CS6501: Topics in Learning and Game Theory
(Fall 2019)

Inherent Trade-Offs in Algorithmic Fairness

Instructor: Haifeng Xu

COMPAS: A Risk Prediction Tool to Criminal Justice

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
  - Used by states of New York, Wisconsin, Cali, Florida, etc.
  - A software that assesses likelihood of a defendant of reoffending
- Still many issues
  - Not interpretable
  - Low accuracy
  - Bias/unfairness

COMPAS: A Risk Prediction Tool to Criminal Justice

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
  - Used by states of New York, Wisconsin, Cali, Florida, etc.
  - A software that assesses likelihood of a defendant of reoffending
- Still many issues
  - Not interpretable
  - Low accuracy
  - Bias/unfairness (this lecture)

COMPAS: A Risk Prediction Tool to Criminal Justice

- In a ProPublica investigation of the algorithm…

"…blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend" -- unequal false positive rate

"… whites are much more likely than blacks to be labeled lower-risk but go on to commit other crimes" -- unequal false negative rate

Algorithms seem unfair!!

Other Examples

➢ Advertising and commercial contents

Search Engines

April 2, 2013
Volume 11, issue 3

PDF

Discrimination in Online Ad Delivery

Google ads, black names and white names, racial discrimination, and click advertising

Latanya Sweeney

Searching names that are likely assigned to black babies generates more ads suggestive of an arrest

Other Examples

- Advertising and commercial contents
  - If a male and female user are equally interested in a product, will they be equally likely to be shown an ad for it?
  - Will women in aggregate be shown ads for lower-paying jobs?
- Medical testing and diagnosis
  - Will treatment be applied uniformly across different groups of patients?
- Hiring or admission
  - Will students or job candidates from different groups be admitted with equal probability?
- …

Why Algorithms May Be "Unfair"?

- Algorithms may encode pre-existing bias
  - E.g., British Nationality act program, designed to automate evaluation of new UK citizens
  - It accurately reflects tenets of the law "a man is the father of only his legitimate children, whereas a woman is the mother of all her children, legitimate or not"

Why Algorithms May Be "Unfair"?

- Algorithms may encode pre-existing bias
  - Easier to handle

Why Algorithms May Be "Unfair"?

- Algorithms may encode pre-existing bias
  - Easier to handle
- Algorithms may create bias when serving its own objective
  - E.g., search engines try to show your favorite contents but not the most fair contents
- Input data are biased
  - E.g., ML may classify based on sensitive features in biased data
  - Can we simply remove these sensitive features during training?
- Biased algorithm may get biased feedback and further strengthen the issue

Why Algorithms May Be "Unfair"?

- Algorithms may encode pre-existing bias
  - Easier to handle
- Algorithms may create bias when serving its own objective
  - E.g., search engines try to show your favorite contents but not the most fair contents
- Input data are biased
  - E.g., ML may classify based on sensitive features in biased data
  - Can we simply remove these sensitive features during training?
- Biased algorithm may get biased feedback and further strengthen the issue

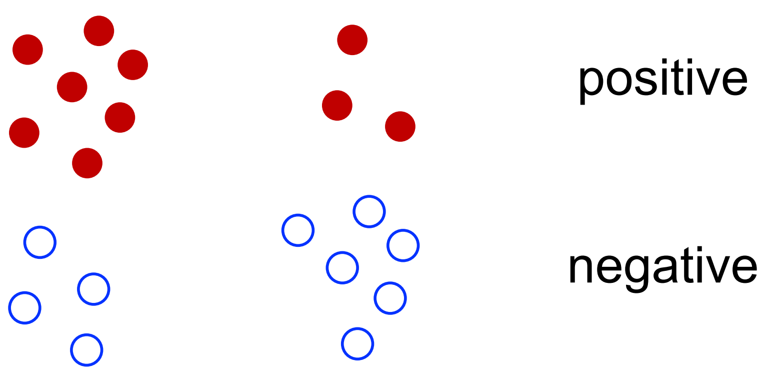
This lecture: there is another reason – some basic definitions of fairness are intrinsically not compatible

The Problem of Predicting Risk Scores

- In many applications, we classify whether people possess some property by predicting a score based on their features
  - Criminal justice
  - Loan lending
  - University admission
- Next: an abstract model to capture this process

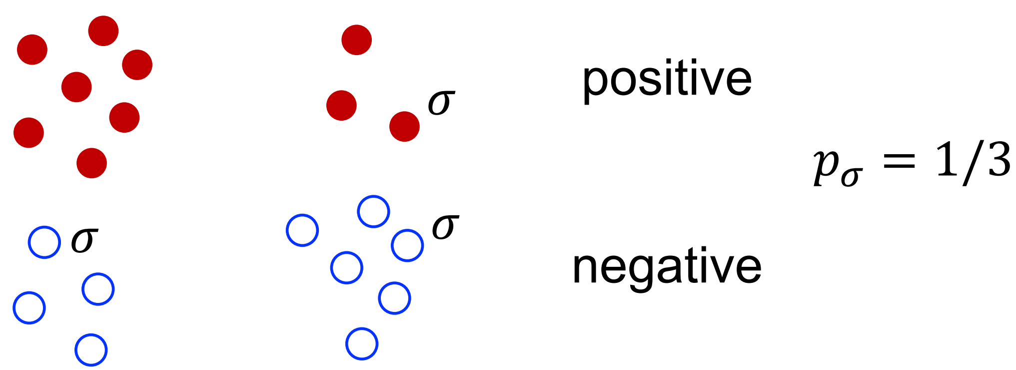
The Problem of Predicting Risk Scores

- There is a collection of people, each of whom is either a positive or negative instance
  - Positive/negative describe the true label of each individual



