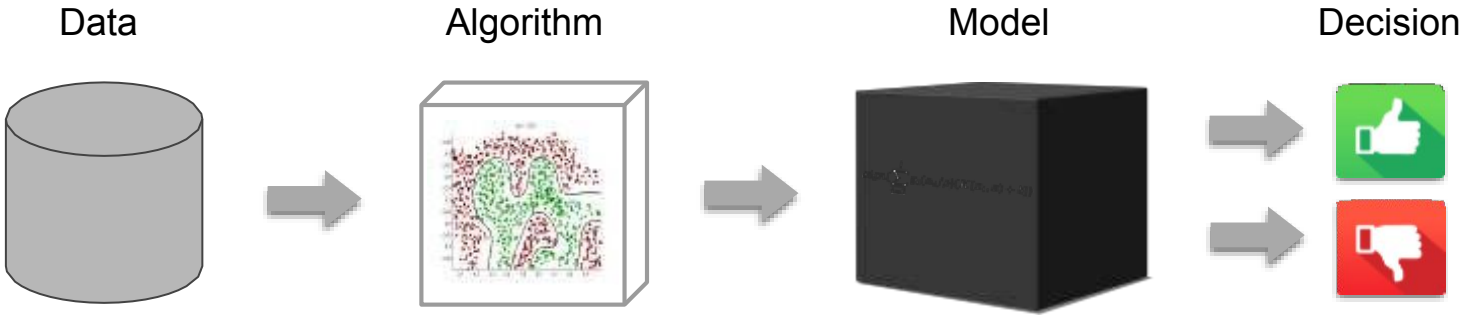
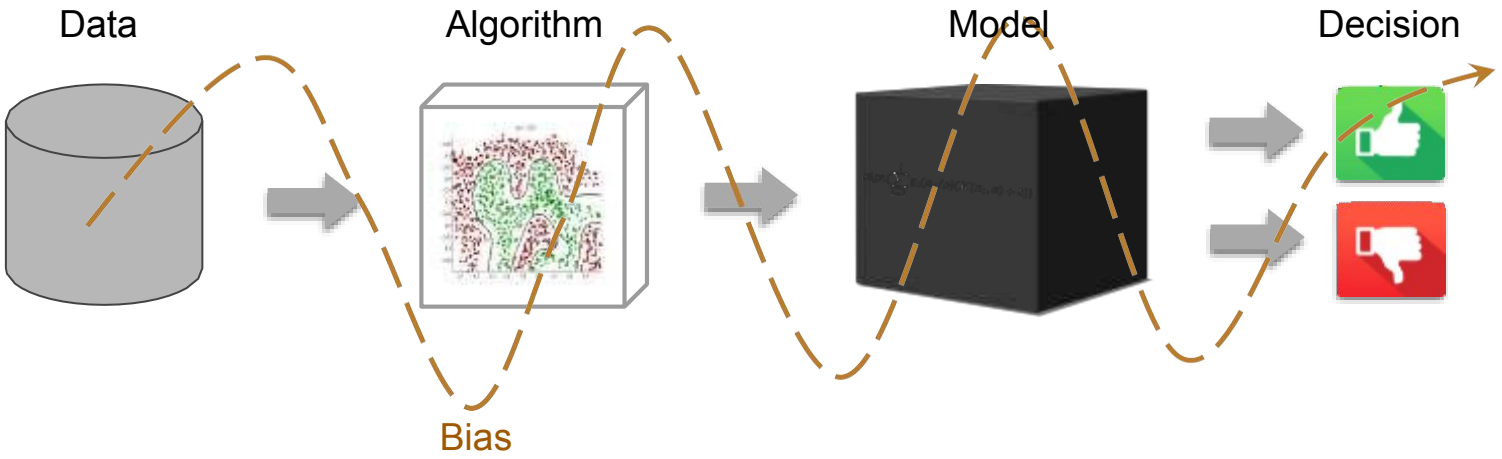
Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining

Fairness-aware data mining

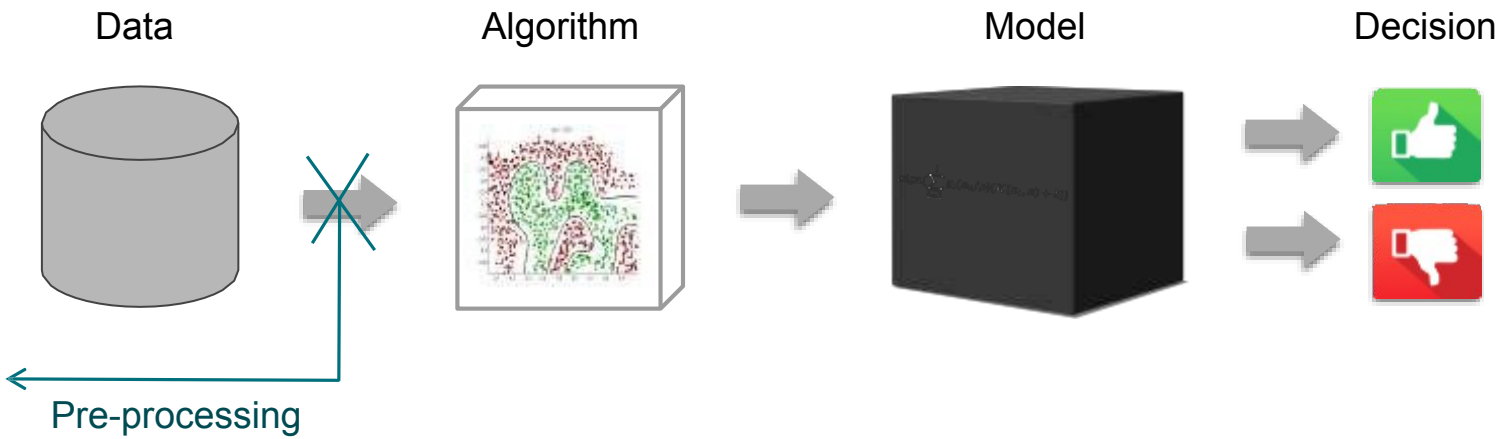
Non-discriminatory data-driven decision-making



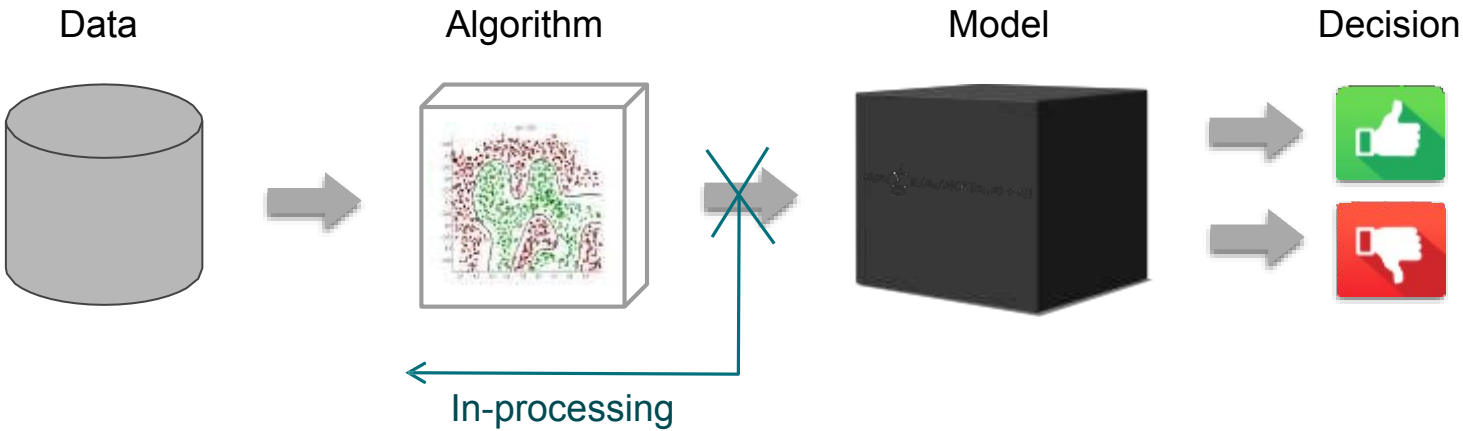
Non-discriminatory data-driven decision-making



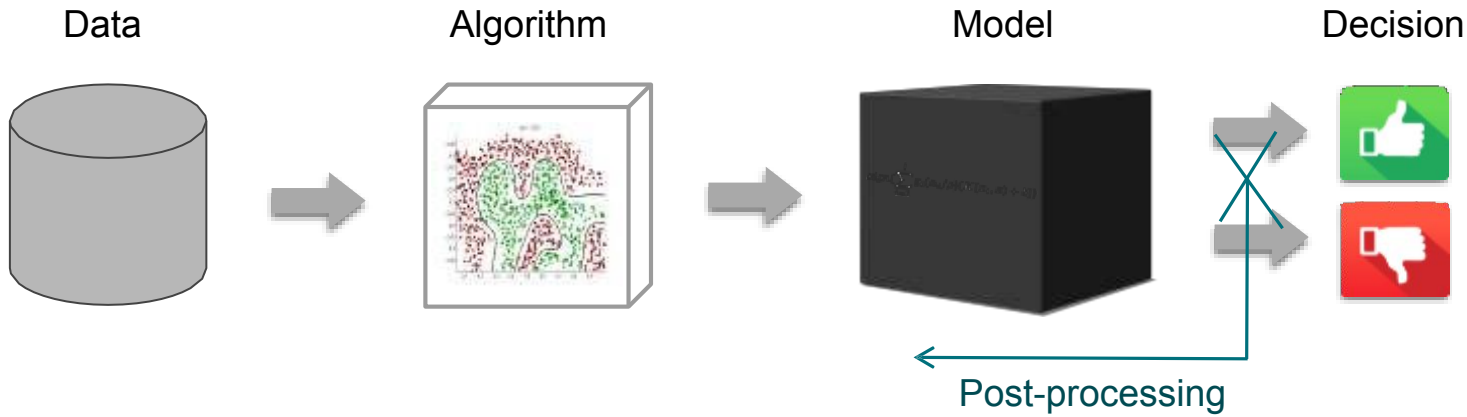
Non-discriminatory data-driven decision-making



Non-discriminatory data-driven decision-making



Non-discriminatory data-driven decision-making



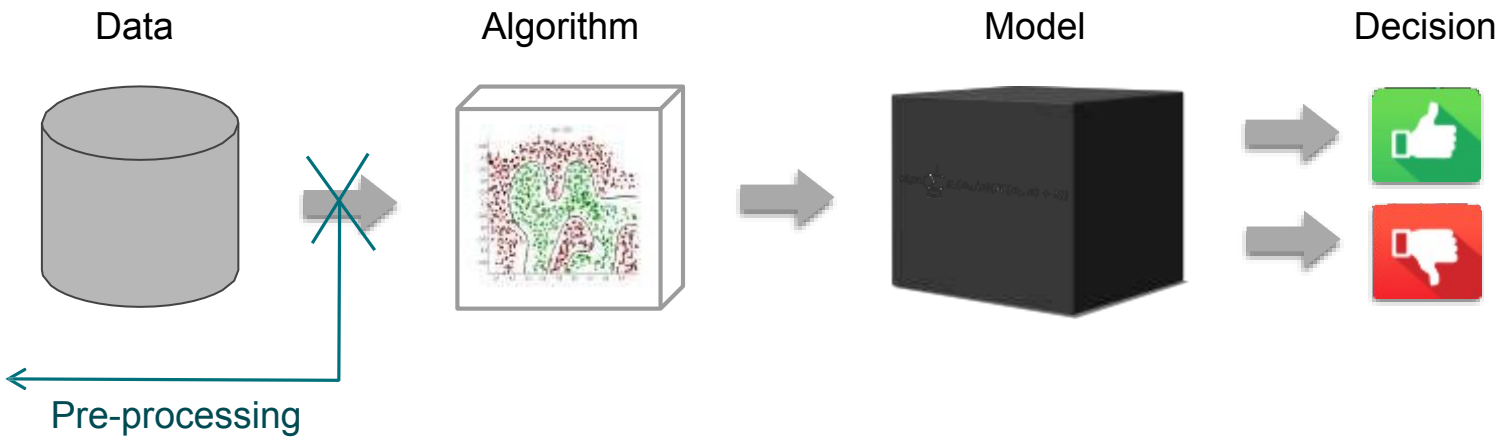
Fairness-aware data mining: common aspects

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:

[Pre_1] M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. *"Certifying and removing disparate impact"*. In KDD, pp. 259-268, 2015.
[Pre_2] F. Kamiran and T. Calders. *"Data preprocessing techniques for classification without discrimination"*. In Knowledge and Information Systems (KAIS), 33(1), 2012.
[Pre_3] S. Hajian and J. Domingo-Ferrer. *"A methodology for direct and indirect discrimination prevention in data mining"*. In IEEE Transactions on Knowledge and Data Engineering (TKDE), 25(7), 2013.
[Pre_4] I. Zliobaite, F. Kamiran and T. Calders. *"Handling conditional discrimination"*. In ICDM, pp. 992-1001, 2011.

Fairness-aware data mining

Pre-processing approaches:

[Pre_1] M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. *"Certifying and removing disparate impact"*. In KDD, pp. 259-268, 2015.

[Pre_2] F. Kamiran and T. Calders. *"Data preprocessing techniques for classification without discrimination"*. In Knowledge and Information Systems (KAIS), 33(1), 2012.

[Pre_3] S. Hajian and J. Domingo-Ferrer. *"A methodology for direct and indirect discrimination prevention in data mining"*. In IEEE Transactions on Knowledge and Data Engineering (TKDE), 25(7), 2013.

[Pre_4] I. Zliobaite, F. Kamiran and T. Calders. *"Handling conditional discrimination"*. In ICDM, pp. 992-1001, 2011.

Disparate Impact

Disparate impact occurs when a selection process has widely different outcomes for different groups, even as it appears to be neutral.

Eg: refusing to hire people because of a poor credit rating, when minorities are disproportionately affected.

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.

D has disparate impact if

$$\frac{\Pr(C=YES \mid X=0)}{\Pr(C=YES \mid X=1)} \leq \tau = 0.8$$

Disparate impact (“80 % rule”)

Confusion Matrix

Outcome	$X = 0$	$X = 1$
$C = \text{NO}$	a	b
$C = \text{YES}$	c	d

80% rule

$$\frac{c/(a+c)}{d/(b+d)} \geq 0.8$$

$$specificity = \frac{a}{a+c}$$

$$sensitivity = \frac{d}{b+d}$$

The likelihood ratio positive

$$LR_+(C,X) = \frac{d/(b+d)}{c/(a+c)} \qquad \frac{sensitivity}{1-specificity}$$

Direct Impact

$$\text{DI} = \frac{1}{\text{LR}_+(C,X)}$$

Disparate impact certification problem

To take some data set D and return a data set D^{-} = (X,Y^{-}, C) that can be certified as not having disparate impact.

Disparate impact removal problem

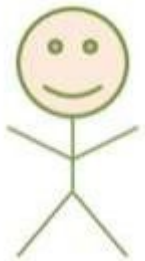
Guarantee that given D, any classification algorithm aiming to predict some C’ from Y would not have disparate impact.

Certifying DI

If Bob cannot predict X given the other attributes of D, then A is fair with respect to Bob on D.

Alice

A: ~~X~~, Y -> C



Bob ~~A~~

Certifying DI

$$\mathrm{BER}(f(Y), X) = \frac{\Pr[f(Y) = 0 | X = 1] + \Pr[f(Y) = 1 | X = 0]}{2}$$

(BER). Let f : Y → X be a predictor of X from Y. The balanced error rate BER of f on distribution D over the pair (X,Y) is defined as the (unweighted) average class conditioned error of f .

Predictability

$$\mathrm{BER}(f(Y), X) \leq \epsilon.$$

Theorem

A data set is (1/2 − β/8)-predictable if and only if it admits disparate impact, where β is the fraction of elements in the minority class (X = 0) that are selected (C = 1).

