
Foundations for Fair Algorithmic Decision Making



Krishna P. Gummadi

Max Planck Institute for Software Systems

Algorithmic decision making

- Refers to **data-driven** decision making
 - By **learning** over data about **past decision outcomes**
- Increasingly influences every aspect of our life

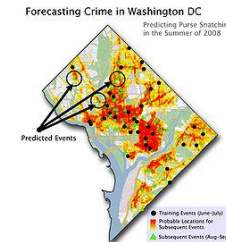
Search, Recommender, Reputation Algorithms



Match / Market-Making Algorithms



Risk Prediction Algorithms



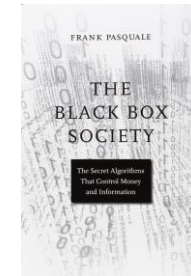
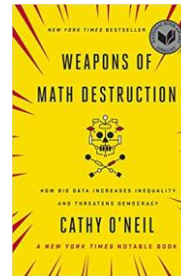
Concerns about their fairness

- ❑ **Discrimination** in predictive risk analytics

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

- ❑ **Opacity** of algorithmic (data-driven) decision making



- ❑ **Implicit biases** in As Germans Seek News, YouTube Delivers Far-Right Tirades

A researcher found the platform's recommendation system had steered viewers to fringe and conspiracy videos on a neo-Nazi demonstration in Chemnitz.



Focus on discrimination

- ❑ Discrimination is a **specific type of unfairness**
 - ❑ Well-studied in **social sciences**
 - ❑ Political science
 - ❑ Moral philosophy
 - ❑ Economics
 - ❑ Law
 - ❑ Majority of countries have anti-discrimination laws
 - ❑ Discrimination recognized in several international human rights laws
 - ❑ But, less-studied from a **computational perspective**
-

**What is a computational perspective?
Why is it needed?**

Defining discrimination

- A first approximate **normative / moralized** definition:

wrongfully impose a **relative disadvantage** on persons **based on** their membership in some **salient social group**
e.g., race or gender

- Challenge: How to **operationalize** the definition?
 - How to make it clearly **distinguishable, measurable, & understandable** in terms of empirical observations
-

Need to operationalize 4 fuzzy notions

1. What constitutes a **relative disadvantage**?
 2. What constitutes a **wrongful imposition**?
 3. What constitutes **based on**?
 4. What constitutes a **salient social group**?
-

Case study: Recidivism risk prediction

- ❑ **COMPAS** recidivism prediction tool
 - ❑ Built by a commercial company, Northpointe, Inc.
 - ❑ Estimates **likelihood** of criminals re-offending in **future**
 - ❑ **Inputs:** Based on a long questionnaire
 - ❑ **Outputs:** Used across US by judges and parole officers
 - ❑ Trained over **big historical recidivism data** across US
 - ❑ **Excluding sensitive feature** info like gender and race
-

COMPAS Goal: Criminal justice reform

- ❑ Many studies show racial biases in human judgments
 - ❑ **Idea:** Nudge subjective human decision makers with objective algorithmic predictions
 - ❑ Algorithms have no pre-existing biases
 - ❑ They simply process information in a consistent manner
 - ❑ Learn to make accurate predictions without race info.
 - ❑ Blacks & whites with same features get same outcomes
 - ❑ No disparate treatment & so non-discriminatory!
-

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Positive Rate: $805 / (805 + 990) = 0.45$

$349 / (349 + 1139) = 0.23$

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Positive Rate: $805 / (805 + 990) = 0.45$

$349 / (349 + 1139) = 0.23$

False Negative Rate: $532 / (532 + 1369) = 0.29$

$461 / (461 + 505) = 0.48$

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Positive Rate: $805 / (805 + 990) = 0.45$ >> $349 / (349 + 1139) = 0.23$

False Negative Rate: $532 / (532 + 1369) = 0.29$ << $461 / (461 + 505) = 0.48$

- ❑ ProPublica: False positive & negative rates are considerably worse for blacks than whites!
- ❑ Constitutes discriminatory **disparate mistreatment**

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

COMPAS study raises many questions

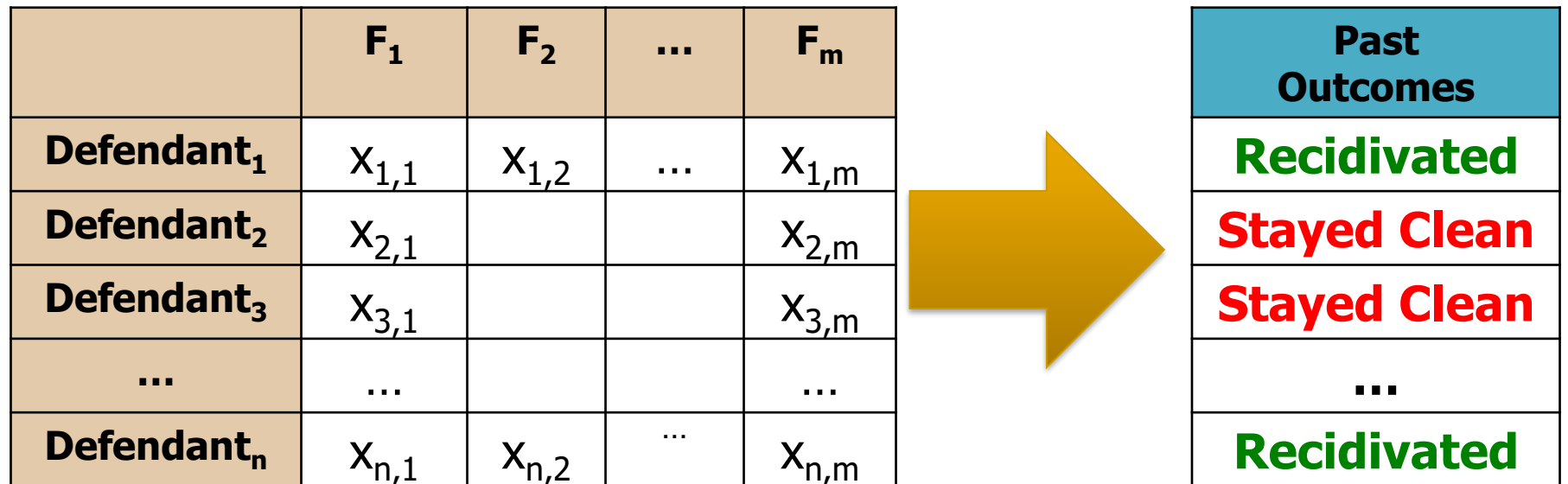
- ❑ Why does COMPAS show high racial FPR/FNR disparity?
 - ❑ Despite being trained without race information
- ❑ Can we train COMPAS to lower racial FPR/FNR disparity?

