Master MVA

Responsible Machine Learning

Nicolas Vayatis

Lecture#3 - On Fairness

Why Fairness in AI?

- "It is the right thing to do"
- Fulfill people expectations (e.g., customers, employees, etc)
- Regulators require it, potential legal risk
- Some popular examples: predictive justice (COMPAS), credit scoring, housing, hiring, advertising...

Fairness in AI: Some Challenges

- How to conceive fair discrimination in machine learning?
- How to measure fairness?
- How to include fairness consideration in training a model?
- What is the trade off with accuracy?
- Can we ensure our fairness measure aligns with law?

Reminder on supervised machine learning

(Plain) binary classification

- Classifcation model : random pair (X, Y) over $\mathbb{R}^d \times \{-1, +1\}$
- Posterior probability: $\eta(x) = \mathbb{P}\{Y \mid X = x\}$, $\forall x \in \mathbb{R}^d$
- Classifier: $g : \mathbb{R}^d \to \{-1, +1\}$
- Classification error: $L(g) = \mathbb{P}\{g(X) \neq Y\}$

Variants of classification error

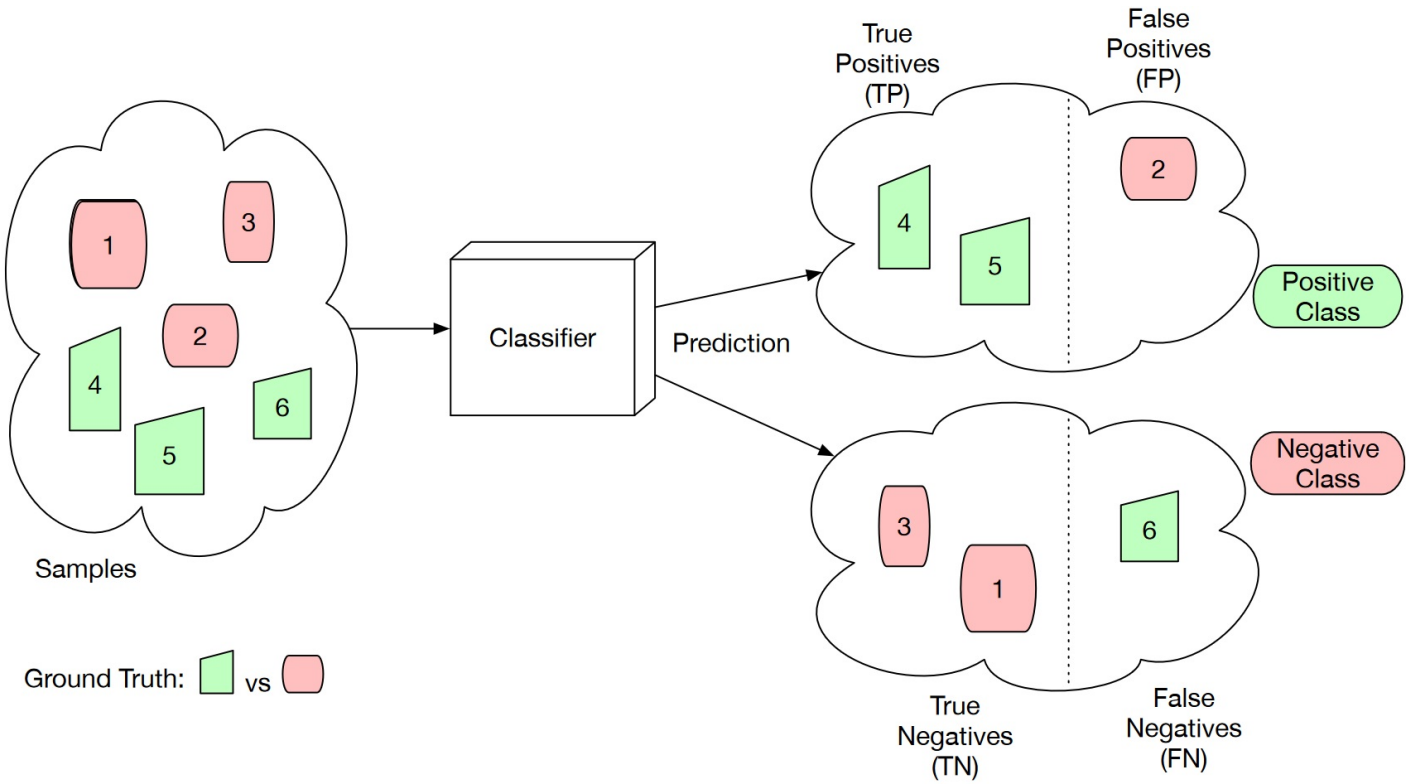
- Asymmetric cost - set $\omega \in (0,1)$,