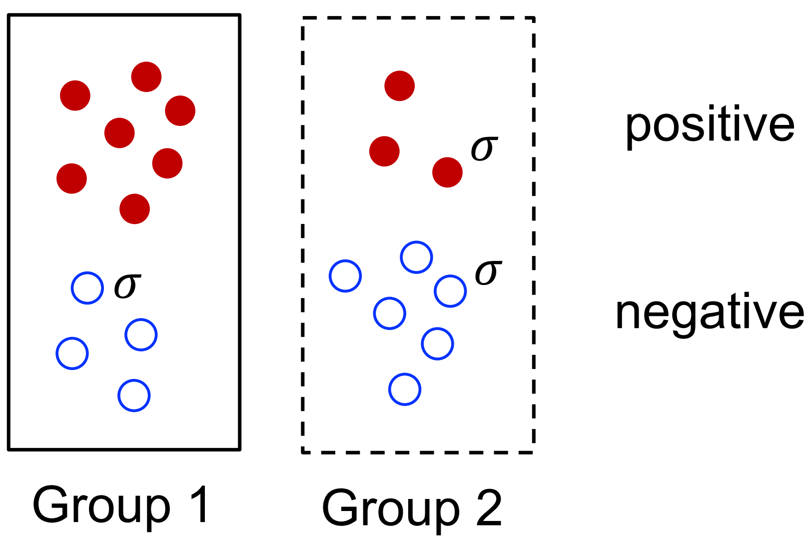
The Problem of Predicting Risk Scores

- There is a collection of people, each of whom is either a positive or negative instance
  - Positive/negative describe the true label of each individual
- Each person has an associated feature vector σ
  - p_σ = fraction of people with σ who are positive



The Problem of Predicting Risk Scores

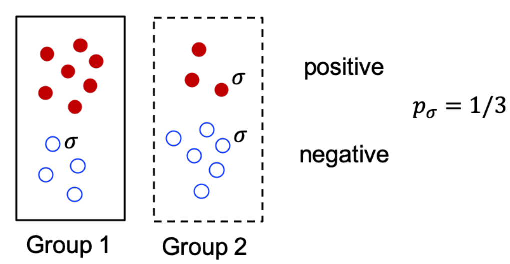
- There is a collection of people, each of whom is either a positive or negative instance
  - Positive/negative describe the true label of each individual
- Each person has an associated feature vector σ
  - p_σ = fraction of people with σ who are positive
- Each person belongs to one of two groups



$p_\sigma = 1/3$ publicly known

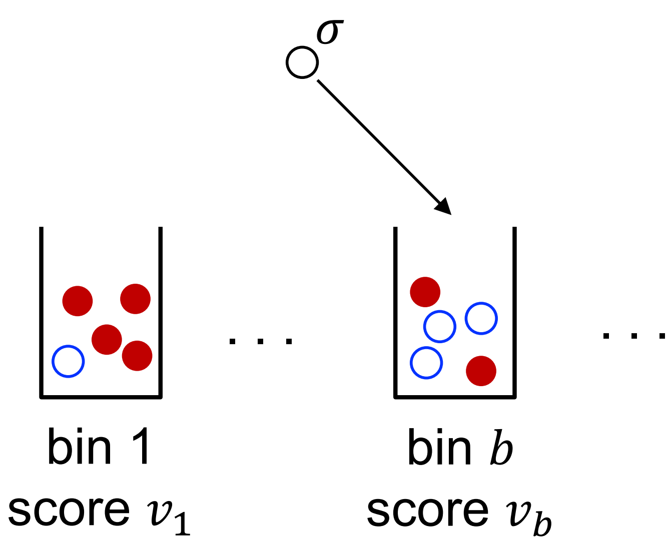
The Problem of Predicting Risk Scores

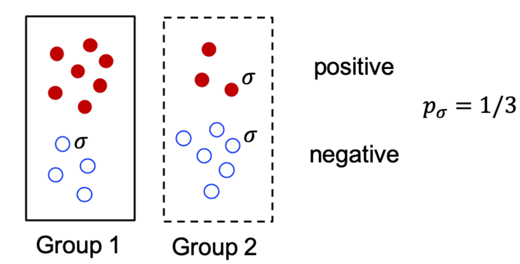
- Task: assign risk score to each individual
- Objective: accuracy (of course) and "fair"
  - Naturally, the score should only depend on σ, not individual's group



