Fairness-Aware
Machine Learning and Data Mining

Toshihiro Kamishima
www.kamishima.net
Updated: 2023-08-03
The spread of machine learning technologies

Machine learning is being increasingly applied for serious decisions
Ex: credit scoring, insurance rating, employment application

Fairness-Aware Machine Learning

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 unfair 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.

Ex:

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 [Athey 17].

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.
The latest version of this slide is distributed at the URL:

Fairness-Aware Machine Learning and Data Mining

http://www.kamishima.net/faml/
Outline
Part I: Backgrounds

- Types of Biases
- Instances of Data Bias
- Instances of Inductive Biases
Part II: Formal Fairness

- Basics of Formal Fairness
- Association-Based Fairness
  - Basics of Associations
  - Criteria
  - Properties
  - Measures
- Counterfactual Fairness
- (Economics-Based Fairness)
Part III: Fairness-Aware Machine Learning

Tasks of Fairness-Aware Machine Learning

Unfairness Discovery

- Discovery from Datasets
  - Association-based fairness
  - (Counterfactual fairness)
- Discovery from Models

Unfairness Prevention

- Classification: Pre-process, In-process, Post-process
- (Regression)
- Recommendation
- Ranking
- (Clustering)
- Other Tasks
Part IV: Other Topics

- Mitigation of a Sample Selection Bias
- Disclosure
- Other Fairness-Aware Machine Learning Topics
- Relation to the Other Machine Learning Topics
- Software
- Evidence-Based Decision Making
Part I
Backgrounds
Types of Biases
Bias on the Web

Activity Bias, or Wisdom of a Few

In 2011, a study by Wu et al. on how people followed other people on Twitter found that the 0.05% of the most popular people attracted almost 50% of all participants; that is, half of the Twitter users in the dataset were following only a few select celebrities. I thus asked myself: What percentage of active Web users generate half the content in a social media website? I did not, however, consider the silent majority of Web users who only watch the Web without contributing to it, which in itself is a form of self-selection bias.

Saez-Trumper and I analyzed four datasets, and as I detail, the results surprised us. Exploring a Facebook dataset from 2009 with almost 40,000 active users, we found 7% of them produced 50% of the posts. In a larger dataset of Amazon reviews from 2013, we found just 4% of the active users. In a very large dataset from 2011 with 12 million active Twitter users, the result was only 2%. Finally, we learned that the first version of half the entries of English Wikipedia was researched and posted by 0.04% of its registered editors, or approximately 2,000 people, indicating only a small percentage of all users contribute to the Web and the notion that it represents the wisdom of the overall crowd is an illusion.

In light of such findings, it did not make sense that just 4% of the people voluntarily write half of all the views in the Amazon dataset. I sensed something else is at play. A month after publication of our results, my hunch was confirmed. In October 2015, Amazon began a corporate campaign against paid fake reviews that continued in 2016 by suing almost 1,000 people accused of writing them. Our analysis also found that if we consider only the reviews that some people find helpful, the percentage decreases to 2.5%, using the positive correlation between the average helpfulness of each review according to users and a proxy of text quality. Although the example of English Wikipedia is the most biased, it represents a positive bias. The 2,000 people at the start of English Wikipedia probably triggered a snowball effect that helped Wikipedia become the vast encyclopedic resource it is today.

Zipf's least-effort principle, also called Zipf's law, maintains that many people do only a little while few people do a lot, possibly helping explain a big part of activity bias. However, economic and social incentives also play a role in yielding this result. For example, Zipf's law can be seen in most Web measures both the growth of the Web and its use.

Here, I explain each of the biases (in red) and classify them by type, beginning with activity bias resulting from how people use the Web and the hidden bias of people without Internet access. I then address bias in Web data and how it potentially taints the algorithms that use it, followed by biases created through our interaction with websites and how content and use recycle back to the Web or to Web-based systems, creating various types of second-order bias.

Consider the following survey of research on bias on the Web, some I was involved with personally, focusing on

Figure 1. The vicious cycle of bias on the Web.

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 software, not by Web users or developers.
Bias on the Web

[Baeza-Yates 18]

- Activity bias
- Second-order bias
- Self-selection bias
- Algorithmic bias
- Interaction bias
- Sampling bias
- Data bias
- Web
- Screen
- Algorithm
- Sample selection bias
- Inductive bias
- Algorithmic bias
- Interaction bias
- Self-selection bias
- Second-order bias
- Data bias
Bias Sources in Machine Learning

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 representatives 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

A Prediction is made by aggregating data

Even if inappropriate data is contained in a given dataset, the data can affect the prediction without correction

Is this an apple?

No, Yes, No, Yes, No

Even if an apple is given, the predictor trained by an inappropriate data set may output “No”
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.

Mismatch between distributions of learned and applied populations.

Simple prediction algorithms cannot learn appropriately from a dataset whose contents depend on contents of the data.
**Example of Sample Selection Bias**

*loan application:* A model is learned from a dataset including only approved applicants, but the model will be applied to applicants including declined applicants ➔ **sample selection bias**

- Population to apply the learned model
- Population to learn a model for prediction

- Mismatch

- Loan application
  - Approved
  - Declined
  - Unknown
- Full payment
- Default

A model is used for the targets different from a learned dataset ➔ **The learned model cannot classify targets correctly**
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
**Occam's Razor**: Entities should not be multiplied beyond necessity

If models can explain a given data at the similar level, the simpler model is preferred

A small number of exceptional samples are treated as noise

The prediction for unseen cases would be more precise in general

Crucial rare cases can cause unexpected behavior

Any prediction, even if it was made by humans, is influenced by inductive biases, because the bias is caused in any generalization
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
Instances of Data Biases
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
- Admission to universities can be influenced by recommendation letters

[Bolukbasi+ 16]
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
  - Arrested? negative ad-text

- European descent’s name
  - Located: neutral ad-text

[Sweeney 13]
Suspective Placement Keyword-Matching Advertisement

Advertised 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

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 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.
Instances of Inductive Biases
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

**Merits and Concerns pointed out by the ProPublica**

- Key decisions in the legal process have been historically affected by personal biases
- Scores can be exploited not for the designed purposes
- **Scores must accurately predict which defendants likely to re-offend, but these are biased**
Defendants of African descents were often predicted to be more risky than they actually were, and vice versa.

*FPR (false positive ratio) = ratio of # of actually non-recidivated to # of people predicted to recidivate*
Rejoinder of US Federal Courts

The merit of risk assessment tool
It might be that the existing justice system is biased against poor minorities … regardless of the degree of bias, risk assessment tools informed by objective data can help reduce racial bias from its current level

Rejoinder to ProPublica's study
1. The COMPAS targets individuals on post-disposition supervision, but the ProPublica analyzed pretrial defendants
2. Collapsing mid- & high-risk categories is problematic
3. Distributions of observations given the predictions should be used, instead of distributions of predictions given observations
4. The standards, such as the federal Post Conviction Risk Assessment (PCRA), are ignored
5. Choosing improper the level of significance
Rejoinder of US Federal Courts

The COMPAS satisfies a fairness condition, sufficiency

- The COMPASS score is designed to satisfy the sufficiency, $Y \perp S \mid \hat{Y}$, following the standard of the federal Post Conviction Risk Assessment (PCRA)
- The chart shows the actual arrest ratios given the predicted risk scores, in the any arrest case
- The Northponte, a COMPAS developer, also pointed out this problem [Dieterich+ 2016]
Algorithms Improve Human Decisions

Pretrial Bail Decisions

- Arrest records in New York City between Nov. 1, 2008 – Nov. 1, 2013
  - male=83.2%, African American=48.8%, Hispanic=33.3%
  - release=73.6% → failure to appear=15.2%, rearrested=25.8%
- Judges decide whether defendants to release or detain, based on a checklist and the information judges see, such as appearance
- Algorithms use the information available to judges and age, but ignore the information judges see

**Algorithms Improve Judges' Decisions**

If defendants were detained based on algorithm prediction until the level that judges of high-detention rate detained, algorithms would achieve:

- at the same crime rate as judges → 48.2% lower detention rate
- at the same detention rate as judges → 75.8% lower crime rate
Algorithms Improve Human Decisions

