Lecture 6
Security and Ethical Aspects of Data

Amaya Nogales Gómez
amaya.nogales-gomez@univ-cotedazur.fr

MSc Data Science & Artificial Intelligence
Université Côte d'Azur

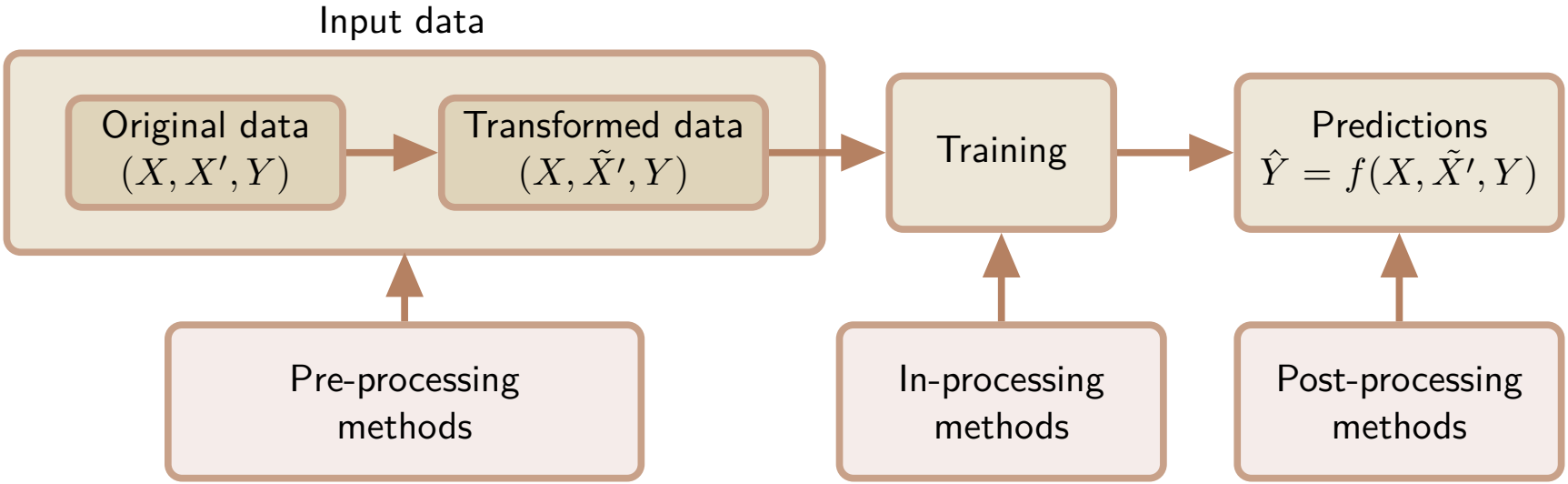
November 29, 2021

Course overview

1. Introduction
   - Preliminaries and Machine Learning basics
   - Motivation for ethics in Artificial Intelligence
2. Sources of unfairness
   - Bias in data
   - Algorithmic unfairness: Examples
3. Fairness criteria
   - Types of discrimination
   - Definitions of fairness
   - Fairness analysis of datasets
4. Algorithms and methods for fair ML
   - Pre-processing methods
   - In-processing methods
   - Post-processing methods
   - Legal and policy perspectives

Methods for Fair Machine Learning

Methods that target fairness in the ML lifecycle fall under three categories:



Pre-processing methods

- Data-based methods try to transform the data so that the underlying discrimination is removed.
- If the algorithm is allowed to modify the training data, then pre-processing can be used.
- Example: reweighting, e.g., up-weighting examples that align with a particular fairness objective.

F. Kamiran and T. Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1-33.

In-processing methods

- Model-based methods try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process.
- If it is allowed to change the learning procedure for a machine learning model, then in-processing can be used.
- Either by incorporating changes into the objective function or imposing new constraint.
- Example: addition of regularization terms or constraints that enforce a particular fairness objective during optimization.

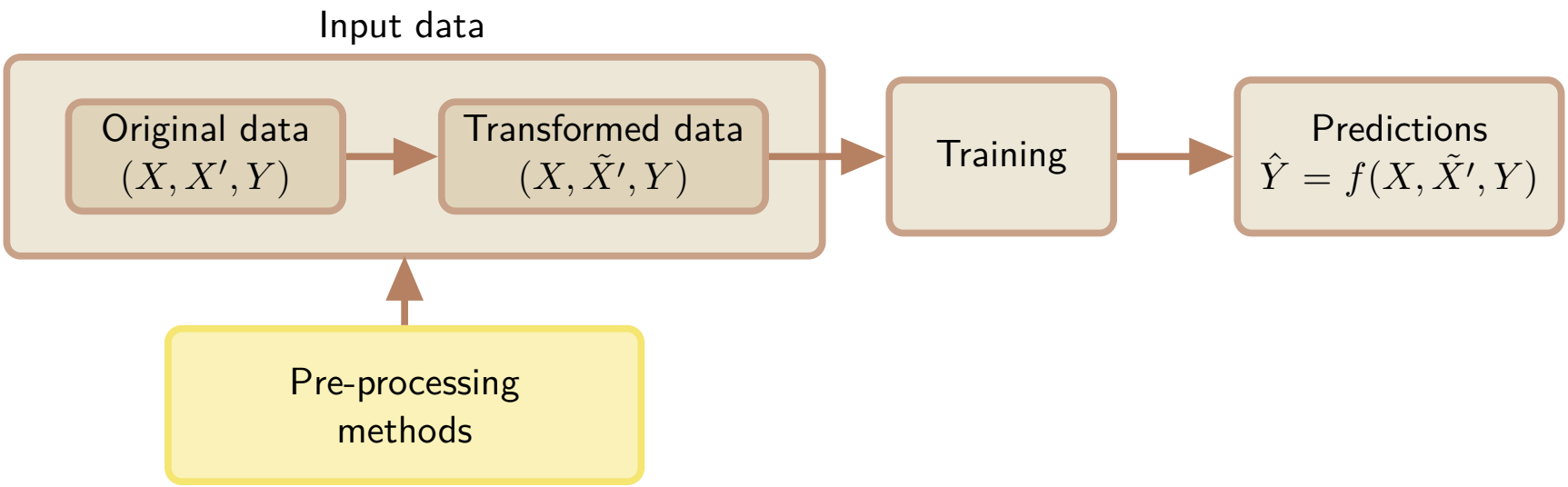
M. B. Zafar, I. Valera, M.-G. Rodriguez, and K.-P.Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:962-970, 2017.