$$L_\omega(g) = 2\mathbb{E}\big((1-\omega)\mathbb{I}\{Y=+1\}\mathbb{I}\{g(X)=-1\} + \omega\mathbb{I}\{Y=-1\}\mathbb{I}\{g(X)=+1\}\big)$$

- Classification with mass constraint - set $u \in (0,1)$

$$\min_g \mathbb{P}(Y \neq g(X)) \quad \text{subject to} \quad \mathbb{P}(g(X)=1) = u$$

Overview of supervised classification



The two types of error for general scoring-based detection

- Consider $f : \mathbb{R}^d \to \mathbb{R}$ a detector response (scoring rule or soft classifier) and a threshold $t \in \mathbb{R}$
- An alarm corresponds to the event $\{g(x) = 1\} = \{f(x) \geq t\}$
- True positive rate and false positive rate:

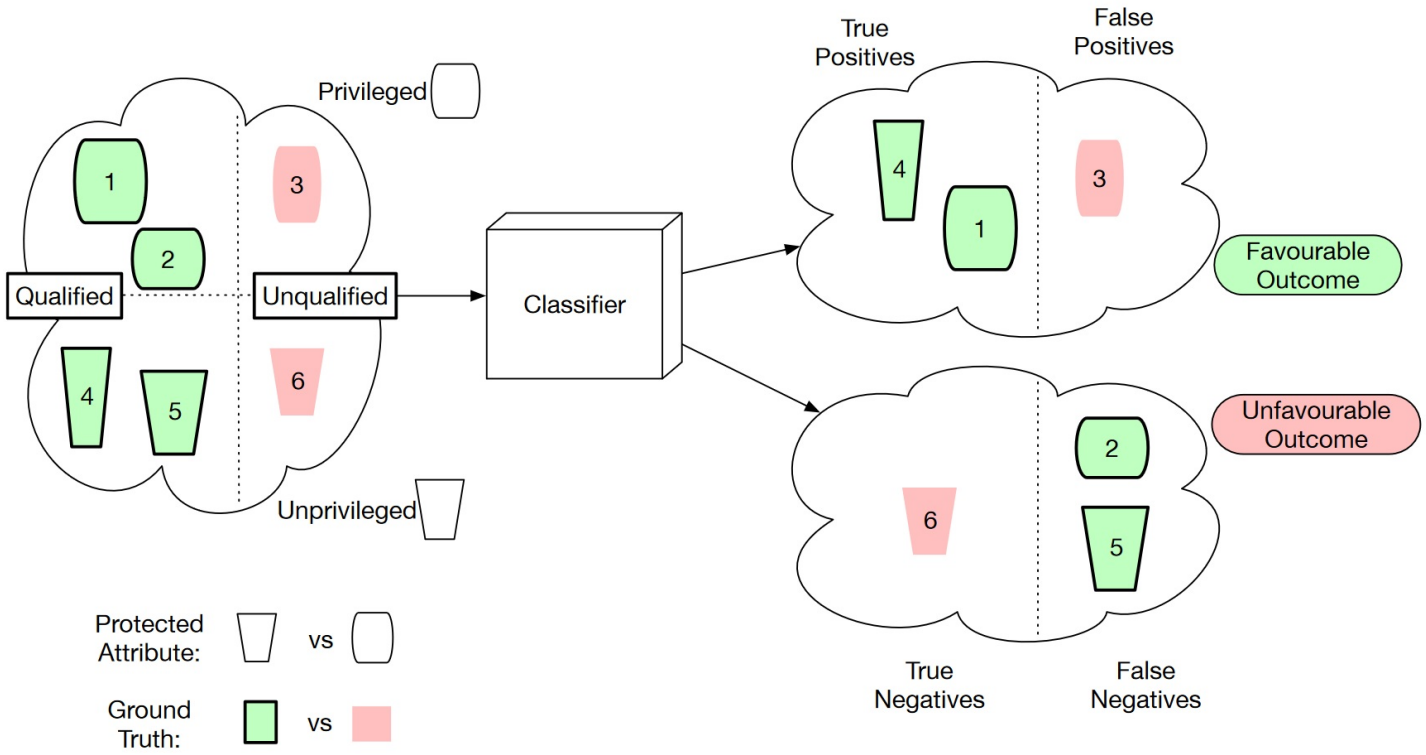
$$\begin{aligned} \beta(f,t) = &\quad \mathbb{P}\{f(X) \geq t \mid Y = +1\} \quad \text{(TPR)} \to \max \\ \alpha(f,t) = &\quad \mathbb{P}\{f(X) \geq t \mid Y = -1\} \quad \text{(FPR)} \to \min \end{aligned}$$