The Problem of Predicting Risk Scores

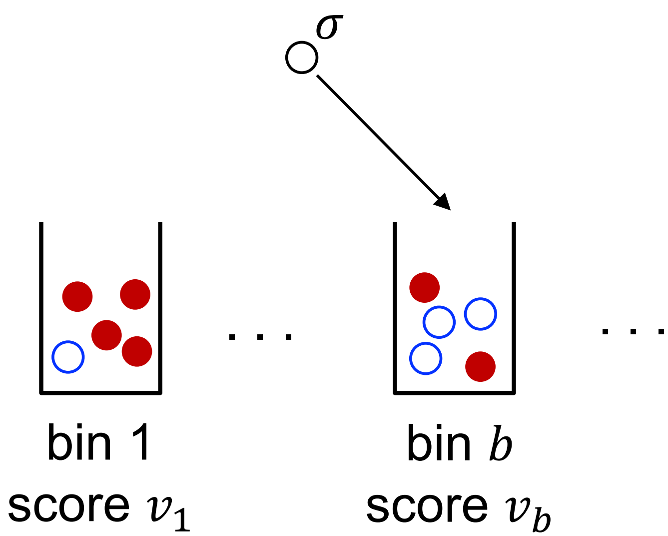
- Task: assign risk score to each individual
- Objective: accuracy (of course) and "fair"
  - Naturally, the score should only depend on σ, not individual's group
- The score assignment process: put σ into bins (possibly randomly)
  - Only depend on σ (label is unknown in advance)

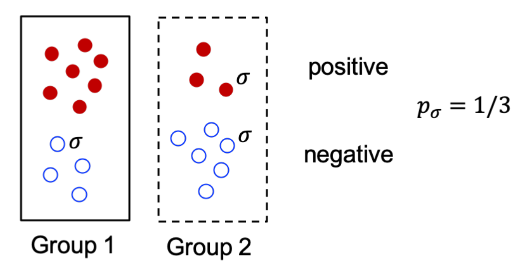




The Problem of Predicting Risk Scores

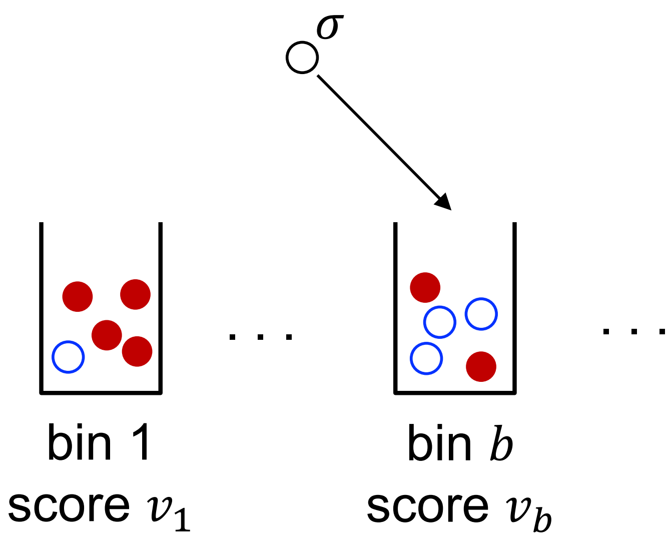
- Task: assign risk score to each individual
- Objective: accuracy (of course) and "fair"
  - Naturally, the score should only depend on σ, not individual's group
- The score assignment process: put σ into bins (possibly randomly)
  - Only depend on σ (label is unknown in advance)
  - Example 1: assign all σ to the same bin; give that bin score p_σ
  - Example 2: assign all people to one bin; give score 1



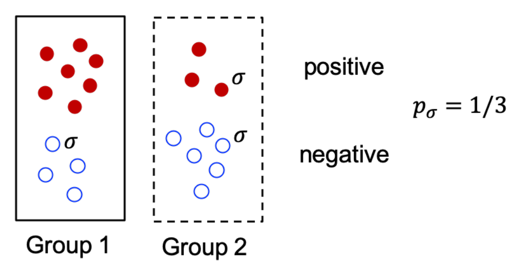


The Problem of Predicting Risk Scores

- Task: assign risk score to each individual
- Objective: accuracy (of course) and "fair"
  - Naturally, the score should only depend on σ, not individual's group
- The score assignment process: put σ into bins (possibly randomly)
  - Only depend on σ (label is unknown in advance)
  - Example 1: assign all σ to the same bin; give that bin score p_σ
  - Example 2: assign all people to one bin; give score 1



Note: may have very bad accuracy but good fairness, as they are different



Well…What Does "Fair" Really Mean?

- A very subjective perception
- Yet, for algorithm design, need a concrete and objective definition
- > 20 different definitions of fairness so far
  - See a survey paper "Fairness Definitions Explained"
- This raises many questions
  - Are they all reasonable? Can we satisfy all of them?
  - Which one/subset of them we should use when designing algorithms?
  - Do I have to sacrifice accuracy to achieve fairness?

Well…What Does "Fair" Really Mean?

- A very subjective perception
- Yet, for algorithm design, need a concrete and objective definition
- >20 different definitions of fairness so far
  - See a survey paper "Fairness Definitions Explained"
- This raises many questions
  - Are they all reasonable? Can we satisfy all of them?
  - Which one/subset of them we should use when designing algorithms?
  - Do I have to sacrifice accuracy to achieve fairness?