Build the infrastructure of DI and BER in terms of β

Step 1. Prove Disparate Impact -> Predictability

Step 2. Prove Predictability -> Disparate Impact

Building the assumptions of proof

Assume $\text{function } g : Y \to C \text{ such that } \mathrm{LR}_+(g(y), c) \geq \frac{1}{\tau}.$

Prediction	$X = 0$	$X = 1$
$g(y) = \text{NO}$	a	b
$g(y) = \text{YES}$	c	d

$$\alpha \triangleq \frac{b}{b+d} \text{ and } \beta \triangleq \frac{c}{a+c}$$

$$\psi : C \to X \text{ such that } \mathrm{BER}(\psi(g(y)), x) < \epsilon \text{ for } (x, y) \in D$$

Thus the combined predictor f = ψ ◦ g satisfies the definition of predictability.

$\phi: Y \to X = \psi \circ g.$

Prediction	$X = 0$	$X = 1$
$\phi(Y) = 0$	a	b
$\phi(Y) = 1$	c	d

$\mathrm{LR}_+(g(y), X) = \frac{1-\alpha}{\beta}$ and $\mathrm{DI}(g) = \frac{\beta}{1-\alpha}$.

$$\mathrm{BER}(\phi) = \frac{\alpha+\beta}{2}.$$

$$\pi_1 = 1 - \alpha \text{ and } \pi_0 = \beta$$

$$\mathrm{DI}(g) = \frac{\pi_0}{\pi_1} \text{ and } \mathrm{BER}(\phi) = \frac{1+\pi_0-\pi_1}{2}$$

Disparate Impact -> Predictability

Fix DI threshold τ, corresponding to the line $\pi_1 = \pi_0/\tau$.

- Notice that the region $\{(\pi_0, \pi_1) \mid \pi_1 \geq \pi_0/\tau\}$ is the region where one would make a finding of disparate impact (for t> = 0.8).
- Now given a classification that admits a finding of disparate impact, we can compute β
- Consider the point $(\beta, \beta/\tau)$ at which the line intersects the DI curve $\pi_1 = \pi_0/\tau$
- This point lies on the BER contour $(1 + \beta - \beta/\tau)/2 = \epsilon$

$$\epsilon = \frac{1}{2} - \frac{\beta}{8}$$

Predictability -> Disparate Impact

Suppose there is a function $f: Y \to X$ such that $\text{BER}(f(y), x) \leq \epsilon$

$\psi^{-1}: X \to C$ be the inverse purely biased mapping

$g: Y \to C = \psi^{-1} \circ f.$ $\pi_1 \geq 1 + \pi_0 - 2\epsilon$

$$\frac{\pi_0}{\pi_1} \leq \frac{\pi_0}{1 + \pi_0 - 2\epsilon} = 1 - \frac{1 - 2\epsilon}{\pi_0 + 1 - 2\epsilon}$$

$\text{DI}(g) \leq 1 - \frac{1-2\epsilon}{\beta+1-2\epsilon} = \tau.$

threshold of $\epsilon = \frac{1}{2} - \frac{\beta}{8}$.

Certifying DI

Algorithm

1. Run a classifier that optimizes BER on the given data set, attempting to predict the protected attributes X from the remaining attributes Y.
2. Suppose the error in this prediction is e.
3. Then using the estimate of β from the data, we can substitute this into the equation above and obtain a threshold e- . If e- < e,
4. Then one can declare the data set free from disparate impact

Removing DI

Once Bob's certification procedure has made a determination of (potential) disparate impact on D, Alice might request a repaired version D' of D, where any attributes in D that could be used to predict X have been changed so that D' would be certified as e-fair.

It is important to change the data in such a way that predicting the class is still possible.

Given protected attribute X and a single numerical attribute Y $Y_x = \Pr(Y|X = x)$ denote the marginal distribution on Y conditioned on X=x

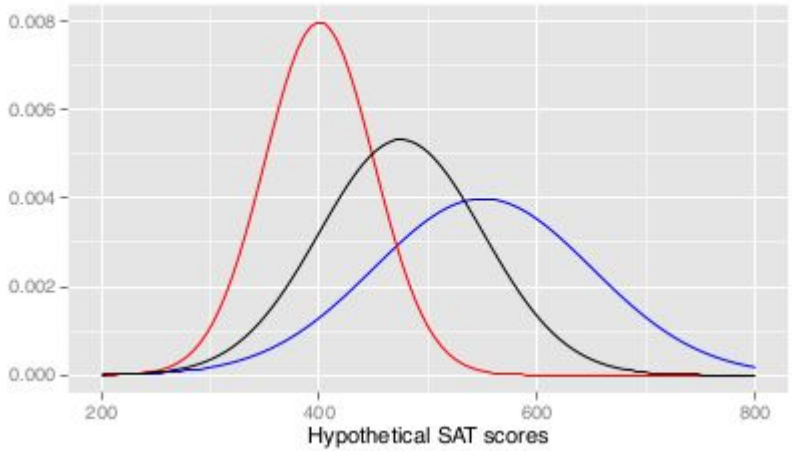
$F_x : Y_x \to [0, 1]$ cumulative distribution function for values y

Let $\bar{Y}$ be the repaired version of Y in $\bar{D}$. We will say that $\bar{D}$ *strongly preserves rank* if

for any $y \in Y_x$ and $x \in X$, its "repaired" counterpart $\bar{y} \in Y_x$ has $F_x(y) = F_x(\bar{y})$

Removing DI

Utility goal: to preserve rank within each marginal distribution P(Y | X = x)



Here the blue curve shows the distribution of SAT scores (Y) for X = female, with μ = 550, σ = 100, while the red curve shows the distribution of SAT scores for X = male, with μ = 400, σ = 50. The resulting fully repaired data is the distribution in black, with μ = 475, σ = 75. Male students who originally had scores in the 95th percentile, i.e., had scores of 500, are given scores of 625 in the 95th percentile of the new distribution in $\bar{Y}$, while women with scores of 625 in $\bar{Y}$ originally had scores of 750.

$$d(P,Q) = \int_0^1 |F_P^{-1}(u) - F_Q^{-1}(u)|du$$

Removing DI

Using the earthmover distance

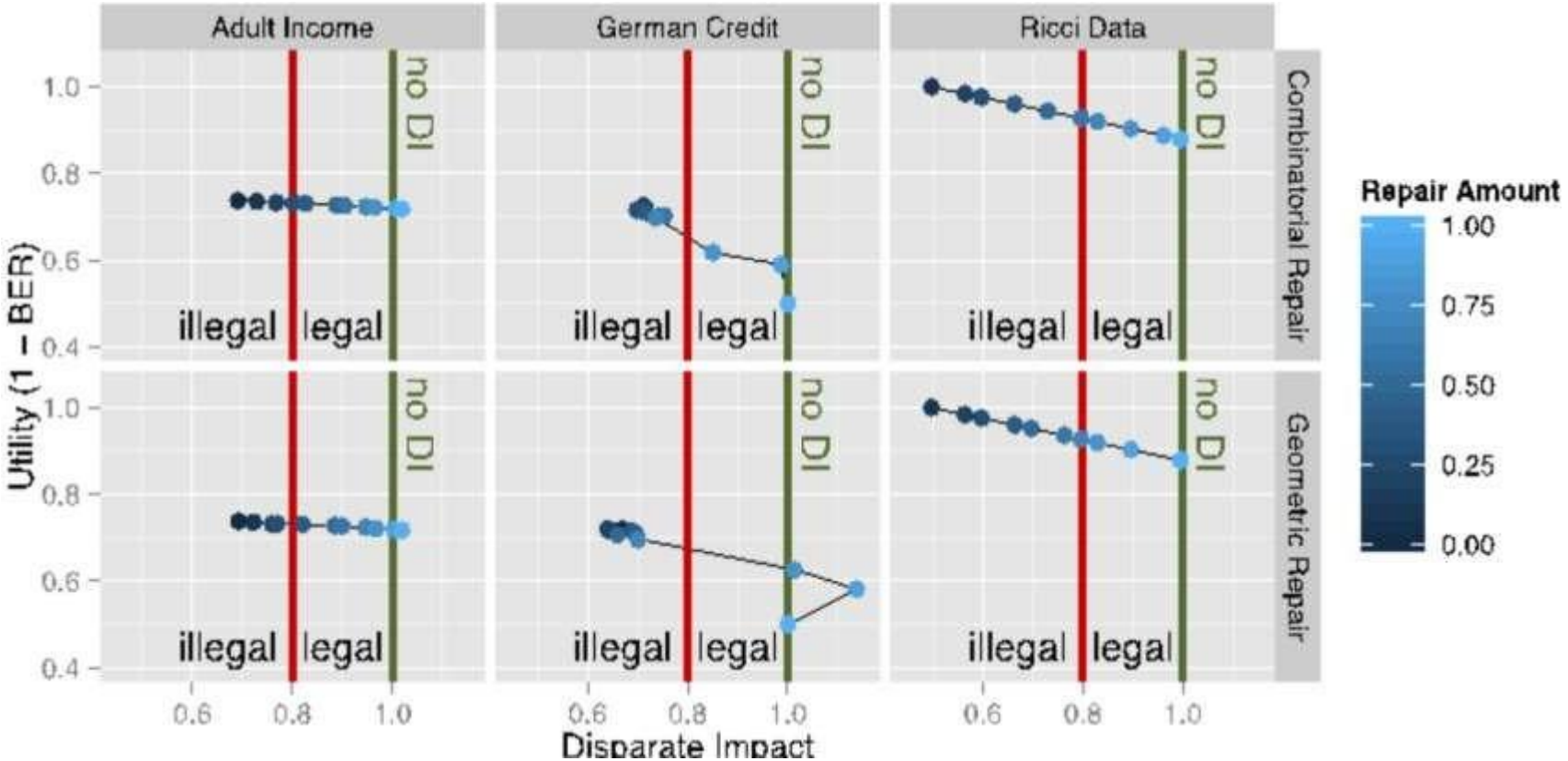
Let $P_i = \Pr(Y = y | X = i)$

$F_i = \text{cdf of } P_i$

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

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

We find a new distribution that is "close" to all conditional distributions.


Adult Income
German Credit
Ricci Data
Combinatorial Repair
Geometric Repair
Utility (1 − BER)
Disparate Impact
illegal
legal
no DI
1.0
0.8
0.6
0.4
0.6
0.8
1.0
Repair Amount
1.00
0.75
0.50
0.25
0.00

Fairness-aware data mining

Pre-processing approaches:

[Pre_1] M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. *"Certifying and removing disparate impact"*. In KDD, pp. 259-268, 2015.

[Pre_2] F. Kamiran and T. Calders. *"Data preprocessing techniques for classification without discrimination"*. In Knowledge and Information Systems (KAIS), 33(1), 2012.

[Pre_3] S. Hajian and J. Domingo-Ferrer. *"A methodology for direct and indirect discrimination prevention in data mining"*. In IEEE Transactions on Knowledge and Data Engineering (TKDE), 25(7), 2013.

[Pre_4] I. Zliobaite, F. Kamiran and T. Calders. *"Handling conditional discrimination"*. In ICDM, pp. 992-1001, 2011.

Data preprocessing techniques for classification without discrimination

1. Training data affects the performance of classifiers
2. Goal: to create discrimination-free classifier for feature classification
3. Input: labelled dataset with one or more sensitive attributes e The quality of the classifier is accuracy and discrimination
4. Restriction: binary sensitive attributes: {b, w } and binary class label: {+, −}

Discrimination measure:

$$disc_{s=b} := P(C(X) = +|X(S) = w) - P(C(X) = +|X(S) = b)$$

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(Change the label of some objects in D to remove discrimination)
3. Reweighting (Instead changing the tuples in the training dataset can be assigned weights)
4. Sampling

Job application example

Sex	Ethnicity	Highest Degree	Job Type	Class
m	native	h. school	board	+
m	native	univ.	board	+
m	native	h. school	board	+
m	non-nat.	h. school	healthcare	+
m	non-nat.	univ.	healthcare	-
f	non-nat.	univ.	education	-
f	native	h. school	education	-
f	native	none	healthcare	+
f	non-nat.	univ.	education	-
f	native	h. school	board	+

Discrimination in labeled dataset

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

The difference of the probability of being in the positive class between the tuples X in D having X(S)=w in D and those having X (S) = b.

Discrimination in classifiers prediction

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

The difference in probability of being assigned the positive class by the classifier between the tuples of D having X(S) = w and those having X (S) = b.

Discrimination aware classification

Given a labeled dataset D, an attribute S ,and a value b $\in$ dom(S) , learn a classifier C such that:
(a) the accuracy of C for future predictions is high; and
(b) the discrimination of new examples classified by C is low.

Accuracy and discrimination trade off

Let C and C' be two classifiers. We say that C dominates C' if the accuracy of C is larger than or equal to the accuracy of C , and the discrimination of C' is at most as high as the discrimination of C.

C strictly dominates C' if at least one of these inequalities is strict.

Given a set of classifiers $\mathcal{C}$, we call a classifier $C \in \mathcal{C}$ *optimal w.r.t. discrimination and accuracy (DA-optimal) in* $\mathcal{C}$ if there is no other classifier in $\mathcal{C}$ that strictly dominates C.

Theorem 1

A classifier C is DA-optimal in $\mathcal{C}_{all}$ iff

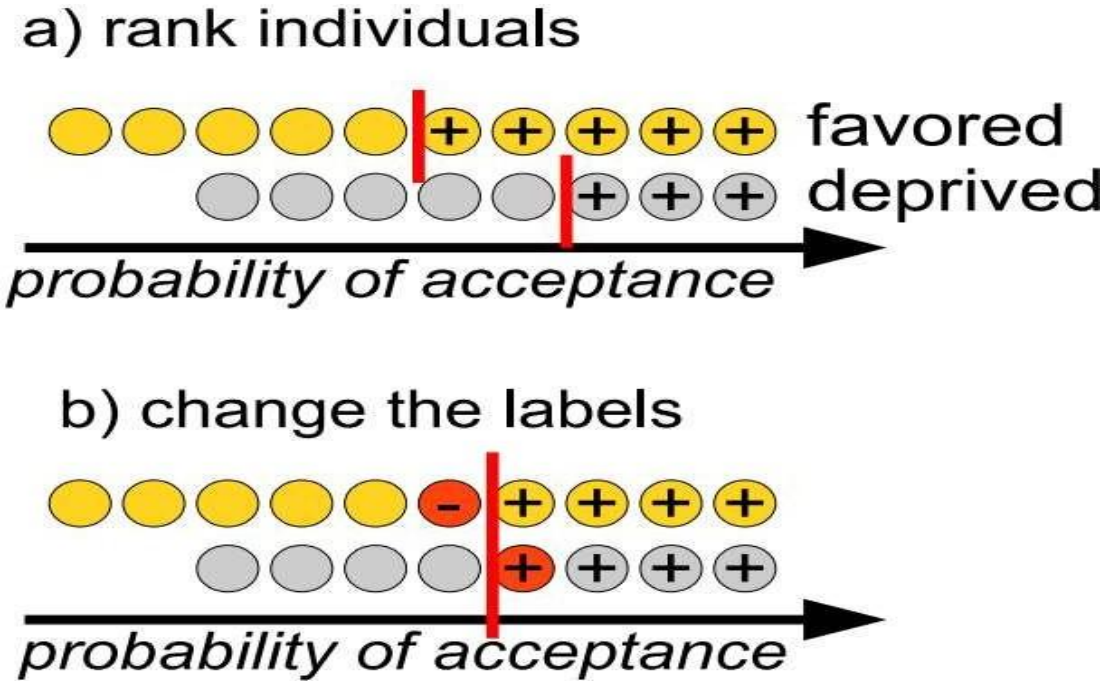
$$acc(C^{Perf}) - acc(C) = \frac{\min(d_b, d_w)}{d}\left(disc(C^{Perf}) - disc(C)\right)$$