- Main point: trade-off required since, for given f,

$$\begin{aligned} \beta(f,t) \to 1 \quad &\text{but} \quad \alpha(f,t) \to 1 \quad \text{when } t \to -\infty \\ \alpha(f,t) \to 0 \quad &\text{but} \quad \beta(f,t) \to 0 \quad \text{when } t \to +\infty \end{aligned}$$

A. Fairness Metrics

Overview of fairness concepts in classification



Main legal concepts related to unfairness

- Disparate treatment:

  Response: ensure equality of opportunity for privileged and unprivileged

- Disparate outcome:

  Response: minimize inequality of outcome (positive and/or negative) for privileged and unprivileged

- Individual vs. group fairness:

  Requirement for individual fairness: individuals with similar characteristics should benefit from the same outcome.

Some examples of simple fairness criteria

- Demographic Parity — "There should be an equal *rate* of positive outcomes in the privileged group and in the unprivileged group"
- Equality of Opportunity — "There should be an equal *true positive rate* in the privileged group and in the unprivileged group"
- Equalized Odds — "There should be an equal *true and false positive rate* in the privileged group and in the unprivileged group"

Setup and notations

- Y target variable
- X features of an individual (browsing history etc.)
- S sensitive attribute (e.g. gender, race, age, etc.)
- $F = f(X, S)$ soft predictor of Y (here, score to show ad or not)

A few notes:

- Random variables are assumed to live in the same probability space (in particular we assume that f is randomized)
- We may call 'Group A' the subset such that $S = a$...
- These notations may themselves convey some bias...

The three fundamental criteria

1. **Independence**: $F \perp\!\!\!\perp S$
2. **Separation**: $F \perp\!\!\!\perp S$ conditional on Y
3. **Sufficiency**: $Y \perp\!\!\!\perp S$ conditional on F

All other criteria are special cases, either equivalent or relaxed versions, of these three.

Fairness criteria (1) - Independence

For any groups a, b and any prediction o then

$$\mathbb{P}\{F = o \mid S = a\} = \mathbb{P}\{F = o \mid S = b\}$$

Variants: demographic parity, statistical parity.

Approximate version for binary predictors (example)

$$|\mathbb{P}\{F = o \mid S = a\} - \mathbb{P}\{F = o \mid S = b\}| \leq \varepsilon \ ,$$

Some Issues:

- May discard perfect predictor $F = Y$ (since Y may be correlated with S)
- One may trade false negatives for false positives
- One may reduce errors only in one group and not in the others

Fairness criteria (2) - Separation

For any groups a, b and any prediction o, any outcome y then

$$\mathbb{P}\{F = o \mid Y = y, S = a\} = \mathbb{P}\{F = o \mid Y = y, S = b\}$$

Variants: equalized odds, equality of opportunity

Some virtues:

- Compatible with perfect predictions $F = Y$
- Ensures error reduction uniformly in all groups

Fairness criteria (3) - Sufficiency

Intuition: no need to access S to predict Y when F is available, in other words, the score F has some semantic meaning:

Note that Bayes rule in least squares regression is:

$$f(x, s) = \mathbb{E}\{Y \mid X = x, S = s\}$$

Method: Sufficiency implied by calibration by group, e.g. for binary classification with soft classifier prediction $o \in [0, 1]$:
$P\{Y = 1 \mid F = o, S = s\} = o$

Calibration by group can be achieved by various standard calibration methods (if necessary, applied for each group).