Some basic definitions of fairness are already not compatible, regardless how much accuracy you are willing to sacrifice

Fairness Def 1: Calibration

Definition [Calibration within groups]. For each bin b, let

- $N_{t,b}$ = # of people assigned to b from group t
- $n_{t,b}$ = # of positive people assigned to b from group t

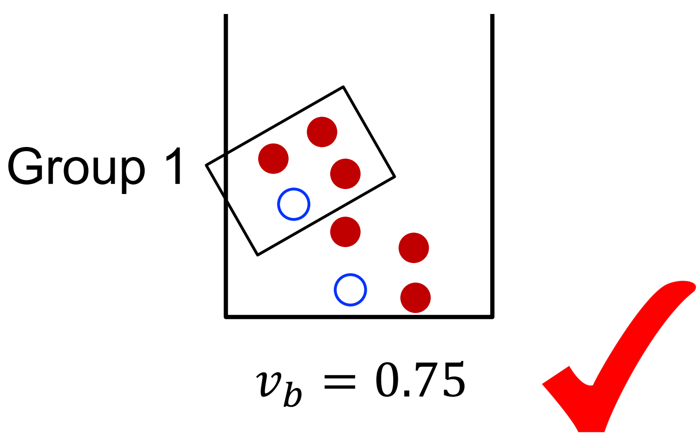
We should have $n_{t,b} = v_b \cdot N_{t,b}$ for each t, b

Fairness Def 1: Calibration

Definition [Calibration within groups]. For each bin b, let

- $N_{t,b}$ = # of people assigned to b from group t
- $n_{t,b}$ = # of positive people assigned to b from group t

We should have $n_{t,b} = v_b \cdot N_{t,b}$ for each t, b

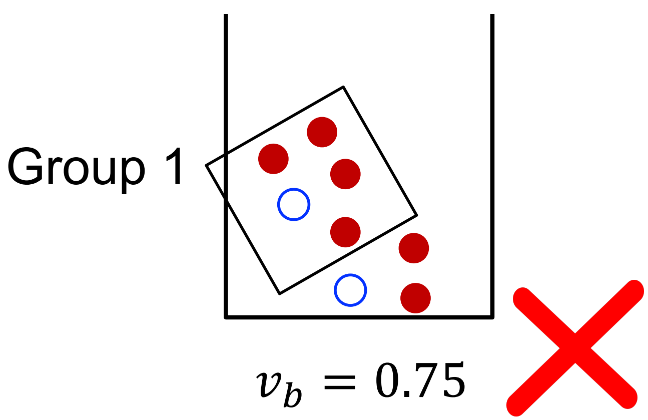


Fairness Def 1: Calibration

Definition [Calibration within groups]. For each bin b, let

- $N_{t,b}$ = # of people assigned to b from group t
- $n_{t,b}$ = # of positive people assigned to b from group t

We should have $n_{t,b} = v_b \cdot N_{t,b}$ for each t, b

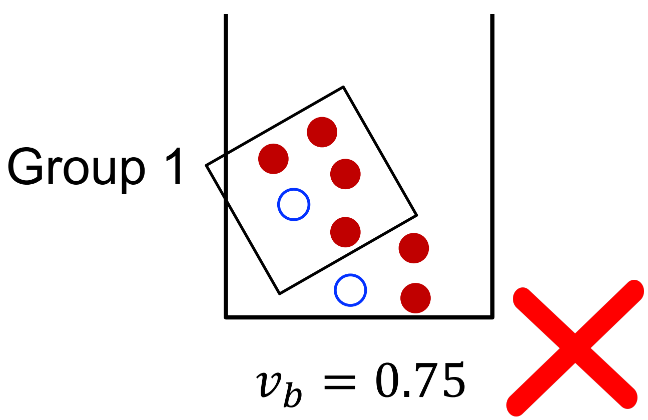


Fairness Def 1: Calibration

Definition [Calibration within groups]. For each bin b, let

- $N_{t,b}$ = # of people assigned to b from group t
- $n_{t,b}$ = # of positive people assigned to b from group t

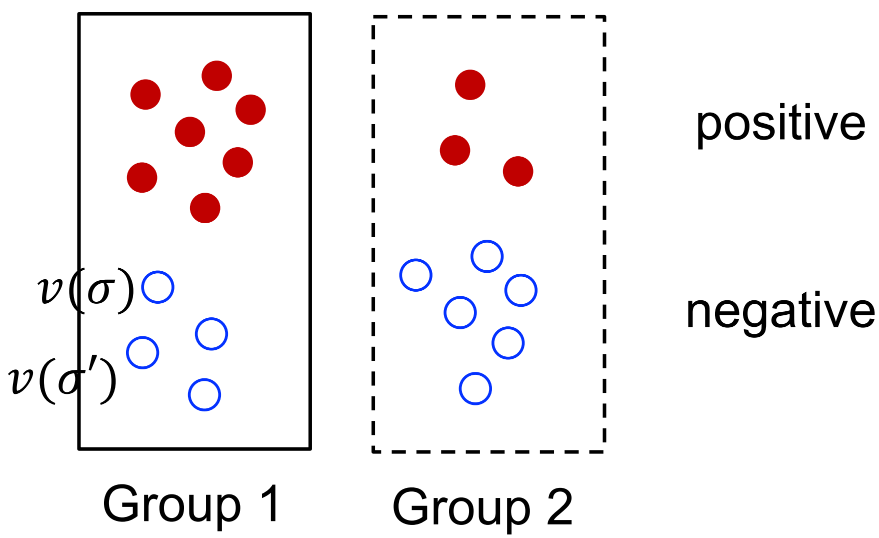
We should have $n_{t,b} = v_b \cdot N_{t,b}$ for each t, b



In practice, we do not know who are positive so cannot check the condition, but the definition still applies