Analysis:

Why does COMPAS classifier show high racial FPR & FNR disparity?

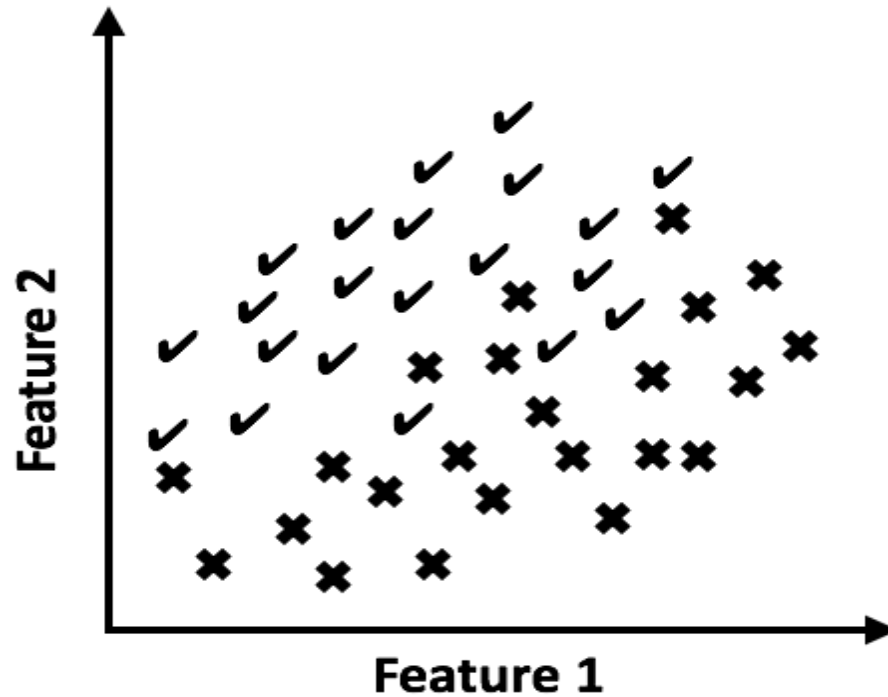
How COMPAS learns who recidivates

- By training over data about past outcomes

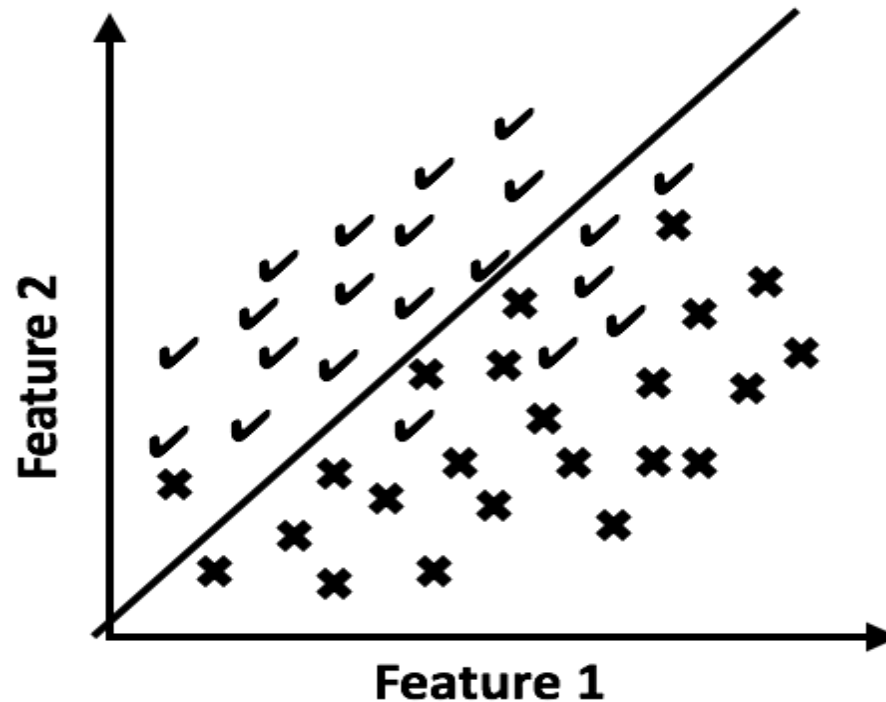


- Challenge:** Learning a decision function over the features that separates the two classes of people

How COMPAS learns who recidivates



How COMPAS learns who recidivates



- By finding the **optimal (most accurate / least loss) linear boundary** separating the two classes
- How does COMPAS find (compute) it?