Theorem 2

If classifier C' is DA-optimal in $\mathcal{C}_C$, then

$$E[acc(C) - acc(C')] = (2acc(C) - 1)\frac{\min(d_b, d_w)}{d}(disc(C) - disc(C'))$$

Massaging


a) rank individuals
favored
deprived
probability of acceptance
b) change the labels
probability of acceptance

Algorithm for data massaging

Algorithm 1: *Learn Classifier on Massaged Data*

Input: Labeled dataset D, sensitive attribute S and value b, desired class $+$

Output: Classifier C, learned on massaged D

1: $(pr, dem) := Rank(D, S, b, +)$

2: $M := \dfrac{disc_{S=b}(D) \times |\{X \in D \mid X(S) = b\}| \times |\{X \in D \mid X(S) = w\}|}{|D|}$

3: Select the top-M of pr

4: Change the class label of the M selected objects to $+$

5: Select the top-M objects of dem

6: Change the class label of the M selected objects to $-$

7: Train a classifier C on the modified D

8: **return** C

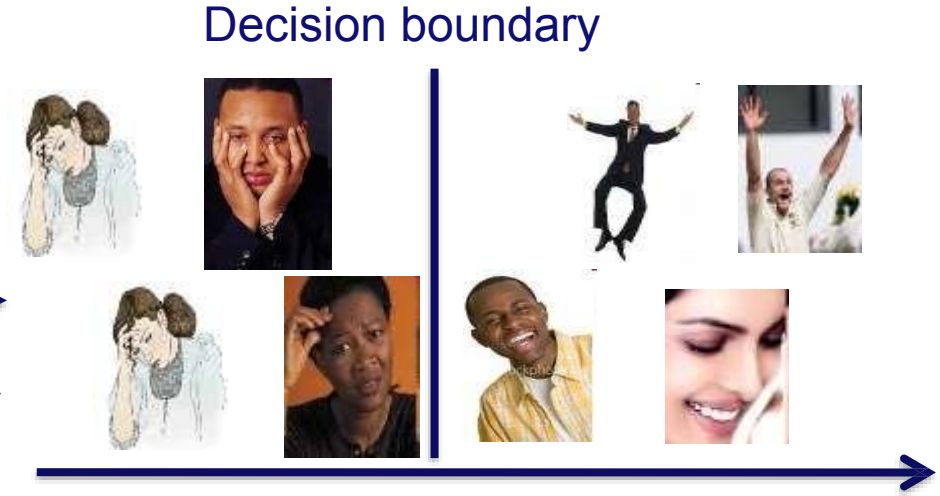
Algorithm for data massaging

Algorithm 2: *Rank*

Input: Labeled dataset D, Sensitive attribute and value S, b, desired class $+$
Output: Ordered promotion list pr and demotion list dem
1: Learn a ranker R for prediction $+$ using D as training data
2: $pr := \{X \in D \mid X(S) = b, X(Class) = -\}$
3: $dem := \{X \in D \mid X(S) = w, X(Class) = +\}$
4: Order pr descending w.r.t. the scores by R
5: Order dem ascending w.r.t. the scores by R
6: **return** (pr, dem)

Learn a

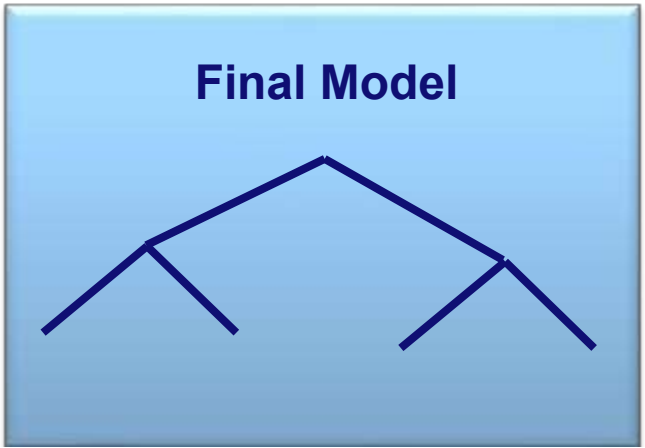
ranker



Relabel

Learn a

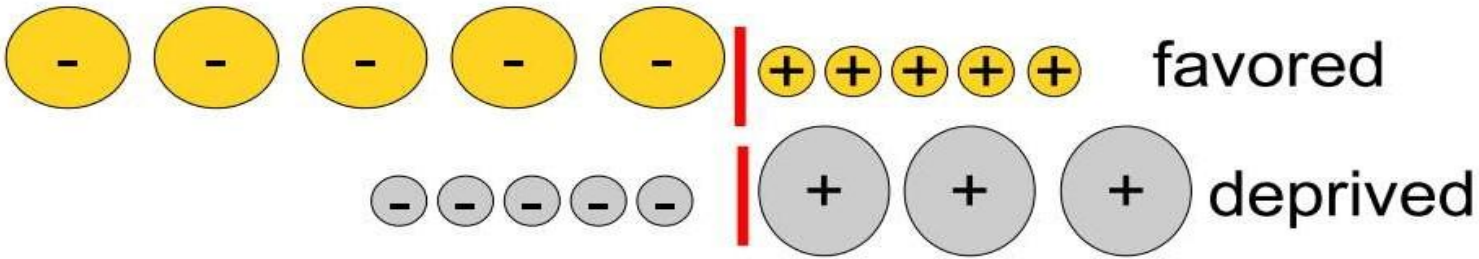
Classifier



Reweighing

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

b) assign weights to make the data impartial



Algorithm for reweighting

Algorithm 3: *Reweighing*

Input: $(D, S, Class)$
Output: Classifier learned on reweighed D

1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: $D_W := \{\}$
7: **for** X in D **do**
8: Add $(X, W(X(S), X(Class)))$ to D_W
9: **end for**
10: Train a classifier C on training set D_W, taking onto account the weights
11: **return** Classifier C

Reweighing

$$W(x(B) = b \mid x(Class) = +) := \frac{P_{exp}(b \wedge +)}{P_{act}(b \wedge +)}$$

$$W(x(B) = b \mid x(Class) = -) := \frac{P_{exp}(b \wedge -)}{P_{act}(b \wedge -)}$$

$$W(x(B) = \overline{b} \mid x(Class) = +) := \frac{P_{exp}(\overline{b} \wedge +)}{P_{act}(\overline{b} \wedge +)}$$

$$W(x(B) = \overline{b} \mid x(Class) = -) := \frac{P_{exp}(\overline{b} \wedge -)}{P_{act}(\overline{b} \wedge -)}$$

Sex	Ethnicity	Highest Degree	Job Type	Cl.	Weight
m	native	h. school	board	+	0.75
m	native	univ.	board	+	0.75
m	native	h. school	board	+	0.75
m	non-nat.	h. school	healthcare	+	0.75
m	non-nat.	univ.	healthcare	-	2
f	non-nat.	univ.	education	-	0.67
f	native	h. school	education	-	0.67
f	native	none	healthcare	+	1.5
f	non-nat.	univ.	education	-	0.67
f	native	h. school	board	+	1.5

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

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

$$W(Sex = f \mid x(Class) = -) = 0.67$$
$$W(Sex = m \mid x(Class) = +) = 0.75$$
$$W(Sex = m \mid x(Class) = -) = 2 \ .$$

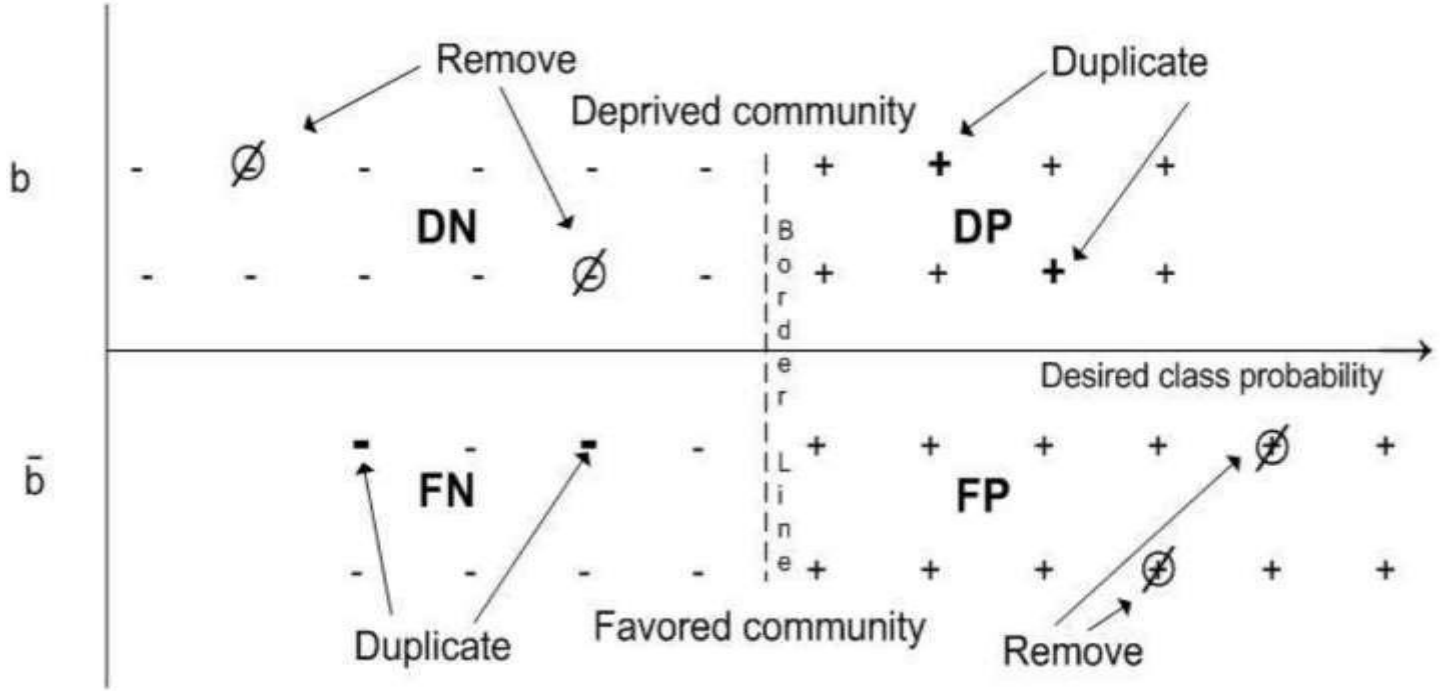
Sampling

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

$$
\begin{aligned}
DP &:= \{x \in D \mid x(B) = b \wedge x(Class) = +\} \\
DN &:= \{x \in D \mid x(B) = b \wedge x(Class) = -\} \\
FP &:= \{x \in D \mid x(B) = \overline{b} \wedge x(Class) = +\} \\
FN &:= \{x \in D \mid x(B) = \overline{b} \wedge x(Class) = -\}
\end{aligned}
$$

Then sample accordingly, possibly duplicating data points.

Uniform Sampling



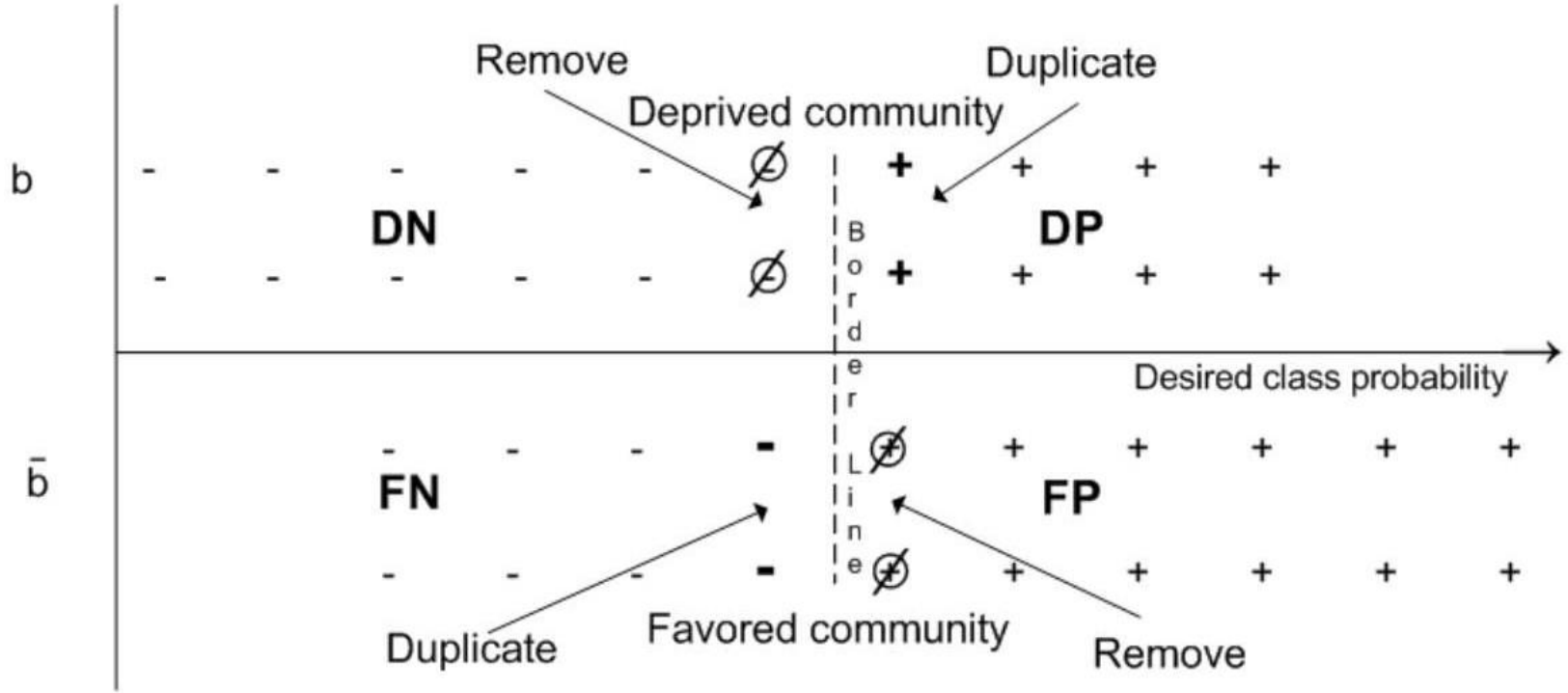
Uniform sampling

Algorithm 4: *Uniform Sampling*

Input: $(D, S, Class)$
Output: Classifier C learned on resampled D

1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: Sample uniformly $W(b, +) \times |DP|$ objects from DP;
7: Sample uniformly $W(w, +) \times |FP|$ objects from FP;
8: Sample uniformly $W(b, -) \times |DN|$ objects from DN;
9: Sample uniformly $W(w, -) \times |FN|$ objects from FN;
10: Let D_{US} be the bag of all samples generated in steps 6 to 9
11: **return** Classifier C learned on D_{US}

Preferential Sampling



Algorithm 5: *Preferential Sampling*

Input: $(D, S, Class)$
Output: Classifier C learned on resampled D

1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: Learn a ranker R for predicting $+$ using D as training set
7: $D_{PS} := \{\}$
8: Add $\lfloor W(b, +) \rfloor$ copies of DP to D_{PS}
9: Add $\lfloor W(b, +) - \lfloor W(b, +) \rfloor \times |DP| \rfloor$ lowest ranked elements of DP to D_{PS}
10: Add $\lfloor W(b, -) \times |DN| \rfloor$ lowest ranked elements of DN to D_{PS}
11: Add $\lfloor W(w, +) \times |FP| \rfloor$ highest ranked elements of FP to D_{PS}
12: Add $\lfloor W(w, -) \rfloor$ copies of FN to D_{PS}
13: Add $\lfloor W(w, -) - \lfloor W(b, -) \rfloor \times |FN| \rfloor$ highest ranked elements of FN to D_{PS}
14: **return** Classifier C learned on D_{PS}

Performance

Preprocess method	Disc (%)	Acc (%)
No	16.4 ± 1.31	86.05 ± 0.29
No_SA	16.6 ± 1.43	86.01 ± 0.31
RW	7.97 ± 1.02	85.62 ± 0.30
US	7.91 ± 2.05	85.35 ± 0.36
PS	3.08 ± 0.79	84.30 ± 0.25
M_NBS	1.77 ± 1.16	83.65 ± 0.24
M_J48	2.49 ± 1.92	83.49 ± 0.47
M_IBk1	7.67 ± 0.86	85.35 ± 0.46
M_IBk2	3.62 ± 0.61	84.44 ± 0.27
M_IBk3	2.40 ± 0.51	83.78 ± 0.43

Fairness-aware data mining

Pre-processing approaches:

[Pre_1] M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. *"Certifying and removing disparate impact"*. In KDD, pp. 259-268, 2015.