Fairness Def 2: Balance of Negative Class

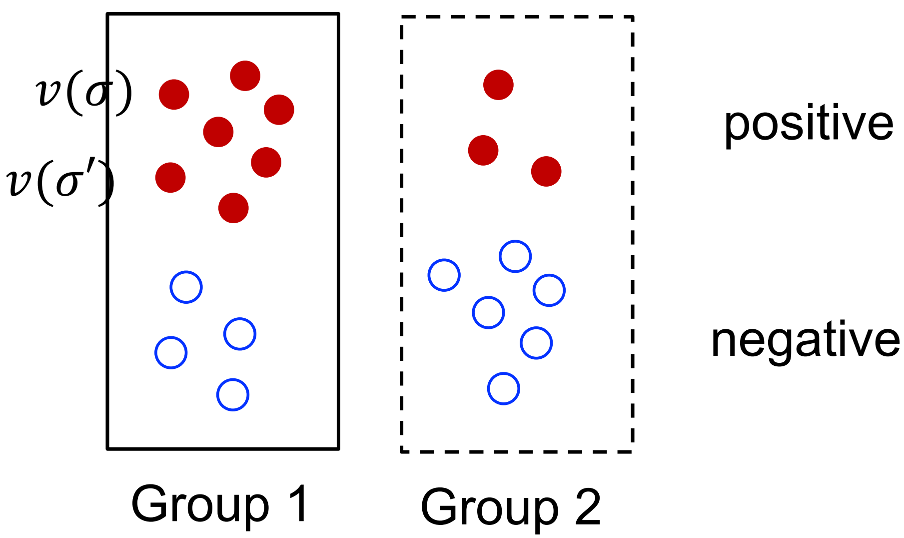
Definition [Balance of Negative Class]. Average scores assigned to people of group 1 who are negative should be the same as average scores assigned to people of group 2 who are negative.



$$E[v(\sigma) \mid \sigma \text{ negative and in group 1}]$$
$$= E[v(\sigma) \mid \sigma \text{ negative and in group 2}]$$

Fairness Def 3: Balance of Positive Class

Definition [Balance of Negative Class]. Average scores assigned to people of group 1 who are positive should be the same as average scores assigned to people of group 2 who are positive.



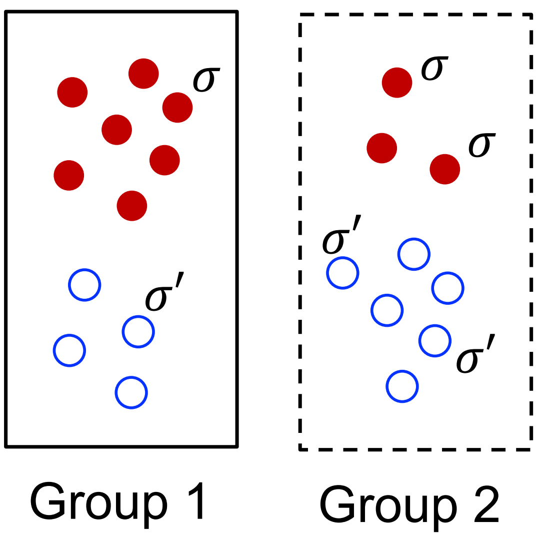
$$E[v(\sigma) |\ \sigma \text{ positive and in group 1}] = E[v(\sigma) |\ \sigma \text{ positive and in group 2}]$$

Is It Possible to Achieve All Three?

Is It Possible to Achieve All Three?

Yes: Example 1

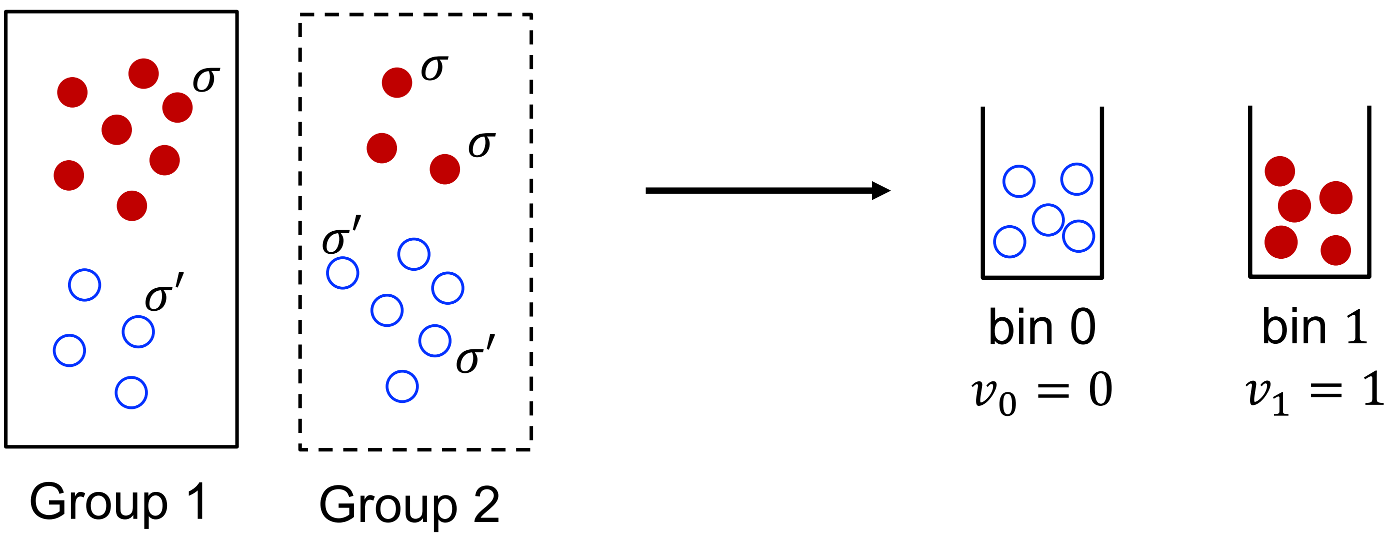
- $p_\sigma = 1\ or\ 0$ for all σ



Is It Possible to Achieve All Three?