The trade-offs of fairness

Main theorem; Any two of the previous criteria are mutually exclusive!

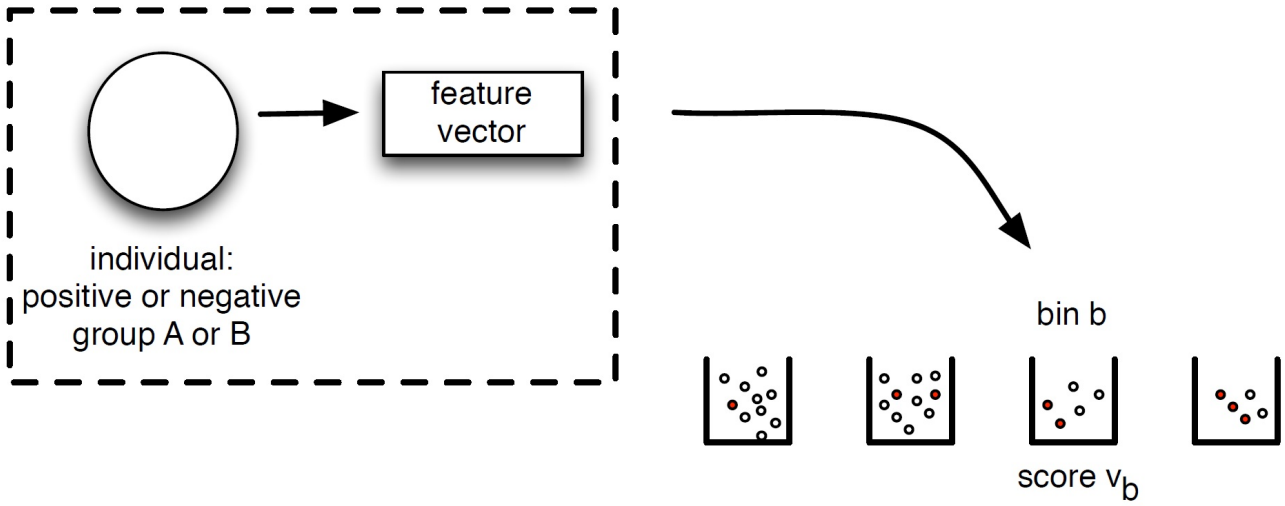
Some straightforward propositions:

- If S and Y are not independent then either independence holds or sufficiency, not both.
- If S and Y are not independent and F and Y are not independent then either independence holds or separation, not both.
- Assume all events in the joint distribution (S, F, Y) have positive probability. If S and Y are not independent then either separation or sufficiency, not both.

The latter has been shown by Choudelchova (2016) and Kleinberg et al. (2016). Many works have studied the case of relaxed and approximate versions of the theoretical criteria.

A model for risk scores

From [Kleinberg et al., 2016]