Learning (computing) the optimal boundary

- Define & optimize a loss (accuracy) function
 - Capturing error (inaccuracy) in individual predictions

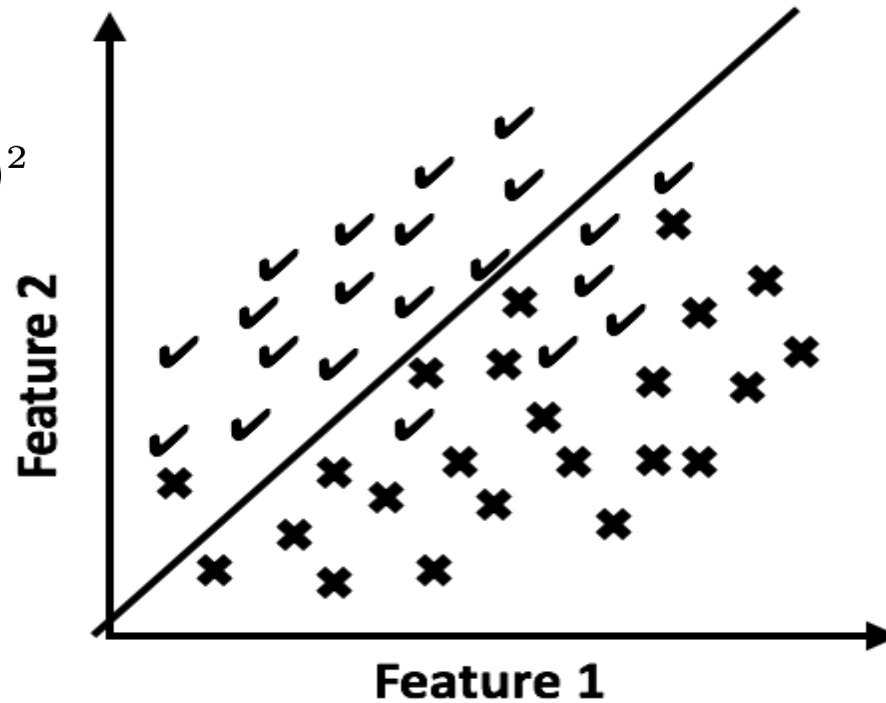
1. Minimized over all examples in training data

$$L(\mathbf{w}) = \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \quad \text{minimize } L(\mathbf{w})$$

1. Functions should allow for efficient optimization
 - Many loss functions used in learning are convex

How COMPAS learns who recidivates

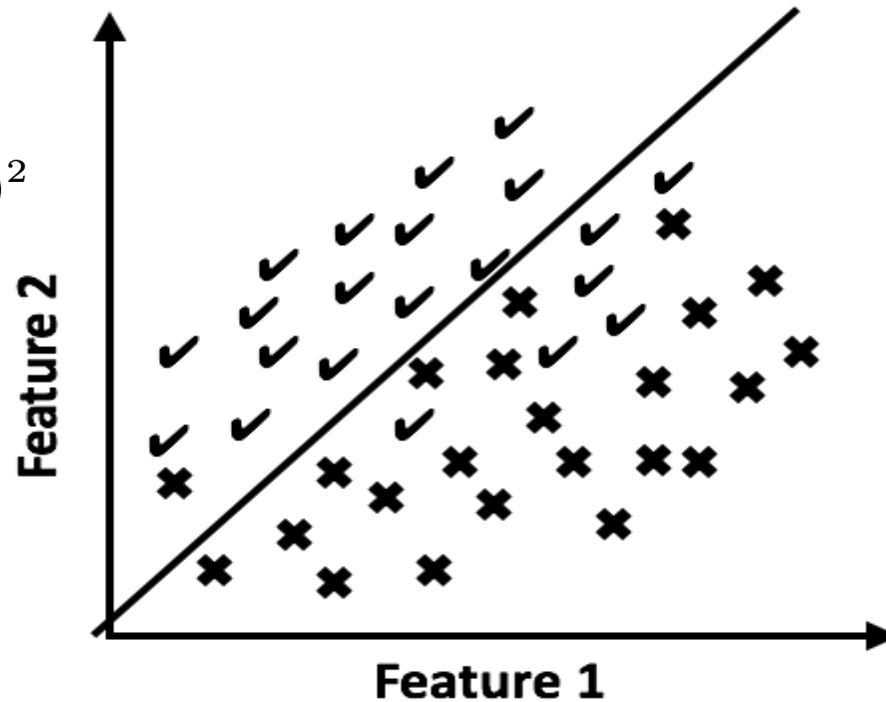
$$\min \sum_{i=1}^N (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$



- How did COMPAS find most accurate linear boundary?

