Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining

Part I: Introduction and context

Part II: Discrimination discovery

Part III: Fairness-aware data mining

Part IV: Challenges and directions for future research

Discussion and further questions
Non-discriminatory data-driven decision-making
Non-discriminatory data-driven decision-making
Non-discriminatory data-driven decision-making
Non-discriminatory data-driven decision-making
Non-discriminatory data-driven decision-making
Fairness-aware data mining: common aspects

**Goal:** develop a non-discriminatory decision-making process while preserving as much as possible the quality of the decision.

**Steps:**
1. Defining anti-discrimination/fairness constraints
2. Transforming data/algorithm/model to satisfy the constraints
3. Measuring data/model utility
Non-discriminatory data-driven decision-making
Fairness-aware data mining

Pre-processing approaches:


Fairness-aware data mining

Pre-processing approaches:


Given $D = (X,Y,C)$ which has been certified having disparate impact potential, where $X$ is protected attribute, $Y$ the remaining attributes, and $C$ is the decision class.

Generate a dataset $D_{\text{new}} = (X, Y_{\text{new}}, C)$ with no disparate impact.

**Utility goal:** to preserve rank within each marginal distribution $P(Y \mid X = x)$.
Using the earthmover distance

Let

\[
P_i = \Pr(Y = y | X = i)
\]

\[
F_i = \text{cdf of } P_i
\]

\[
P_* = \arg \min \sum_i d_{EM}(P, P_i)
\]

\[
F_*^{-1}(\lambda) = \text{median } F_i^{-1}(\lambda)
\]

We find a new distribution that is “close” to all conditional distributions.
Fairness-aware data mining

Pre-processing approaches:


“Data preprocessing techniques for classification without discrimination” (Kamiran and Calders, 2012)

**Discrimination measure:**

\[ P(C = + | B = b_1) - P(C = + | B = b_2) \]

**Goal:** minimize discrimination, while maximizing accuracy

**Techniques for removing dependencies from the input data:**

1. Suppression (baseline, just remove B and the top-k attributes most correlated with B)
2. Massaging
3. Reweighing
4. Sampling
## Job application example

<table>
<thead>
<tr>
<th>Sex</th>
<th>Ethnicity</th>
<th>Highest Degree</th>
<th>Job Type</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>native</td>
<td>h. school</td>
<td>board</td>
<td>+</td>
</tr>
<tr>
<td>m</td>
<td>native</td>
<td>univ.</td>
<td>board</td>
<td>+</td>
</tr>
<tr>
<td>m</td>
<td>native</td>
<td>h. school</td>
<td>board</td>
<td>+</td>
</tr>
<tr>
<td>m</td>
<td>non-nat.</td>
<td>h. school</td>
<td>healthcare</td>
<td>+</td>
</tr>
<tr>
<td>m</td>
<td>non-nat.</td>
<td>univ.</td>
<td>healthcare</td>
<td>-</td>
</tr>
<tr>
<td>f</td>
<td>non-nat.</td>
<td>univ.</td>
<td>education</td>
<td>-</td>
</tr>
<tr>
<td>f</td>
<td>native</td>
<td>h. school</td>
<td>education</td>
<td>-</td>
</tr>
<tr>
<td>f</td>
<td>native</td>
<td>none</td>
<td>healthcare</td>
<td>+</td>
</tr>
<tr>
<td>f</td>
<td>non-nat.</td>
<td>univ.</td>
<td>education</td>
<td>-</td>
</tr>
<tr>
<td>f</td>
<td>native</td>
<td>h. school</td>
<td>board</td>
<td>+</td>
</tr>
</tbody>
</table>
Massaging

a) rank individuals

b) change the labels
Input dataset

Job=No  Job=Yes

Learn a ranker

Decision boundary

Final Model

Learn a Classifier

Relabel
Reweighing

a) calculate weights for the objects to neutralize the discriminatory effects from data

b) assign weights to make the data impartial

- - - - - - + + + + + favored
- - - - - - + + + deprived
Reweighing

\[ W(x(B) = b \mid x(Class) = +) := \frac{P_{exp}(b \land +)}{P_{act}(b \land +)} \]

\[ W(x(B) = b \mid x(Class) = -) := \frac{P_{exp}(b \land -)}{P_{act}(b \land -)} \]

\[ W(x(B) = \overline{b} \mid x(Class) = +) := \frac{P_{exp}(\overline{b} \land +)}{P_{act}(\overline{b} \land +)} \]

\[ W(x(B) = \overline{b} \mid x(Class) = -) := \frac{P_{exp}(\overline{b} \land -)}{P_{act}(\overline{b} \land -)} \]