[Pre_2] F. Kamiran and T. Calders. *"Data preprocessing techniques for classification without discrimination"*. In Knowledge and Information Systems (KAIS), 33(1), 2012.

[Pre_3] S. Hajian and J. Domingo-Ferrer. *"A methodology for direct and indirect discrimination prevention in data mining"*. In IEEE Transactions on Knowledge and Data Engineering (TKDE), 25(7), 2013.

[Pre_4] I. Zliobaite, F. Kamiran and T. Calders. *"Handling conditional discrimination"*. In ICDM, pp. 992-1001, 2011.

Discrimination: direct or indirect.

- Direct discrimination: decisions are made based on sensitive attributes.
- Indirect discrimination (redlining): decisions are made based on nonsensitive attributes which are strongly correlated with biased sensitive ones.
- Decision rules

Definitions

Dataset – collection of records

Item - attribute with its value, e.g., Race = black

Item set - collection of items

$$\{Foreign\ worker = Yes;\ City = NYC\}$$

Classification rule $X \rightarrow C \in \{yes/no\}$

$$\{Foreign\ worker = Yes;\ City = NYC\} \rightarrow Hire = no$$

Definitions

support, supp(X) - fraction of records that contain X

confidence, conf(X→ C) - how often C appears in records that contain X
conf X → C = supp(X,C) / supp(X)

frequent classification rule:
supp (X, C) > s
conf (X → C) > c

negated item set: X = {Foreign worker = Yes}
¬X = {Foreign worker = No}

Classification rules

- DI - predetermined discriminatory items
  DI = {Foreign worker = Yes; Race = Black}
- X → C - potentially discriminatory (PD)
  X = A, B with A ⊆ DI , B ⊈ DI
  {Foreign worker = Yes; City = NYC} → Hire = No
- X → C - potentially nondiscriminatory (PND)
  X = D, B with D ⊈ DI s , B ⊈ DI s
  {Zip = 10451; City = NYC} → Hire = No

Direct Discrimination Measure

- extended lift (*elift*):

$$elift(A, B \rightarrow C) = \frac{conf(A,B \rightarrow C)}{conf(B \rightarrow C)}$$

- $A \in DI_S$
- $A, B \rightarrow C$ is α-protective, if and $elift(A, B \rightarrow C) < \alpha$
- $A, B \rightarrow C$ is α-discriminatory, if $elift(A, B \rightarrow C) \geq \alpha$

Indirect Discrimination Measure

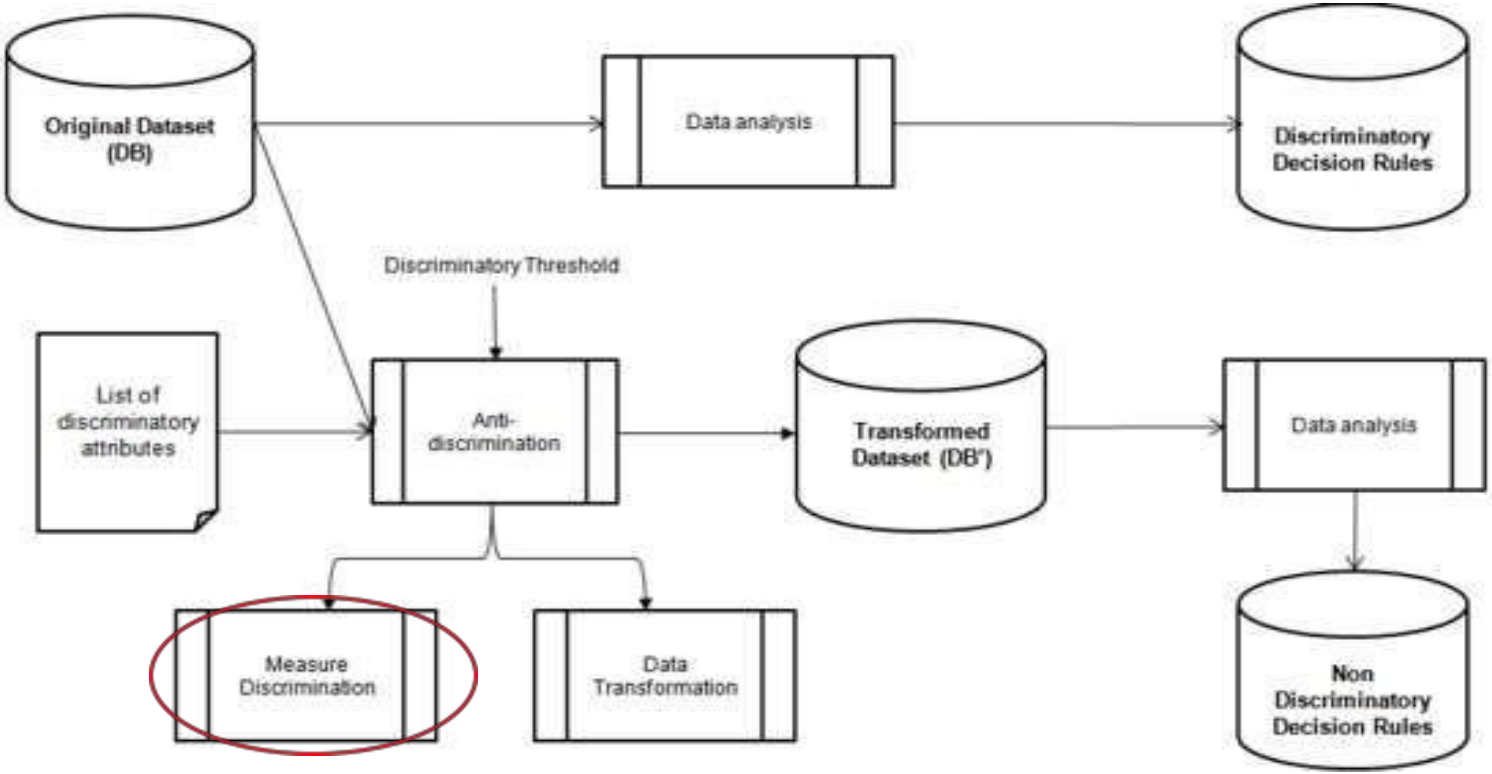
- **Theorem**: Let $r: D, B \rightarrow C$ is PND;
- $\gamma = conf(r: D, B \rightarrow C)$ and $\delta = conf(B \rightarrow C) > 0$
- $A \subseteq DI_s$,
- $conf(r_{b1}: A, B \rightarrow D) \geq \beta_1$, $conf(r_{b2}: D, B \rightarrow A) \geq \beta_2 > 0$
- $f(x) = \frac{\beta_1}{\beta_2}(\beta_2 + x - 1)$
- $elb(x, y) = \begin{cases} \frac{f(x)}{y}, if\ f(x) > 0 \\ 0, \quad otherwise \end{cases}$
- Then if $elb(\gamma, \delta) \geq \alpha$,

  then PD $r': A, B \rightarrow C$ is α-discriminatory

The Approach

- Discrimination measurement:
  - Find PD and PND
  - Direct discrimination:
    - In PD find α-discriminatory by $elif()$
  - Indirect discrimination:
    - In PND find redlining by $elb()$ + background knowledge
- Data transformation:
  - Alter dataset and remove discriminatory biases
  - Minimum impact on data and legitimate rules

A framework for direct and indirect discrimination prevention in data mining



Classification rule: $r = A, B \rightarrow C$

B	C	$\neg C$	
A	a_1	$n_1 - a_1$	n_1
$\neg A$	a_2	$n_2 - a_2$	n_2

$a_1 = supp(A, B, C)$
$a_2 = supp(\neg A, B, C)$
$n_1 = supp(A, B)$
$n_2 = supp(\neg A, B)$

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

$$elift(r) = \frac{p_1}{p}, \quad elift_d(r) = p_1 - p, \quad elift_c(r) = \frac{1 - p_1}{1 - p}$$

$$slift(r) = \frac{p_1}{p_2}, \quad slift_d(r) = p_1 - p_2, \quad slift_c(r) = \frac{1 - p_1}{1 - p_2}$$

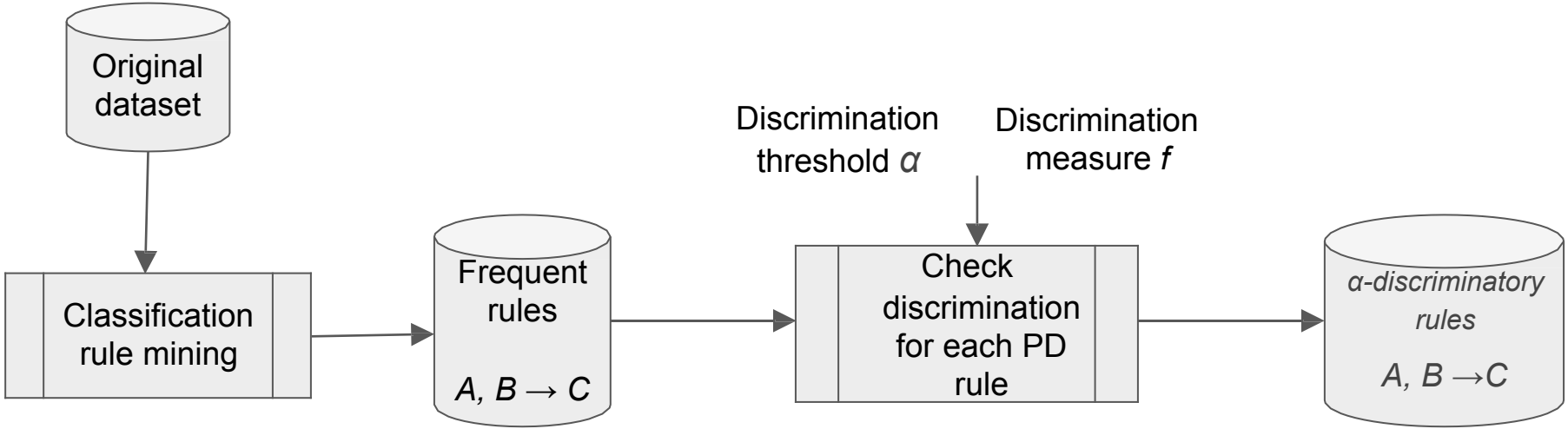
Based on direct discriminatory measures $f \in$ {*elift, slift*, ...}, a PD classification rule r: $A, B \rightarrow C$ is:

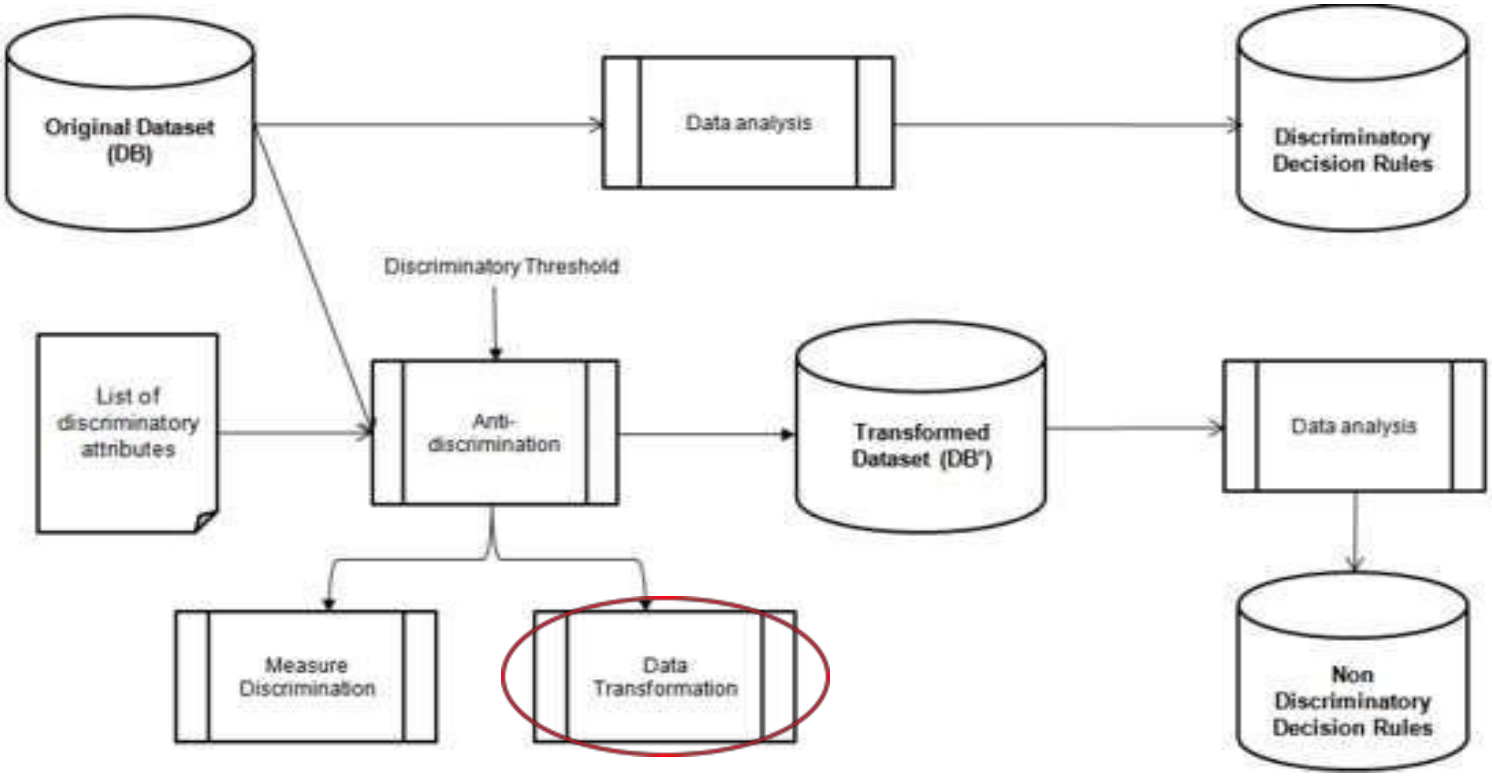
α-discriminatory if $f(r) \geq \alpha$; *or* α-protective if $f(r) < \alpha$

α states an acceptable level of discrimination according to laws and regulations

e.g. U.S. Equal Pay Act: This amounts to using *slift* with = 1.25.

Measure discrimination




Original Dataset (DB)
Data analysis
Discriminatory Decision Rules
Discriminatory Threshold
List of discriminatory attributes
Anti-discrimination
Transformed Dataset (DB')
Data analysis
Measure Discrimination
Data Transformation
Non Discriminatory Decision Rules

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 α-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=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

Data transformation methods for direct rule protection

	Direct Rule Protection
DTM 1	$\neg A, B \rightarrow \neg C \Rightarrow A, B \rightarrow \neg C$
DTM 2	$\neg A, B \rightarrow \neg C \Rightarrow \neg A, B \rightarrow C$

Which records should be change and how?