[Kleinberg+ 18]

Judges Release High-Risk Defendants
The riskiest 1% of defendants in prediction:
If released, fail to appear=57.3%, rearrested=62.7%

Judges release 48.5% of them

Algorithms Are Fairer Than Judges
If a distribution of detained races is constrained to satisfy a fairness condition, algorithms reduce crime rate relative to judges:
- no constraint ➞ 24.68%
- match a distribution that judges detain ➞ 24.64%
- match a distribution of defendants (= statistical parity) ➞ 23.02%
- match lower of a distribution of defendants or a distribution that judges detain ➞ 22.74%
Bias in Image Recognition

- Auditing the image recognition API's for predicting a gender from facial images
- Available benchmark datasets of facial images is highly skewed to the images of males with lighter skin
- Pilot Parliaments Benchmark (PPB) is a new dataset balanced in terms of skin types and genders
  - Skin types are lighter or darker based on the Fitzpatrick skin type
  - Perceived genders are male or female
- Facial-image-recognition API's by Microsoft, IBM, and Face++ are tested on the PPB dataset

[Buolamwini+ 18]
Bias in Image Recognition

Error rates (1 - TPR) in a gender prediction from facial images

<table>
<thead>
<tr>
<th></th>
<th>darker male</th>
<th>darker female</th>
<th>lighter male</th>
<th>lighter female</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Microsoft</strong></td>
<td>6.0%</td>
<td><strong>20.8%</strong></td>
<td>0.0%</td>
<td>1.7%</td>
</tr>
<tr>
<td><strong>IBM</strong></td>
<td>12.0%</td>
<td>34.7%</td>
<td>0.3%</td>
<td>7.1%</td>
</tr>
<tr>
<td><strong>Face++</strong></td>
<td>0.7%</td>
<td>34.5%</td>
<td>0.8%</td>
<td>7.1%</td>
</tr>
</tbody>
</table>

Error rates for **darker females** are generally worse than **lighter males**
IBM have improved the performance by new training dataset and algorithm, before Buolamwini's presentation,:

<table>
<thead>
<tr>
<th></th>
<th>darker male</th>
<th>darker female</th>
<th>lighter male</th>
<th>lighter female</th>
</tr>
</thead>
<tbody>
<tr>
<td>old IBM</td>
<td>12.0%</td>
<td><strong>34.7%</strong></td>
<td>0.3%</td>
<td>7.1%</td>
</tr>
<tr>
<td>new IBM</td>
<td>2.0%</td>
<td><strong>3.5%</strong></td>
<td>0.3%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Error rates for darker females are improved.
**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

- 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.
Part II
Formal Fairness
Basics of Formal Fairness
In fairness-aware machine learning, we maintain the influence:

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

- 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
- $Y = \text{observed/true}$, $\hat{Y} = \text{predicted}$, $Y^\circ = \text{fairized}$

- $Y=1$ advantageous decision / $Y=0$ disadvantageous decision

- **$S$** 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

\[ \mathcal{D} = \{ y_i, s_i, x_i \}_{i=1}^2 \]  
\textit{dataset}

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

\[ \mathcal{D}^{(s)} = \{ y_i, s_i, x_i \}_{i=1}^{n(s)} \text{ s.t. } s_i = s \]  
\textit{sensitive group}

a group consisting of the same sensitive value
If \( s_i = 0 \) indicates a minority individual to protect, \( \mathcal{D}^{(0)} \), is called a \textit{protected group}, and the rest of dataset, \( \mathcal{D}^{(1)} \), is called a \textit{non-protected group}

\[ E / \bar{E} \]  
\textit{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
Type of Formal Fairness

**association-based fairness**
- defined based on statistical association, namely correlation and independence
- mathematical representation of ethical notions, such as distributive justice

**counterfactual fairness**
- causal effect of the sensitive information to the outcome
- maintaining a counterfactual situation if the sensitive information was changed

**economics-based fairness**
- using a notion of a fairness in collaborative game theory or welfare economics
Why an instance of discrimination is bad?

- **harm-based account:** Discrimination makes the discriminatees worse off
- **disrespect-based account:** Discrimination involves disrespect of the discriminatees and it is morally objectionable
  - An act or practice is morally disrespectful of \( X \)
    - It presupposes that \( X \) has a lower moral status than \( X \) in fact has

**Techniques of Fairness-Aware Machine Learning**

based on the harm-based account

The aim of FAML techniques remedy the harm of discriminatees
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, EEOC)**
- 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.
Title VII of the Civil Rights Act of 1964

- Prohibit discrimination due to race, religion, gender, and ethnicity


- Evidence of long-lasting and **gross disparity** between the composition of a workforce and that of the general population thus may be significant even though § 703(j) makes clear that Title VII imposes no requirement that a workforce mirror the general population

- Where **gross statistical disparities** can be shown, they alone may, in a proper case, constitute *prima facie* proof

**Gross Statistical Disparity:** Discrimination in employment is determined whether the ratio of protected and non-protected groups of employees is diverged from the corresponding ratio in general population
A harm-based account requests a baseline for determining whether the discriminatees have been made worse off.

- **Ideal outcome:** the discriminatees are in just, or the morally best
  - **association-based fairness:** letting predictors get ideal outcomes

- **Counterfactual:** the discriminatees had not been subjected to the discrimination
  - **counterfactual fairness:** comparing with the counterfactuals that a status of a sensitive feature was different
Association-Based Fairness: Basics of Associations
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*
(Unconditional) Independence: $Y \perp S$

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

**dependent**

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

**independent**

$\Pr[Y=1 \mid S=0]$  $\Pr[Y=1 \mid S=1]$  $\Pr[Y=1 \mid S=0]$  $\Pr[Y=1 \mid S=1]$
Conditional Independence:

\[ Y \perp S \mid X \]

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

**Dependent**

\[ \Pr[Y=1 \mid S=s, X=0] \neq \Pr[Y=1 \mid X=0] \]

\[ \Pr[Y=1 \mid S=s, X=1] \neq \Pr[Y=1 \mid X=1] \]

**Independent**

\[ \Pr[Y=1 \mid S=s, X=0] = \Pr[Y=1 \mid X=0] \]

\[ \Pr[Y=1 \mid S=s, X=1] = \Pr[Y=1 \mid X=1] \]
Unconditional & Conditional Independence

Conditional independence does not imply unconditional independence in general.

\[ S \perp Y \mid X \rightarrow S \perp Y \]

**Conditionally Independent**

<table>
<thead>
<tr>
<th>S=0</th>
<th>S=1</th>
<th>Y=0</th>
<th>Y=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>7</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>12</td>
<td>30</td>
</tr>
</tbody>
</table>

\[ 14 + 8 = 22 \]

\[ 6 + 12 = 18 \]

**Unconditionally Dependent**

<table>
<thead>
<tr>
<th>S=0</th>
<th>S=1</th>
<th>Y=0</th>
<th>Y=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>30</td>
<td>7</td>
<td>20</td>
</tr>
</tbody>
</table>

\[ 3 + 30 = 33 \]

\[ 7 + 20 = 27 \]
Unconditional & Conditional Independence

Inversely, unconditional independence does not imply conditional independence in general.

\[ S \perp Y \mid X \quad \text{versus} \quad S \perp Y \]

Conditionally Dependent

Unconditionally Independent
Simpson's Paradox: Numerical facts that the results obtained from a whole dataset is processed are contradicted with the results obtained when a dataset is grouped or stratified

Admission to the Univ. of California, Berkeley, for the fall 1973 quarter
Aggregated data for the campus
- Admission rate: male=44% female=35%  ➔ discriminative

Grouped by the departments
- Among 85 departments, females are fewer in 4 departments and males are fewer in 6 departments  ➔ non-discriminative

This case is not discriminative, because more females were applied to the department whose admission rate was lower

even the naive question could not answered adequately without recourse to sophisticated methodology and careful examination of underlying process
"Cholesterol" and "exercise" are **positively correlated**, if all data are aggregated.

