S$
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 $I(S; Y)$
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
- medicine $\rightarrow$ acceptance=low
- computer $\rightarrow$ acceptance=high

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

**conditional statistical independence:** $Y \perp S \mid E$
the degree of conditional independence between $Y$ and $S$ given $E$ 

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

We can exploit additional information $I(S; Y; E)$ to obtain outcomes
Fairness Indexes
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 \( X=x \)
**Difference-based Measures**

**risk difference / mean difference**  [Calders+ 10, Pedreschi 09]
Difference of receiving advantageous decisions between groups

\[
RD = \Pr[\hat{Y} = 1|S = 1] - \Pr[\hat{Y} = 1|S = 0]
\]

- \(RD \rightarrow 0 \Rightarrow Y \independent S\)
- Equivalent to the total causal effect of changing \(S\) on \(\hat{Y}\)

**balanced error ratio**  [Feldman+ 15]
Mean of the probability of the disadvantageous decision for a non-protected group and the probability of the advantageous decision for protected group

\[
BER = \frac{\Pr[\hat{Y} = 0|S = 1] + \Pr[\hat{Y} = 1|S = 0]}{2} = \frac{1 - RD}{2}
\]

- \(BER \rightarrow 1/2 \Rightarrow Y \independent S\)
Measures from Contingency Table

\[
\begin{array}{|c|c|c|}
\hline
 & \hat{Y} = 0 & \hat{Y} = 1 \\
S = 0 & a_1 & n_1 - a_1 \\
S = 1 & a_2 & n_2 - a_2 \\
\hline
\end{array}
\]

\[
p_0 = \Pr[\hat{Y} = 0 \mid S = 0] = \frac{a_0}{n_0}
\]

\[
p_1 = \Pr[\hat{Y} = 0 \mid S = 1] = \frac{a_1}{n_1}
\]

\[
p = \Pr[\hat{Y} = 0] = \frac{a_0 + a_1}{n_0 + n_1}
\]

\[
p_0 - p_1 = \text{risk difference} / \text{mean difference} / \text{slift}_d
\]

\[
p_0 - p = \text{extended risk difference} / \text{elift}_d
\]

\[
p_0/p_1 = \text{risk ratio} / \text{relative risk} / \text{slift}
\]

\[
(1 - p_0)/(1 - p_1) = \text{relative chance}
\]

\[
p_0/p = \text{extended risk ratio} / \text{elift}
\]

\[
p_0(1 - p_1)
\]

\[
p_1(1 - p_0) = \text{odds ratio} / \text{olift}
\]
Tasks of Fairness-Aware Data Mining
Formal Tasks of Fairness-Aware ML

- **Unfairness Discovery**
  - finding unfair treatments
  - Discovery from Datasets
    - finding unfair data or subgroups in a dataset
  - Discovery from Models
    - finding unfair outcomes of a blackbox model

- **Unfairness Prevention**
  - predictor or transformation leading fair outcomes
  - Taxonomy by Process
    - pre-process, in-process, post-process
  - Taxonomy by Tasks
    - classification, regression, recommendation, etc…
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
Supervised Learning

**Learning**

- **Training data**
- Is this an apple?
  - Yes
  - No
  - No

**Classification rule**

A pattern between inputs and decisions (labels)

**Inference**

- Is this an apple?
  - Yes
  - Inference based on the learned rule
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
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 \(\mathbb{1}\), and then the learned classifier is modified to satisfy a fairness constraint \(\mathbb{5}\)

- This approach adopts the rather restrictive assumption, obliviousness [Hardt+ 16], under which 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 Discovery:
Discovery from Datasets
**Association Rule**

\[ X \Rightarrow Y \]

\( X \) : antecedent, \( Y \) : consequent

If \( X \) is satisfied, \( Y \) is also satisfied with a high probability

**Ex.**

( milk \( \in \) Item ) \( \land \) ( bread \( \in \) Item ) \( \Rightarrow \) ( egg \( \in \) 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**