How COMPAS learns to discriminate

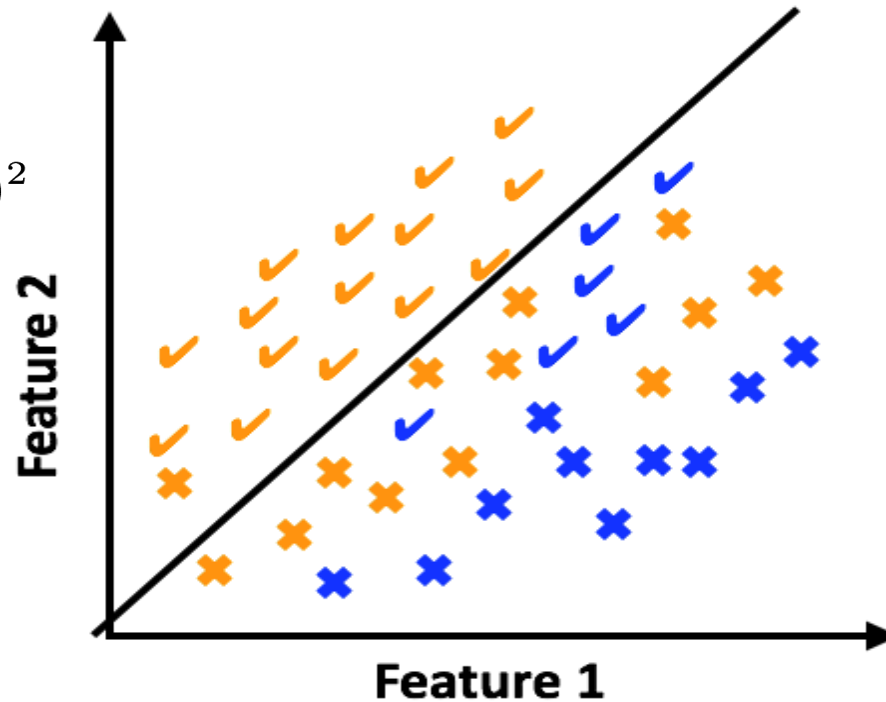
$$\min \sum_{i=1}^N (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$



- Observe the most accurate linear boundary

How COMPAS learns to discriminate

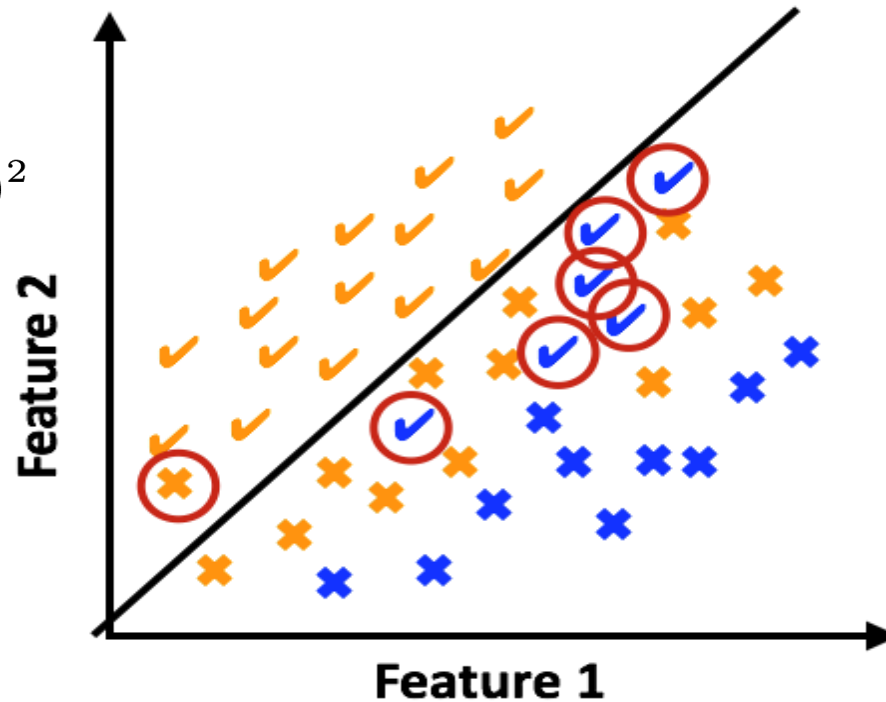
$$\min \sum_{i=1}^N (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$



- Observe the most accurate linear boundary

How COMPAS learns to discriminate

$$\min \sum_{i=1}^N (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$



- ❑ Observe the most accurate linear boundary
- ❑ Makes **few errors for yellow, lots of errors for blue!**
 - ❑ Causes **disparate mistreatment** – inequality in error rates

The cause of error rate disparity

- To minimize overall error, classifiers minimize **sum of individual-level errors**

$$\min P(y_{\text{pred}} \neq y_{\text{true}}) \approx \min \sum_{i=1}^N (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$

- Which doesn't guarantee **equal avg. group-level errors**

Overall Error Rate:

$$P(y_{\text{pred}} \neq y_{\text{true}} \mid \text{race}=\text{B}) \neq P(y_{\text{pred}} \neq y_{\text{true}} \mid \text{race}=\text{W})$$

False Positive Rate:

$$P(y_{\text{pred}} \neq y_{\text{true}} \mid y_{\text{true}} = +1, \text{race}=\text{B}) \neq P(y_{\text{pred}} \neq y_{\text{true}} \mid y_{\text{true}} = +1, \text{race}=\text{W})$$

False Negative Rate:

$$P(y_{\text{pred}} \neq y_{\text{true}} \mid y_{\text{true}} = -1, \text{race}=\text{B}) \neq P(y_{\text{pred}} \neq y_{\text{true}} \mid y_{\text{true}} = -1, \text{race}=\text{W})$$

Synthesis:

**How to train non-discriminatory
classifiers?** [WWW '17]

How to learn to avoid discrimination

- ❑ Specify **discrimination** measures as learning constraints
- ❑ Optimize for **accuracy under those constraints**

$$\min P(y_{\text{pred}} \neq y_{\text{true}})$$

$$\text{s.t. } P(y_{\text{pred}} \neq y_{\text{true}} \mid \text{race}=\text{B}) = P(y_{\text{pred}} \neq y_{\text{true}} \mid \text{race}=\text{W})$$

- ❑ The constraints **embed ethics & values** when learning
 - ❑ **No free lunch**: Additional constraints lower accuracy!
 - ❑ **Need race info** in training to avoid disp. mistreatment!
-

The technical challenge

- How to **learn efficiently** under these constraints?