If grouped by "age", they are **negatively correlated**, because cholesterol of aged people tends to be higher.
Information-Theoretic Interpretation

Sensitive: $S$

$$H(S) \quad H(S \mid Y) \quad I(S; Y)$$

Target: $Y$

$$H(Y) \quad H(Y \mid S)$$

Statistical parity, $S \perp Y$, implies zero mutual information: $I(S; Y) = 0$

If the information about $Y$ is known, no information about $S$ cannot be gained, and vice versa.
Information Theoretic Interpretation

Sensitive: $S$

- $H(S | Y, E)$
- $I(S; Y | X)$
- $H(S)$

Non-Sensitive: $X$

- $I(S; E | Y)$
- $I(S; Y; X)$
- $I(Y; X | S)$

Target: $Y$

- $H(Y)$
- $H(Y | S, X)$
- $H(X | S, Y)$
- $H(X)$

Mutual information, $I(S; Y | X)$, shows the information gained by knowing about $Y$ in the information about $S$ by knowing $X$ ($= H(S | X)$).
Markov network: undirected graphical model for probabilistic distribution

Pr[A, B, C, D, E] = f(A, B, D)f(B, C, D)f(B, E)/Z

Variables, A and C, are separated by removing B and D

conditional independence: A ⊥⊥ C | B, D
Correlation

Correlation Coefficient

\[
\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}
\]

\* \(\bar{x}\) is a sample mean of \(x\). \(\text{Var}(X)\) and \(\text{Cov}(X, Y)\) are a variance and covariance, respectively.

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

\[
\text{independence} \implies \text{no-correlation}
\]

Continuous Variable

- If \(X\) and \(Y\) follow 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

\[
\implies \text{If } X \text{ and } Y \text{ are binary, no-correlation implies independence}
\]
Partial Correlation

Partial Correlation Coefficient

\[
\rho_{xy,z} = \frac{\text{Cov}(\Delta_{xz}, \Delta_{yz})}{\sqrt{\text{Var}(\Delta_{xz})} \sqrt{\text{Var}(\Delta_{yz})}} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{1 - \rho_{xz}^2} \sqrt{1 - \rho_{yz}^2}}
\]

- \(\theta_{xz}\): a regression coefficient from \(z\) to \(x\).
- \(\Delta_{xz}^{(i)} = x_i - \theta_{xz} z_i\)
- \(\rho_{xy}\): correlation coefficient between \(x\) and \(y\).

- The \(\rho_{xy,z}\) (the partial correlation between \(x\) and \(y\) given \(z\)) is the correlation between \(x\) and \(y\) while removing the influence of \(z\) to \(x\) and \(y\), respectively.
Association-Based Fairness: Criteria
Criteria of Association-Based Fairness

Fairness through Unawareness — Fairness through Awareness
- Prohibition to access sensitive information during the process of learning and inference

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

Statistical Parity
- Satisfying the equality of outcome

Equalized Odds / Sufficiency
- Equalizing biases of prediction from observed data

Context-Sensitive Independence
- Fairness in Specific Contexts

Correlation-based Fairness
- Sensitive information correlates with a target variable
# Association-Based Fairness

<table>
<thead>
<tr>
<th></th>
<th>fairness through unawareness</th>
<th>statistical parity</th>
<th>equalized odds</th>
<th>sufficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{Y} \perp S \mid X$</td>
<td>$\hat{Y} \perp S$</td>
<td>$\hat{Y} \perp S \mid Y$</td>
<td>$Y \perp S \mid \hat{Y}$</td>
</tr>
<tr>
<td>awareness</td>
<td>unaware</td>
<td></td>
<td>aware</td>
<td></td>
</tr>
<tr>
<td>unit</td>
<td>individual</td>
<td></td>
<td>group</td>
<td></td>
</tr>
<tr>
<td>wordview</td>
<td>WAE</td>
<td></td>
<td>WYSIWYG</td>
<td></td>
</tr>
<tr>
<td>comments</td>
<td>treat like cases alike</td>
<td>equality of outcomes</td>
<td>equality of false positive and false negative rates</td>
<td>equality of positive and negative predictive values</td>
</tr>
<tr>
<td></td>
<td>alias: situation testing</td>
<td>alias: demographic parity, independence</td>
<td>alias: separation</td>
<td></td>
</tr>
</tbody>
</table>
**Fairness through Unawareness:** Prohibiting to access individuals' sensitive information during the process of learning and inference.

This is a kind of procedural fairness, in which a decision is fair, if it is made by following pre-specified procedure.

A **unfair model** is trained from a dataset including sensitive and non-sensitive information.

A **fair model** is trained from a dataset eliminating sensitive information.

A unfair model, $\Pr[\hat{Y}|X,S]$, is replaced with a fair model, $\Pr[\hat{Y}|X]$.

$$
\Pr[\hat{Y},X,S] = \Pr[\hat{Y}|X,S] \Pr[S|X] \Pr[X] \quad \Rightarrow \quad \Pr[\hat{Y}|X] \Pr[S|X] \Pr[X]
$$

**Fairness through Unawareness:** $\hat{Y} \perp S|X$
Fairness through Unawareness

A kind of procedural fairness → Fairness through Unawareness

These gaps indicate unfair decision

A learned model directly access sensitive information

\( \hat{Y} \perp S \mid X \)

\[
\Pr[\hat{Y}, S \mid X] = \Pr[\hat{Y} \mid X] \Pr[S \mid X]
\]
Group Fairness / Individual Fairness

Target unit for which a fairness condition is satisfied

Group Fairness

- **Individuals are equally treated as a group**
- Instantiation of the ethical notion “distributive fairness”
- Implemented by match the aggregated statistics, such as means or errors, between groups
- **Ex:** statistical parity, equalized odds, sufficiency

Individual Fairness

- **Individuals are treated alike regardless of group membership**
- Instantiation of the principle “treat like cases alike”
- Implemented by conditioning on individuals, usually represented by \( X \), in a case of association-based fairness
- **Ex:** individual fairness
Group Fairness

**Group Fairness:** Outcomes of a target variable are equal for all sensitive groups as a whole

- **statistical parity:** equal share between groups
  \[
  \Pr[\hat{Y} \mid S = s] = \Pr[\hat{Y}], \forall s \in \text{Dom}(S) \Rightarrow \hat{Y} \perp S
  \]

- **equalized odds:** equal errors between group
  \[
  \Pr[\hat{Y} \mid S = s, Y] = \Pr[\hat{Y} \mid Y], \forall s \in \text{Dom}(S) \Rightarrow \hat{Y} \perp S \mid Y
  \]

Limitations of Group Fairness

- **Individuals are differently treated in each group**
  \[\Rightarrow\text{some protected individual may receive disadvantageous decision}\]

- **Reverse Tokenism:** justify unfair treatment for members of a protected group by sacrificing a few superior members of a non-protected group
  \[\Rightarrow\text{This cannot be prevented by achieving group fairness}\]

[Dwork+ 12]
**Individual Fairness**: Implementation of the principle of “Treat like cases alike”

Distributions of a target variable are equal for all possible sensitive groups given a specific non-sensitive values

\[
\Pr[\hat{Y} \mid S, X=x] = \Pr[\hat{Y} \mid X=x], \quad \forall x \in \text{Dom}(X) \implies \hat{Y} \perp S \mid X
\]

**Conditioning fairness criteria by X can be considered as individual fairness**

- **Simple individual fairness and fairness through unawareness are the same in a mathematical form, \(\hat{Y} \perp S \mid X\), but not in their semantics**

**Ex:** To satisfy individual fairness simultaneously with equalized odds, sensitive information must be observed, and this violates a condition of fairness through unawareness

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

[Luong+ 11]
Detection of Individual Fairness

Probability distributions must be estimated for all non-sensitive values

\[ Pr[Y | S, X=x] = Pr[Y | X=x], \forall x \in \text{dom}(X) \Leftrightarrow Y \perp S | X \]

To test individual fairness, it is practically impossible to observe data whose non-sensitive values are exactly same

aggregate information of its neighbors

- A probability distribution, \( Pr[Y | S, X=x] \), is estimated from a dataset composed of the k-nearest neighbor of the point, \( x \)

estimate its counterfactual case

- Given a factual case in which \( X = x \) and \( S = s \), its counterfactual case in which \( X = x \) and \( S = s' \) is estimated by assuming the underlying causal relations