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

**Confidence**

\[
\text{conf}(X, Y) = \frac{\# \text{ of data that satisfy both } X \text{ and } Y}{\# \text{ of data that satisfy } X} = \Pr[Y \mid X]
\]
Unfair Association Rules

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)**

\[ \text{elift} = \frac{\text{conf}(A ∧ B ⇒ C)}{\text{conf}(A ⇒ C)} \]

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 \( \text{elift} = \frac{\text{conf}(b)}{\text{conf}(a)} = 3 \)

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

\[ \Pr[\text{loan=deny} | \text{city=NYC, race=African}] \gg \Pr[\text{loan=deny} | \text{city=NYC}] \]
Unfair Association Rules

[Pedreschi+ 08, Ruggieri+ 10]

**Indirect unfairness:** a target variable depends on a sensitive feature through a non-sensitive feature

A target ‘loan’ does not directly depends 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

[Miettinen+ 16]

Ex. A literal ‘city=NYC \land ZIP=10451’ is a redescription of ‘city=NYC \land race=African’
**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.

\[ \text{Condition of situation testing is} \]
\[ \Pr[ Y \mid E, S=0 ] = \Pr[ Y \mid E, S=1 ] \equiv Y \perp S \mid E \ (E : \text{explainable variables}) \]
Unfairness Discovery: Discovery from Models
Gradient Feature Auditing

**Direct Influence:** comparing outputs when changing $S$

$(X_i, S)$ original data

Blackbox Model $\rightarrow Y$

ignore the influence if a feature in $X_i$ is correlated with $S$

$(X_i, S')$ sensitive is perturbed

Blackbox Model $\rightarrow Y'$

**Indirect Influence:** the influence of features correlated with $S$

$(X_i, S)$ original data

Blackbox Model $\rightarrow Y$

measure the influence of features in $X_i$ correlated with $S$

$(X'_i, S)$ non-sensitive is perturbed

Blackbox Model $\rightarrow Y'$

$X_i$ is perturbed so as not to predict $S$ from the perturbed data $X'_i$
Unfairness Prevention:
Classification (pre-process)
Massaging: Pre-process type method

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

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 Representation**
\[
\begin{align*}
\text{min loss function} \\
\text{s.t. fairness constraint}
\end{align*}
\]

**Fairness Constraint:**
- **Lipschitz Condition:** similar original data are mapped to similar archetypes
  - Distance between archetypes \( D(M(x), M(y)) \) \leq d(x, y) 
  - Distance between original data
- **Statistical Parity:** protected and non-protected groups are equally treated
  - Mean over sensitive archetypes \( D(\mu_S, \mu_T) \) \leq \epsilon
  - Mean over all archetypes
Learning Fair Representations

original data

\( x_i \)

prototypes

\( \nu_k \)

data are probabilistically assigned to each prototype

prototype is the mean of assigned data

class

\( \hat{y}_i \)

classes are predicted from prototypes

Requirements for Prototypes

- Probabilities assigned to each prototype is equal between groups
  \[ L_z = \sum_k |M_k^{S=0} - M_k^{S=1}| \]
- Original data should be close to the data recovered from prototypes
  \[ L_x = \sum_n (x_n - \hat{x}_n)^2 \]
- Classes predicted from prototypes should close to original classes
  \[ L_y = \sum_n -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n) \]

Maps to prototypes are learned so as to maximize these requirements
Distributions of the $j$-th feature are matched between datasets whose sensitive feature is $S=0$ and $S=1$.

Feature values are modified so as to minimize the sum of the L1 distances the modified cumulative distribution function (CDF) from original CDFs.
Unfairness Prevention: Classification (in-process)
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

  \[
  \text{[fairness index]} / \text{[decrease in accuracy]}
  \]
Prejudice Remover Regularizer

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 \(s \in \text{Dom}(S)\)

\[\Pr[Y=1 \mid x, s] = \text{sig}(w^{(s)^\top}x)\]

As a prejudice remover regularizer, we adopt a mutual information between a target and a sensitive feature, \(I(Y; S)\)
Unfairness Prevention: Classification (post-process)
Unfair decisions are modeled by introducing the dependence of $X$ on $S$ as well as on $Y$.