A suitable data transformation with minimum information loss to make each *α*-discriminatory rule *α*-protective.

we should enforce the following inequality for each *α*-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 *α*-discriminatory rule *r* α-protective w.r.t. *f* do not generate new *α*-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 *α*-discriminatory rule *α*-protective.

we should enforce the following inequality for each *α*-discriminatory rule *r*

f(r: A, B → C) < α, where f = elift

DTM1: Taking ∆elift equal to the ceiling of the right-hand side of Equation (below) suffices to make *α*-discriminatory rule *r*, α-protective w.r.t. f = elift.

$$\Delta_{elift} > \frac{\alpha \times supp(A,B) \times supp(B,C) - supp(A,B,C) \times supp(B)}{supp(A,B,C) - \alpha \times supp(A,B)}$$

In which order records should be changed?

DTM1: perturb the discriminatory itemset from ~A (male) to A (female) in the subset *Dc* of all records of the original data set which completely support the rule ~A, B →~C and have minimum impact on other rules

$\mathcal{D}_c \leftarrow$ All records completely supporting $\neg A, B \rightarrow \neg C$
for each $db_c \in \mathcal{D}_c$ **do**
  Compute $impact(db_c) = |\{r_a \in \mathcal{FR} | db_c \text{ supports the premise of } r_a\}|$
end for
Sort $\mathcal{D}_c$ by ascending impact

Fairness-aware data mining

Pre-processing approaches:

[Pre_1] M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. *"Certifying and removing disparate impact"*. In KDD, pp. 259-268, 2015.

[Pre_2] F. Kamiran and T. Calders. *"Data preprocessing techniques for classification without discrimination"*. In Knowledge and Information Systems (KAIS), 33(1), 2012.

[Pre_3] S. Hajian and J. Domingo-Ferrer. *"A methodology for direct and indirect discrimination prevention in data mining"*. In IEEE Transactions on Knowledge and Data Engineering (TKDE), 25(7), 2013.

[Pre_4] I. Zliobaite, F. Kamiran and T. Calders. *"Handling conditional discrimination"*. In ICDM, pp. 992-1001, 2011.

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

	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate	20%	20%	40%	40%
accepted (+)	160	40	80	320

- 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

	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate	15%	25%	35%	45%
accepted (+)	120	50	70	360

Traditional method:

discr. = P(+ | m) – P(+ | f)

= (20% x 25% + 80% x 45%)

- (80% x 15% + 20% x 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*(+|Fac) for faculty Fac?

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

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

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

$$P(+ \mid m,Fac) = P(+ \mid f,Fac) = P^*(+ \mid Fac)$$

$D_{bad} = D_{all} - D_{expl}$

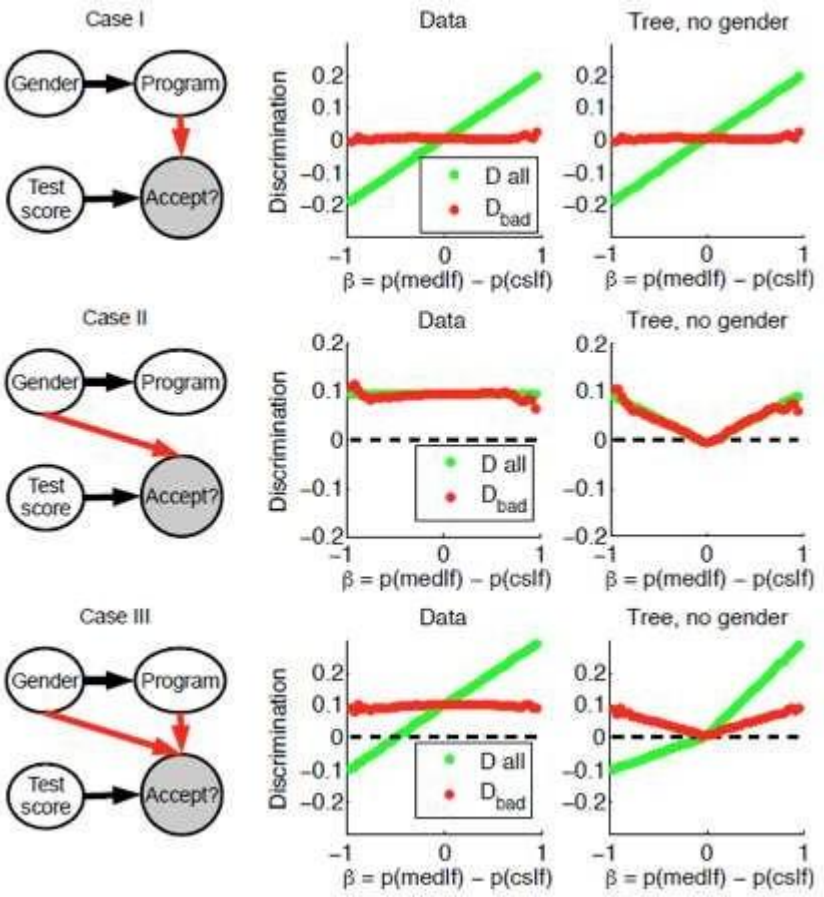
Analysis of explainable discrimination

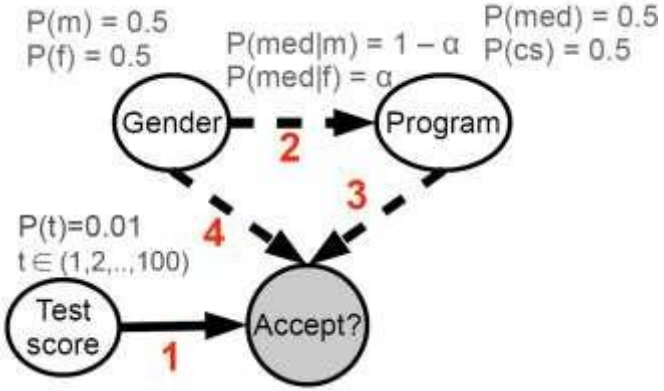
	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate (Example 2)	15%	25%	35%	45%
corrected acceptance rate	20%		40%	
accepted explainable	160	40	80	320

D_{expl} = (20% x 20% + 80% x 40%) - (80% x 20% + 20% x 40%) = 12%

$D_{bad} = D_{all}$ – 12% = 22% - 12% = 10%

Simulation Experiments





$$y = \delta\left[\left(t + a(-1)^{\delta[med]} + b(-1)^{\delta[f]}\right) > 70\right]$$

- 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

	$P(t)$	a	b	$P(med\|f)$
Case I, only explainable	0.01	10	0	α
Case II, only bad	0.01	0	5	α
Case III, explainable and bad	0.01	10	5	α

Solution: Locally change input data

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

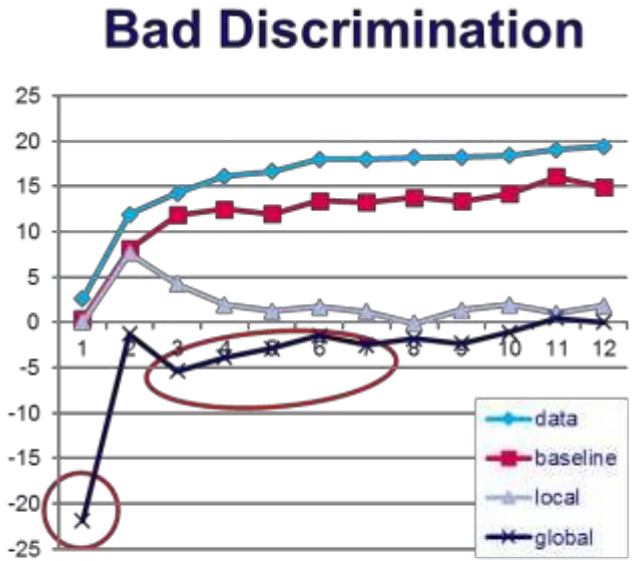
2. Estimate $P^*(+|e_i)$ for all partitions e_i

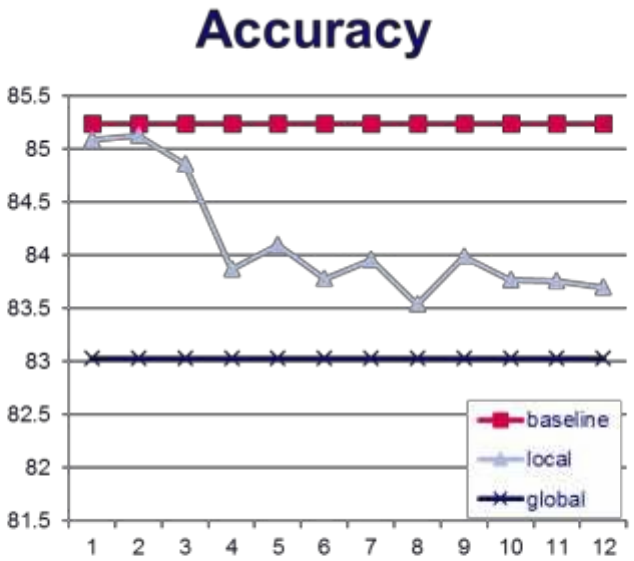
3. Apply local techniques on partition e_i so that

$P(+ |e_i,f) = P(+ | e_i,m) = P^*(+ | e_i)$ becomes true

- Local massaging
- Local preferential sampling

Experiments: Discrimination after Massaging

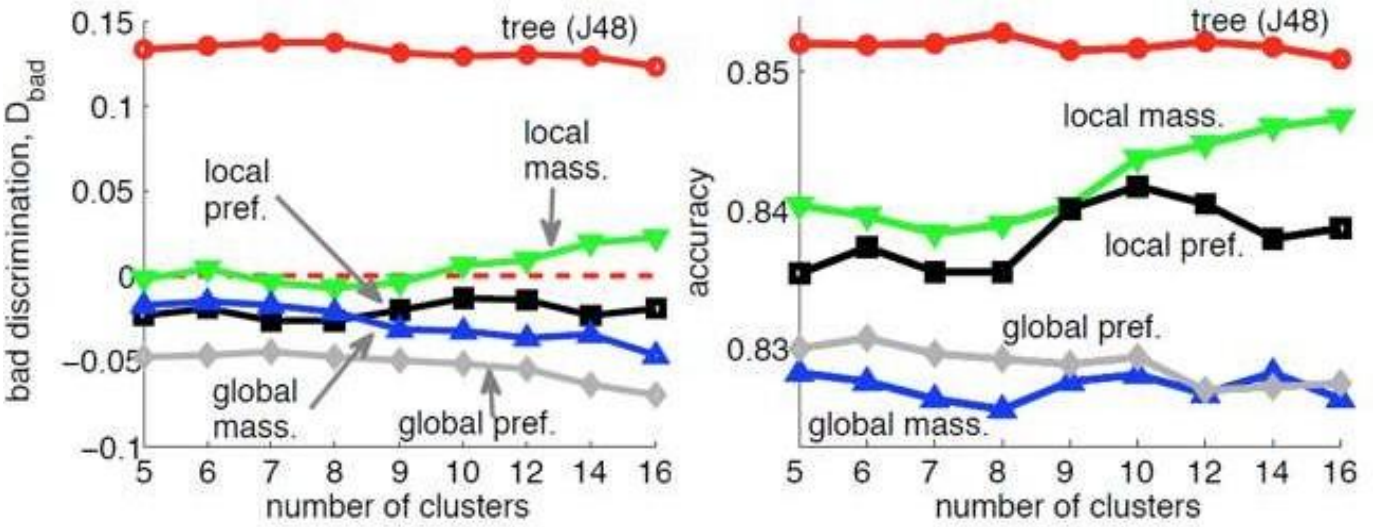




- **Global techniques tend to overshoot when large part of the discrimination can be explained**

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

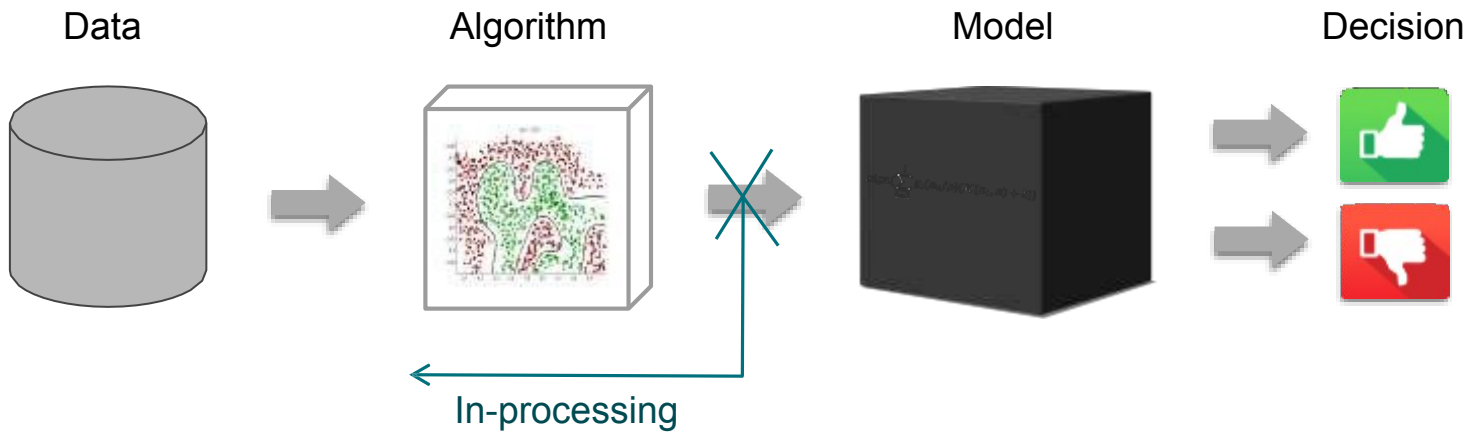
Pre-processing approaches: **(not covered here)**

[Pre_5] S. Ruggieri. *"Using t-closeness anonymity to control for non-discrimination"*. Transactions on Data Privacy, 7(2), pp.99-129, 2014.

[Pre_6] S. Hajian, J. Domingo-Ferrer, and O. Farras. *"Generalization-based privacy preservation and discrimination prevention in data publishing and mining"*. Data Mining and Knowledge Discovery, 28(5-6), pp.1158-1188, 2014.

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:

[In_1] F. Kamiran, T. Calders and M. Pechenizkiy. *"Discrimination aware decision tree learning"*. In ICDM, pp. 869-874, 2010.

[In_2] T. Calders and S. Verwer. *"Three Naïve Bayes Approaches for Discrimination-Free Classification"*. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[In_3] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. *"Fairness Constraints: A Mechanism for Fair Classification"*. arXiv preprint arXiv:1507.05259, 2015.

[In_4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. *"Fairness through awareness"*. In ITCS 2012, pp. 214-226, 2012.

[In_5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. *"Learning fair representations"*. In ICML, pp. 325-333, 2013.

Fairness-aware data mining

In-processing approaches: **(not covered here)**

[In_6] T. Kamishima, S. Akaho, H. Asoh and J. Sakuma. *"Fairness-aware classifier with prejudice remover regularizer"*. In PKDD, pp. 35-50, 2012.

[In_7] K. Fukuchi, J. Sakuma, T. Kamishima. *"Prediction with Model-Based Neutrality"*. ECML/PKDD (2) 2013: 499-514.

[In_8] B. Fish, J. Kun and A.D. Lelkes. *"A Confidence-Based Approach for Balancing Fairness and Accuracy"*
arXiv preprint:1601.05764, 2015.

many more… (probably)

Fairness-aware data mining

In-processing approaches:

[In_1] F. Kamiran, T. Calders and M. Pechenizkiy. *"Discrimination aware decision tree learning"*. In ICDM, pp. 869-874, 2010.

[In_2] T. Calders and S. Verwer. *"Three Naïve Bayes Approaches for Discrimination-Free Classification"*. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[In_3] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. *"Fairness Constraints: A Mechanism for Fair Classification"*. arXiv preprint arXiv:1507.05259, 2015.

[In_4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. *"Fairness through awareness"*. In ITCS 2012, pp. 214-226, 2012.

[In_5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. *"Learning fair representations"*. In ICML, pp. 325-333, 2013.