[Luong+ 11]
**Worldview and Bias**

**Worldview** is an assumption about mapping from construct space to observed space:
- **construct space**: underlying ideal features and decisions
- **observed space**: observed features and decisions

---

**We're All Equal Worldview**
Instances in different groups are mapped differently

**What You See Is What You Get Worldview**
Mapping while keeping relative positions between groups

---

![Diagram](image-url)
Statistical Parity / Independence

equality of outcome: Goods are distributed by following pre-specified procedure
In a context of FAML, the predictions are distributed so as to be proportional to the sizes of sensitive groups

\[
\frac{\Pr[Y = y_1, S = s_1]}{\Pr[Y = y_2, S = s_2]} = \frac{\Pr[S = s_1]}{\Pr[S = s_2]} \quad \forall y_1, y_2 \in \text{Dom}(Y), \forall s_1, s_2 \in \text{Dom}(S)
\]

Statistical Parity / Independence: \( \hat{Y} \perp S \)

- **Worldview:** “We're All Equal” worldview is assumed, and so it is used for mitigating a data bias
- **Information theoretic view:**
  \( \hat{Y} \perp S \iff I(\hat{Y}; S) = 0 \Rightarrow \hat{Y} \) has no information about \( S \)
Statistical Parity / Independence

equality of outcome $\Rightarrow$ Statistical Parity / Independence

$\hat{Y} = 1$

$\hat{Y} = 0$

$\hat{Y} = 1$

$\hat{Y} = 0$

Ratios between positives and negatives in prediction should be matched among all sensitive groups

$\hat{Y} \perp S$
Statistical Parity / Independence

equality of outcome $\Rightarrow$ Statistical Parity / Independence

\[
\begin{align*}
\Pr[\hat{Y} = 1 | S = 0] &= \Pr[\hat{Y} = 0 | S = 1] \\
\Pr[\hat{Y} = 0 | S = 0] &= \Pr[\hat{Y} = 1 | S = 1]
\end{align*}
\]

$\hat{Y} \perp S$

$\Pr[\hat{Y}, S] = \Pr[\hat{Y}] \Pr[S]$

This gap indicates unfair decision

Ratios between positives and negatives in prediction should be matched among all sensitive groups

[Calders+ 10, Dwork+ 12, Barocas+ 19]
Equalized Odds / Separation

Removing inductive bias: calibrating inductive errors to observation

- True positive rates should be matched among all sensitive groups
  \[ \Pr[\hat{Y}=1 \mid Y=1, S=s_1] = \Pr[\hat{Y}=1 \mid Y=1, S=s_2] \quad \forall s_1, s_2 \in \text{Dom}(S) \]
- False positive rates should be matched among all sensitive groups
  \[ \Pr[\hat{Y}=1 \mid Y=0, S=s_1] = \Pr[\hat{Y}=1 \mid Y=0, S=s_2] \quad \forall s_1, s_2 \in \text{Dom}(S) \]

Equalized Odds / Separation: \( \hat{Y} \perp S \mid Y \)

- **Worldview:** “What You See Is What You Get” worldview is assumed, and so it is used for mitigating an inductive bias

[Hardt+ 16, Zafar+ 17]
Equalized Odds

Removing inductive bias $\Rightarrow$ Equalized Odds / Separation

\[
\begin{align*}
Y &= 0 & \hat{Y} = 0 & S = 1 \\
& & \hat{Y} = 1 & \\
& & \hat{Y} = 0 & \\
& & \hat{Y} = 1 & S = 0 \\
Y &= 0 & \hat{Y} = 0 & \\
& & \hat{Y} = 1 & \\
& & \hat{Y} = 0 & \\
& & \hat{Y} = 1 & \\
\end{align*}
\]

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

false positive ratio and true positive ratio should be matched among all sensitive groups

[Hardt+ 16, Zafar+ 17, Barocas+ 19]
Equalized Odds

Removing inductive bias ➞ Equalized Odds / Separation

\[
\begin{align*}
\Pr[Y = 0] & = \Pr[\hat{Y} = 0 | S = s, Y = 0] \\
\Pr[Y = 1] & = \Pr[\hat{Y} = 1 | S = s, Y = 1] \\
\Pr[\hat{Y} = 0 | S = s, Y = 0] & = \Pr[\hat{Y} = 0 | Y] \Pr[S = s | Y] \\
\Pr[\hat{Y} = 1 | S = s, Y = 1] & = \Pr[\hat{Y} = 1 | Y] \Pr[S = s | Y]
\end{align*}
\]

These gaps indicate unfair decision
False positive ratio (FPR) and true positive ratio (TPR) should be matched among all sensitive groups

[Hart+ 16, Zafar+ 17, Barocas+ 19]
**Sufficiency / Calibration**

**Removing inductive bias:** calibrating inductive errors to observation

- Positive predictive values should be matched between any groups
  \[ \Pr[Y=1 \mid \hat{Y}=1, S=s_1] = \Pr[Y=1 \mid \hat{Y}=1, S=s_2] \quad \forall s_1, s_2 \in \text{Dom}(S) \]
- Positive predictive values should be matched between any groups
  \[ \Pr[Y=0 \mid \hat{Y}=0, S=s_1] = \Pr[Y=0 \mid \hat{Y}=0, S=s_2] \quad \forall s_1, s_2 \in \text{Dom}(S) \]

**Sufficiency / Calibration:** \( Y \perp S \mid \hat{Y} \)

- **Worldview:** “What You See Is What You Get” worldview is assumed, and so it is used for mitigating an inductive bias
- In psychology or education disciplines, this criterion is accepted as a fairness condition

[Chouldechova 17]
Removing inductive bias $\rightarrow$ **Sufficiency / Calibration**

Positive and Negative Predictive Values (PPV & NPV) should be matched among all sensitive groups
Removing inductive bias $\rightarrow$ **Sufficiency / Calibration**

These gaps indicate unfair decision

Precisions for positive and negative classes should be matched among all sensitive groups

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

$\Pr[Y, S \mid \hat{Y}] = \Pr[Y \mid \hat{Y}] \Pr[S \mid \hat{Y}]$

$\star \hat{Y}$ and $Y$ are exchanged from the separation case
Context-Specific Independence: \( Y \) and \( S \) are independent, if \( X \) are fixed to specific values, \( x \)

\( \alpha \)-protection

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

- \( \alpha \)-protection is the context-specific independence, \( \hat{Y} \perp S \mid X=x \)

Equalized Odds / Equal Opportunity

- Equalized odds is conditional independence, \( \hat{Y} \perp S \mid Y \)
- Equal Opportunity is context-specific independence, \( \hat{Y} \perp S \mid Y=1 \)

Sufficiency / Predictive Parity

- Sufficiency is conditional independence, \( Y \perp S \mid \hat{Y} \)
- Predictive Parity is context-specific independence, \( Y \perp S \mid \hat{Y}=1 \)
Correlation-Based Fairness

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
- Conditional Independence
- Discovery & Prevention

Statistics
- Correlation
- Partial Correlation
- Discovery only

Statistical Parity / Independence
- Darlington (1971) criterion 4

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

Sufficiency / Calibration
- Darlington (1971) criterion (2)
Association-Based 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 Discrimination — Indirect Discrimination**
- Sensitive information influences targets directly, or indirectly

**Type of Biases to Remove**
- Fairness criteria are designed to remove a specific type of bias

**Relation between Fairness Criteria**
- One criterion implies or conflicts with other criterion

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

[Barocas+ 17, Feldman+ 15]

**Legal notions about fairness**

**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
Direct Discrimination & Indirect Discrimination

Direct Discrimination
discrimination on the basis of sensitive information

Indirect Discrimination
discrimination on the basis of other features resulting in direct discrimination

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

Technical notions about fairness

These technical notions are often expressed by legal terms

[Pedreschi+ 08, Žliobaitė+ 16]
Red-Lining Effect: Simple elimination of a sensitive features from training dataset fails to remove the influence of sensitive information to a target

Eliminating sensitive information is equivalent to replacing an unfair model, $\Pr[Y \mid X, S]$ with a fair model, $\Pr[Y \mid X]$

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

This corresponds to conditional independence: $\hat{Y} \perp\!\!\!\!\perp S \mid X$ (not $\hat{Y} \perp\!\!\!\!\perp S$)

$S$ still influences $Y$ through $X$
Red-Lining Effect:

Elimination of a sensitive information from training dataset fails to remove the influence of the 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 fairness through unawareness.
Types of Bias to Remove

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 remove
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:
- $\hat{Y} \perp S$: statistical parity
- $\hat{Y} \perp S \mid X$: fairness through unawareness
- $\hat{Y} \perp S \mid X=x$: $Y$ and $S$ are context-sensitive independent given $X=x$
**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
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 \( S \).