<table>
<thead>
<tr>
<th>Sex</th>
<th>Ethnicity</th>
<th>Highest Degree</th>
<th>Job Type</th>
<th>Cl.</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>native</td>
<td>h. school</td>
<td>board</td>
<td>+</td>
<td>0.75</td>
</tr>
<tr>
<td>m</td>
<td>native</td>
<td>univ.</td>
<td>board</td>
<td>+</td>
<td>0.75</td>
</tr>
<tr>
<td>m</td>
<td>native</td>
<td>h. school</td>
<td>board</td>
<td>+</td>
<td>0.75</td>
</tr>
<tr>
<td>m</td>
<td>non-nat.</td>
<td>h. school</td>
<td>healthcare</td>
<td>+</td>
<td>0.75</td>
</tr>
<tr>
<td>m</td>
<td>non-nat.</td>
<td>univ.</td>
<td>healthcare</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>f</td>
<td>non-nat.</td>
<td>univ.</td>
<td>education</td>
<td>-</td>
<td>0.67</td>
</tr>
<tr>
<td>f</td>
<td>native</td>
<td>h. school</td>
<td>education</td>
<td>-</td>
<td>0.67</td>
</tr>
<tr>
<td>f</td>
<td>native</td>
<td>none</td>
<td>healthcare</td>
<td>+</td>
<td>1.5</td>
</tr>
<tr>
<td>f</td>
<td>non-nat.</td>
<td>univ.</td>
<td>education</td>
<td>-</td>
<td>0.67</td>
</tr>
</tbody>
</table>

\[ P_{exp}(Sex = f \mid x(Class) = +) = 0.5 \times 0.6 \]

\[ W(Sex = f \mid x(Class) = +) = \frac{0.5 \times 0.6}{0.2} = 1.5 \]

\[ W(Sex = f \mid x(Class) = -) = 0.67 \]

\[ W(Sex = m \mid x(Class) = +) = 0.75 \]

\[ W(Sex = m \mid x(Class) = -) = 2 \]
Sampling

Similarly to reweighing, compare the expected size of a group with its actual size, to define a sampling probability.

\[
DP := \{x \in D \mid x(B) = b \land x(Class) = +\}
\]

\[
DN := \{x \in D \mid x(B) = b \land x(Class) = -\}
\]

\[
FP := \{x \in D \mid x(B) = \overline{b} \land x(Class) = +\}
\]

\[
FN := \{x \in D \mid x(B) = \overline{b} \land x(Class) = -\}
\]

Then sample accordingly, possibly duplicating data points.
Uniform Sampling
Preferential Sampling

Diagram showing the process of preferential sampling with categories DN (Deprived Not), DP (Desired Positive), FN (Favored Not), and FP (Favored Positive). The diagram illustrates the removal or duplication of communities based on their classification.
## Performance

<table>
<thead>
<tr>
<th>Preprocess method</th>
<th>Disc (%)</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>16.4 ± 1.31</td>
<td>86.05 ± 0.29</td>
</tr>
<tr>
<td>No-SA</td>
<td>16.6 ± 1.43</td>
<td>86.01 ± 0.31</td>
</tr>
<tr>
<td>RW</td>
<td>7.97 ± 1.02</td>
<td>85.62 ± 0.30</td>
</tr>
<tr>
<td>US</td>
<td>7.91 ± 2.05</td>
<td>85.35 ± 0.36</td>
</tr>
<tr>
<td>PS</td>
<td>3.08 ± 0.79</td>
<td>84.30 ± 0.25</td>
</tr>
<tr>
<td>M-NBS</td>
<td>1.77 ± 1.16</td>
<td>83.65 ± 0.24</td>
</tr>
<tr>
<td>M-J48</td>
<td>2.49 ± 1.92</td>
<td>83.49 ± 0.47</td>
</tr>
<tr>
<td>M-IBk1</td>
<td>7.67 ± 0.86</td>
<td>85.35 ± 0.46</td>
</tr>
<tr>
<td>M-IBk2</td>
<td>3.62 ± 0.61</td>
<td>84.44 ± 0.27</td>
</tr>
<tr>
<td>M-IBk3</td>
<td>2.40 ± 0.51</td>
<td>83.78 ± 0.43</td>
</tr>
</tbody>
</table>
Fairness-aware data mining

Pre-processing approaches:


A framework for direct and indirect discrimination prevention in data mining
Based on direct discriminatory measures $f \in \{\text{elift, slift, ...}\}$, a PD classification rule $r: A, B \rightarrow C$ is:

- $\alpha$-discriminatory if $f(r) \geq \alpha$; or
- $\alpha$-protective if $f(r) < \alpha$

$\alpha$ states an acceptable level of discrimination according to laws and regulations

* e.g. U.S. Equal Pay Act: This amounts to using $\text{slift}$ with $\alpha = 1.25$. 