$$\min \quad P(y_{\text{pred}} \neq y_{\text{true}}) \quad \approx \quad \min \quad \sum_{i=1}^N (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$

$$\text{s.t.} \quad P(y_{\text{pred}} \neq y_{\text{true}} \mid \text{race}=\text{B}) = P(y_{\text{pred}} \neq y_{\text{true}} \mid \text{race}=\text{W})$$

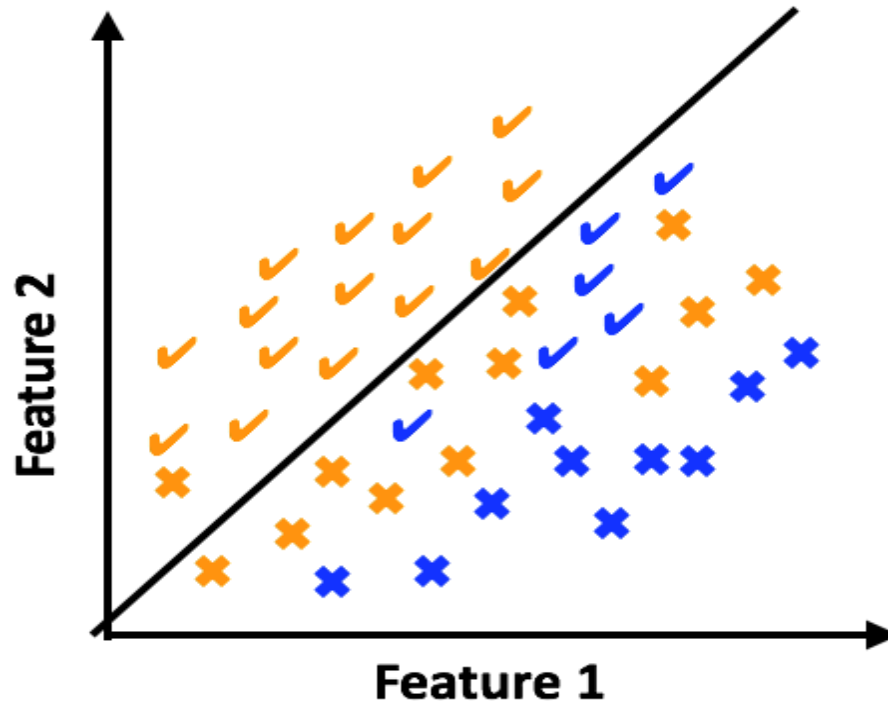
- Problem: The above formulations are **not convex!**
 - Can't learn it efficiently
- Need **to rewrite the constraints**

Rewriting mistreatment constraints

$$\min \sum_{i=1}^N (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$

$$\text{s.t.} \quad P(y_{\text{true}} \neq y_{\text{pred}} \mid \text{race}=\text{B}) = P(y_{\text{true}} \neq y_{\text{pred}} \mid \text{race}=\text{W})$$

Rewriting mistreatment constraints



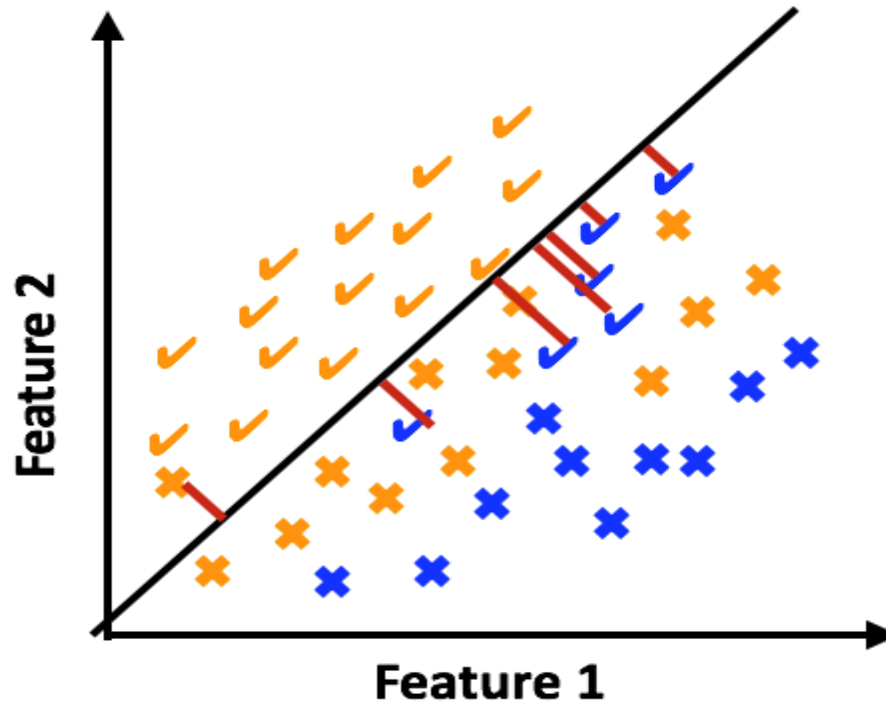
Idea: Avg. misclassification distance from boundary for both groups should be the same

Rewriting mistreatment constraints

$$\min(0, y_i d_{\mathbf{w}}(\mathbf{x}_i))$$

Concave

($d_{\mathbf{w}}(\mathbf{x})$ is affine)



Idea: Avg. misclassification distance from boundary for both groups should be the same

Rewriting mistreatment constraints

$$\begin{aligned} \min \quad & \sum_{i=1}^N (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2 \\ \text{s.t.} \quad & -\epsilon \leq \frac{1}{|\mathcal{M}|} \sum_{\mathcal{M}} \min(0, y_i d_{\mathbf{w}}(\mathbf{x}_i)) - \frac{1}{|\mathcal{F}|} \sum_{\mathcal{F}} \min(0, y_i d_{\mathbf{w}}(\mathbf{x}_i)) \leq \epsilon \end{aligned}$$

Concave

Concave

$P(y_{\text{true}} \neq y_{\text{pred}} \mid \text{race}=\mathbf{B})$ $P(y_{\text{true}} \neq y_{\text{pred}} \mid \text{race}=\mathbf{W})$

- Can be solved **efficiently**
 - Using **Disciplined Convex-Concave Programming**
 - DCCP [*Shen, Diamond, Gu, Boyd, 2016*]

Evaluation: Do our constraints work?