\( \hat{Y} \perp S \mid Y \) : Equalized Odds / Separation

Empirical errors of \( \hat{Y} \) over sample outcomes, \( Y \), are equal for all groups consist of the same sensitive values.
Satisfiability between Fairness Criteria

\[ \hat{Y} \perp S \]

equality of outcome

\[ Y \perp S | \hat{Y} \]

Calibrating predictive values

equalized odds

\[ \hat{Y} \perp S | Y \]

Calibrating prediction errors to observation

statistical parity

\[ \hat{Y} \perp S \]

fairness through unawareness

\[ \hat{Y} \perp S | X \]

unaware sensitive info

sufficiency

[Žliobaitė+ 16]

[Kleinberg+ 16, Chouldechova 17]

mutually exclusive criteria

simultaneously satisfiable criteria

group fairness
Fairness through Unawareness & Statistical Parity

Satisfying **fairness through unawareness**, \( S \perp \hat{Y} \mid X \)

To simultaneously satisfy **statistical parity**, \( S \perp \hat{Y} \), a condition of \( S \perp X \) OR \( \hat{Y} \perp X \) must be satisfied

\( S \perp X \): a sensitive feature and non-sensitive features are independent

- **unrealistic** \( X \) and \( S \) are uncontrollable, and \( X \) is high-dimensional

\( \hat{Y} \perp X \): a sensitive feature and a target variable are independent

- **meaningless** \( \hat{Y} \) must be random guess

Simultaneous satisfaction of individual fairness and statistical parity is unrealistic or meaningless
**Equalized Odds & Statistical Parity**

**Equalized odds,** \( S \perp \hat{Y} \mid Y \), is satisfied

To simultaneously satisfy **statistical parity,** \( S \perp \hat{Y} \), a condition of \( S \perp Y \) OR \( \hat{Y} \perp Y \) must be satisfied

\( S \perp Y \): a observed class and non-sensitive features are independent

- **violating an assumption** \( \Rightarrow \) observed classes are already fair

\( \hat{Y} \perp Y \): a sensitive feature and a target variable are independent

- **meaningless** \( \Rightarrow \) \( Y \) depends on \( X \) and \( \hat{Y} \) must be random guess

**Simultaneously satisfying equalized odds and statistical parity is meaningless**
Impossibility between Sufficiency and Equalized Odds

[Kleinberg+ 16]

Well-calibration (= sufficiency):
True class distribution given the prediction is independent from groups

\[ Pr[Y | \hat{Y} = \hat{y}] = Pr[Y | \hat{Y} = \hat{y}, S = s], \forall \hat{y}, s \]

Balance for the positive and negative classes (= equalized odds):
TPR and NPR are equal between sensitive groups

\[ \Pr[\hat{Y} = 1 | Y = y, s = 0] = \Pr[\hat{Y} = 1 | Y = y, s = 1], \forall y \]

Perfect prediction: \( \Pr[Y = 1 | x] \in \{0,1\}, \forall x \in \text{Dom}(X) \)

Equal base rates: \( \Pr[Y = 1 | S = 0] = \Pr[Y = 1 | S = 1] \equiv Y \perp S \)

Satisfying sufficiency and equalized odds implies distributions of true class must be either perfect prediction or equal base rates

Sufficiency and Equalized odds cannot be satisfied simultaneously in general
Individual Fairness & Equalized Odds

- Equalized odds, \( \hat{Y} \perp S \mid Y \), and individual fairness, \( \hat{Y} \perp S \mid X \), can be simultaneously satisfiable.

- The resultant combined condition is:
  \[
  \Pr[\hat{Y}, Y, S, X] = \frac{\Pr[\hat{Y} \mid X] \Pr[S \mid X] \Pr[X]}{\Pr[\hat{Y} \mid Y] \Pr[S \mid Y] \Pr[Y]}
  \]

- A condition, \( \hat{Y} \perp S \mid X, Y \), is weaker than the combined condition, but what the two criteria are intended to accomplish is fulfilled.
Explainable Variable / Legally-grounded 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.

- **Sensitive feature** $S$
- **Target** $Y$
- **Explainable variable** $E$

**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.

[Žliobaitė+ 11, Kamiran+ 13]
Fair Determination

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: $Y$
- admission
  - admit / not admit

Fair determination: the gender does not influence the acceptance

(unconditional) independence: $Y \perp\!\!\!\!\!\!\!\!\!\!\perp S$
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:** $Y$
- admission
- admit / not admit

**Explainable feature:** $E$
- (confounding feature)
- program
- medicine / computer

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

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

**Conditional statistical independence:** $Y \perp S \mid E$
Association-Based Fairness: Measures
**Difference-based Measures**

### Risk Difference / Mean Difference

[Calderas+ 10, Pedreschi 09]

**Definition:** Difference of receiving advantageous decisions between groups

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

- **RD → 0** ⟹ \(Y \perp S\)
- Equivalent to the total causal effect of changing \(S\) on \(\hat{Y}\)

### Balanced Error Ratio

[Feldman+ 15]

**Definition:** 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 → 1/2** ⟹ \(Y \perp S\)
elift (extended lift) = \frac{\text{conf}(X=x, S=0 \Rightarrow Y=0)}{\text{conf}(X=x \Rightarrow Y=0)}

the ratio of the confidence of a rule with a sensitive condition,
to that of a rule without the condition

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
Measures from Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>( \hat{Y} = 0 )</th>
<th>( \hat{Y} = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S = 0 )</td>
<td>( a_1 )</td>
<td>( n_1 - a_1 )</td>
</tr>
<tr>
<td>( S = 1 )</td>
<td>( a_2 )</td>
<td>( n_2 - a_2 )</td>
</tr>
</tbody>
</table>

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

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

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

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

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

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

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

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

\[
p_0(1 - p_1)
\]

\[
p_1(1 - p_0)
\]

\[
= \text{odds ratio / olift}
\]
Counterfactual Fairness
Pearl's Ladder of Causation

**Counterfactuals**

**Activity:** Imaging, Retrospection, Understanding

**Questions:** What if I had done …? Why?

**Examples:** Was it the aspirin that stopped my headache?

**Intervention**

**Activity:** Doing, Intervening

**Questions:** What if I do …? How?

**Examples:** If I take aspirin, will my head ache be cured

**Association**

**Activity:** Seeing, Observing

**Questions:** What if I see …?

**Examples:** What does a symptom tell me about a disease?
Counterfactual Fairness

**Observations (Facts):** If a sensitive feature is $S = s$ and the corresponding non-sensitive features, $X_{S=s}$, are given, an outcome, $Y = y$, is observed.

![Diagram](image)

**Counterfactuals:** Even if a sensitive feature was changed so that $S = s'$ while holding the non-sensitive features fixed, it was fair if an outcome of a predictor is unchanged

![Diagram](image)

**Ethical viewpoint** [Lippert-Rasmussen 06]

- **Harm-based Account with counterfactual baseline:**
  the discriminatees had not been subjected to the discrimination

- To establish a disparate-treatment claim under this plain language, a plaintiff must prove that age was the but-for cause of the employer's adverse decision.
- A plaintiff must prove by a preponderance of the evidence (which may be direct or circumstantial), that age was the but-for cause of the challenged employer decision.

The but-for cause: After occurring $X$ and $Y$, if $X$ was not occurred, whether or not $Y$ would be occur?