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$\neg C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>$n_1 - a_1$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>$a_2$</td>
<td>$n_2 - a_2$</td>
<td>$n_2$</td>
</tr>
</tbody>
</table>

$p_1 = a_1/n_1$, $p_2 = a_2/n_2$, $p = (a_1 + a_2)/(n_1 + n_2)$

$\text{elift}(r) = \frac{p_1}{p}$, $\text{elift}_d(r) = p_1 - p$, $\text{elift}_c(r) = \frac{1 - p_1}{1 - p}$

$\text{slift}(r) = \frac{p_1}{p_2}$, $\text{slift}_d(r) = p_1 - p_2$, $\text{slift}_c(r) = \frac{1 - p_1}{1 - p_2}$
Measure discrimination

Original dataset

Classification rule mining

Frequent rules $A, B \rightarrow C$

Discrimination threshold $\alpha$

Discrimination measure $f$

Check discrimination for each PD rule

$\alpha$-discriminatory rules $A, B \rightarrow C$
Data transformation

The purpose is transform the original data $D$ in such a way to remove direct and/or indirect discriminatory biases, with minimum impact on the data, and on legitimate decision rules.

Different metrics and algorithms have been developed to specify

Which records (and in which order) should be changed?

How many records should be changed?

How those records should be changed during data transformation?

Metrics for measuring data utility and discrimination removal
Which records should be change and how?

We need to enforce the following inequality for each $\alpha$-discriminatory rule $r$

$$f(r: A, B \rightarrow C) < \alpha, \text{ where } f \in \{\text{elift, slift, ...}\}$$

Data transformation method to enforce the above inequality where $f=\text{elift}$

**DTM1:** Changes the discriminatory itemset e.g., gender changed from male to female in the records with granted credits

**DTM 2:** Changes the class item e.g., from grant credit to deny credit in the records with male gender
Which records should be change and how?

A suitable data transformation with minimum information loss to make each $\alpha$-discriminatory rule $\alpha$-protective.

we should enforce the following inequality for each $\alpha$-discriminatory rule $r$

$$f(r: A, B \rightarrow C) < \alpha, \text{ Where } f \in \{\text{elift, slift, ...}\}$$

**Theorem**: DTM1 and DTM2 methods for making each $\alpha$-discriminatory rule $r$ $\alpha$-protective w.r.t. $f$ do not generate new $\alpha$-discriminatory rules as a result of their transformations.
How many records should be changed?

A suitable data transformation with minimum information loss to make each $\alpha$-discriminatory rule $\alpha$-protective.

we should enforce the following inequality for each $\alpha$-discriminatory rule $r$

$$f(r: A, B \rightarrow C) < \alpha, \text{ where } f = \text{elift}$$

DTM1: Taking $\Delta_{\text{elift}}$ equal to the ceiling of the right-hand side of Equation (below) suffices to

$$\Delta_{\text{elift}} > \frac{\alpha \times \text{supp}(A, B) \times \text{supp}(B, C) - \text{supp}(A, B, C) \times \text{supp}(B)}{\text{supp}(A, B, C) - \alpha \times \text{supp}(A, B)}$$
In which order records should be changed?

DTM1: perturb the discriminatory itemset from \( \sim A \) (male) to \( A \) (female) in the subset \( D_c \) of all records of the original data set which completely support the rule \( \sim A, B \rightarrow \sim C \) and have minimum impact on other rules.

\[
D_c \leftarrow \text{All records completely supporting } \sim A, B \rightarrow \sim C
\]

for each \( db_c \in D_c \) do

Compute \( \text{impact}(db_c) = |\{ r_a \in FR | db_c \text{ supports the premise of } r_a \}| \)

end for

Sort \( D_c \) by ascending impact
Fairness-aware data mining

Pre-processing approaches:


Handling conditional discrimination
(Zliobaite et al., 2011)

Previous pre-processing techniques aimed at removing all discrimination

However:

- Some parts may be explainable;
- Leads to reverse discrimination
Example of fully explainable discrimination