Post-processing methods

- Post-hoc methods are performed after training by accessing a set which was not involved during the training of the model.
- If the algorithm can only treat the learned model as a black box, then only post-processing can be used.
- Labels assigned by the black-box model initially get reassigned based on a function during the post-processing phase.
- Example: identifying different decision thresholds for different groups based on a predicted score, e.g., in order to equalize false positive rates.

M. Hardt, E. Price, and N Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Pre-processing methods



F. Kamiran and T. Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1-33.

Pre-processing methods

We need to define discrimination in a labeled dataset

$$disc(\Omega) = \frac{|\{x_i \in \Omega | a_i = 0, y_i = +1\}|}{|\{x_i \in \Omega | a_i = 0\}|} - \frac{|\{x_i \in \Omega | a_i = 1, y_i = +1\}|}{|\{x_i \in \Omega | a_i = 1\}|}$$

The following methods[1] for incorporating non-discrimination constraints into the classifier construction process are based on preprocessing the dataset after which the normal classification tools can be used to learn a classifier:

- Supression
- Reweighting
- Massaging

[1] F. Kamiran and T. Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1-33.

Supression

To reduce the discrimination between the class labels and the attribute A:

- Remove A from the dataset Ω.
- Find the attributes that correlate most with the sensitive attribute A.
- Remove the most correlated attributes with A.

Reweighting

- The tuples (x_i, a_i, y_i) in the training dataset are assigned weights.
- By carefully choosing the weights, the training dataset can be made discrimination-free w.r.t. A without having to change any of the labels.
- The weights on the tuples can be used directly in any classification method.

Massaging the dataset

To remove the discrimination from the input data:

- The labels of some objects in the dataset are flipped.
  - The labels of some objects x_i with $a_i = 1$ from -1 to $+1$.
  - The same number of objects with $a_i = 0$ from $+1$ to -1.
- A good selection of which labels to change is essential.
- A ranker is used to select the best candidates for relabeling.

Pre-processing method I: Massaging the dataset

- A ranker R for ranking the objects according to their positive class probability is learned.
- Higher scores indicate a higher chance to be in the positive class.

Promotion candidates

$$\{x_i \in \Omega, a_i = 1, y_i = -1\}$$

Demotion candidates

$$\{x_i \in \Omega, a_i = 0, y_i = +1\}$$

- Promotion candidates are sorted in descending order.
- Demotion candidates are sorted in ascending order.
- The first the top-M elements will be chosen: the objects closest to the decision border are selected first to be relabeled.
- This modification of the training data is continued until the discrimination becomes zero.

Pre-processing method I (cont.)

The number M of pairs needed to be modified to make a dataset Ω discrimination-free can be calculated as follows. If we modify M pairs, the resulting discrimination will be:

$$\frac{p_{\bar{a}} - M}{|\Omega_{\bar{a}}|} - \frac{p_a + M}{|\Omega_a|} = disc(\Omega) - M\left(\frac{1}{|\Omega_a|} + \frac{1}{|\Omega_{\bar{a}}|}\right) = disc(\Omega) - \left(M\frac{|\Omega|}{|\Omega_a||\Omega_{\bar{a}}|}\right)$$

And to reach zero discrimination, we hence have to make:

$$M = \frac{disc(\Omega) \times |\Omega_a| \times |\Omega_{\bar{a}}|}{|\Omega|},$$

with p_a and $p_{\bar{a}}$ the number of positive objects with $a = 1$ and $a = 0$ respectively.

Pre-processing method I (cont.)

Algorithm: Rank

Input: Dataset $\Omega = \{(x_i, a_i, y_i)\}_{i=1}^{n}$

1: Learn a ranker R in Ω
2: $pr := \{x_i \in \Omega | a_i = 1, y_i = -1\}$
3: $dem := \{x_i \in \Omega | a_i = 0, y_i = +1\}$
4: Order pr descending w.r.t. the scores by R
5: Order dem ascending w.r.t. the scores by R

Output: (pr, dem), ordered promotion and demotion list

Pre-processing method I (cont.)

Algorithm: Learn Classifier on Massaged data

Input: Dataset $\Omega = \{(x_i, a_i, y_i)\}_{i=1}^{n}$

1: $(pr, dem) := Rank(X, A, Y)$
2: $M = \frac{disc(\Omega) \times |\{x_i \in \Omega | a_i = 1\}| \times |\{x_i \in \Omega | a_i = 0\}|}{|\Omega|}$
3: Select the top-M objects of pr
4: Change the class label of the M objects to $+1$
5: Select the top-M objects of dem
6: Change the class label of the M objects to -1

Output: Classifier learned on massaged dataset $\tilde{\Omega}$

Pre-processing method II

Fair Cluster Support Vector Machines (FCLSVM)

- A methodology to build an SVM-type classifier with categories clustered around their peers
- Clustering the K_j categories of categorical feature j into L_j clusters, $\forall j$.