- In a causal inference context, this is interpreted as probability of necessity, that is the probability of the counterfactual, $Y_{X=0}$, is 0 given facts $X = 1$ and $Y = 1$.

$$\Pr[Y_{X=0} = 0 \mid X = 1, Y = 1]$$
A counterfactual outcome, $Y_{S=s'}$, if a sensitive feature, $S$, was changed from $s$ to $s'$, could be estimated by conditioning by non-sensitive features (confoundings), $X$

**Stable Unit Treatment Value Assumption (SUTVA)**
- Each individual will have the same effect of treatment regardless of what treatment the other individuals receive.

**Consistency**
- The same effect that is observed by experimental design will be observed in a real world.

**Ignorability**
- The potential outcome, $Y_s$, is independent of the treatment actually received, $S$, given the values of a certain set of confoundings, $X$. 

[Pearl+ 18]
Ignorability

Example of the violation of ignorability

An Employee's salary, $Y$, depends on their “Years of Education”, $S$, and “Years of Experience”, $X$.

If an employee had received longer years of education, how much would their salary be?

Due to the conditioning of $X$, the influence of $S$ on $Y$ through $X$ is distorted.

The violation of ignorability.
Association-based fairness and counterfactual fairness can commonly cope with the influence of sensitive information / cause to a target / effect.

**Objective:** estimating the degree of the effect between variables

**Approach:** based on a causal graph representing strict causal relationships

**Objective:** finding or removing direct influence from a sensitive to a target

**Approach:** Modifying models so as to satisfy pre-specified associative constraints
Causality and Association

Total causal effect of changing a sensitive feature, $S$, on a target, $Y$

Intervention on a sensitive feature $S = 0$, $S = 1$ and change decision $Y = 0$, $Y = ?$

Any direct and indirect causal paths are considered.

Total causal effect:

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

Total causal effect is equal to the risk difference:

$$TE(S = 1, S = 0) = \Pr[Y = 1 \mid S = 1] - \Pr[Y = 1 \mid S = 0]$$
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:

\[
Y \perp S \mid E \quad \rightarrow \quad Y \perp S \mid e(S)
\]

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

<table>
<thead>
<tr>
<th>Strata</th>
<th>Propensity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strata 1</td>
<td>( e(S) \in [0, 1/3) )</td>
</tr>
<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>
Part III
Fairness-Aware Machine Learning
Tasks of Fairness-Aware Machine Learning
Tasks of Fairness-Aware ML

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

A pattern between inputs and decisions (labels)

classification rule

Inference

Is this an apple?

Yes

inference based on the learned rule
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

![Diagram showing the transformation from potentially unfair data to fair data with a standard classifier applied. The diagram includes labeled points and lines indicating the flow from the input space to the fair sub-space.](image-url)
Unfairness Prevention:  
In-Process Approach

**In-Process:** a fair model is learned directly from a potentially unfair dataset ③

- 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 ⁴, and then the learned classifier is modified to satisfy a fairness constraint 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 in Association-Based Fairness Cases
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**

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

**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

[Pedreschi+ 08, Ruggieri+ 10]

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 α

ex: rule (b) isn’t α-protected if a = 2, because elift = \( \frac{\text{conf(b)}}{\text{conf(a)}} \) = 3

Direct Discrimination: a target directly depends on a sensitive feature
Pr[ loan=deny | city=NYC, race=African ] ⇒ Pr[ loan=deny | city=NYC ]
**Indirect Discrimination:** a target 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

Ex. A literal ‘city=NYC ∧ ZIP=10451’ is a redescription of ‘city=NYC ∧ race=African’
Situation Testing by $k$-NN

**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.

$k$-nearest neighbors of a protected member

- The statistics of decisions in $k$-nearest neighbors of data points in a protected group
- Condition of situation testing is

$$\Pr[ Y \mid E, S=0 ] = \Pr[ Y \mid E, S=1 ] \equiv Y \perp S \mid E \ (E: \text{explainable variables})$$

[Luong+ 11]
Unfairness Discovery: Discovery from Models
Gradient Feature Auditing

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

$(X_i, S)$ original data

Blackbox Model $\rightarrow Y$

$(X_i, S')$ sensitive is perturbed

Blackbox Model $\rightarrow Y'$

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

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

$(X_i, S)$ original data

Blackbox Model $\rightarrow Y$

$(X_i', S)$ non-sensitive is perturbed

Blackbox Model $\rightarrow Y'$

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

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

**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 (Individual Fairness)

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

**Vendor (data user)**
- Original data \(M(x)\)
- Archtype
- Loss function \(L(x, y)\)
- Loss function representing utilities for the vendor
- Distance between archtypes
- Distance between original data
- Lipschitz condition: similar data are mapped to similar archtypes

**Individual Fairness**: Treat like cases alike
1. Map original data to archtypes so as to satisfy Lipschitz condition
2. Make prediction referring the mapped archetypes

**Lipschitz condition**: similar data are mapped to similar archtypes

\[
D(M(x_1), M(x_2)) \leq d(x_1, x_2)
\]
**Statistical Parity:** protected group, $S$, and non-protected group, $\bar{S}$, are equally treated

Mean of protected archetypes and mean of non-protected archetypes should be similar

$$D(\mu_S, \mu_{\bar{S}}) \leq \epsilon$$

- If original distributions of both groups are similar, Lipschitz condition implies statistical parity
- If not, statistical parity and individual fairness cannot be satisfied simultaneously
- To satisfy statistical parity, protected data are mapped to similar non-protected data while the mapping is as uniform as possible
Learning Fair Representations

Requirements for Prototypes

- Probabilities assigned to each prototype is equal between groups
  \[ L_z = \sum_k |M_i^{S=0} - M_i^{S=1}| \]
- 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
Removing Disparate Impact

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)
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

\[ - \sum_s \sum_{\mathcal{D}(s)} \ln \Pr[y \mid x; \Theta^{(s)}] + \frac{\lambda}{2} \sum_s \|\Theta^{(s)}\| + \eta I(Y; S) \]

fairness parameter to adjust a balance between accuracy and fairness

- A class distribution, \( \Pr[Y \mid X; \Theta^{(s)}] \), is modeled by a set of logistic regression models, each of which corresponds to \( s \in \text{Dom}(S) \)
  \[
  \Pr[Y = 1 \mid x; \Theta^{(s)}] = \text{sig}(w^{(s)\top} x)
  \]

- As a prejudice remover regularizer, we adopt a mutual information between a target and a sensitive feature, \( I(Y; S) \)
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 \( \Pr[Y] \Pr[S] \Pr[X | Y, S] \)

**Not Independent in general**
Bayes optimal Labels are generated by a deterministic decision rule:
\[
y^* \leftarrow \arg \max_y \Pr[y|x, s]
\]
**Model-based Independence:** Class labels are assumed to be generated probabilistically

\[
\hat{Y} \perp S, \text{ where } (\hat{Y}, S) \sim \Pr[\hat{Y}, S]
\]

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

\[
\tilde{Y} \perp S, \text{ where } (\tilde{Y}, S) \sim \Pr[\tilde{Y}, S] = \sum_s \Pr[s] \frac{1}{n} \sum_{x \in \mathcal{D}_s} \Pr[\tilde{Y} | x, s]
\]

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

satisfy actual independence instead of model-based independence

Fairness in class labels can be greatly improved
Correlation-based Fairness

Quantify unfairness by covariance, proportional to correlation

\[
\text{Cov}(Y, S) = \mathbb{E}[YS] - \mathbb{E}[Y] \mathbb{E}[S] \\
= \mathbb{E}[d_\theta(x)(s - \bar{S})] - \mathbb{E}[d_\theta(x)]E[s - \bar{S}] \\
= \frac{1}{N} \sum_{i}^{N} (s_i - \bar{S}) d_\theta(x)
\]

This constraint is convex, helpful for the easy optimization

- \(d_\theta(x)\) is a signed distance from \(x\) to the decision boundary, and is equal to

\[
d_\theta(x) = \theta^\top x
\]

in a linear model with a parameter \(\theta\)

minimize accuracy loss under fairness constraints

\[
\min_{\theta} \text{loss}(\theta) \text{ s.t. } |\text{Cov}(Y(\theta), S)| \leq \eta
\]

maximize fairness under accuracy constraints

\[
\min_{\theta} |\text{Cov}(Y(\theta))| \text{ s.t. } \text{loss}(\theta) \leq (1 + \eta) \text{loss}(\theta^*)
\]
Adversarial Learning

Gradient-based learner for fairness-aware prediction

Predictor minimizes $\text{loss}_P(Y, \hat{Y}; \Theta)$, to predict outputs as accurately as possible while preventing adversary's objective.