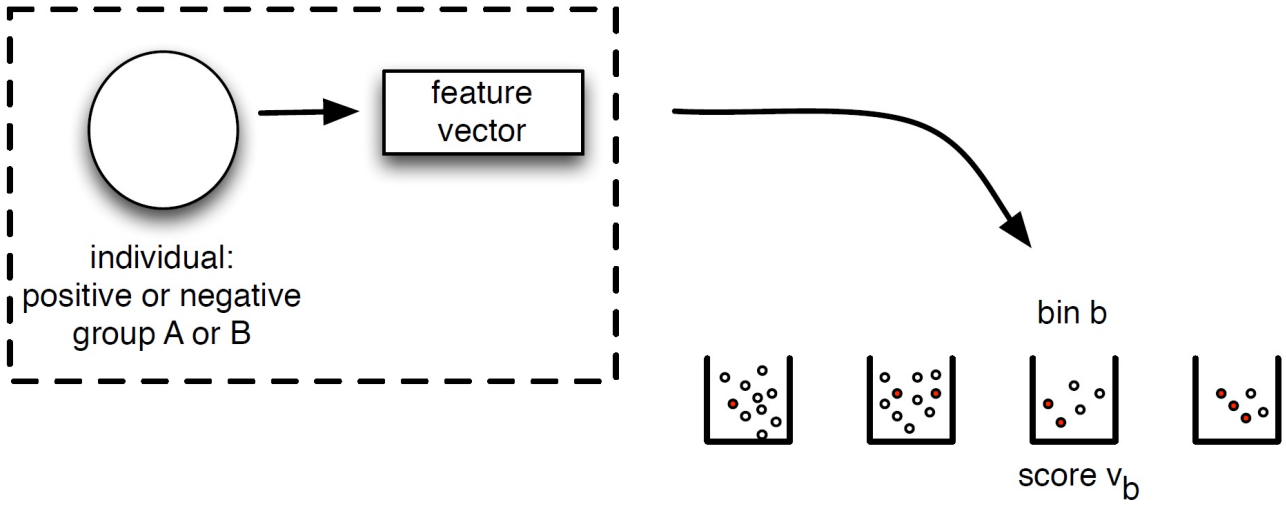


Basic model for assigning scores as probability estimates.

- Individuals are either positive or negative (exhibit the behavior or not).
- Each individual belongs to group A or B.
- Each individual has a set of features, with the data we have access to.
- A risk score is a function mapping individuals to discrete "bins," where everyone in bin b is assigned a score of v_b.

Fairness criteria

From [Kleinberg et al., 2016]



Desired properties:

- Calibration within groups: For each group, a v_b fraction of people in bin b are positive.
- Balance for the positive class: Average score of positive members in group A equals average score of positive members in group B.
- Balance for the negative class: Average score of negative members in group A equals average score of negative members in group B.

Impossibility theorem

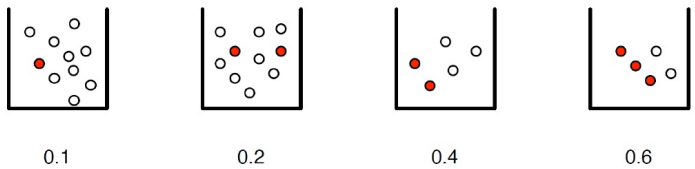
Can achieve all three properties in two simple cases:

- Perfect prediction: for each feature set, either everyone is in the negative class or everyone is in the positive class. (Then we can assign scores of 0 or 1 to everyone.)
- Equal base rates: the groups have the same fraction of positive instances. (Then there's a trivial risk score equal to this base rate for everyone.)

Theorem [Kleinberg-Mullainathan-Raghavan 2016]. In any instance of risk score assignment where all three properties can be achieved, we must have either perfect prediction or equal base rates.

Impossibility theorem - sketch of proof

From [Kleinberg et al., 2016]



Let N_t be the number of people in group t.
Let k_t be the number of people in the positive class in group t.

The calibration condition implies:

- The total score of all group-t people in bin b equals the expected number of group-t people in the positive class in bin b.

Summing over all bins:

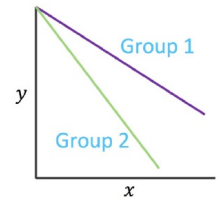
- The total score of all group-t people equals the expected number of group-t people in the positive class.

Let N_t be the number of people in group t.
Let k_t be the number of people in the positive class in group t.
(By calibration, k_t is also the total score in group t.)

Let x be the average score of a person in the negative class.
Let y be the average score of a person in the positive class.
(Note: independent of which group t we're talking about.)

Total score in group t is
$(N_t - k_t)x + k_t y = k_t$.

Rearranging:
$x = (1 - y)\dfrac{k_t}{N_t - k_t}$.