Pre-processing method II

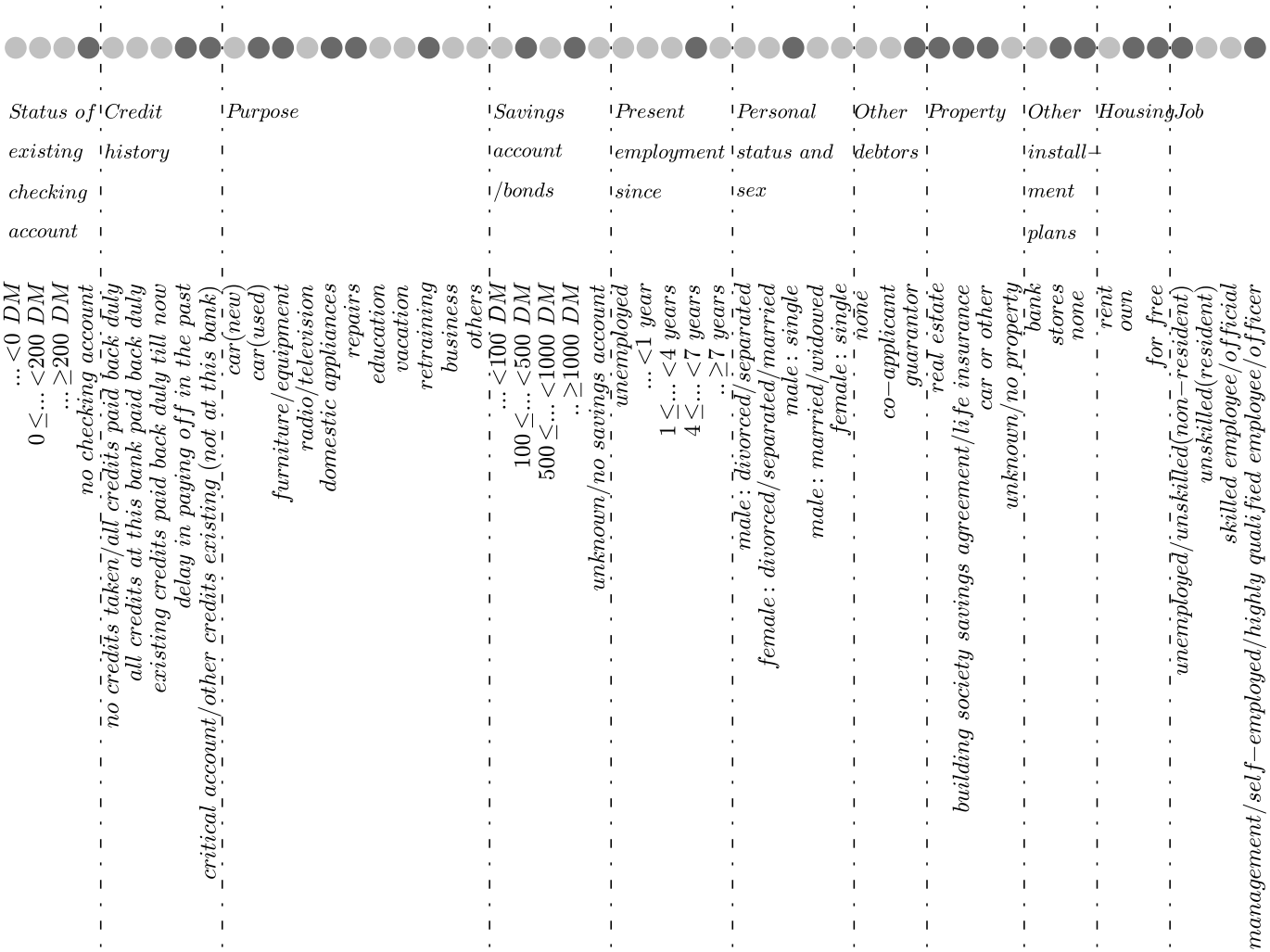
Fair Cluster Support Vector Machines (FCLSVM)

- A methodology to build an SVM-type classifier with categories clustered around their peers
- Clustering the K_j categories of categorical feature j into L_j clusters, $\forall j$.

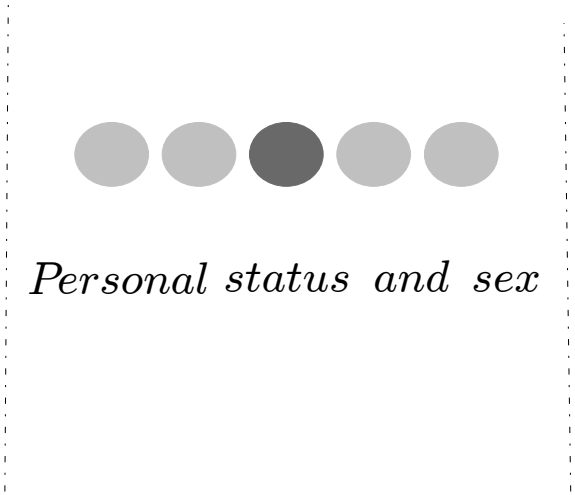
With the pursue of:

- Reducing the number of relevant features, i.e., reducing the complexity of the SVM classifier
- Improving fairness (equal opportunity).
- Without important loss in accuracy

The german dataset, $L_j = 2$



Pre-processing method II (cont.)



- Categorical feature Personal status and sex has $K_j = 5$ categories
- The $K_j = 5$ categories have been clustered into $L_j = 2$ clusters
  - 1st cluster associated with Personal status and sex in light grey
  - 2nd cluster associated with Personal status and sex in dark grey

Pre-processing method II (cont.)