<table>
<thead>
<tr>
<th></th>
<th>medicine</th>
<th></th>
<th>computer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>female</td>
<td>male</td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>number of applicants</td>
<td>800</td>
<td>200</td>
<td>200</td>
<td>800</td>
</tr>
<tr>
<td>acceptance rate</td>
<td>20%</td>
<td>20%</td>
<td>40%</td>
<td>40%</td>
</tr>
<tr>
<td>accepted (+)</td>
<td>160</td>
<td>40</td>
<td>80</td>
<td>320</td>
</tr>
</tbody>
</table>

• 36% of males accepted, 24% of females accepted
• However, the difference is fully explainable by the fact that females applied to the more competitive program (medicine).
• Similar to the famous University of California, Berkeley 1973 case.
Some explainable + some bad discrimination

<table>
<thead>
<tr>
<th></th>
<th>medicine</th>
<th></th>
<th>computer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>female</td>
<td>male</td>
<td>female</td>
</tr>
<tr>
<td>number of applicants</td>
<td>800</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>acceptance rate</td>
<td>15%</td>
<td>25%</td>
<td>35%</td>
</tr>
<tr>
<td>accepted (+)</td>
<td>120</td>
<td>50</td>
<td>70</td>
</tr>
</tbody>
</table>

\[
= (20\% \times 25\% + 80\% \times 45\%) \\
- (80\% \times 15\% + 20\% \times 35\%) \\
= 41\% - 19\% = 22\%
\]

Part of this discrimination can be explained, although not all of it.
Analysis of explainable discrimination

How much discrimination can be explained?

What should have been the acceptance rate $P^*(+|\text{Fac})$ for faculty Fac?

(1) $P^*(+|\text{Fac}) = P_{\text{obs}}(+ | \text{Fac}) \rightarrow$ leads to redlining

(2) $P^*(+ | \text{Fac}) = \frac{[P_{\text{obs}} (+ | \text{Fac}, \text{m}) + P_{\text{obs}} (+ | \text{Fac}, \text{f})]}{2}$

$D_{\text{expl}} =$ discrimination when it would be true that:

$P(+ | \text{m,Fac}) = P(+ | \text{f,Fac}) = P^*(+ | \text{Fac})$

$D_{\text{bad}} = D_{\text{all}} - D_{\text{expl}}$
Analysis of explainable discrimination

\begin{align*}
D_{\text{expl}} &= (20\% \times 20\% + 80\% \times 40\%) - (80\% \times 20\% + 20\% \times 40\%) = 12\% \\
D_{\text{bad}} &= D_{\text{all}} - 12\% = 22\% - 12\% = 10\%
\end{align*}
Simulation Experiments

- $t =$ test score (integer in [1,100] uniform at random);
- $a =$ effect on acceptance decision due to program;
- $b =$ effect on acceptance decision due to gender bias

| Case                | $P(t)$ | $a$ | $b$ | $P(\text{med}|f)$ |
|---------------------|--------|-----|-----|--------------------|
| Case I, only explainable | 0.01   | 10  | 0   | $\alpha$          |
| Case II, only bad   | 0.01   | 0   | 5   | $\alpha$          |
| Case III, explainable and bad | 0.01   | 10  | 5   | $\alpha$          |
Solution: Locally change input data

1. Divide the dataset according to the explanatory attribute(s)

2. Estimate $P^*(+|e_i)$ for all partitions $e_i$

3. Apply local techniques on partition $e_i$ so that $P(+|e_i,f) = P(+|e_i,m) = P^*(+|e_i)$ becomes true

- Local massaging
- Local preferential sampling
Experiments: Discrimination after Massaging

- Global techniques tend to overshoot when large part of the discrimination can be explained.
Experiments with multiple explanatory attributes

If there are multiple explanatory attributes: create groups of individuals by clustering based upon explanatory attributes (e.g., working hours and experience when determining salary).
Fairness-aware data mining

Pre-processing approaches: (not covered here)


These two papers (together with others) deal with simultaneously with privacy and anti-discrimination. This new promising family of approaches will be discussed in Part 4 of the tutorial.
Non-discriminatory data-driven decision-making
Fairness-aware data mining

In-processing approaches:


Fairness-aware data mining

In-processing approaches: (not covered here)


many more… (probably)
Fairness-aware data mining

In-processing approaches:


Problem definition

• Given: dataset D, an attribute B, a value \( b \in \text{dom}(B) \)
• Find a classifier M that:
  • Minimizes discrimination w.r.t. \( B=b \)
  • Maximizes predictive accuracy

First attempt: decision tree
  - Change split criterion
  - Leaf Relabeling
Change split criterion