Yes: Example 1

- $p_\sigma = 1\ or\ 0$ for all σ
- Two bins with $v_0 = 0$ and $v_1 = 1$; assign all σ with $p_\sigma = 0$ to bin 0 and all σ with $p_\sigma = 1$ to bin 1

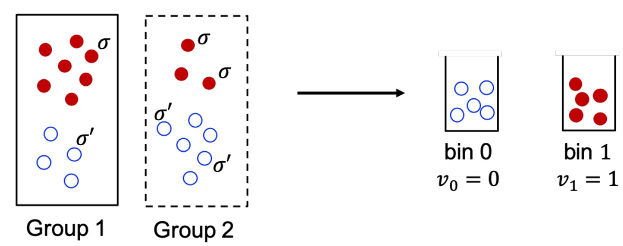


Is It Possible to Achieve All Three?

Yes: Example 1

- $p_\sigma = 1\ or\ 0$ for all σ
- Two bins with $v_0 = 0$ and $v_1 = 1$; assign all σ with $p_\sigma = 0$ to bin 0 and all σ with $p_\sigma = 1$ to bin 1

Claim: This score assignment satisfies all 3 fairness defs.



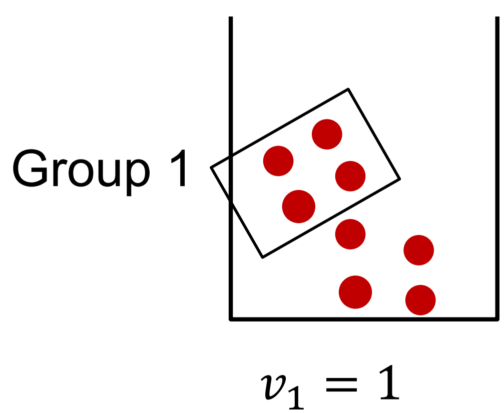
Is It Possible to Achieve All Three?

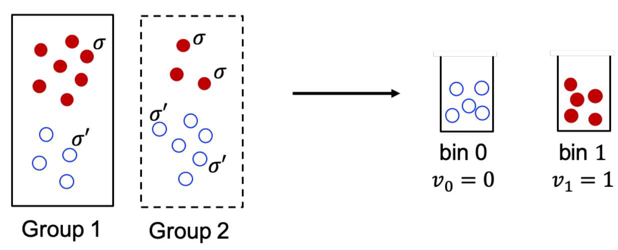
Yes: Example 1

- $p_\sigma = 1 \; or \; 0$ for all σ
- Two bins with $v_0 = 0$ and $v_1 = 1$; assign all σ with $p_\sigma = 0$ to bin 0 and all σ with $p_\sigma = 1$ to bin 1

Claim: This score assignment satisfies all 3 fairness defs.

- Calibration: yes, all the ratio is 1 or 0 for each group





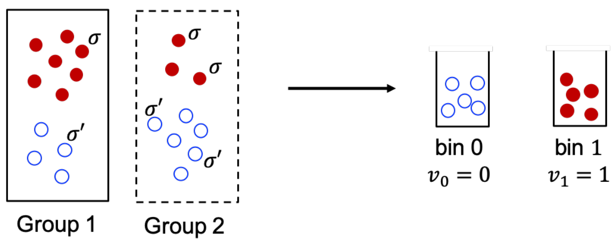
Is It Possible to Achieve All Three?

Yes: Example 1

- $p_\sigma = 1 \; or \; 0$ for all σ
- Two bins with $v_0 = 0$ and $v_1 = 1$; assign all σ with $p_\sigma = 0$ to bin 0 and all σ with $p_\sigma = 1$ to bin 1

Claim: This score assignment satisfies all 3 fairness defs.

- Calibration: yes, all the ratio is 1 or 0 for each group
- Balance of positive class: yes, both groups have average score 1
- Balance of negative class: yes, both groups have average score 0



Is It Possible to Achieve All Three?

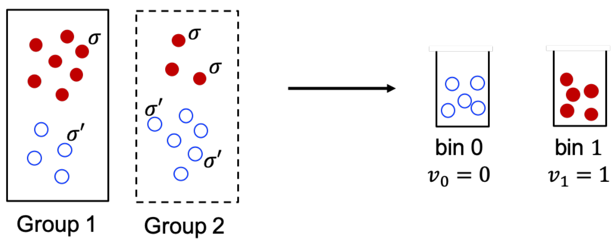
Yes: Example 1

- $p_\sigma = 1\ or\ 0$ for all σ
- Two bins with $v_0 = 0$ and $v_1 = 1$; assign all σ with $p_\sigma = 0$ to bin 0 and all σ with $p_\sigma = 1$ to bin 1

Claim: This score assignment satisfies all 3 fairness defs.