Discrimination metric

Let us assume now that we have only one categorical feature a protected, having K_a categories, $\mathcal{A} = \{a\} \subset \{0,1\}^{K_a}$, and with only one protected category a_{k^*}. Let us define the following general bias metric, based on the equal opportunity definition:

$$\delta_{a,k} = \frac{P\left(\widehat{Y}=+1|a_k=1,Y=+1\right)}{P\left(\widehat{Y}=+1|a_{k^*}=1,Y=+1\right)}$$

$\forall k = 1,\ldots,K_a, k \neq k^*$.

Since by definition k^* is the discriminated category, this measure takes a value $\delta \geq 1$.

Pre-processing method II (cont.)

And for any *non-sensitive* x_j:

$$\delta_{j,k} = \frac{P\left(\widehat{Y}=+1|a_{k^*}=0,x_{j,k}=1,Y=+1\right)}{P\left(\widehat{Y}=+1|a_{k^*}=1,x_{j,k}=1,Y=+1\right)}$$

$\forall j = 1,\ldots,J, \forall k = 1,\ldots,K_j, j \neq a.$

Pre-processing method II (cont.)

FCLSVM algorithm

For each β, C,

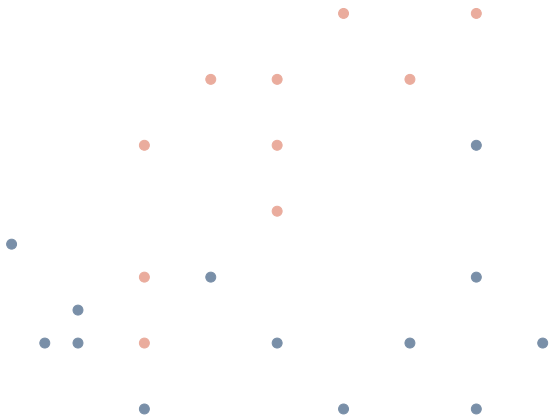
1: Solve the *SVM*, obtain solution ω.
2: Compute vector δ.
3: For each j, cluster the K_j categories of feature j into L_j clusters solving a K-means for $\nu = \beta\omega_{j\cdot} + (1-\beta)\delta_{j\cdot}$, obtaining the assignment vector $z^*_{j\cdot\cdot}$.
4: Obtain the clustered dataset $\tilde{\Omega}$

$$(y_i, x_i, x'_i) \rightarrow (y_i, \tilde{x}_i, x'_i)$$

where $\tilde{x}_i = (\tilde{x}_{i,j,\ell})$ and $\tilde{x}_{i,j,\ell} = \sum_{k=1}^{K_j} z^*_{j,k,\ell} x_{i,j,k}$

5: Solve the *SVM* for $\tilde{\Omega}$, and return this as the FCLSVM classifier.

An example: Large-scale fair classification

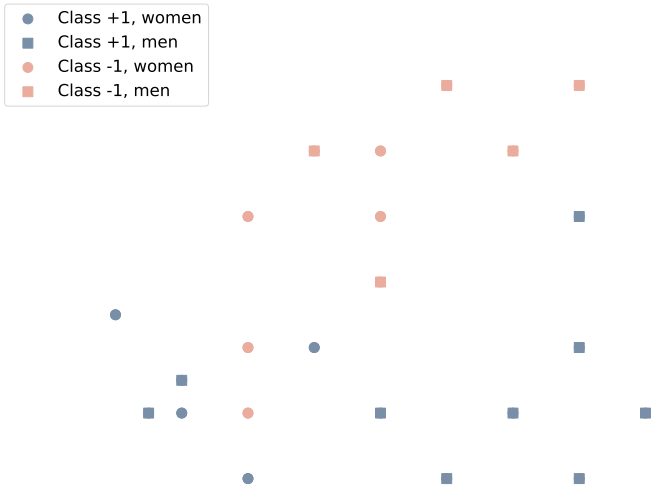


Each individual is represented by

$$(x, x', y) \begin{cases} x, \text{ vector of categorical features} \\ x', \text{ vector of continuous features} \\ y \in \{-1, +1\}, \text{ class membership} \end{cases}$$

$$x = (x^1, \ldots, x^d) = (\underbrace{x^1, \ldots, x^k}_{\text{non-sensitive}}, \underbrace{x^{k+1}, \ldots, x^d}_{\text{sensitive}})$$

An example: Large-scale fair classification



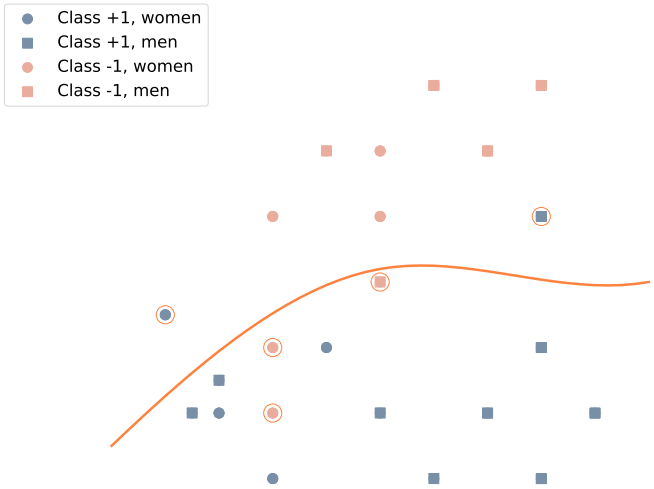
Each individual is represented by

$$(x, x', y) \begin{cases} x, \text{ vector of categorical features} \\ x', \text{ vector of continuous features} \\ y \in \{-1, +1\}, \text{ class membership} \end{cases}$$

$$x = (x^1, \ldots, x^d) = (\underbrace{x^1, \ldots, x^k}_{\text{non-sensitive}}, \underbrace{x^{k+1}, \ldots, x^d}_{\text{sensitive}})$$

