Fairness in Machine Learning: Part I

Privacy & Fairness in Data Science

CS848 Fall 2019

Outline

- High Level View
  - Recap Supervised Learning: Binary Classification
  - Survey of Approaches to Fairness in Supervised Learning

- Warmup: Fairness Through Awareness (Dwork, et. al, ITCS 2012)
  - Definitions
  - Linear Programming and Differential Privacy

- Certifying and Removing Disparate Impact (Feldman et. al, KDD 2015)
  - Definitions
  - Certifying Disparate Impact
  - Removing Disparate Impact
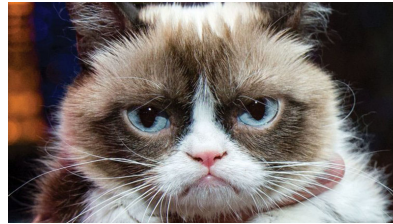 - Limitations

High Level View
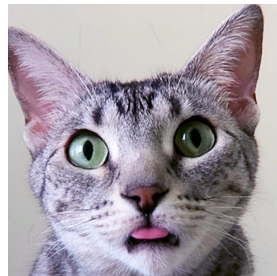
Binary Classification

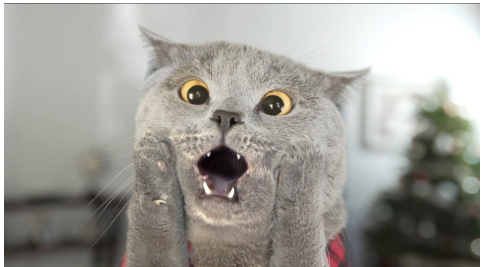- Suppose we want a cat classifier. We need **labeled training data**.

= cat

= cat

= cat

!= cat

Binary Classification

- We learn a **binary classifier**, which is a function *f* from the input space (pictures, for example) to a **binary class** (e.g., 1 or 0).

- To classify a new data point, apply the function to make a prediction. Ideally, we get:

- *f* = 1.

Justice

- Fact: sometimes we make errors in prediction. So what?
- In the cases we consider, prediction = judgement, and impacts lives of real people. (in binary classification, one is a good judgement, one bad.)
  - Recidivism prediction for granting bail
  - Predicting credit worthiness to give loans
  - Predicting success in school/job to decide on admission/hiring
- Big question of justice: are people being treated as they deserve?
  - ("Justice is the constant and perpetual wish to render every one their due." – Corpus Juris Civilis, Codex Justinianus, 534).
- This seems hard. Potentially any error is an injustice to that person.

Fairness

- Smaller question of fairness: are people being treated equally?
- Is our classifier working as well for black cats as white cats?

- Accompanying question: what is the relevant sense of "treated equally?"

Survey of Approaches to Fairness in Supervised Learning

- Individual Fairness
  - Fairness Through Awareness: Similar individuals should be treated similarly.
- Group Fairness: Statistical Parity
  - Disparate Impact: We should make predictions at the same rate for both groups.
  - Equality of Opportunity: We should make predictions at the same rate for both groups, conditioned on the ground truth.
  - Predictive Value Parity: Of those we predicted as 1, the same fraction should really be 1's (ground truth) for both groups.
- Causal Inference
  - We should make the same prediction in a counterfactual world where the group membership is flipped.

Outline

- ~~High Level View~~
  - ~~Recap Supervised Learning: Binary Classification~~
  - ~~Survey of Approaches to Fairness in Supervised Learning~~

- Warmup: Fairness Through Awareness (Dwork, et. al, ITCS 2012)
  - Definitions
  - Linear Programming and Differential Privacy

- Certifying and Removing Disparate Impact (Feldman et. al, KDD 2015)
  - Definitions
  - Certifying Disparate Impact
  - Removing Disparate Impact
  - Limitations

Fairness Through Awareness

- What does it mean to be fair in binary classification?

- According to Fairness Through Awareness: **Similar data points should be classified similarly**.

- In pictures, it's unfair to classify   as a cat, but
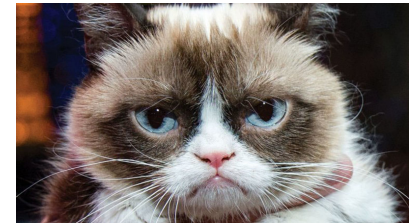
classify as not a cat.