Unless slopes are the same, only intersect at (0, 1)

Impossibility theorem - sketch of proof

back to our notations

- Calibration condition: for any group t, for any score bin b,

$$v_b \mathbb{P}(f = v_b \mid S = t) = \mathbb{P}(Y = 1, f = v_b \mid S = t)$$

- Summing over all bins: for any group t,

$$\sum_b v_b \mathbb{P}(f = v_b \mid S = t) = \mathbb{P}(Y = 1 \mid S = t)$$

then

$$\mathbb{E}(f \mid S = t) = \mathbb{P}(Y = 1 \mid S = t) \overset{\text{def}}{=} p_t$$

but we have:

$$\mathbb{E}(f \mid S = t) = p_t \mathbb{E}(f \mid S = t, Y = 1) + (1 - p_t) \mathbb{E}(f \mid S = t, Y = 0)$$

An overview of fairness criteria

From [Barocas-Hardt-Narayanan, 2021]

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

Alignment with Law

Consider two types of fairness in ML approaches:

- Bias Preserving
- Bias Transforming/Correcting

Which approaches fall in each category?

Check [Wachter, Mittelstadt, Russell (2021)]
`https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772`

B. Incorporating Fairness in Machine Learning Models

Three levels to ensure fairness

1. Preprocessing: discrimination-free training data
   - Relabeling
   - Resampling
   - Fair representations
2. Postprocessing : correcting biased predictors
   - output correction
   - input correction
   - classifier correction
3. **Fairness-aware ML algorithms**

Starting point: Equality of opportunity

Simplified setup: Outcome Y is binary $\{0, 1\}$ and predictor F is binary-valued.

Equal opportunity means we have the same true positive rate for each group: for any groups a, b then:

$$\mathbb{P}\{F = 1 \mid Y = 1, S = a\} = \mathbb{P}\{F = 1 \mid Y = 1, S = b\}$$

(same true positive rates. Equalized odds extend the concept to both the positive and the negative class.)

Learning while imposing fairness

Theoretical formulation - Minimize classification error with fairness constraints over real-valued f:

$$\min_f L(f)$$

such that:

$$\mathbb{P}\{f(X,S) > 0 \mid Y = 1, S = a\} = \mathbb{P}\{f(X,S) > 0 \mid Y = 1, S = b\}$$

(equality of opportunity constraint)

From [Donini, Oneto, Ben-David, Shawe-Taylor, Pontil (NeurIPS, 2018)]

Relaxed formulation

Fair empirical risk minimization

- Set :

$$\begin{aligned} L^{+,a} &= \mathbb{P}\{f(S,S) > 0 \mid Y = 1, S = a\} \ , \\ L^{+,b} &= \mathbb{P}\{f(X,S) > 0 \mid Y = 1, S = b\} \end{aligned}$$

- Relaxed fairness constraint: for some $\varepsilon \in [0,1]$

$$|L^{+,a} - L^{+,b}| \leq \varepsilon$$

- FERM with empirical counterparts:

$$\min_f \widehat{L}(f)$$

such that:

$$|\widehat{L}^{+,a} - \widehat{L}^{+,b}| \leq \varepsilon$$

Making things work with kernels

- Using the binary loss in the criterion and in the fairness constraint makes the problem non convex.
- Using the hinge loss in the criterion and the linear loss in the fairness constraint makes the problem convex.

Fair kernel methods
Optimization problem

- Introducing a Hilbert space $\mathcal{H}$ such that $x \mapsto \phi(x) \in \mathcal{H}$, we consider that candidate functions f are of the form: $f(x) = \langle w, \phi(X) \rangle$
- The constraint becomes:

$$|\langle w, u \rangle| \leq \varepsilon$$

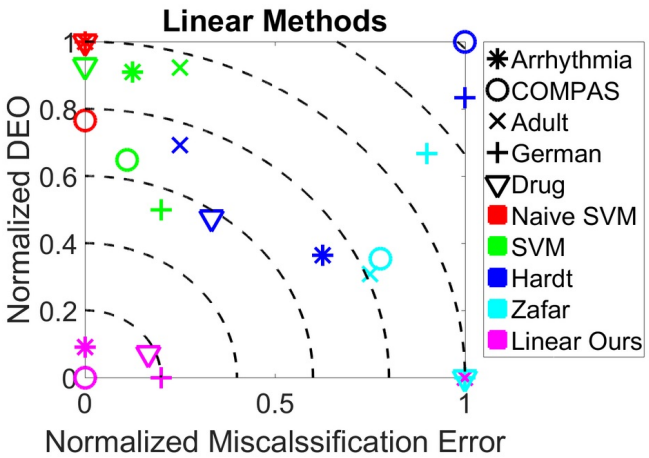
where $u = u_a - u_b$ and $u_a = \frac{1}{n^{+,a}} \sum_{i \in \mathcal{I}^{+,a}} \phi(x_i)$