	Sensitive feature	non-sensitive features		
	Gender & status	income	credit history	decision
Applicant 1	male married	1.5k	5	$\surd$
Applicant 2	female single	2.5k	3	$\surd$

An example: Large-scale fair classification



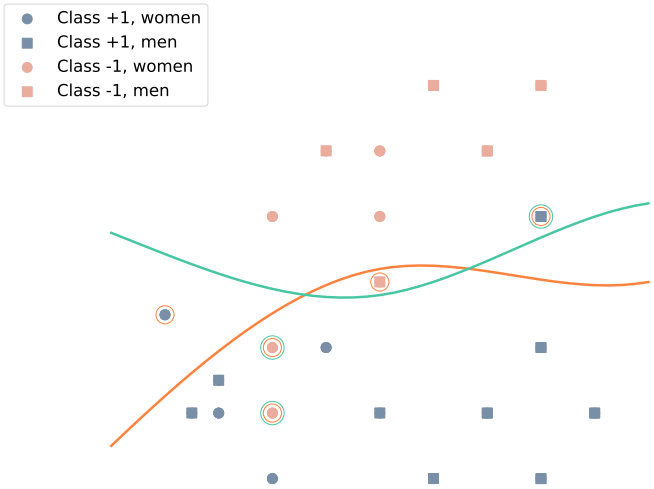
Each individual is represented by

$$(x, x', y) \begin{cases} x, \text{ vector of categorical features} \\ x', \text{ vector of continuous features} \\ y \in \{-1, +1\}, \text{ class membership} \end{cases}$$

$$x = (x^1, \ldots, x^d) = (\underbrace{x^1, \ldots, x^k}_{\text{non-sensitive}}, \underbrace{x^{k+1}, \ldots, x^d}_{\text{sensitive}})$$

Clustering	Accuracy	Equal Opportunity
male single, male married, male divorced female single, female married, female divorced	78.3%	84.4%

An example: Large-scale fair classification



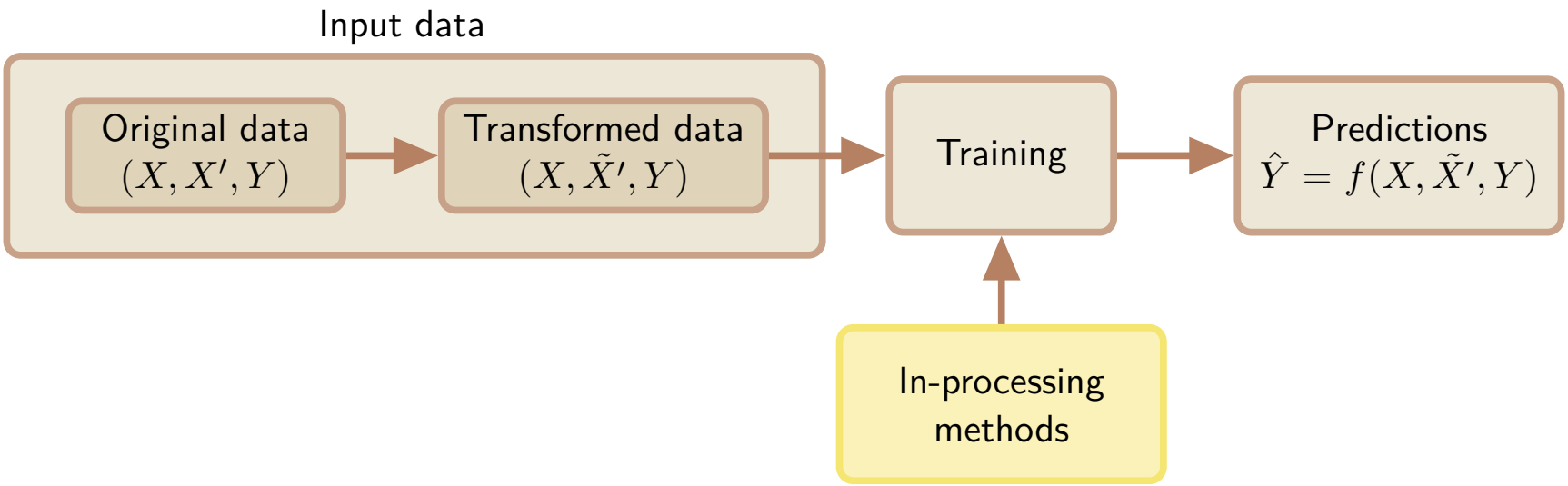
Each individual is represented by

$$(x, x', y) \begin{cases} x, \text{ vector of categorical features} \\ x', \text{ vector of continuous features} \\ y \in \{-1, +1\}, \text{ class membership} \end{cases}$$

$$x = (x^1, \ldots, x^d) = (\underbrace{x^1, \ldots, x^k}_{\text{non-sensitive}}, \underbrace{x^{k+1}, \ldots, x^d}_{\text{sensitive}})$$

Clustering	Accuracy	Equal Opportunity
male single, male married, male divorced female single, female married, female divorced	78.3%	84.4%
male married female single, female married, female divorced, male single, male divorced	87.0%	112.5%

In-processing methods



M. B. Zafar, I. Valera, M.-G. Rodriguez, and K.-P.Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:962-970, 2017.

In-processing method

Maximizing accuracy under fairness constraints

- **Goal**: design classifiers (logistic regression and SVM) that avoid both disparate treatment and disparate impact.
- Measure of decision boundary (un)fairness: the covariance between the sensitive attributes and the (signed) distance between the objects and the decision boundary.

M. B. Zafar, I. Valera, M.-G. Rodriguez, and K.-P.Gummadi. Fairness Constraints: Mechanisms for Fair Classification.

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54:962-970, 2017.