Problem definition

- Given: dataset D, an attribute B, a value $b \in dom(B)$
- 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 overresultant 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: $GINI_{split}(Class) / GINI_{split}(B)$

Objective: $GINI_{split}(Class) - GINI_{split}(B)$

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:

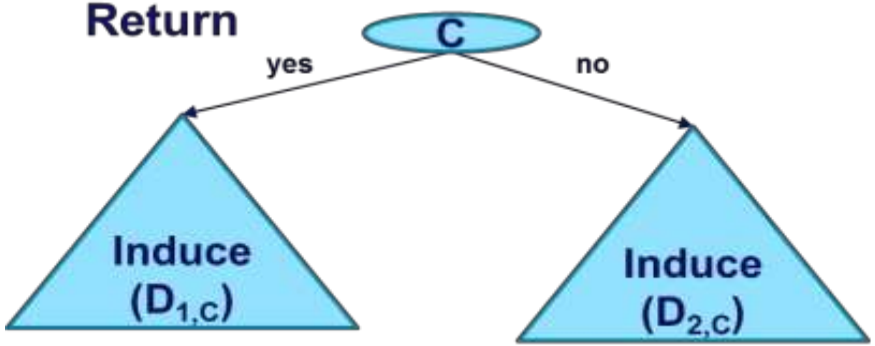
$D_{1,C}$ = { t in D | t satisfies C}

$D_{2,C}$ = D - D_1

Measure Quality(D_1 ,D_2) ← $GINI_{split}(Class) / GINI_{split}(G)$

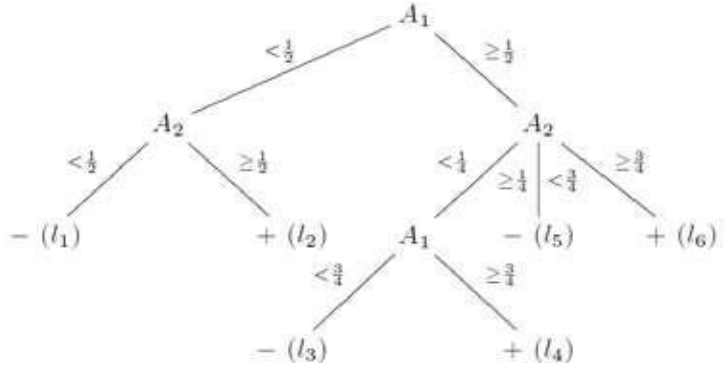
Let C be the best split

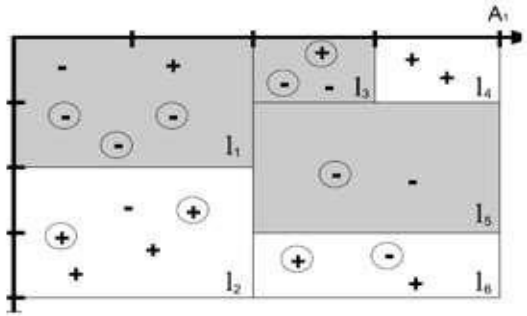
Return



Leaf relabeling

Decision trees divide up the decision space





Labels are assigned according to the majority class

Disc $_T$ = p(M = + | B ≠ b) - p(M = + | B = b) = 6/10 - 4/10 = 0.2 or 20 %

Relabel some leaves to reduce the discrimination

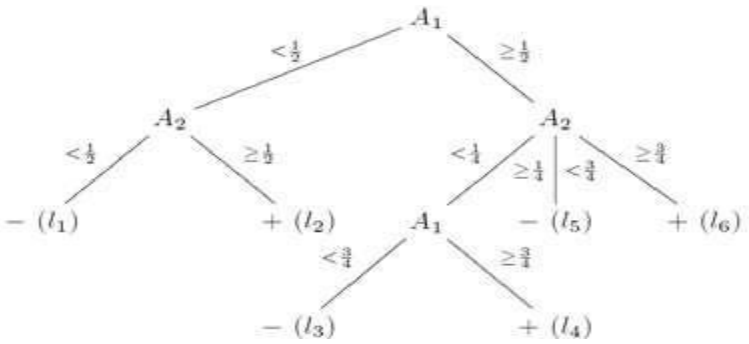
Leaf relabeling

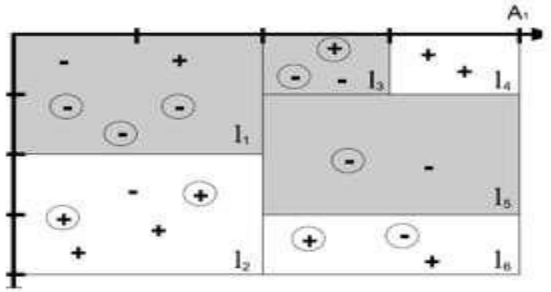
E.g.: Relabel node l_1 from – to +

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

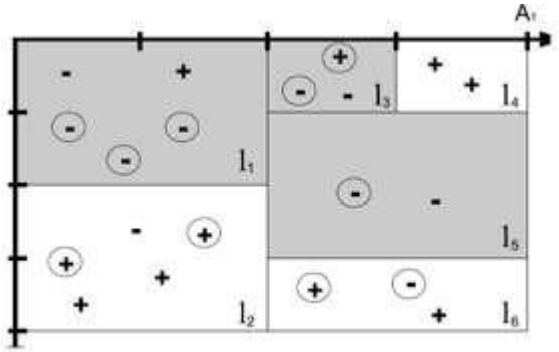
Knapsack problem:

Select nodes NOT relabeled

Dacc: weight Ddisc: size

K = 30% (that can stay)

Outcome: relabel l_4



$Node$	Δacc	$\Delta disc$
l_1	-15%	-10%
l_2	-15%	-10%
l_3	-5%	-10%
l_4	-10%	-20%
l_5	-10%	0%
l_6	-5%	10%

Fairness-aware data mining

In-processing approaches:

[In_1] F. Kamiran, T. Calders and M. Pechenizkiy. *"Discrimination aware decision tree learning"*. In ICDM, pp. 869-874, 2010.

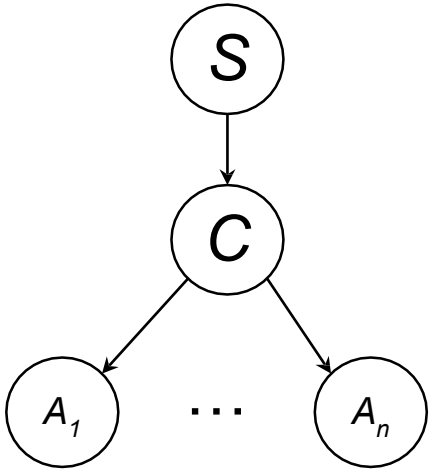
[In_2] T. Calders and S. Verwer. *"Three Naïve Bayes Approaches for Discrimination-Free Classification"*. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[In_3] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. *"Fairness Constraints: A Mechanism for Fair Classification"*. arXiv preprint arXiv:1507.05259, 2015.

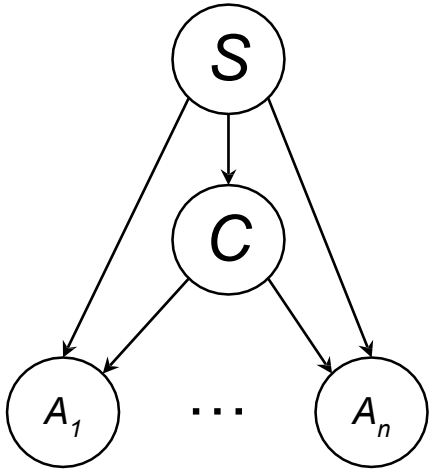
[In_4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. *"Fairness through awareness"*. In ITCS 2012, pp. 214-226, 2012.

[In_5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. *"Learning fair representations"*. In ICML, pp. 325-333, 2013.

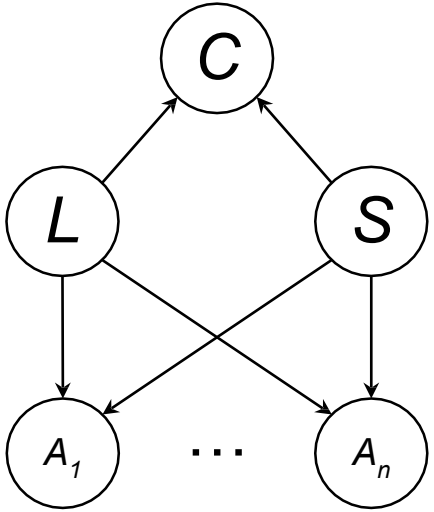
Three Naive Bayes Approaches for Discrimination-Free Classification



Approach 1:
Modified Naive Bayes

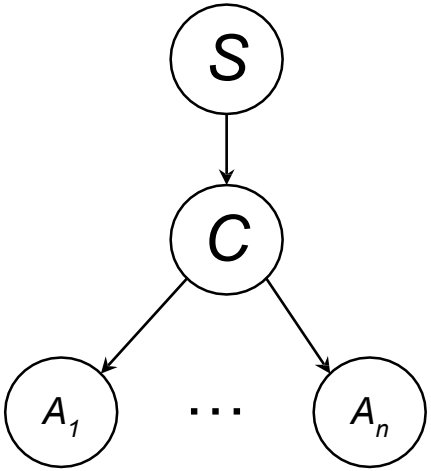


Approach 2:
Two Naive Bayes models



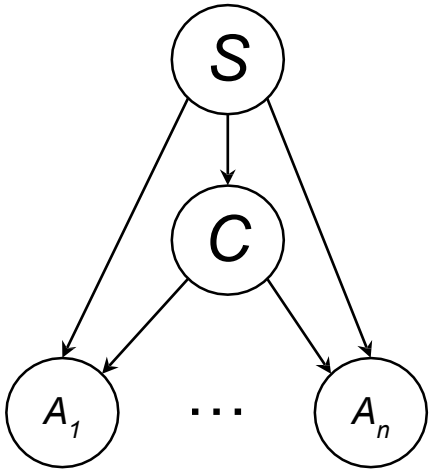
Approach 3:
Latent variable model

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.

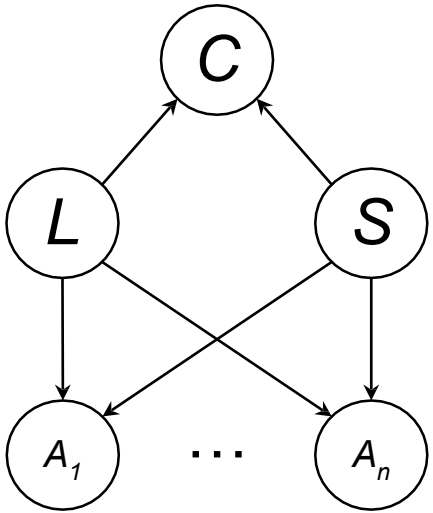
Approach 2: Two Naive Bayes models



- 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 → L is discirmination-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:

[In_1] F. Kamiran, T. Calders and M. Pechenizkiy. *"Discrimination aware decision tree learning"*. In ICDM, pp. 869-874, 2010.

[In_2] T. Calders and S. Verwer. *"Three Naïve Bayes Approaches for Discrimination-Free Classification"*. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[In_3] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. *"Fairness Constraints: A Mechanism for Fair Classification"*. arXiv preprint arXiv:1507.05259, 2015.

[In_4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. *"Fairness through awareness"*. In ITCS 2012, pp. 214-226, 2012.

[In_5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. *"Learning fair representations"*. In ICML, pp. 325-333, 2013.

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

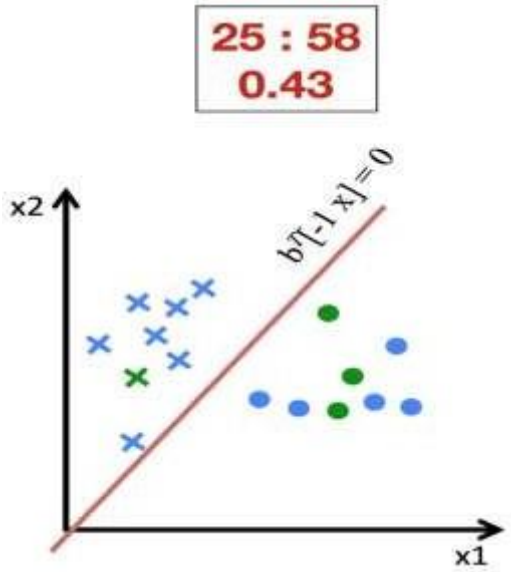
Fair classifier

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

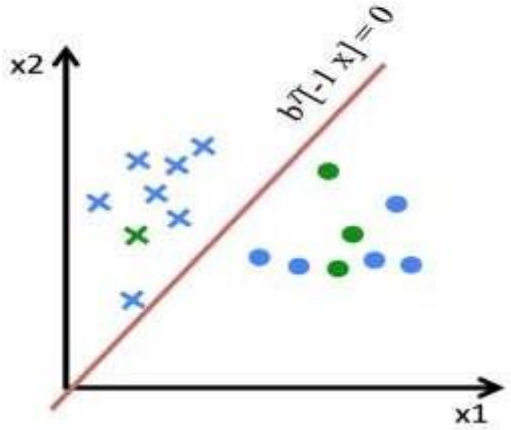
Fairness constraint



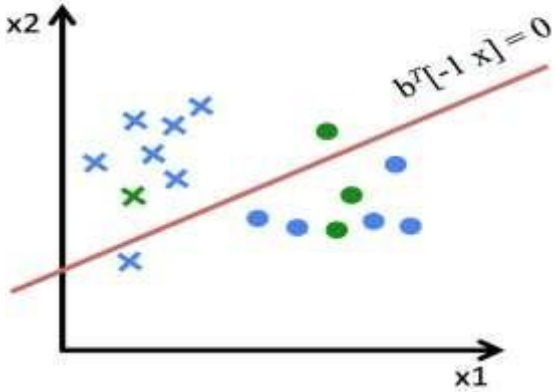
Fairness constraint

$$\left| \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}}) \, \mathbf{b}^T [-1 \ \mathbf{x}_i] \right| \leq \mathbf{c}$$

25 : 58
0.43



50 : 50
1.0



Modifying the logistic regression classifier

$$p(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-b_0 + \sum_j b_j x_{ij}}}$$

$$\text{maximize} \quad \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i)$$

Modifying the logistic regression classifier

$$p(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-b_0 + \sum_j b_j x_{ij}}}$$

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i) \\ \text{subject to} \quad & \tfrac{1}{N} \textstyle\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T [-1 \ \mathbf{x}_i] \leq \mathbf{c}, \\ & \tfrac{1}{N} \textstyle\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T [-1 \ \mathbf{x}_i] \geq -\mathbf{c} \end{aligned}$$