- ❑ Gathered a recidivism history dataset
 - ❑ Broward Country, FL for 2013-14
 - ❑ **Features:** arrest charge, #prior offenses, age,...
 - ❑ **Class label:** 2-year recidivism

 - ❑ **Traditional classifiers** without constraints
 - ❑ Acc.: **67%** FPR Disparity: **+0.20** FNR Disparity: **-0.30**

 - ❑ Training classifiers **with fairness constraints**
 - ❑ Acc.: **66%** FPR Disparity: **+0.03** FNR Disparity: **-0.11**
-

Lessons from the COMPAS story

Take-aways for ethical machine learning

High-level insight: Ethics & Learning

- ❑ Learning objectives **implicitly embody ethics**
 - ❑ By how they explicitly define **trade-offs in decision errors**
 - ❑ Traditional objective accuracy reflects **utilitarian ethics**
 - ❑ The rightness of decisions is a **function of individual utilities**
 - ❑ The desired function is **maximizing sum of individual utilities**
 - ❑ Lots of scenarios where utilitarian ethics fall short
 - ❑ **Change learning objectives** for other ethical considerations
 - ❑ E.g., non-discrimination requires equalizing group-level errors
-

Three challenges with ethical learning

❑ Operationalization:

- ❑ How to formally interpret fairness principles in different algorithmic decision making scenarios?

❑ Synthesis:

- ❑ How to design efficient learning mechanisms for different fairness interpretations?

❑ Analysis:

- ❑ What are the trade-offs between the learning objectives?
-

Two operationalizations of discrimination:
disparate treatment & disparate mistreatment

Are they sufficient for all scenarios?

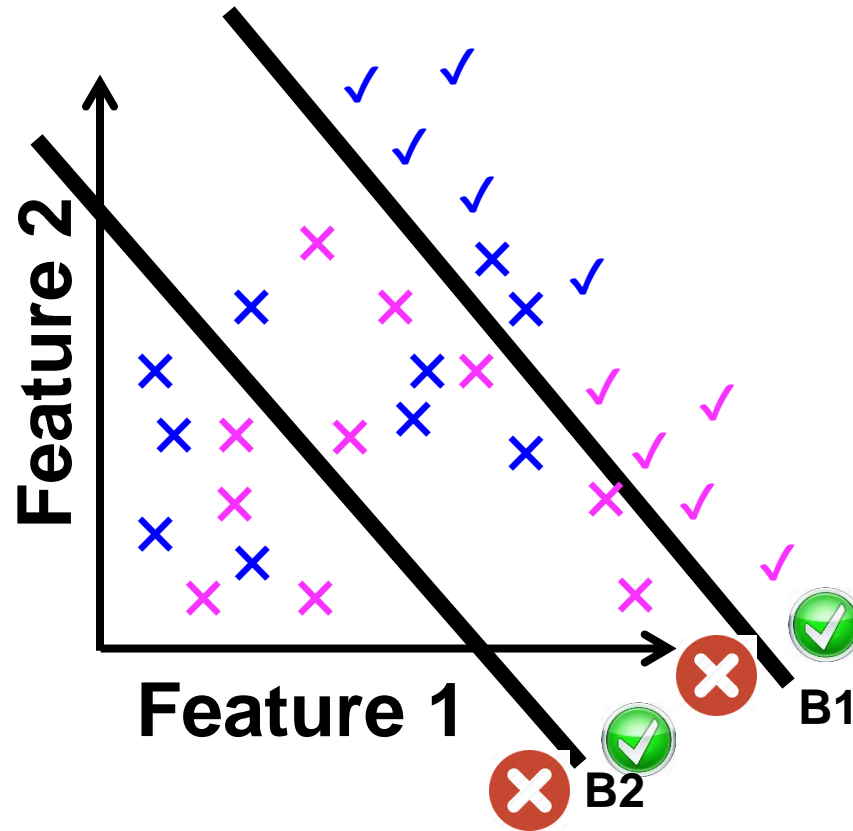
Discrimination in different scenarios

- What if **training data labels** were biased?
 - Require **equal group acceptance error rates** [AISTATS '17]
- Can requiring parity result in **all groups being worse-off**?