- Dataset $\{(x_i, a_i, y_i)\}_{i=1}^n$
- $d_{\omega,b}(x)$: signed distance from x to $\omega^\top x + b = 0$
- $f_{\omega,b}(x_i) = 1$ if $d_{\omega,b}(x_i) \geq 0$
- $f_{\omega,b}(x_i) = -1$ if $d_{\omega,b}(x_i) \leq 0$

$$\begin{aligned} Cov(a, d_{\omega,b}(x)) &= \mathbb{E}[(a - \bar{a})d_{\omega,b}(x)] - \mathbb{E}[a - \bar{a}]\bar{d}_{\omega,b}(x) \\ &= \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})d_{\omega,b}(x_i), \end{aligned}$$

and since $\mathbb{E}[a - \bar{a}] = 0$, the term $\mathbb{E}[(a - \bar{a})]\bar{d}_{\omega,b}(x)$ cancels out.

M. B. Zafar, I. Valera, M.-G. Rodriguez, and K.-P.Gummadi. Fairness Constraints: Mechanisms for Fair Classification.

In-processing method (cont.)

- Dataset $\{(x_i, a_i, y_i)\}_{i=1}^n$
- $d_{\omega,b}(x)$: signed distance from x to $\omega^\top x + b = 0$.

$$Cov(a, d_{\omega,b}(x)) = \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})d_{\omega,b}(x_i)$$

$$\min_{\omega,\, b} \mathcal{L}(\omega, b)$$

s.t.

$$\frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})d_{\omega,b}(x_i) \leq c \qquad \forall i = 1, \ldots, n$$

$$\frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})d_{\omega,b}(x_i) \geq -c \qquad \forall i = 1, \ldots, n$$

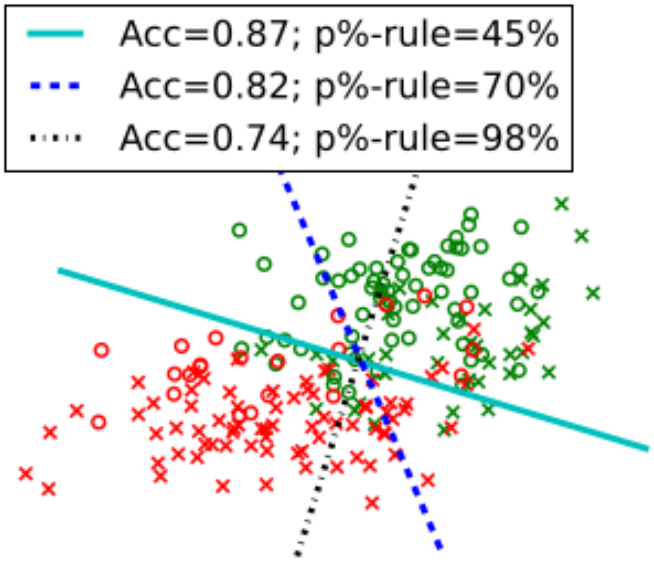
$$\omega \in \mathbb{R}^d$$

$$b \in \mathbb{R}.$$

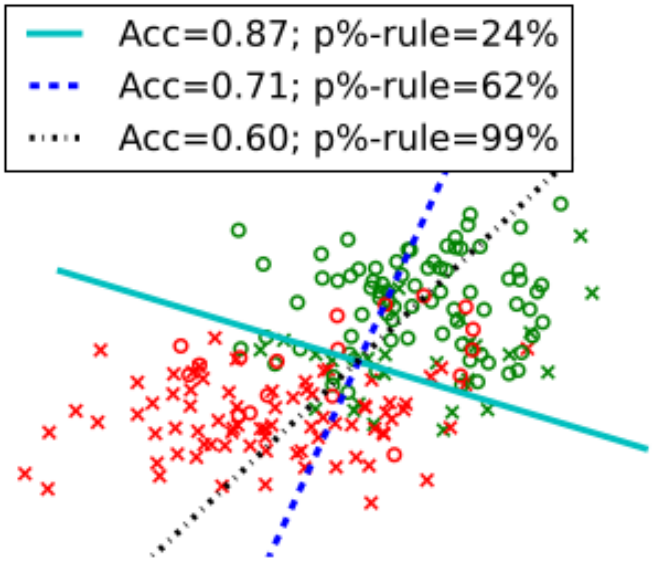
M. B. Zafar, I. Valera, M.-G. Rodriguez, and K.-P.Gummadi. Fairness Constraints: Mechanisms for Fair Classification.

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54:962-970, 2017.

In-processing method (cont.)