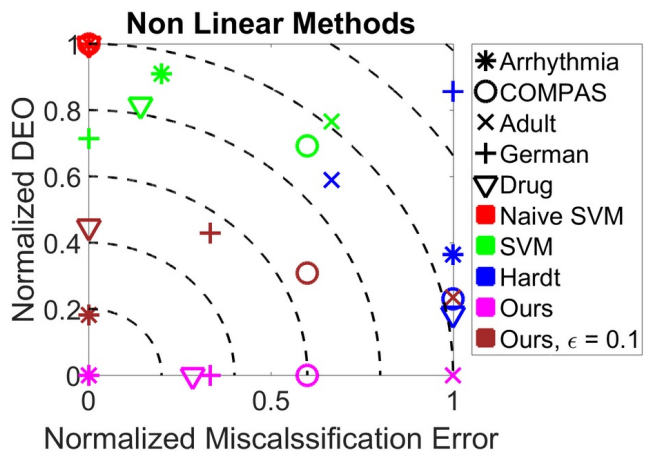
- Optimization problem (in feature space):

$$\min_{w \in \mathcal{H}} \sum_{i=1}^{n} \ell(\langle w, \phi(X_i) \rangle, y_i) + \lambda \|w\|_{\mathcal{H}}^2 \text{ , such that } |\langle w, u \rangle| \leq \varepsilon$$

Fair kernel methods
Results from Donini et al.

Method	Arrhythmia ACC	Arrhythmia DEO	COMPAS ACC	COMPAS DEO	Adult ACC	Adult DEO	German ACC	German DEO	Drug ACC	Drug DEO
					s inside $\boldsymbol{x}$					
Naïve Lin. SVM	0.79 ± 0.06	0.14 ± 0.03	0.76 ± 0.01	0.17 ± 0.02	0.81	0.14	0.71 ± 0.06	0.17 ± 0.05	0.81 ± 0.02	0.44 ± 0.03
Lin. SVM	0.78 ± 0.07	0.13 ± 0.04	0.75 ± 0.01	0.15 ± 0.02	0.80	0.13	0.69 ± 0.04	0.11 ± 0.10	0.81 ± 0.02	0.41 ± 0.06
Hardt	0.74 ± 0.06	0.07 ± 0.04	0.67 ± 0.03	0.21 ± 0.09	0.80	0.10	0.61 ± 0.15	0.15 ± 0.13	0.77 ± 0.02	0.22 ± 0.09
Zafar	0.71 ± 0.03	0.03 ± 0.02	0.69 ± 0.02	0.10 ± 0.06	0.78	0.05	0.62 ± 0.09	0.13 ± 0.11	0.69 ± 0.03	0.02 ± 0.07
Lin. Ours	0.79 ± 0.07	0.04 ± 0.03	0.76 ± 0.01	0.04 ± 0.03	0.77	0.01	0.69 ± 0.04	0.05 ± 0.03	0.79 ± 0.02	0.05 ± 0.03
Naïve SVM	0.79 ± 0.06	0.14 ± 0.04	0.76 ± 0.01	0.18 ± 0.02	0.84	0.18	0.74 ± 0.05	0.12 ± 0.05	0.82 ± 0.02	0.45 ± 0.04
SVM	0.78 ± 0.06	0.13 ± 0.04	0.73 ± 0.01	0.14 ± 0.02	0.82	0.14	0.74 ± 0.03	0.10 ± 0.06	0.81 ± 0.02	0.38 ± 0.03
Hardt	0.74 ± 0.06	0.07 ± 0.04	0.71 ± 0.01	0.08 ± 0.01	0.82	0.11	0.71 ± 0.03	0.11 ± 0.18	0.75 ± 0.11	0.14 ± 0.08
Ours	0.79 ± 0.09	0.03 ± 0.02	0.73 ± 0.01	0.05 ± 0.03	0.81	0.01	0.73 ± 0.04	0.05 ± 0.03	0.80 ± 0.03	0.07 ± 0.05