Parity outcomes are not pareto-optimal

$$\begin{aligned} \text{DM (B1)} \\ &= (15 - 13)/15 \\ &= 0.13 \end{aligned}$$

$$\begin{aligned} \text{DM (B2)} \\ &= (9 - 9)/15 \\ &= 0 \end{aligned}$$

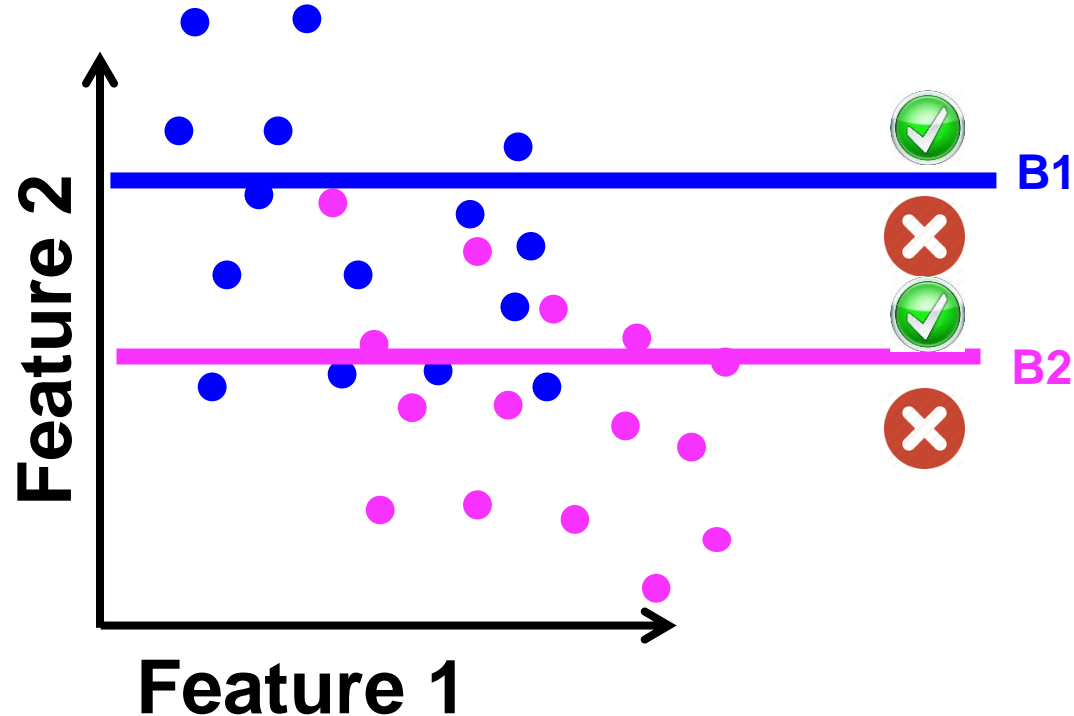


Both groups **are worse off** with **parity boundary B2!**
Both groups **prefer pareto-optimal B1** over **B2**

Discrimination in different scenarios

- ❑ What if **training data labels** were biased?
 - ❑ Require **equal group acceptance error rates** [AISTATS '17]
 - ❑ Can requiring parity result in **all groups being worse-off**?
 - ❑ **Yes!** Parity outcomes are **non pareto-optimal** [NIPS '17]
 - ❑ Allow disparity when **no groups is worse-off than parity**
 - ❑ Why not pick **group-specific decision boundaries**?
-

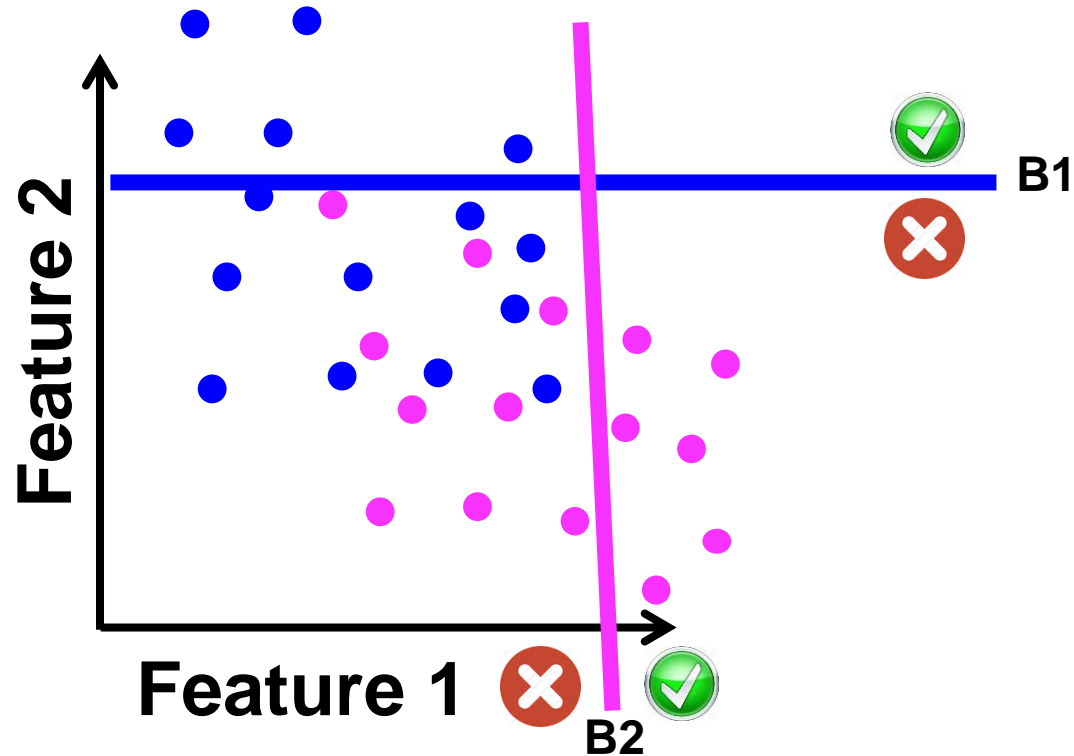
Reverse discrimination by group-specific boundaries



Both groups prefer B2 over B1

Blue group is envious of pink group; claims reverse discrimination

Envy-free group-specific boundaries



Blue group prefers **B1** and pink group prefers **B2**

No group is **envious** of another; **NO** reverse discrimination!