Purity with respect to *Class* attribute

Impurity with respect to sensitive attribute *B*

Guarantee over resultant discrimination level on training data; e.g., not more than 3%

E.g.: Information gain maximal w.r.t. class and minimal w.r.t. B

Objective: $\frac{\text{GINI}_{\text{split}}(\text{Class})}{\text{GINI}_{\text{split}}(B)}$

Objective: $\text{GINI}_{\text{split}}(\text{Class}) - \text{GINI}_{\text{split}}(B)$
Input: Dataset D
Output: Decision tree t

Induce(D):
    If all tuples t in D have label + then return +
    If all tuples t in D have label – then return –
    For all split criteria C:
        $D_{1,c} = \{ t \in D \mid t \text{ satisfies } C \}$
        $D_{2,c} = D - D_{1}$
        Measure Quality($D_{1,c}, D_{2,c}$)
    Let C be the best split
    Return C

    yes

    Induce ($D_{1,c}$)

    no

    Induce ($D_{2,c}$)
Leaf relabeling

Decision trees divide up the decision space

![Decision Tree Diagram]

Labels are assigned according to the majority class

\[ \text{Disc}_T = p( M = + | B \neq b) - p(M = + | B = b) = \frac{6}{10} - \frac{4}{10} = 0.2 \text{ or } 20\% \]

Relabel some leaves to reduce the discrimination
Leaf relabeling

E.g.: Relabel node $l_1$ from $-$ to $+$

Influence on accuracy: -15%

Influence on discrimination: 20% - 30% = -10%

Change in accuracy and discrimination independent of changes in other leaves

Task: find the optimal relabeling of the nodes
Leaf relabeling

Optimal Leaf Relabeling is equivalent to the Knapsack problem

Given:
- A knapsack of size $K$
- A set of objects $O$
- A weight and a size for every object

Find:
- A subset of objects that fits in the knapsack and maximizes the weight

This problem is known to be NP-complete
Yet it has good approximations; e.g., the greedy algorithm
Leaf Relabeling = Knapsack

Do not consider relabelings:
that reduce accuracy
without lowering discrimination

Current discrimination = 20%
Relabeling all: -50%
Hence, 30% can stay

Knapsack problem:
Select nodes \textbf{NOT} relabeled

\begin{align*}
\text{Dacc: weight} \\
\text{Ddisc: size} \\
\text{K = 30\% (that can stay)}
\end{align*}

Outcome: relabel $l_4$
Fairness-aware data mining

In-processing approaches:


Three Naive Bayes Approaches for Discrimination-Free Classification

Approach 1: Modified Naive Bayes
Approach 2: Two Naive Bayes models
Approach 3: Latent variable model
Approach 1: Modified Naive Bayes

- Use $P(C,S,A_1,\ldots,A_n) = P(S)P(C|S)P(A_1|C)\ldots P(A_n|C)$ instead of $P(C,S,A_1,\ldots,A_n) = P(C)P(S|C)P(A_1|C)\ldots P(A_n|C)$

- Alterate distribution $P(C|S)$ until there is no more discrimination.

- It creates a discrimination free Naive Bayes classifier but does not avoid red-lining effect due to attributes $A_s$ correlated with $S$. 

$S \rightarrow C \rightarrow A_1 \ldots A_n$
Approach 2: Two Naive Bayes models

- How to remove correlation between attributes $A_s$ and $S$?
- Simply remove attributes $A_s$ → big loss in accuracy!

- Remove the fact that attributes $A_s$ can be used to decide $S$, by splitting the learning in two, w.r.t. the value of $S$. For instance if $S$ is gender, build one model for male and one model for female.
Approach 3: Latent variable model

- Try to discover the actual class labels that the dataset should have had if it was discrimination-free.
- This is modeled by a latent variable $L$.

- Assumptions:
  1. $L$ is independent from $S \rightarrow L$ is discrimination-free;
  2. $C$ is determined by discriminating $L$ using $S$ uniformly at random.

- Fit $L$ by means of Expectation-Maximization (EM)
Fairness-aware data mining

In-processing approaches:


Defining fairness

Applying doctrine of disparate impact: 80% rule

If 50% of male applicants get selected for the job, at least 40% of females should also get selected.