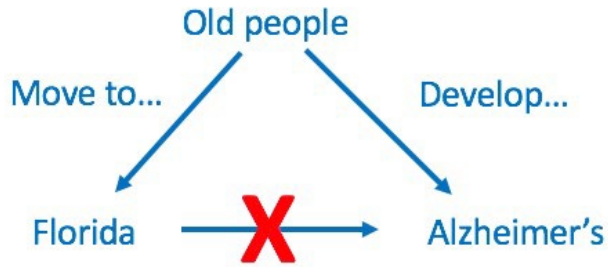
C. Issues to Consider

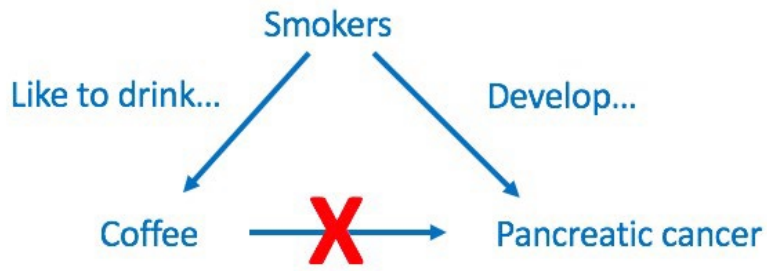
Sources of bias

Biases in data or models - 23 different types of bias have been reported but most important are:

- Selection bias such as sampling bias
- Information bias such as missing factors (Simpson's paradox)
- **Confounding factors**

Example of confounding factors





Methodology to address confounding factors in healthcare studies

- Step 1: Measure and report all potential confounders
- Step 2: Routinely assess the role of confounding factors and adjust for them in analyses
- Step 3: Report adjusted and crude estimates of association and discuss limitations of the study that may be due to confounding and the magnitude of the influence

Skelly, A. C., Dettori, J. R., Brodt, E. D. (2012). Assessing bias: the importance of considering confounding. Evidence-based spine-care journal, 3(1), 9–12.

Two-sample comparison
An add-on to detect and assess bias

At a level of significance $\alpha \in (0,1)$, the following test is performed:

$$"\mathcal{H}_0 : \ W_\phi^* = \int_0^1 \phi" \ \textit{versus} \ "\mathcal{H}_1 : \ W_\phi^* < \int_0^1 \phi".$$

Testing procedure. Let $\{\mathcal{D}_1, \ \mathcal{D}_2\}$ disjoint partition of the initial dataset $\mathcal{D}_0$.

1. Ranking.

(i) Find the optimal scoring rule $\hat{s} := \hat{s}_{n_1,m_1}$ on $\mathcal{D}_1$,

(ii) Compute the sequence of ranks $\{R(\hat{s}(\mathbf{X}_i))\}_{n_1<i\leq n}$ over the second pooled sample $\mathcal{D}_{n_2,m_2}$.

2. Rank-sum statistical test. Reduce and center the statistic $\widehat{W}_{n_2,m_2}(\hat{s})$ to perform the test $\eta(\hat{s}; \mathbf{X}_1, \ldots, \mathbf{X}_{n_2}, \mathbf{Y}_1, \ldots, \mathbf{Y}_{m_2}) := \mathbb{I}\{\widehat{W}_{n_2,m_2}(\hat{s}) > q_\phi^{1-\alpha}(\hat{s})\}$, where $q_\phi^{1-\alpha}$ is the $(1-\alpha)$-quantile of the distribution under $\mathcal{H}_0$ of the statistic used depending on ϕ, considered as test's *threshold*.

Result. If step (1) leads to a universally consistent *scoring* rule in the W-ranking performance measure sense, the score-based rank-sum statistical test is universally consistent at level α as $N \to \infty$.

Further topics

- Fairness with optimal transport
  [Eustasio del Barrio, Fabrice Gamboa, Paula Gordaliza, Jean-Michel Loubes (2019). Obtaining Fairness using Optimal Transport Theory. Proceedings of PMLR.]
- How to unveil sensitive variables?
  [Jens Ludwig, Sendhil Mullainathan (2023). Machine Learning as a Tool for Hypothesis Generation. NBER 31017.]