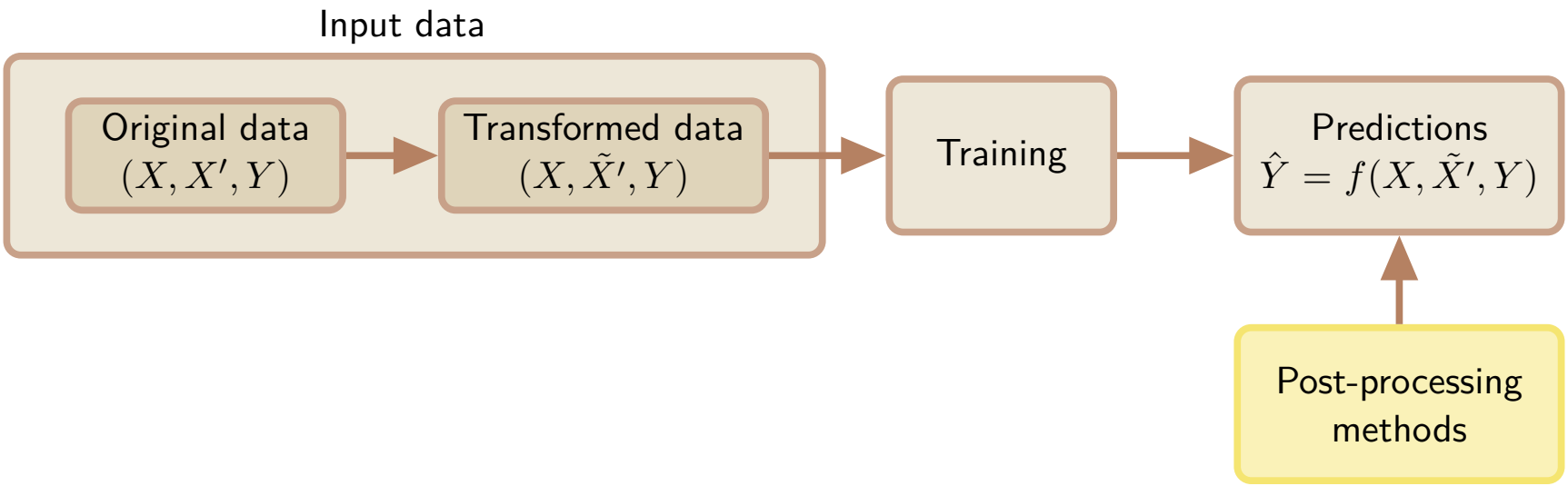
$\phi = \pi/4$



$\phi = \pi/8$

- The solid lines show the decision boundaries without fairness constraints.
- The dashed lines show the decision boundaries trained to maximize accuracy under fairness constraints.
- Circles represent the sensitive feature and each figure corresponds to a dataset, with different correlation value between sensitive attribute values and class labels.

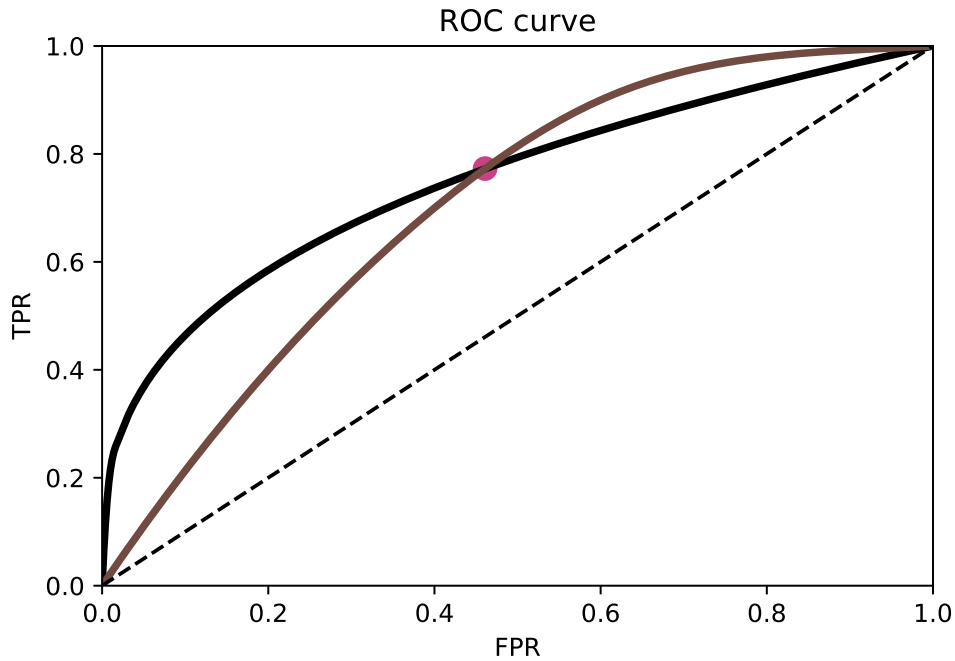
Post-processing methods



M. Hardt, E. Price, and N Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Post-processing method

- We can achieve separation by post-processing a given score function without the need for retraining: using the ROC curve.
- A binary classifier that satisfies separation must achieve the same true positive rates and the same false positive rates in all groups.