$S$ and $X$ are conditionally independent given $Y$.

Non-sensitive features in $X$ 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 method.
parameters are initialized by the corresponding sample distributions

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

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

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

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

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

if \# of data classified as “1” < \# of “1” samples in original data then

increase \(\text{Pr}[Y=1, S=0]\), decrease \(\text{Pr}[Y=0, S=0]\)

else

increase \(\text{Pr}[Y=0, S=1]\), decrease \(\text{Pr}[Y=1, S=1]\)

reclassify samples using updated model \(\text{Pr}[Y, S]\)

update the joint distribution so that its fairness is enhanced
Unfairness Prevention: Classification (additional topics)
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.

\[
\text{Difference} = \Pr[Y, S] - \Pr[Y] \Pr[S]
\]

- **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

\[ \tilde{Y} \perp S, \text{ where } (\tilde{Y}, S) \sim \hat{P}_r[\tilde{Y}, S] \]

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

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

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

satisfy actual independence instead of model-based independence

Fairness in class labels can be greatly improved
Experiment: Actual vs Model-based

Accuracy (Acc)

Fairness (NMI)

more accurate

fairer

fairness parameter $\eta$ : the larger value more enhances the fairness

Accuracy and fairness has the trade-off relation

By satisfying actual independence, instead of model-based independence the trade-off was drastically improved
Unfairness Prevention: Recommendation
Recommender System

Recommenders: Tools to help identify worthwhile stuff

Find Good Items

Predicting Ratings

Ranking items according to users' preference, to help for finding at least one target item

Presenting items with predicted ratings for a user, to help for exploring items

* Screen-shots are acquired from Amazon.co.jp and Movielens.org on 2007-07-26
Collaborative filtering is a major approach for predicting users' preference in a word-of-mouth manner recommending items liked by those who having similar preferences

There are other approaches: content-based filtering or knowledge-based filtering
A recommendation service must be managed while adhering to laws and regulations.

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

Socially discriminative treatments must be avoided.

Sensitive feature = users’ demographic information.

Legally or socially sensitive information can be excluded from the inference process of recommendation.
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 treatment in recommendation

A hotel booking site should not abuse their position to recommend hotels of its group company

**sensitive feature** = a content provider of a candidate item

Information about who provides a candidate item can be ignored, and providers are treated fairly
Exclusion of Unwanted Information


Information unwanted by a user is excluded from recommendation

**Filter Bubble:** To fit for Pariser’s preference, conservative people are eliminated from his friend recommendation list in Facebook

Sensitive feature = a political conviction of a friend candidate

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
Recommendation Independence

Recommendation Independence
statistical independence
between a recommendation outcome, $R$, and a sensitive feature, $S$

$$\Pr[R \mid S] = \Pr[R] \equiv R \perp S$$

- No information about a sensitive feature influences the outcome
- The status of the sensitive feature is explicitly excluded from the inference of the recommendation outcome

Independence-Enhanced Recommendation
Preferred items are predicted
so as to satisfy a constraint of recommendation independence
Probabilistic Matrix Factorization

Probabilistic Matrix Factorization Model
predict a preference rating of an item \( y \) rated by a user \( x \)
well-performed and widely used

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

Prediction Function

Objective Function

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
Independence Terms

**Mutual Information with Histogram Models** [Kamishima+ 12]
- computationally inefficient

**Mean Matching** [Kamishima+ 13]
- matching means of predicted ratings for distinct sensitive groups
- improved computational efficiency, but considering only means

**Mutual Information with Normal Distributions** [Kamishima+ 18]

\[- ( \text{mean} \left( \mathcal{D}^{(0)} \right) - \text{mean} \left( \mathcal{D}^{(1)} \right) )^2 \]

**Distribution Matching with Bhattacharyya Distance** [Kamishima+ 18]

\[- \left( \text{ln} \int \sqrt{\text{Pr}[r|S=0] \text{Pr}[r|S=1]} dr \right) \]

- These two terms can take both means and variances into account, and are computationally efficient
Effect of Independence Enhancement