Discrimination in different scenarios

- ❑ What if **training data labels were biased**?
 - ❑ Require **equal group acceptance error rates** [AISTATS '17]
 - ❑ Can requiring parity result in **all groups being worse-off**?
 - ❑ **Yes!** Parity outcomes are **non pareto-optimal** [NIPS '17]
 - ❑ Allow disparity when **no groups is worse-off than parity**
 - ❑ Why not pick **group-specific decision boundaries**?
 - ❑ Need to avoid **reverse-discrimination** [NIPS '17]
 - ❑ Allow group-specific boundaries only **when they are envy-free**
-

Looking Forward:

**From Non-Discrimination To
Fair Algorithmic Decision Making**

Social Welfare Theory

Moral Philosophy

Social Choice Theory

Law

Behavioral Economics

Communication & Media Studies

Learning Non-Discriminatory Classification

Regression

Set Selection

Ranking

Matching

Clustering

Social Welfare Theory [KDD'18, NIPS'18] [WWW'18, AAI'18] **Moral Philosophy**

Social Choice Theory [FAT* '19]

[ICML'18, NDSS '18] **Law**

Behavioral Economics [AIES '19]

[ICWSM '18] **Comm. & Media Studies**

Learning Non-Discriminatory Classification

Regression

Set Selection

[SIGIR'18] **Ranking**

Matching

Clustering

My agenda

Foundations for Fair Algorithmic Decision Making

- View fairness principles through a computational lens
 - Operationalize the principles in learning-based decision making
 - Key challenges: **Interpretation, Synthesis and Analysis**
-

BACKUP SLIDES

Social Welfare Theory

Moral Philosophy

Social Choice Theory

Law

Behavioral Economics

Communication & Media Studies

Learning Fair Algorithmic Decision Making

Regression

Set Selection

Ranking

Matching

Clustering

Social Welfare Theory

Moral Philosophy

Social Choice Theory

Law

Behavioral Economics

Communication & Media Studies

Learning Classification with Non-Discriminatory Outcomes

Regression

Set Selection

Ranking

Matching

Clustering

Procedural

Informational

Social Welfare Theory

Moral Philosophy

Social Choice Theory

Law

Behavioral Economics

Communication & Media Studies

Learning Classification with Non-Discriminatory Outcomes

Regression

Set Selection

Ranking

Matching

Clustering

Procedural

Informational

Social Welfare Theory [KDD '18, NIPS '18] [WWW '18] **Moral Philosophy**

Social Choice Theory [ICML '18] **Law**

Behavioral Economics [ICWSM '18] **Communication & Media Studies**

Learning Classification with Non-Discriminatory Outcomes

Regression [AAAI '18] **Procedural**

Set Selection [NDSS '18] **Informational**

Ranking [SIGIR '18]

Matching

Clustering **Human vs. Machine**

Representation **Machine-assisted Human**

Beyond disparate mistreatment:

**Is there more to discrimination than
equalizing error rates?**

The non-discrimination principle

- A first approximate **normative** definition:

wrongfully impose a **relative disadvantage** on persons **based on** their membership in some **salient social group**
e.g., race or gender

- Challenge: How to **operationalize** the definition?
 - How to make it clearly **distinguishable, measurable, & understandable** in terms of empirical observations
-

Operationalizing four fuzzy notions

- ❑ What constitutes a **salient social group**?
 - ❑ What constitutes **based on**?
 - ❑ What constitutes a **relative disadvantage**?
 - ❑ What constitutes a **wrongful imposition**?
-

Operationalizing four fuzzy notions

- What constitutes a **salient social group**?
 - What constitutes **based on**?
 1. Using group info. in **training or deployment [COMPAS]**
 2. Using group info. in **deployment, but not training [WWW '17]**
 - What constitutes a **relative disadvantage**?
 1. **Disparity in outcomes for similar users** across groups **[COMPAS]**
 2. **Additionally, disparity in error rates** across groups **[WWW '17]**
 - What constitutes a **wrongful imposition**?
-

Ethics & Algorithmic decision making

- ❑ **Societal need: Ethics for algorithms**
 - ❑ All algorithms err, but **not all errors the same**

class: 793
label: n04209133 shower cap
certainty: 99.7%



Turkish - detected

o bir aşçı
o bir mühendis
o bir doktor
o bir hemşire

English

she is a cook
he is an engineer
he is a doctor
she is a nurse

- ❑ Ethical errors make use of algorithms **untenable**
- ❑ **Scientific curiosity: Ethics through algorithmic lens**
 - ❑ **New interpretations** of fairness principles
 - ❑ Better **understanding of trade-offs** between interpretations
 - ❑ **Building** learning systems & **computing their consequences**

Computational perspective of ethics

- **Physical symbol system hypothesis:**
 - **A physical symbol system has the necessary and sufficient means for general intelligent action**
 - Simon & Newell
 - Two physical symbol systems: **Humans & Machines**
 - **Hypothesis about ethics:**
 - **Ethical actions are a form of intelligent actions**
 - Goal: Explore the limits of the ethics hypothesis
 - Both for societal benefits and scientific curiosity
-

So far, explored discrimination ethics

- ❑ Showed that it is possible to capture many nuanced interpretations in computational decision making
 - ❑ Computational interpretations raise new scenarios
 - ❑ previously overlooked by human decision makers
 - ❑ Many of which are beyond cognitive abilities of humans
-

Collaborators within MPG

Social Welfare Theory

Moral Philosophy

Social Choice Theory

Law

Behavioral Economics

Communication & Media Studies

Learning Non-Discriminatory Classification

Regression

Set Selection

Ranking

Matching

Clustering