A fair system might not always be 80:100

In certain scenarios, the prescribed proportion could be 50:10

The goal is to enable a range of "fair" proportions.
A classifier whose output achieves a given proportion of items (in positive class) with different values of sensitive feature
Fairness constraint

**Key Idea:** Limit the cross-covariance between sensitive feature value and distance from decision boundary
Fairness constraint
Fairness constraint

\[ \left| \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) b^T [-1 \ x_i] \right| \leq c \]
Modifying the logistic regression classifier

\[ p(y_i = 1 | x_i) = \frac{1}{1 + e^{-b_0 + \sum_j b_j x_{ij}}} \]

maximize \[ \sum_{i=1}^{N} \log p(y_i | x_i) \]
Modifying the logistic regression classifier

\[
p(y_i = 1 | x_i) = \frac{1}{1 + e^{-b_0 + \sum_j b_j x_{i,j}}}
\]

**Key point:** possible to solve this problem efficiently
Modifying the Hinge loss classifier

minimize \[ \sum_{i=1}^{N} \max(0, y_i (b^T[-1 \ x_i])) \]
subject to \[ \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) b^T[-1 \ x_i] \leq c, \]
\[ \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) b^T[-1 \ x_i] \geq -c, \]
Modifying the SVM classifier

\[
\begin{align*}
\text{minimize} & \quad \|b\|^2 + C \sum_{i=1}^n \xi_i \\
\text{subject to} & \quad y_i (b^T [-1 \ x_i]) \geq 1 - \xi_i, \ \forall i \in \{1, \ldots, n\} \\
& \quad \xi_i \geq 0, \ \forall i \in \{1, \ldots, n\}, \\
& \quad \frac{1}{N} \sum_{i=1}^N (z_i - \bar{z}) b^T [-1 \ x_i] \leq c, \\
& \quad \frac{1}{N} \sum_{i=1}^N (z_i - \bar{z}) b^T [-1 \ x_i] \geq -c.
\end{align*}
\]
Tightening the constraints increases fairness
Fairness vs accuracy trade-off

**Random:** takes the output of the unconstrained classifier and shuffles labels randomly until satisfying the given $c$.

**Ad-hoc:** takes the output of the unconstrained classifier and change females to +ve until satisfying the given $c$. 
Fairness vs accuracy trade-off
Fairness for Multiple Features
Fairness-aware data mining

In-processing approaches:


User visits capitalone.com

Capital One uses tracking information provided by the tracking network \([x+1]\) to personalize offers

**Concern:** *Steering* minorities into higher rates (illegal)
\[ M : V \rightarrow O \]

Ad network 
\( (x+1) \)

Vendor 
(capital one)

\[ f : O \rightarrow A \]

\( V \): Individuals

\( O \): outcomes

\( A \): actions

\( x \)

\( M(x) \)
The goal

Achieve fairness in the classification step
Fairness through Blindness

• Ignore all irrelevant/protected attributes

• Point of **failure**: Redundant encodings
  – Machine learning: You don’t need to see the label to be able to predict it
Group Fairness

- Equalize two groups $S$, $T$ at the level of outcomes
  - E.g. $S =$ minority, $T = S^c$
  - $\Pr[\text{outcome } o \mid S] = \Pr[\text{outcome } o \mid T]$

- **Insufficient** as a notion of fairness
  - Has some good properties, but can be abused
  - **Example:** Advertise burger joint to carnivores in $T$ and vegans in $S$. 
Lesson: Fairness is task-specific

- Fairness requires understanding of classification task
  - Cultural understanding of protected groups
  - Awareness
Individual fairness

Treat *similar* individuals *similarly*

- Similar for the purpose of the classification task
- Similar distribution over outcomes
• Assume *task-specific similarity metric*
  – Extent to which two individuals are similar w.r.t. the classification task at hand
• Ideally captures *ground truth*
  – Or, society’s best approximation
• Open to public discussion, refinement
  – In the spirit of Rawls
• Typically, does not suggest classification!
Examples

• Financial/insurance risk metrics
  – Already widely used (though secret)

• **AALIM health care metric**
  – health metric for treating similar patients similarly

• Roemer’s relative effort metric
  – Well-known approach in Economics/Political theory

Maybe not so much science fiction after all...
Formal setup

Classification

$M : V \rightarrow \Delta(O)$

$M(x)$

$V$: Individuals  $O$: outcomes
Metric $d : V \times V \rightarrow \mathbb{R}$

Lipschitz condition $\|M(x) - M(y)\| \leq d(x, y)$
Utility Maximization

Vendor can specify arbitrary utility function

\[ U : V \times O \rightarrow \mathbb{R} \]

Can efficiently maximize vendor’s expected utility subject to Lipschitz condition

\[
\max_{x \in V} \mathbb{E}_{o \sim M(x)} U(x, o)
\]

s.t. \( M \) is \( d \)-Lipschitz
More contributions

• Several examples showing the inadequacy of group fairness (or statistical parity)

• Connection between individual and group fairness: the Lipschitz condition implies statistical parity between two groups if and only if the Earthmover distance between two groups is small.