Adversary minimizes $\text{loss}_A(S, \hat{S}; W, V)$, to violate fairness condition.

For accurate prediction & not beneficial for adversary:

- $\nabla_{\Theta} \text{loss}_P - \text{proj}_{\nabla_{\Theta} \text{loss}_A} \nabla_{\Theta} \text{loss}_P - \eta \nabla_{\Theta} \text{loss}_A$

Beneficial for adversary's objective:

- $\nabla_{\Theta} \text{loss}_P - \text{proj}_{\nabla_{\Theta} \text{loss}_A} \nabla_{\Theta} \text{loss}_P$
To prevent the prediction of $S$, gradients from a classifier is propagated straightforward, but those from an adversary is multiplied by $-1$ in backpropagation.
An embedding $Z$ is generated so that

- minimize the reconstruction error between $X$ and $X'$
- minimize the prediction error of the classifier
- maximize the prediction error of the optimized adversary
Unfairness Prevention: Classification (post-process)
Calders-Verwer’s 2-Naive-Bayes

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.
Calders-Verwer’s 2-Naive-Bayes

parameters are initialized by the corresponding sample distributions

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

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

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

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

\[\text{while} \quad \Pr[Y=1 \mid S=1] - \Pr[Y=1 \mid S=0] > 0 \]

\[\text{if} \quad \# \text{ of data classified as “1” } < \# \text{ of “1” samples in original data} \quad \text{then}\]

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

\[\text{else}\]

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
Hardt's Method

Given unfair predicted class, $\hat{Y}$, and a sensitive feature, $S$, a fair class, $Y^*$, is predicted **maximizing accuracy** under an **equalized odds** condition.

* True class, $Y$, cannot be used by this predictor.

The perfectly accurate point satisfying an equalized odds condition is:

- $\Pr[Y^*=1 \mid \hat{Y}=1, S=1] = 1.0$
- $\Pr[Y^*=1 \mid \hat{Y}=0, S=1] = 0.0$

The feasible region for $S=0$ is:

- $\Pr[Y^*=1 \mid \hat{Y}=1, S=1] = 0.0$
- $\Pr[Y^*=1 \mid \hat{Y}=0, S=1] = 0.0$

The feasible region for $S=1$ is:

- $\Pr[Y^*=1 \mid \hat{Y}=1, S=1] = 1.0$
- $\Pr[Y^*=1 \mid \hat{Y}=0, S=1] = 1.0$

The false positive ratio (FPR) $\Pr[Y=1 \mid \hat{Y}=1, S=0]$ and true positive ratio (TPR) $\Pr[Y=1 \mid \hat{Y}=1, S=1]$ can be matched by the points where $S=0$. The points where $S=1$ satisfy the equalized odds condition.
Unfairness Prevention: Recommendation
Recommender System

Recommenders: Tools to help identify worthwhile stuff

[Konstan+ 03]

Find Good Items

Predicting Ratings

[Herlocker+ 04, Gunawardana+ 09]

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.

Any good sushi restaurant?

The "Taro" is awesome

They like the "Taro" restaurant

I like the "Taro"

I'll go to the "Taro"

* 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
Multistakeholder in Recommendation

Utilities of multiple stakeholders
example cases in job recommendation

**Consumer**: End-users who receive recommendation
- Applicants want to be highly evaluated their own experience or skills

**Provider**: Entities that supply recommended objects
- Employers should be exposed frequently

**System**: A platform who manages a recommender system
- Increasing job-matchings is beneficial for the system owner

*These fairness constraints might conflict*
- Equal exposure of employers
- Employers can be recommended less matched employers frequently
- Less matches reduces the profit of the system owner

[Abdollahpouri+ 20]
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 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
user-dependent bias
item-dependent bias
cross effect of users and items

Objective Function

\[
\sum_{D} (r_i - \hat{r}(x_i, y_i))^2 + \lambda \| \Theta \|_2^2
\]

squared loss function
regularization parameter
$L_2$ regularizer

For a given training dataset, model parameters are learned by
minimizing the squared loss function with an $L_2$ regularizer
Independence Enhanced PMF

[Kamishima+ 12, Kamishima+ 13, Kamishima+ 18]

Prediction Function

a prediction function is selected according to a sensitive value

\[
\hat{r}(x, y, s) = \mu^{(s)} + b^{(s)}_x + c^{(s)}_y + p^{(s)}_x q^{(s)}_y \tag{6.17}
\]

Objective Function

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

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

**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]

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

- matching means of predicted ratings for distinct sensitive groups
- improved computational efficiency, but considering only means

Mutual Information with Normal Distributions [Kamishima+ 18]

\[
-\left( H(R) - \sum_s \Pr[s] H(R|s) \right)
\]

Distribution Matching with Bhattacharyya Distance [Kamishima+ 18]

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

- These two terms can take both means and variances into account, and are computationally efficient
Latent Class Model: A probabilistic model for collaborative filtering

Latent topic variable

basic topic model, pLSA

A rating value can be predicted by the expectation of ratings

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
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 \( Z \)

Experimental results show that the performance of these two models are nearly equal
Unfairness Prevention: Ranking
**Ranking**

**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](image)

![Irrelevant Items](image)
**Fair Ranking**: for each rank $i = 1, \ldots, k$, the ratio between two sensitive groups must not diverge 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 the fair ranking condition.

_Merged Ranking list_  |  _Ranking list within each sensitive group_
---|---
1.0  |  1.0
1.0  |  1.0
0.7  |  0.9
0.9  |  0.7
0.7  |  0.5
0.9  |  0.3
0.3  |

This item is less relevant, but it is prioritized to maintain fairness.
Singh's Method

Singh's method is an in-process type ranking algorithm

**Step 1:** optimize $P$ by solving the linear programming problem

$$
\min_P \sum_{d_i \in \mathcal{D}} \sum_{j=1}^{N} u(d_i \mid q) P_{i,j} v_j
$$

subject to: $P$ satisfies the constraints of probabilities and the following fairness constraint (statistical parity)

$$
\sum_{d_i \in \mathcal{D}} \sum_{j=1}^{N} \left( \frac{I(d_i \in \mathcal{D}_0)}{|\mathcal{D}_0|} - \frac{I(d_i \in \mathcal{D}_1)}{|\mathcal{D}_1|} \right) P_{i,j} v_j = 0
$$

**Step 2:** By applying the Birkhoff-von Neumann decomposition to $P$, get probability masses of corresponding rankings
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

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

Debiasing Embeddings
- **neutralize**: non-gender words are uncorrelated to gender vector
- **equalize**: equal distance from occupational words to gender words
Fairness GAN: Fair Data Generator

Generative adversarial network for fair data generation

\[ \text{Dataset} \xrightarrow{(X_r, Y_r, S_r)} \text{Discriminator} \]

\[ \begin{align*}
\text{Discriminator predict real or fake?} \\
\text{Pr}[D_X | X] & \quad \text{Pr}[D_{XY} | X, Y] \\
\text{Pr}[S | X] & \quad \text{Pr}[S | Y]
\end{align*} \]

\[ \text{Random seed} \xrightarrow{(S, \text{Rnd})} \text{Generator} \xrightarrow{(X_f, Y_f)} \text{Discriminator} \]

**Likelihood to maximize**

**Discriminator**

\[ \mathcal{L}(D_X | X_{r,f}) + \mathcal{L}(D_{XY} | X_{r,f}, Y_{r,f}) + \mathcal{L}(S | X_r) + \mathcal{L}(S | Y_r) \]

**Generator**

\[ - \left( \mathcal{L}(D_X | X_f) + \mathcal{L}(D_{XY} | X_f, Y_f) \right) + \mathcal{L}(S | X_f) - \mathcal{L}(S | Y_f) \]

Discriminator predicts whether real or fake, but generator prevents it generating high-quality data.

Data conditioned on input sensitive value preventing to predict \( S \) from \( Y \).