Key point: possible to solve this problem efficiently

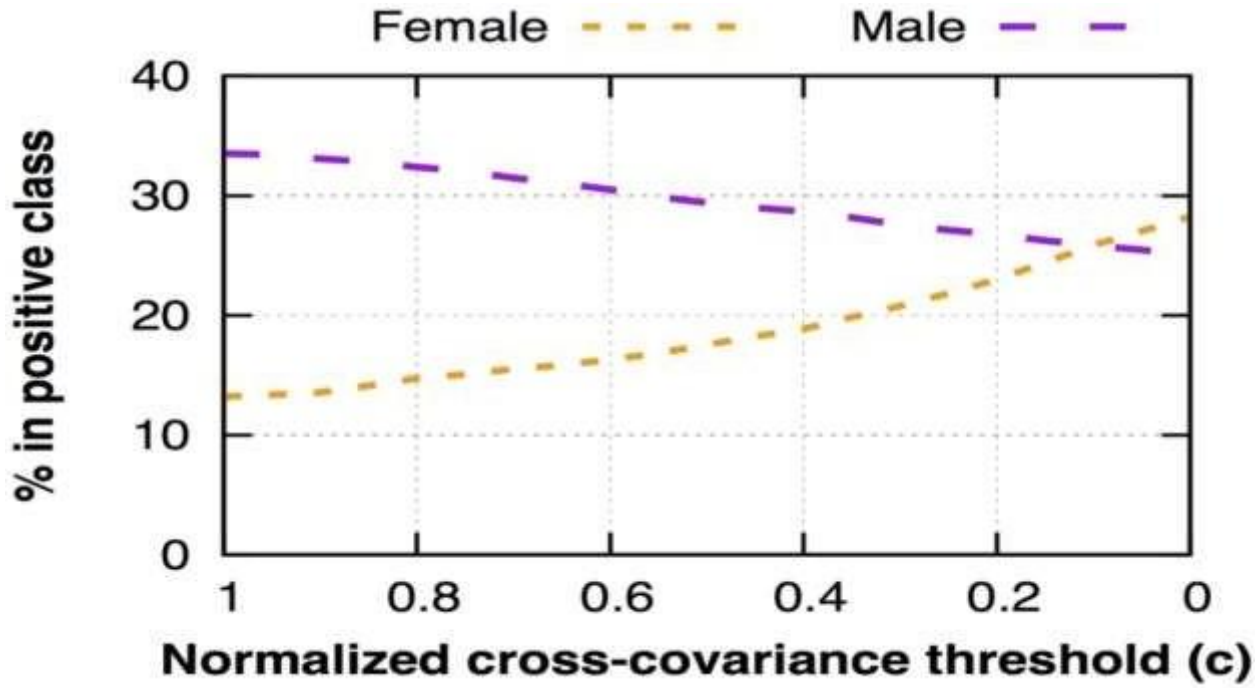
Modifying the Hinge loss classifier

$$
\begin{aligned}
\text{minimize} \quad & \textstyle\sum_{i=1}^{N} max(0, y_i(\mathbf{b}^T[-1\ \mathbf{x}_i])) \\
\text{subject to} \quad & \tfrac{1}{N}\textstyle\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\,\mathbf{b}^T[-1\ \mathbf{x}_i] \leq \mathbf{c}, \\
& \tfrac{1}{N}\textstyle\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\,\mathbf{b}^T[-1\ \mathbf{x}_i] \geq -\mathbf{c},
\end{aligned}
$$

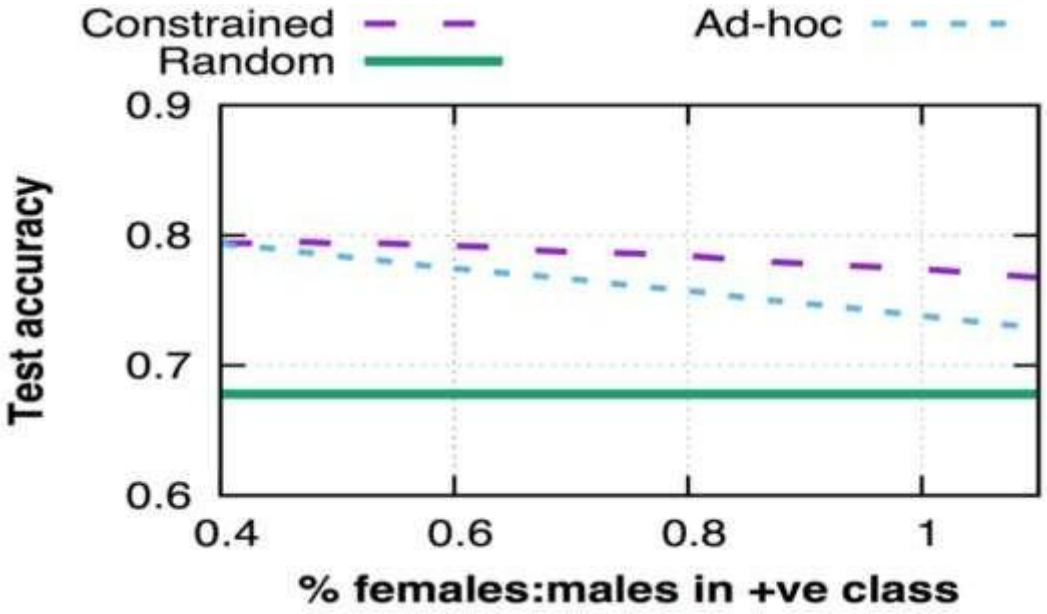
Modifying the SVM classifier

$$
\begin{aligned}
\text{minimize} \quad & \|\mathbf{b}\|^2 + C \sum_{i=1}^{n} \xi_i \\
\text{subject to} \quad & y_i(\mathbf{b}^T[-1\ \mathbf{x}_i]) \geq 1 - \xi_i, \forall i \in \{1, \ldots, n\} \\
& \xi_i \geq 0, \forall i \in \{1, \ldots, n\}, \\
& \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \leq \mathbf{c}, \\
& \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \geq -\mathbf{c}.
\end{aligned}
$$

Tightening the constraints increases fairness



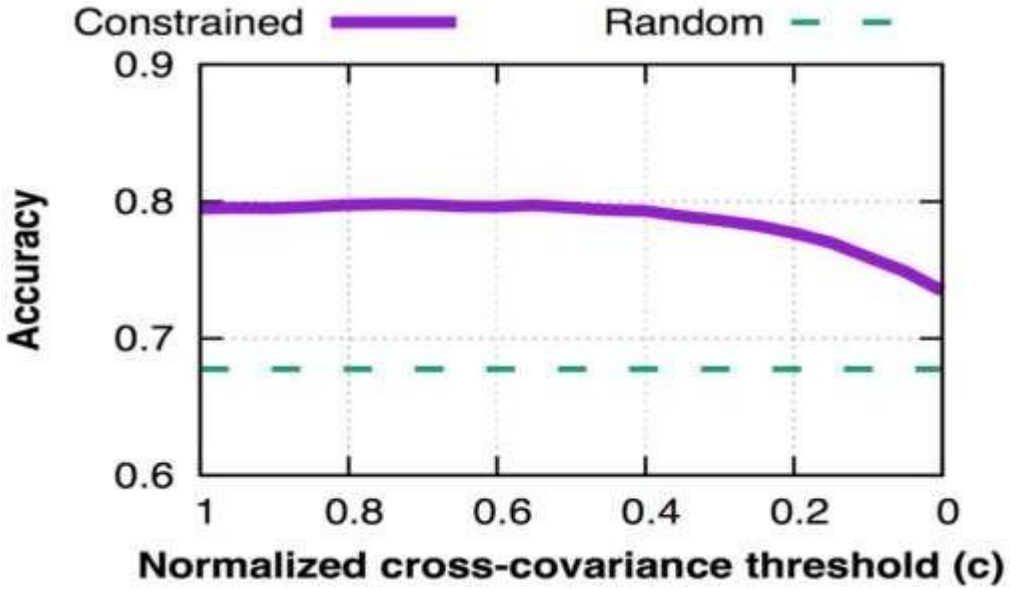
Fairness vs accuracy trade-off



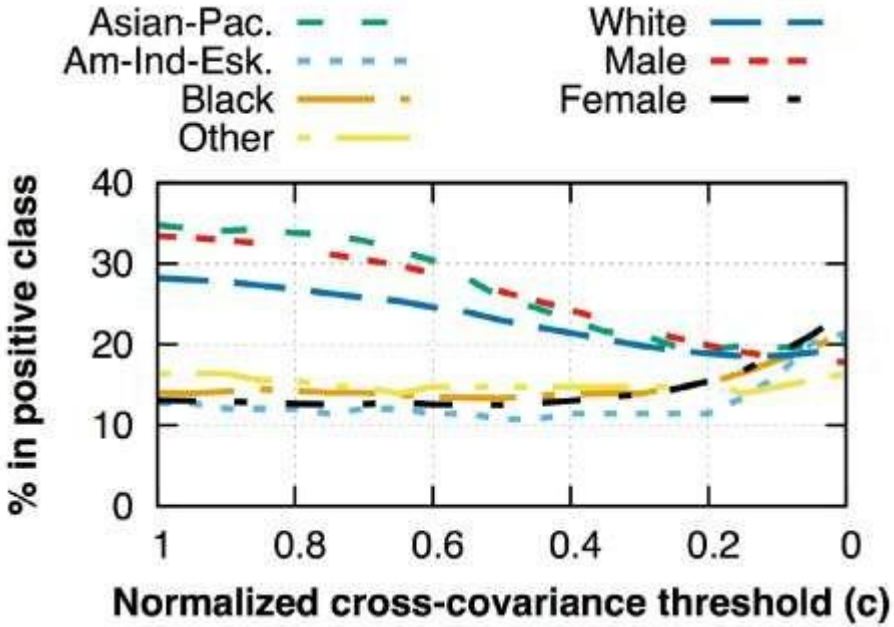
Random: takes the output of the uncontrained classifier and shuffles labels randomly until satisfying the given c.

Ad-hoc: takes the output of the uncontrained 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:

[In_1] F. Kamiran, T. Calders and M. Pechenizkiy. *"Discrimination aware decision tree learning"*. In ICDM, pp. 869-874, 2010.

[In_2] T. Calders and S. Verwer. *"Three Naïve Bayes Approaches for Discrimination-Free Classification"*. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[In_3] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. *"Fairness Constraints: A Mechanism for Fair Classification"*. arXiv preprint arXiv:1507.05259, 2015.

[In_4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. *"Fairness through awareness"*. In ITCS 2012, pp. 214-226, 2012.

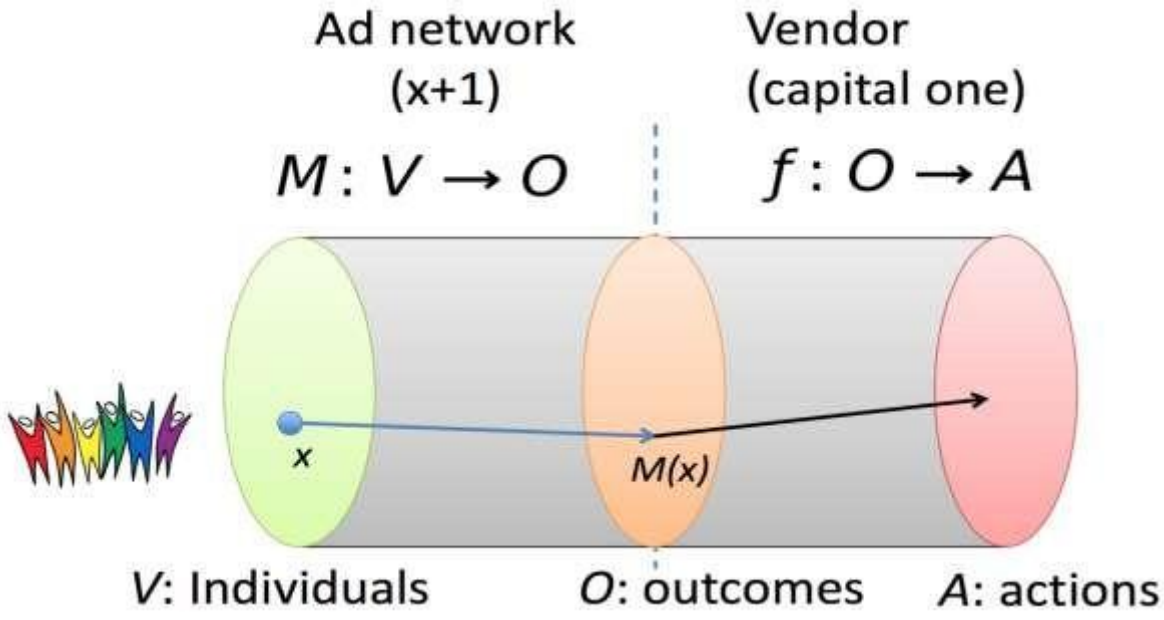
[In_5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. *"Learning fair representations"*. In ICML, pp. 325-333, 2013.

Credit Application (WSJ 8/4/10)

User visits capitalone.com

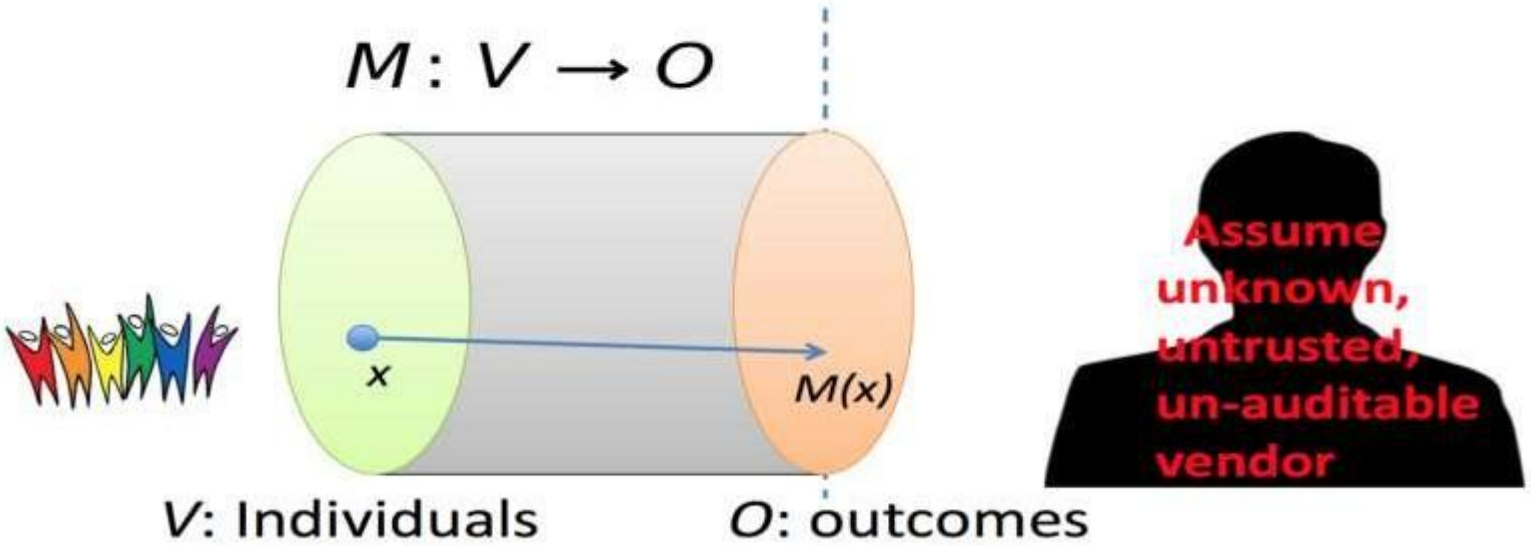
Capital One uses tracking information provided by the tracking network [x+1] to personalize offers

Concern: *Steering* minorities into higher rates (illegal)


Ad network
(x+1)
Vendor
(capital one)
$M : V \rightarrow O$
$f : O \rightarrow A$
x
M(x)
V: Individuals
O: outcomes
A: actions

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

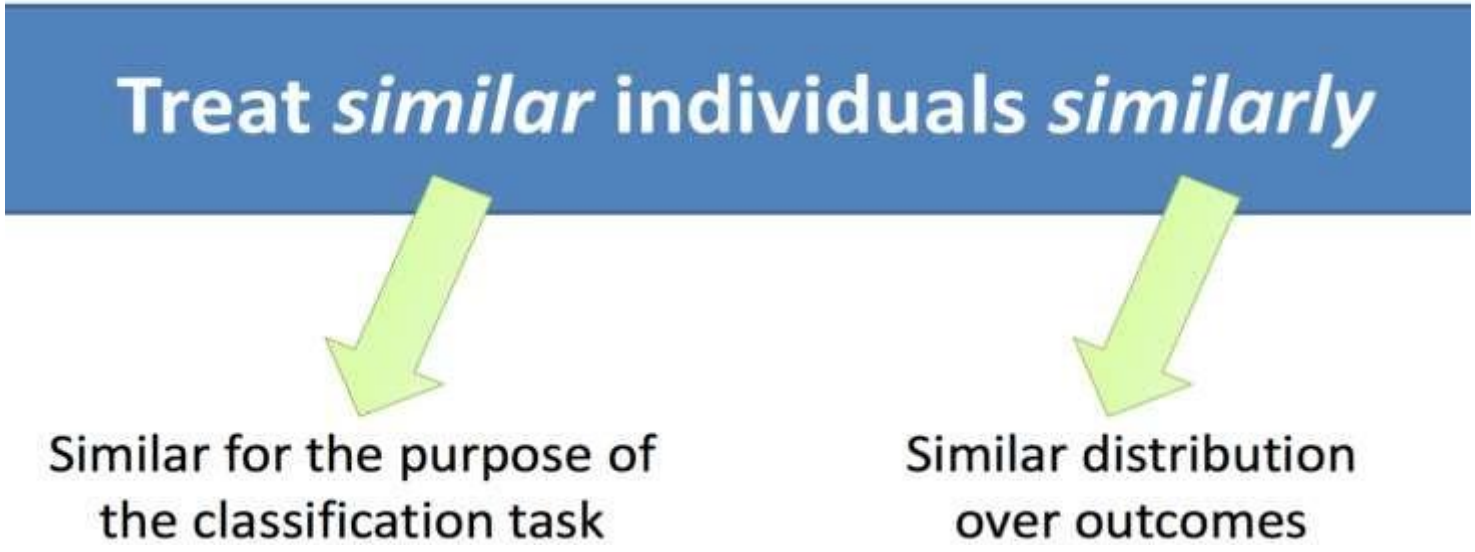
- Equalize two groups S, T at the level of outcomes
  - E.g. S = minority, $T = S^c$
  - Pr[outcome o | S] = Pr [outcome o | T]

- **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



- 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 classificiation!