• Fair affirmative action. Provide techniques for forcing statistical aprity when it is not implied by the Lipschtiz condition, while preserving as much fariness for the individuals as possible.

• Relationship with privacy: the proposed definition of fairness is a generalization of the notion of differential privacy.
Fairness-aware data mining

In-processing approaches:


Main limitations of “Fairness through awareness”

1. The problem of fairness in classification is reduced to the problem of establishing a fair distance metric. The *distance metric* that defines the similarity between the individuals is *assumed to be given*. This might be unrealistic in certain settings.

2. Their framework is *not formulated as a learning framework*: it gives a mapping for a given set of individuals, but it doesn't provide any mean to generalize to novel unseen data (new individuals).
“Learning fair representations” (Zemel et al. ICML 2013)

... extends “Fairness through awareness” in several important ways.

1. It develops a learning framework: learn a general mapping, applies to any individual.

2. Learns a restricted form of a distance function as well as the intermediate representation. No longer needed a distance function given a-priori.

3. Achieves both group fairness and individual fairness.

4. The intermediate representation can be used for other classification tasks (i.e., transfer learning is possible).

5. Experimental assessment.
Main idea [sketch]

• Map each individual (a data point in the input space) to a probability distribution in a new representation space.

• The aim of the new representation is to lose any information that can reconstruct whether the individual belongs to the protected subgroups, while maintaining as much other information as possible.

• Fairness becomes an optimization problem of finding the intermediate representation that best encodes the data while obfuscating membership to the protected subgroups.

• Tool: probabilistic mapping to a set of prototypes (it can be seen as a form of discriminative clustering model). [Details omitted]
Non-discriminatory data-driven decision-making
Fairness-aware data mining

Post-processing approaches:


Fairness-aware data mining

Post-processing approaches:


Privacy and anti-discrimination should be addressed together

Suppose to publish frequent pattern (support > k, for k-anonymity) extracted from personal data for credit approval decision making.

Privacy protection only

sex=female → credit-approved=no (support 126)

Discrimination protection only

job =veterinarian, salary =low → credit-approved=no (support 40)
job = veterinarian → credit-approved=no (support 41)
Support > k doesn’t imply k-anonymity

job = veterinarian, salary = low → credit-approved = no (support 40)
job = veterinarian → credit-approved = no (support 41)

Supp(job = veterinarian, salary = high, credit-approved = no) = 1

In the dataset there is only one veterinarian with high salary.
If somebody knows a veterinarian with high salary, can imply that he/she got credit denied.
Overall post-processing approach
Formal results

**Theorem:** Anti-discrimination pattern sanitization for making $F(D,s)$ a-protective does not generate new discrimination as a result of its transformation.

**Theorem:** Using anti-discrimination pattern sanitization for making $F(D,s)$ a-protective cannot make $F(D,s)$ non-$k$-anonymous.

**Theorem:** Using privacy pattern sanitization for making $F(D,s)$ $k$-anonymous can make $F(D,s)$ more or less a-protective.
Evaluation

Pattern distortion scores to make the Adult dataset $\alpha$-protective $k$-anonymous
Discrimination-and privacy-aware patterns

Shows that privacy pattern sanitization methods based on either $k$-anonymity or differential privacy can work against fairness.

Proposes new anti-discrimination pattern sanitization methods that do not interfere with a privacy-preserving sanitization based on either $k$-anonymity or differential privacy.

Shows that the utility loss caused by simultaneous anti-discrimination and privacy protection is only marginally higher than the loss caused by each of those protections separately.
Part I: Introduction and context

Part II: Discrimination discovery

Part III: Fairness-aware data mining

Part IV: Challenges and directions for future research

Discussion and further questions
Challenges: the ground-truth problem

- The trade-off between data utility and discrimination avoidance

- Utility based on potentially biased training data!
- Hard to assess the quality of the results
- Lack of datasets and benchmarks
Challenges: definitions of discrimination

- Unlike for privacy, anti-discrimination legal concepts are diverse and vague
  - Direct vs indirect discrimination
  - Individual vs group fairness
  - Affirmative actions
  - Explainable vs unexplainable discrimination
  - ...