Ensuring statistical parity.
**Fair Bandit**

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

- **player**
- **bandits**

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
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
Part IV
Other Topics
Mitigation of a Sample Selection Bias
Zadrony's Theorem

\[(x, y) \text{ is sampled and observed if } z = 1; \text{ it is not sampled if } z = 0\]

- \((x, y) \perp z\) : i.i.d. data \(\Rightarrow\) no problem
- \(x \perp z \mid y\) : sampled depending on \(y\) \(\Rightarrow\) replacing prior \(\text{Pr}[Y]\)
- \(y \perp z \mid x\) : sampled depending on \(x\)
  \(\Rightarrow\) assumption of this theory: The values of \(X\) influence whether or not a datum is observed, but those of \(y\) do not

\[\text{Under the assumption of } y \perp z \mid x \text{ and } \text{Pr}[x] > 0, \text{ local learners are NOT affected by a sample selection bias, but global learners are}\]

- **Local:** the output of learner depends only on \(\text{Pr}[y \mid x]\)
  - full Bayes, logistic regression, hard-margin SVM
- **Global:** the output of learner depends on both \(\text{Pr}[y \mid x]\) and \(\text{Pr}[x]\)
  - naïve Bayes, decision trees, soft-margin SVM
Zadrony's Theorem

Only $x$ is observed for both $z=0$ and $z=1$ data.

- Under the assumption of $y \perp z \mid x$ and $\Pr[x] > 0$, a likelihood function, $\Pr[y \mid x]$, is unbiased, even if it is learned only from approved data.
- $\Pr[z \mid x]$ can be estimated from all applicants data.

A leaner free from a sample selection bias can be trained by maximizing the weighted log-likelihood

$$\max_\Theta \sum_{z=1} \frac{\Pr[z = 1]}{\Pr[z = 1 \mid x]} \log \Pr[y \mid x; \Theta]$$
Predictors might be applied to data distributed differently from a distribution that it has been trained

**Covariate Shift:** $\Pr[X, S]$ is different between test and training, but $\Pr[y \mid X, S]$ is same

A distribution of $S$ in training is $\Pr[S]$, and that in test is $\tilde{\Pr}[S]$

Given a joint distribution of $X$ and $S$ in training, $\Pr[X \mid S]$ and , that in test is:

$$\tilde{\Pr}[X, S] = \sum_s \frac{\tilde{\Pr}[S]}{\Pr[S]} \Pr[X \mid S]$$

Under the covariate shift assumption, a predictor maximizing the weighted log-likelihood is unbiased

$$\max_{\Theta} \sum_{x,s,y} \frac{\tilde{\Pr}[x, s]}{\Pr[x, s]} \log \Pr[y \mid x; \Theta]$$
Disclosure
Paul Zilly heard his COMPAS score, and his lawyer agreed to a plea deal of one year imprisonment, in a court in Barron County, Wisconsin.

Judge James Babler had seen Zilly's high-risk score, and the judge overturned the deal and imposed two year imprisonment.

In an appeal hearing, the developer of the COMPAS, Brennan, testified that the COMPAS was designed not for sentencing.

Zilly’s sentence was reduced to 1.5 years imprisonment.

In theory, the COMPAS is designed to determine which defendants are eligible for probation or treatment programs.

Like this case, the disclose of the design intent of the model is important for correcting such a misuse.
For a Proper Use of the ML

How to use ML techniques as a tool properly

Quality Control as in Other Industrial Products

- **Design**: datasets, algorithms
- **Test**: performance test, explainable ML
- **Maintenance**: monitoring, model updation

Given a fairness criterion, an algorithm meets to the criterion can be built

Disclosing which criterion the algorithm is designed to satisfy, and why the criterion is proper for the target task

* In a case of the COMPAS, the US court adopts the sufficiency criterion based on the federal Post Conviction Risk Assessment
Model Card: standardizing ethical practice and reporting

Model Details (developer, date, version, ...)

Intended Use

Factors (features, evaluation factors, ...)

Metrics (summary statistics, performance)

Training Data

Ethical Considerations

Caveats and Recommendation
Datasheet for Datasets

- Standardized process for documenting datasets
- Intended to consider potential risks and underlying assumptions

Dataset creators should answer the 57 questions at 7 stages of creating a dataset

**motivation**
- purpose, creator, funding

**composition**
- content, size, sampling, features, missing info, splits, noises, external datasets, confidentiality, offensiveness, demographics, identity, sensitive info

**collection process**
- method, instruments, sampling, data operators, collection period, ethical review, directly collected, consent, cancel agreement, influence

**Uses**
- use cases, repository, possible use cases, influence of preprocess, prohibited cases

**preprocessing / cleaning / labeling**
- methods, raw data, software

**distribution**
- distributor, method, date, license, limitation, regulation

**maintenance**
- maintainer, contact info, errata, updates, restrictions by subjects, older version, third-party updates
Other Fairness-Aware Machine Learning Topics
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
Relation to Other Machine Learning 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

Difference from PPDM
- introducing randomness is occasionally inappropriate for severe decisions, such as job application
- disclosure of identity isn’t problematic in FAML, generally
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
\[
\frac{C(1 \mid 0)}{C(0 \mid 1)}
\]
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.

* 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
- Machine learning 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
Software
Non-enterprise Software

- AI Fairness 360 (IBM)
- Fairlearn (Microsoft)
- What-If Tool, ML-fairness-gym (Google)

Commercial Packages: DataRobot, Fiddler AI

Non-commercial Packages: FairTest, Fairness Measures, Aequitas, Fairkit-learn

Enterprise Software

- LinkedIn Fairness Toolkit (LinkedIn)
- Amazon SageMaker (Amazon)
**AI Fairness 360 (AIF360):** [https://github.com/Trusted-AI/AIF360](https://github.com/Trusted-AI/AIF360)

- Software packages for measuring and mitigating fairness
- Developed by IBM, implemented in Python

**Dataset class:** In addition to the information required for standard ML algorithms, the sensitive information is maintained, and dealing with CSV files or a Pandas DataFrame

**Metric class:** Evaluate the achievement of the target fairness criteria

**Explainer class:** Report fairness metric in a text or JSON format, including Web interface

**Bias Mitigating Algorithms:** 4 pre-processing, 2 in-processing, and 3 post-processing algorithms

* These documented specifications might be updated in the latest version
**Fairlearn:** [https://fairlearn.org/](https://fairlearn.org/)

- Developed by Microsoft, implemented in Python
- Mitigating allocation harms and quality-of-service harms

**Interactive visualization dashboard**
- Visualize the disparities between sensitive groups

**Unfairness mitigation algorithms**
- **Hardt's method:** Tuning decision boundaries for each sensitive group to minimize the disparity between the groups
- **Reduction algorithms:** Iterate re-weighting data points and re-training models, to minimize the disparity between sensitive groups

* These documented specifications might be updated in the latest version
LinkedIn Fairness Toolkit (Lift): [https://github.com/linkedin/LiFT](https://github.com/linkedin/LiFT)

- Enterprise software for measuring and mitigating fairness
- Developed by LinkedIn
- Implemented in Scala, parallel computation using the Apache Spark

Check whether a collected dataset represents original population before training

Tune hyperparameters to satisfy fairness criteria while training

Check the trained model and mitigate unfairness

Watch the performance of deployed model to avoid model or data drifts
Evidence-Based Decision Making
Once proper regulation is in place, better algorithms can help to ensure equitable treatment in our society. The data for training and test are carefully stored, and regulatory agency with trained auditors process data.
The Three I's Problem

3I: Ideology, Ignorance, Inertia

why policies fail and why aid does not have the effect it should

The nurses’ workload was based on an ideology that wants to see nurses as dedicated social workers, designed in ignorance of the conditions on the ground, that lives on, mostly just on paper, because of inertia.

If we resist the kind of lazy, formulaic thinking that reduces every problem to the same set of general principles; … if we accept the possibility of error and subject every idea, …, to rigorous empirical testing, then we will be able not only to construct a toolbox of effective policies but also to better understand why the poor live the way they do.


M. Hildebrandt. Rude awakenings from behaviourists dreams. the methodological integrity and the gdpr. The 13th ACM Conf. on Recommender Systems, Keynote, 2019.