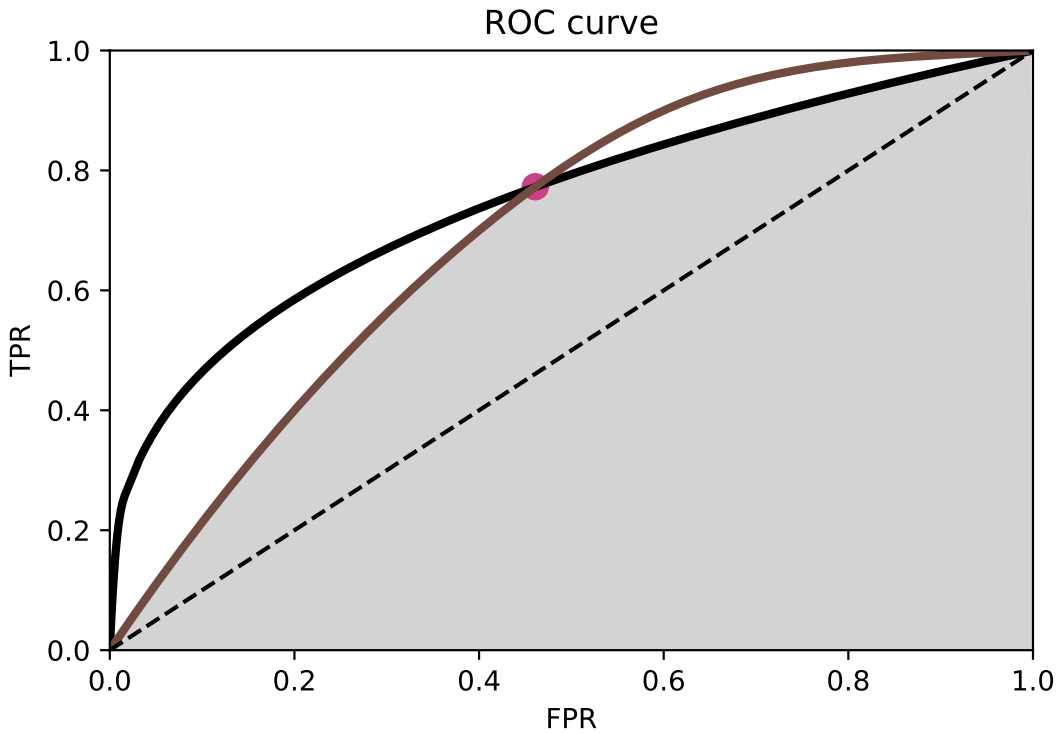


In black: ROC curve for non-protected group.

In brown: ROC curve for protected group.

M. Hardt, E. Price, and N Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Post-processing method (cont.)



- This condition corresponds to taking the intersection of all group-level ROC curves.
- Within this constraint region, we can then choose the classifier that minimizes the given cost.

Post-processing method (cont.)

- A score function obeys equalized odds if and only if the ROC curves for the protected and non-protected groups coincide for all decision threshold r. That is:

$$P(r(X, A = 1) > r|Y = y, A = 1) = P(r(X, A = 0) > r|Y = y, A = 0)$$

- Let us denote the two ROC curves for $A = 1$ and $A = 0$ as:

$$f_1(r) = (TPR_1(r), FPR_1(r))$$
$$f_0(r) = (TPR_0(r), FPR_0(r))$$

- The intersection between the two curves: $f_1(r_1) = f_0(r_2)$ for some r_1, r_2.
- Then if we choose different thresholds for the protected and non-protected groups, we can achieve equalized odds:

$$P(r(X, A = 1) > r_1|Y = y, A = 1) = P(r(X, A = 0) > r_2|Y = y, A = 0)$$

Legal and policy perspectives

Within the scope of application of the Treaty establishing the European Community and of the Treaty on European Union, and without prejudice to the special provisions of those treaties, any discrimination on grounds of nationality shall be prohibited.

Legally recognized protected features: Europe

Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.

Legal aspects: Europe

This right is enshrined in article 21 of the Charter of Fundamental Rights. There are different directives:

1. Against discrimination on grounds of race and ethnic origin.
2. Against discrimination at work on grounds of religion or belief, disability, age or sexual orientation.
3. Towards equal treatment for men and women in matters of employment and occupation.
4. Towards equal treatment for men and women in the access to and supply of goods and services.
5. Against discrimination based on age, disability, sexual orientation and religion or belief beyond the workplace.

https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/know-your-rights/equality/non-discrimination_en

Legal aspects: US

Legally recognized protected features

Race, color, sex, religion, national origin, citizenship, age, pregnancy, disability status, genetic information, veteran status, familial status.

Regulated domains

- Credit (Equal Credit Opportunity Act)
- Education (Civil Rights Act of 1984)
- Employment(Civil Rights Act of 1984)
- Housing (Fair Housing Act)

Example I: The US Equal Pay Act

The US Equal Pay Act

- Requires that men and women in the same workplace be given equal pay for equal work.
- The jobs need not be identical, but they must be substantially equal.
- This law covers all forms of pay including salary, overtime pay, bonuses, stock options, profit sharing and bonus plans, life insurance, etc.
- This act aimed at abolishing wage disparity based on sex.
- According to the US Bureau of Labor Statistics, women's salaries compared to men's have risen dramatically since the enactment of this equal pay act, from 62% in 1970 to 80% in 2004.
- This real-world case illustrates a scenario where our historical data are discriminatory due to a biased data generation process, but where classifiers learned on the data are forced to be discrimination-free by law.

Example II: The Australian Sex Discrimination Act 1984

The Australian Sex Discrimination Act 1984

- It prohibits discrimination in work, education, services, accommodation, land, pregnancy or potential pregnancy, and family responsibilities.
- This act defines sexual harassment and other discriminatory practices on different grounds and declares them unlawful.
- This law also prohibits indirect and unintentional discrimination.
- It is the responsibility of the accused party to prove that his/her intention was not to discriminate the aggrieved party: the burden of proving that an act does not constitute discrimination lies on the person who did the act.
- Notice that under this law it is insufficient to remove the sex attribute from a dataset before learning; also indirect discrimination on the basis of a "characteristic that appertains generally to persons of the sex of the aggrieved person" is disallowed.

Example III: The US Equal Credit Opportunity Act 1974

The US Equal Credit Opportunity Act 1974

Declares unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction, on the basis of race, color, religion, national origin, sex or marital status, or age.

Example IV: European Council Directive 2004

European Council Directive 2004

- Even though there is clear historical evidence showing higher accident rates for male drivers in traffic, insurance companies are no longer allowed to discriminate based on gender in many countries.
- The European Court of Justice decided (in 2011) that it will no longer be legal under EU law to charge women less for insurance than men.
- The verdict means that different priced premiums for men and women drivers will now be considered to be in breach of the EU's anti-discrimination rules.

Final remarks

- All of the anti-discriminatory laws prohibit discriminatory practices in future (or present).
- If we are interested in applying ML algorithms, and our available historical data is biased, it is simply **illegal** to use traditional algorithms without taking the fairness aspect into account.
- Because of the above mentioned laws and due to ethical concerns, restricting ourselves to the single use of traditional ML techniques is unacceptable.

Bibliography

- M. Hardt, E. Price, and N Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.
- M. B. Zafar, I. Valera, M.-G. Rodriguez, and K.-P.Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:962-970, 2017.
- F. Kamiran and T. Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1-33.