Caveats

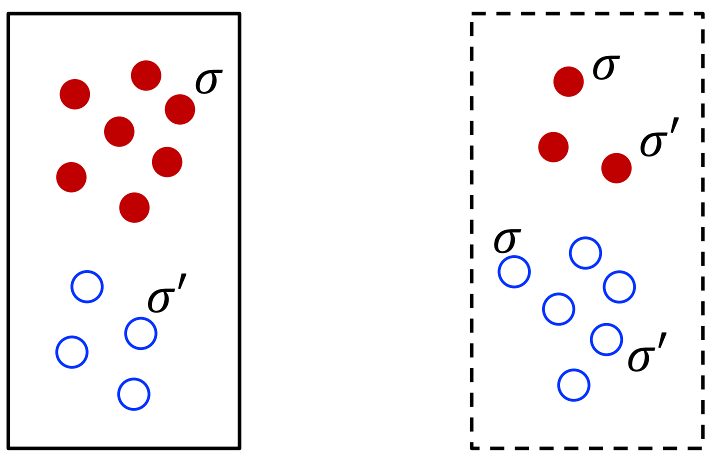
- But, this is not really a realistic setting…
- $p_\sigma = 0\ or\ 1$ means we know for sure each individual's label



Is It Possible to Achieve All Three?

Yes: Example 2

- Average p_σ (over σ's) is the same among two groups



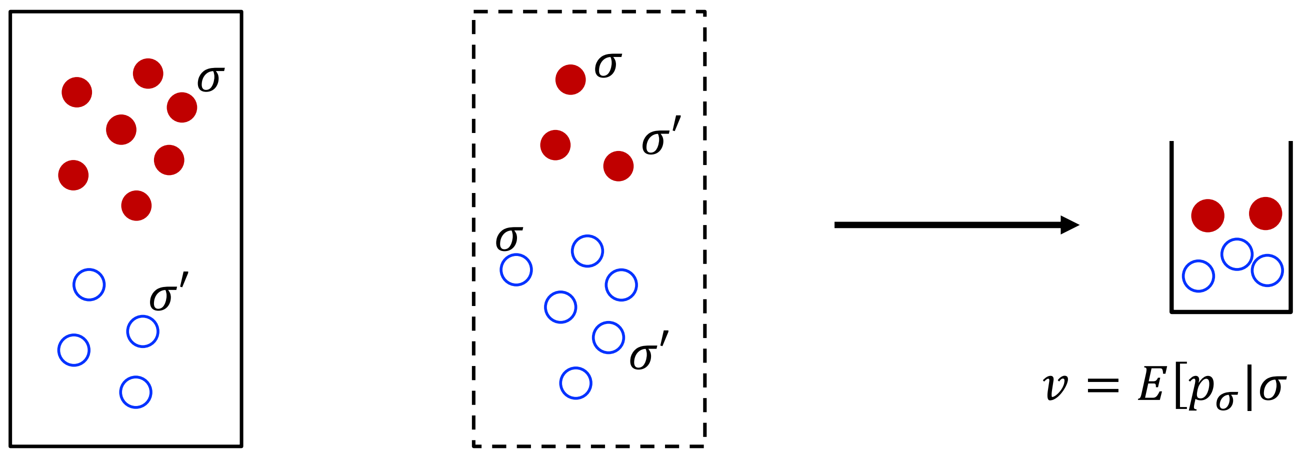
Group 1 Group 2

$$E[p_\sigma|\sigma \in \text{Group 1}] \quad = \quad E[p_\sigma|\sigma \in \text{Group 2}]$$

Is It Possible to Achieve All Three?

Yes: Example 2

- Average p_σ (over σ's) is the same among two groups
- One bin, with v equal the above average p_σ



$v = E[p_\sigma | \sigma \in \text{Group 1}]$

Group 1　　　Group 2

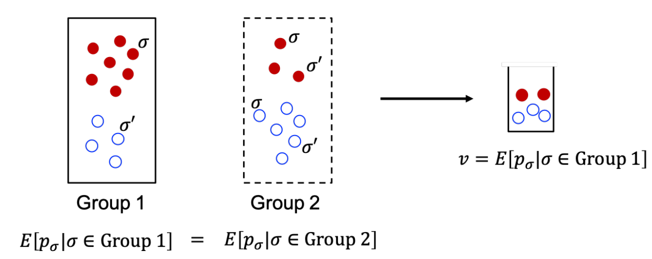
$E[p_\sigma | \sigma \in \text{Group 1}] \quad = \quad E[p_\sigma | \sigma \in \text{Group 2}]$

Is It Possible to Achieve All Three?

Yes: Example 2

- Average p_σ (over σ's) is the same among two groups
- One bin, with v equal the above average p_σ

Claim: This score assignment satisfies all 3 fairness defs.



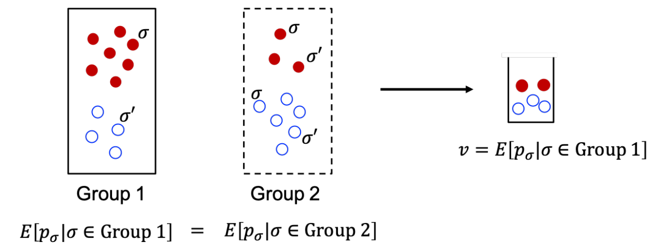
Is It Possible to Achieve All Three?

Yes: Example 2

- Average p_σ (over σ's) is the same among two groups
- One bin, with v equal the above average p_σ

Claim: This score assignment satisfies all 3 fairness defs.