Standard

two distributions are largely diverged

Independence-enhanced

two distributions become closer

a sensitive feature = whether a movie is newer or older

* each bin of histograms of predicted scores for older and newer movies

The bias that older movies were rated higher could be successfully canceled by enhancing independence

[Kamishima+ 12, Kamishima+18]
Latent Class Model: A probabilistic model for collaborative filtering

A basic topic model, pLSA

Latent topic variable

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:

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

the \( r \)-th rating value

A rating value can be predicted by the expectation of ratings
Independence-Enhanced LCM

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
Unfairness Prevention: Other Tasks
**Word Embedding:** vector representing semantics of words
The differences of vectors reflect analogy of the corresponding words
\[ he - she = king - queen \]

### Occupational stereotype
Occupational words whose embeddings are the 10 nearest from the word embeddings of **she** or **he**

- **Extreme she**
  1. homemaker
  2. nurse
  3. receptionist
  4. librarian
  5. socialite
  6. hairdresser
  7. nanny
  8. bookkeeper
  9. stylist
  10. housekeeper

- **Extreme he**
  1. maestro
  2. skipper
  3. protege
  4. philosopher
  5. captain
  6. architect
  7. financier
  8. warrior
  9. broadcaster
  10. magician

Word embeddings are unfair due to the gender bias in the training corpus

### Debiassing Embeddings
- **neutralize:** non-gender words are uncorrelated to gender vector
- **equalize:** equal distance from occupational words to gender words
Non-Redundant Clustering

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

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.
**Ranking**: select $k$ items and rank them according to the relevance to users' need
A fundamental task for information retrieval and recommendation

---

**Step 1: Calculate Relevance Score**

**Relevance Score**: the degree of relevance to user's need
- **Information Retrieval**: relevance to the user's query
- **Recommendation**: user's preference to the item

**Step 2: Rank Items**

sort according to their relevance scores

select top-$k$ items

relevant items \[ \begin{array}{cccccc} 1.0 & 1.0 & 0.9 & 0.7 & 0.5 & 0.3 & 0.1 \end{array} \] irrelevant items

106
**Fair Ranking:** for each rank $i = 1, \ldots, k$, the ratio between two sensitive groups must not diverged from the ratio in the entire candidate set

1. Generate ranking lists for each sensitive group
2. Merge two ranking lists so as to satisfy fair ranking condition

This item is less relevant, but it is prioritized to maintain fairness
Relation to Other Research Topics
Privacy-Preserving Data Mining

Fairness in Machine Learning
the independence between an objective $Y$ and a sensitive feature $S$

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

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 FAML, generally
Causal Inference

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

FAML 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

```
cause          outcome

confounding
```

**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

```
sensitive   target

explainable
```
Causal Inference

[Zhang+ 18]

total causal effect of changing a sensitive feature, \( S \), on a target, \( Y \)

intervention on a sensitive change decision?

any direct and indirect causal paths are considered

![Diagram](image)

**total causal effect**

\[
TE(S = 1, S = 0) = \Pr[Y = 1|do(S = 1)] - \Pr[Y = 1|do(S = 0)]
\]

**total causal effect is equal to the risk difference**

\[
TE(S = 1, S = 0) = \Pr[Y = 1|S = 1] - \Pr[Y = 1|S = 0]
\]
**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:

\[
Y \perp S \mid E \Rightarrow Y \perp S \mid e(S)
\]

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

![Diagram showing the division of a training dataset into strata based on propensity scores.](image)

- **strata 1**: \( e(S) \in [0, 1/3) \)
- **strata 2**: \( e(S) \in [1/3, 2/3) \)
- **strata 3**: \( e(S) \in [2/3, 1] \)
Cost-Sensitive Learning: learning classifiers so as to optimize classification costs, instead of maximizing prediction accuracies

FAML 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 FML 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
Fairness-Aware Machine Learning & Data Mining

http://www.kamishima.net/faml/
References
References


S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: from discrimination discovery to fairness-aware data mining. The 22nd ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining, Tutorial, 2016.