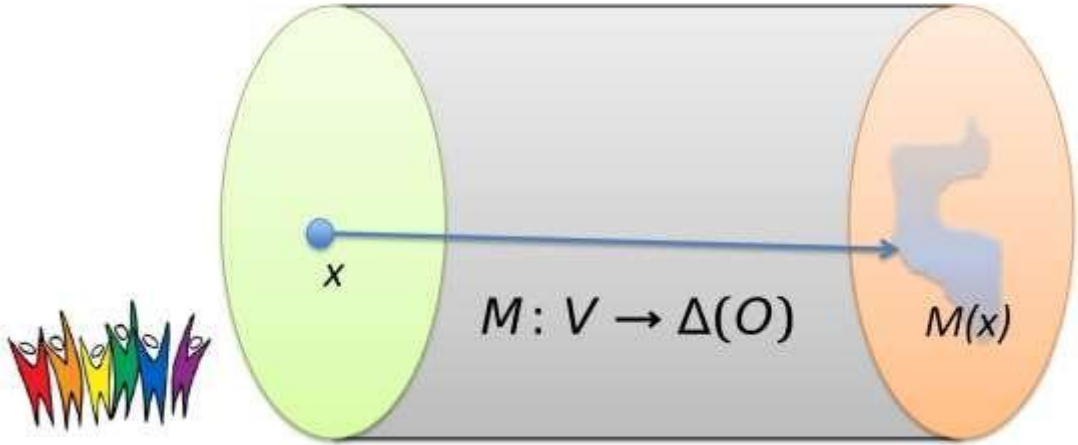
Examples

- 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

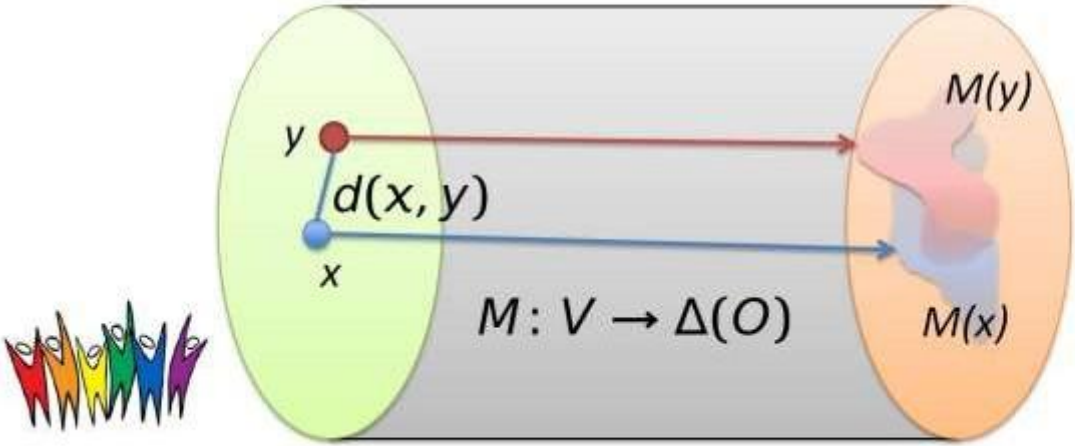


V: Individuals

O: outcomes

Metric $d : V \times V \to \mathbb{R}$

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



V: Individuals

O: outcomes

Utility Maximization

Vendor can specify **arbitrary utility function**

$$U : V \times O \to \mathbb{R}$$

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

$$\max \mathop{\mathbb{E}}_{x \in V} \mathop{\mathbb{E}}_{o \sim M(x)} U(x, o)$$

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:

[In_1] F. Kamiran, T. Calders and M. Pechenizkiy. *"Discrimination aware decision tree learning"*. In ICDM, pp. 869-874, 2010.

[In_2] T. Calders and S. Verwer. *"Three Naïve Bayes Approaches for Discrimination-Free Classification"*. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[In_3] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. *"Fairness Constraints: A Mechanism for Fair Classification"*. arXiv preprint arXiv:1507.05259, 2015.

[In_4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. *"Fairness through awareness"*. In ITCS 2012, pp. 214-226, 2012.

[In_5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. *"Learning fair representations"*. In ICML, pp. 325-333, 2013.

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)

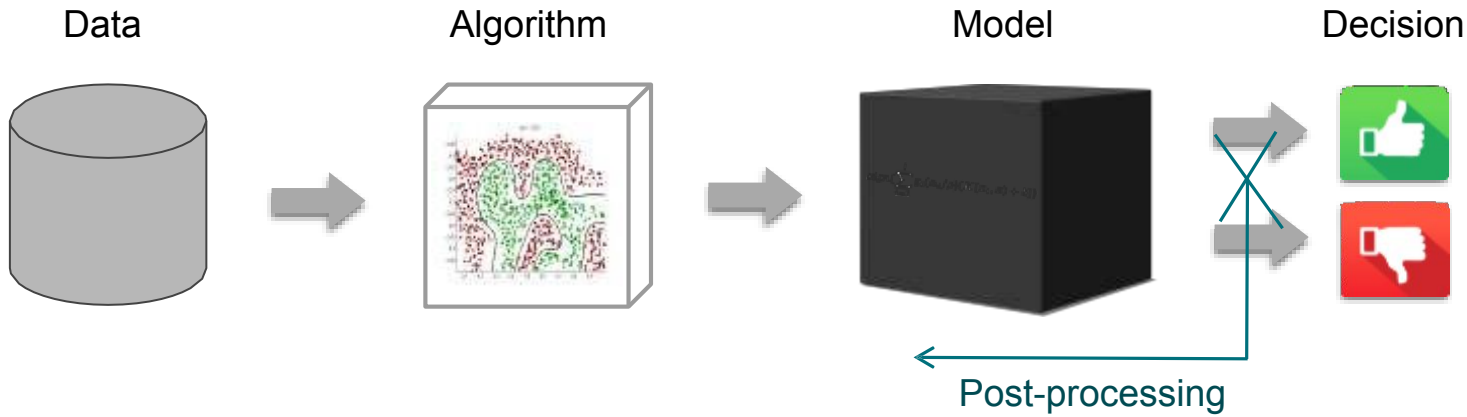
… extends “Fairness through awareness” in several important ways.

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:

[Post_1] F. Kamiran, T. Calders and M. Pechenizkiy. *"Discrimination aware decision tree learning"*. In ICDM, pp. 869-874, 2010. **(already covered)**

[Post_2] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti. *"Discrimination-and privacy-aware patterns"*. In Data Mining and Knowledge Discovery, 29(6), 2015.

[Post_3] F. Kamiran, A. Karim, and X. Zhang. *"Decision Theory for discrimination-aware classification"*. In ICDM, pp. 924-929, 2012. **(not covered)**

Fairness-aware data mining

Post-processing approaches:

[Post_1] F. Kamiran, T. Calders and M. Pechenizkiy. *"Discrimination aware decision tree learning"*. In ICDM, pp. 869-874, 2010. **(already covered)**

[Post_2] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti. *"Discrimination-and privacy-aware patterns"*. In Data Mining and Knowledge Discovery, 29(6), 2015.

[Post_3] F. Kamiran, A. Karim, and X. Zhang. *"Decision Theory for discrimination-aware classification"*. In ICDM, pp. 924-929, 2012. **(not covered)**

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.

Sex	Job	Credit_history	Salary	Credit_approved
Male	Writer	No-taken	... €	Yes
Female	Lawyer	Paid-duly	... €	No
Male	Veterinary	Paid-delay	...€	Yes
...	...	...	...	...

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

Atzori, Bonchi, Giannotti, Pedreschi. *"Anonymity Preserving Pattern Discovery"* VLDB Journal 2008

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

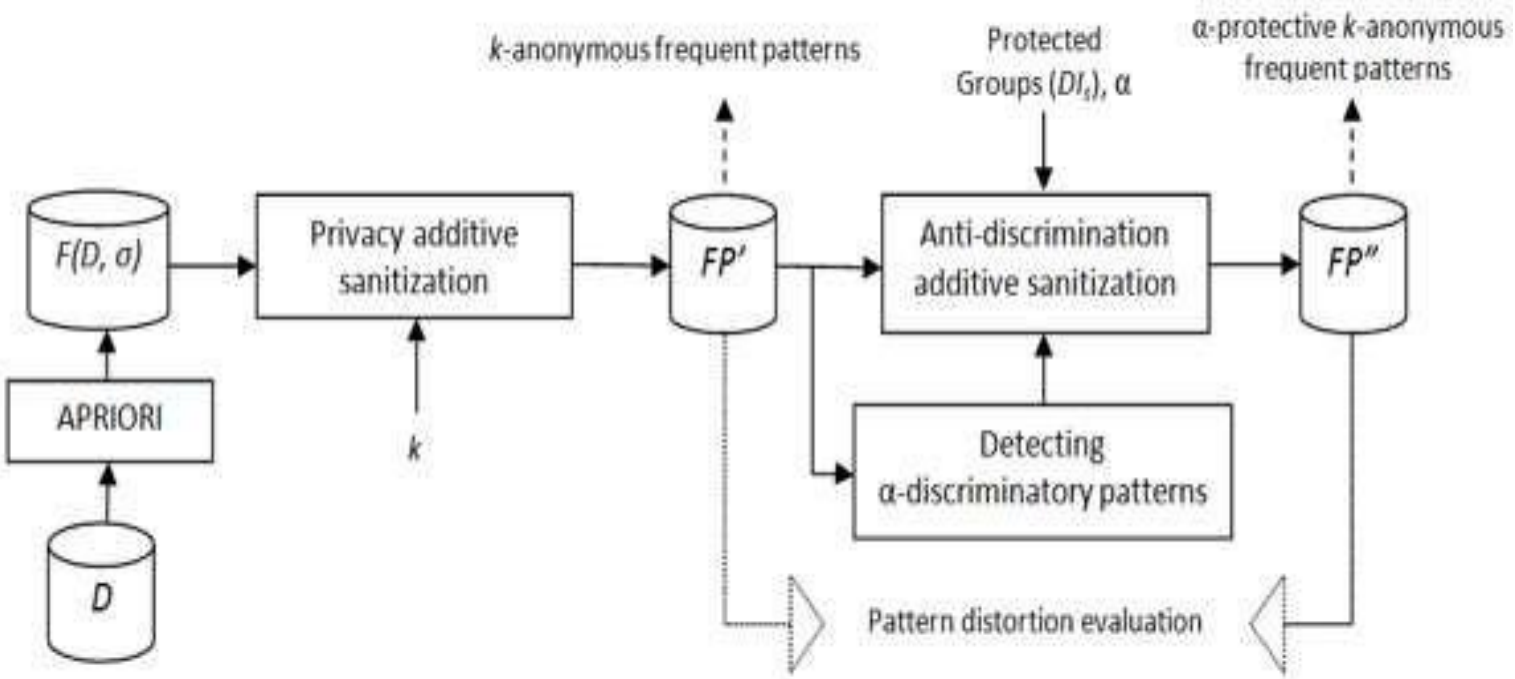
job = veterinarian → credit-approved = no (support 41)

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

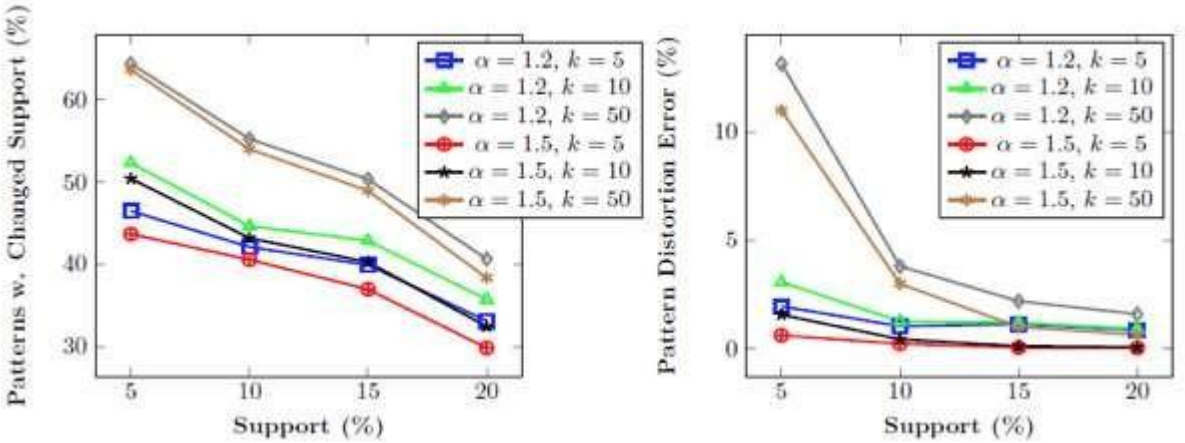
Theorem: Anti-discrimination pattern sanitization for making *F(D,s)* *a*-protective does not generate new discrimination as a result of its transformation.

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 *α*-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.

B. Goodman and S. Flaxman (2016): EU regulations on algorithmic decision-making and a" right to explanation". arXiv preprint arXiv:1606.08813.

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

PPDM	FADM
Measuring disclosure risk	Measuring potential discrimination
Data, algorithm or model transformation to protect privacy	Data, algorithm or model transformation to prevent discrimination
Measuring data/model utility	Measuring data/model utility
Trade-off between privacy and utility	Trade-off between fairness and utility
…	…

A promising direction

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

[pre-processing] S. Ruggieri. *"Using t-closeness anonymity to control for non-discrimination"*. Transactions on Data Privacy, 7(2), pp.99-129, 2014.

[pre-processing] S. Hajian, J. Domingo-Ferrer, and O. Farras. *"Generalization-based privacy preservation and discrimination prevention in data publishing and mining"*. Data Mining and Knowledge Discovery, 28(5-6), pp.1158-1188, 2014.

[in-processing] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. *"Fairness through awareness"*.
In ITCS 2012, pp. 214-226, 2012.

[post-processing] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti. *"Discrimination-and privacy-aware patterns"*. In Data Mining and Knowledge Discovery, 29(6), 2015.

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

I. Žliobaitė (2015): A survey on measuring indirect discrimination in machine learning. arXiv pre-print.

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)

Fariness 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

Slides will be made available at the tutorial webpage:
http://francescobonchi.com/algorithmic_bias_tutorial.html