- Calibration: yes, since v = average p_σ is exactly the probability of positive instances in both groups
- Balance of positive class: trivial, as scores are the same
- Balance of negative class: trivial as well



Is It Possible to Achieve All Three?

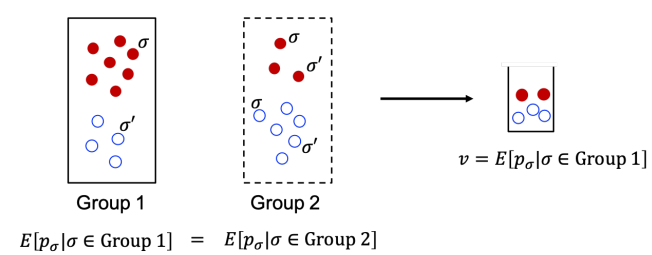
Yes: Example 2

- Average p_σ (over σ's) is the same among two groups
- One bin, with v equal the above average p_σ

Claim: This score assignment satisfies all 3 fairness defs.

Caveats

- But, this score assignment is not useful and has low accuracy
- There may exist a more accurate score assignment in this case that still satisfy three definitions
  - Bad news: it is NP-hard to find



Inherent Trade-offs of Algorithmic Fairness

Theorem: For the problem of risk score assignment, if there is a risk assignment that satisfies all the three fairness definitions before, the problem must be one of the previous two example cases.

- The two (degenerated) examples are the only cases where you can possibly satisfy all three fairness definitions

Proof Sketch

- Assume there is a score assignment satisfying all three defs
- Will derive contradictions, unless the instance is the previous degenerated settings

Proof Sketch

Notations

- N_t = total number of people in group t
- n_t = total number of positive people in group t

Calibration condition implies

- Total score of all group-t people in bin b (i.e., $v_b \cdot N_{t,b}$) equal expected number of positive group-t people in bin b (i.e., $n_{t,b}$)

Definition [Calibration]. For each bin b, let

- $N_{t,b}$ = # of people assigned to b from group t
- $n_{t,b}$ = # of positive people assigned to b from group t

We should have $n_{t,b} = v_b \cdot N_{t,b}$ for each t, b

Proof Sketch

Notations
- N_t = total number of people in group t
- n_t = total number of positive people in group t

Calibration condition implies
- Total score of all group-t people in bin b (i.e., $v_b \cdot N_{t,b}$) equal expected number of positive group-t people in bin b (i.e., $n_{t,b}$)
- Summing over all bins → total score of all group-t people equals expected number of positive group-t people

Proof Sketch

Notations

- N_t = total number of people in group t
- n_t = total number of positive people in group t

Another way to calculate total scores

- x = average score of a person in negative class
- y = average score of a person in positive class

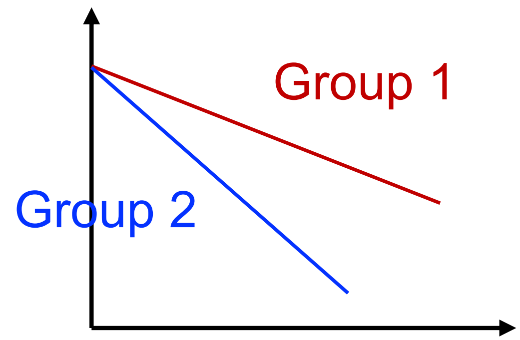
Proof Sketch

Notations

- N_t = total number of people in group t
- n_t = total number of positive people in group t

Another way to calculate total scores

- x = average score of a person in negative class
- y = average score of a person in positive class
- Total score in group t is $y(N_t - n_t) + xn_t = n_t$ by calibration
- Re-arranging $x = (1 - y)\frac{n_t}{N_t - n_t}$



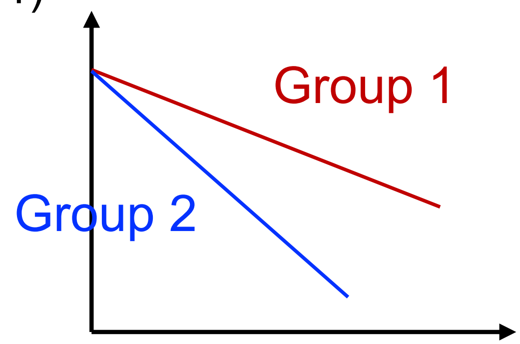
Proof Sketch

Notations

- N_t = total number of people in group t
- n_t = total number of positive people in group t

Another way to calculate total scores

- x = average score of a person in negative class
- y = average score of a person in positive class
- Total score in group t is $y(N_t - n_t) + xn_t = n_t$ by calibration
- Re-arranging $x = (1-y)\frac{n_t}{N_t - n_t}$
- To make sure x, y are the same for both groups, the two lines must intersect
  - Unless slopes are the same, only intersect at (0,1)



Can Achieve Two Definitions

- "Equality of Opportunity in Supervised Learning [NeurIPS'16]"
  - Can achieve balance of positive and negative class, but no requirement for calibration
  - Objective: find most accurate prediction subject to fairness constraints
- "On Fairness and Calibration [NeurIPS'17]"
  - Can achieve calibration and any linear combination of balance of positive and negative class

Similar Negative Results

"Fair prediction with disparate impact: A study of bias in recidivism prediction instruments"

"Algorithmic decision making and the cost of fairness"

- Show similar negative results, but for classification

Happy Thanksgiving