- Current methods in fairness-aware data mining used different definitions of discrimination/fairness
  - No single agreed-upon measure for discrimination/fairness

- How different definitions of fairness affect algorithm design?
Challenges: interaction with law and policies

• As for research in privacy preservation, there is an interaction between the research on algorithmic fairness and the anti-discrimination regulations:
  • Laws give us the rules of the game: definitions, objective functions, constraints
  • New technical developments need to be taken in consideration by legislators

• However, the communication channel is not clear:
  • Is my data transformation algorithm legal?
  • Can my discrimination-detection algorithm be useful in a real-world case law?

• Wide variety of cases and different interpretations: difficult for a CS to navigate
  • Importance of multidisciplinarity

• As usual, many differences between USA and EU regulation
General Data Protection Regulation (GDPR)  
(Regulation (EU) 2016/679)

• Aims to strengthen and unify data protection for individuals within the EU, as well as setting rules about the export of personal data outside the EU.

• Primary objectives of the GDPR are to give citizens back the control of their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EU.

• It deals with concept such as consent, responsibility, accountability, right to be forgotten, etc.

• The regulation was adopted on 27 April 2016. It enters into application 25 May 2018 after a two-year transition period and, unlike a Directive it does not require any enabling legislation to be passed by governments.

• When the GDPR takes effect it will replace the data protection directive (officially Directive 95/46/EC) from 1995.
Right to explanation (GDPR 2018)

• It will restrict automated decision-making which “significantly affect” individuals.

• An individual can ask for an explanation of an algorithmic decision.

• This law will pose **large challenges for industry**
  • There is a gap between the legislators’ aspirations and technical realities
  • Intentional concealment on the part of corporations or other institutions, where decision making procedures are kept from public scrutiny
  • A “mismatch between the mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of interpretation”

• It highlights **opportunities for machine learning researchers** to take the lead in designing algorithms and evaluation frameworks which avoid discrimination.

Fairness and privacy

• Privacy-preservation
  - How do we prevent sensitive information from being leaked?

• Discrimination-prevention
  - How do we prevent sensitive information from being abused?

• Sensitive features in these two contexts might overlap or not
  - One may not mind other people knowing about their ethnicity, but would strenuously object to be denied a credit or a grant if their ethnicity was part of that decision

• Hiding sensitive information from data due to privacy, might also hide the presence of discriminatory patterns
A promising direction…

Dealing with privacy-preserving data mining (PPDM) and fairness-aware data mining (FADM) *jointly*…

- Share common challenges
- Share common techniques
- Sometimes one can help the other

<table>
<thead>
<tr>
<th>PPDM</th>
<th>FADM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measuring disclosure risk</td>
<td>Measuring potential discrimination</td>
</tr>
<tr>
<td>Data, algorithm or model transformation to protect privacy</td>
<td>Data, algorithm or model transformation to prevent discrimination</td>
</tr>
<tr>
<td>Measuring data/model utility</td>
<td>Measuring data/model utility</td>
</tr>
<tr>
<td>Trade-off between privacy and utility</td>
<td>Trade-off between fairness and utility</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
A promising direction

Dealing with privacy-preserving data mining (PPDM) and fairness-aware data mining (FADM) *jointly*…


Future work: beyond binary classification

So far …

mostly binary classification problems such as "HIRE" vs "DON'T HIRE"

Future …

Multi-class and multi-label classification settings
Regression settings
Noisy input data
Multiple protected characteristics
Potentially missing protected characteristics

Future work: beyond classification

General theory of algorithmic fairness

Fairness in recommender systems and personalization

Fairness in network data (e.g., hire/don’t hire based on social network)

Fairness in text data (e.g., automatically detect sexist or racist text)

Tools for discovering discrimination practices in different online settings

E.g., google image search, Airbnb hosts with racist behavior, price discrimination (see e.g., $heriff tool), ads targeting discrimination (see e.g., Adfisher tool)
Acknowledgments

• Ricardo Baeza-Yates, Toon Calders, Ciro Cattuto, Faisal Kamran, Salvatore Ruggieri, Suresh Venkatasubramanian and many many other researchers in the area which (more or less consciously) donated their slides to made this tutorial possible.

• This tutorial was partially supported by
  ▪  **TYPES**: European Union’s Horizon 2020 innovation action programme under grant agreement No 653449
  ▪  **FARADAY**: Partially supported by the Catalonia Trade and Investment Agency (Agència per la competitivitat de l’empresa, ACCIÓ).
Thank you!

Sara Hajian
Francesco Bonchi  @francescobonchi
Carlos Castillo   @chatox