Fairness Through Awareness

- We have a set V of data points. Let $C = \{0, 1\}$ be a binary class. Let $t(x)$ be the true binary class of x in V.

- Let $f: V \to \Delta C$ be a **randomized classifier**, where ΔC is the set of distributions over C.

- We have two notions of "distance" given as input.
  - $d: V \times V \to [0, 1]$ is a measure of distance between data points.
    - Assume $d(x, y) = d(y, x) \geq 0$ and $d(x, x) = 0$.
  - $D: \Delta C \times \Delta C \to \mathbb{R}$ is a measure of distance between distributions.
    - E.g., total variation distance $D_{TV}(X, Y) = \frac{|X(0) - Y(0)| + |X(1) - Y(1)|}{2}$

- f is fair if it satisfies the **Lipshitz condition**:
$$\forall x, y \in V, \qquad D\big(f(x), f(y)\big) \leq d(x, y).$$

Fairness Through Awareness

- **Claim.** There always exists a fair classifier.
- **Proof.** Let f be a constant function. Then

$$\forall x, y \in V, \qquad D\big(f(x), f(y)\big) = 0. \; \square$$

Fairness Through Awareness

- **Claim**. Assuming _____, the only fair deterministic classifier is a constant function.

- **Proof**. Assume there exist data points x and y with $d(x, y) < 1$ and $t(x) \neq t(y)$.

- If f is fair, then $D\big(f(x), f(y)\big) < 1$. Since f is deterministic, $D\big(f(x), f(y)\big) \in \{0,1\}$, so it must be that $D\big(f(x), f(y)\big) = 0$. $\square$

- **Corollary** (loosely stated)…
  - Deterministic classifiers that are fair in this sense are useless.

- Make you think of differential privacy?

Fairness Through Awareness

- To quantify the utility of a classifier, we need a loss function. For example, let $L(f, V) = \frac{1}{|V|} \sum_{x \in V} |\mathbb{E}[f(x)] - t(x)|$.

- Then the problem we want to solve is:

$Min. \quad L(f, V)$
$s.t. \quad D\big(f(x), f(y)\big) \leq d(x, y) \quad \forall x, y \in V$

- Can we do this efficiently?

Fairness Through Awareness

- We can write a linear program!

$$Min. \quad \frac{1}{|V|} \sum_{v \in V} |z_1(x) - t(x)|.$$

$$s.t. \quad \frac{|z_0(x) - z_0(y)| + |z_1(x) - z_1(y)|}{2} \leq d(x, y) \quad \forall x, y \in V$$

$$z_0(x) + z_1(x) = 1 \quad \forall x \in V$$

Fairness Through Awareness: Caveats

- Where does the distance metric d come from?
  - Note that for any classifier f, there exists d such that f is fair.
  - d might actually be **more difficult** to learn accurately than a good f!

- f is only fair *ex ante*, and this is necessary.

- Fairness in this sense makes no promises of group parity.
  - If individuals of one racial group are, on average, a large distance from those of another, a "fair" algorithm is free to discriminate between the groups.
 - For more on this, see sections 3 and 4 of the paper.

Outline

- High Level View
  - Recap Supervised Learning: Binary Classification
  - Survey of Approaches to Fairness in Supervised Learning

- Warmup: Fairness Through Awareness (Dwork, et. al, ITCS 2012)
  - Definitions
  - Linear Programming and Differential Privacy

- Certifying and Removing Disparate Impact (Feldman et. al, KDD 2015)
  - Definitions
  - Certifying Disparate Impact
  - Removing Disparate Impact
  - Limitations

Recap: Disparate Impact

- Suppose we are contracted by Waterloo admissions to build a machine learning classifier that predicts whether students will succeed in college. For simplicity, assume we admit students who will succeed.

Gender	Age	GPA	SAT
0	19	3.5	1400
1	18	3.8	1300
1	22	3.3	1500
1	18	3.5	1500
…	…	…	…
0	18	4.0	1600

Succeed
1
0
0
1
…
1
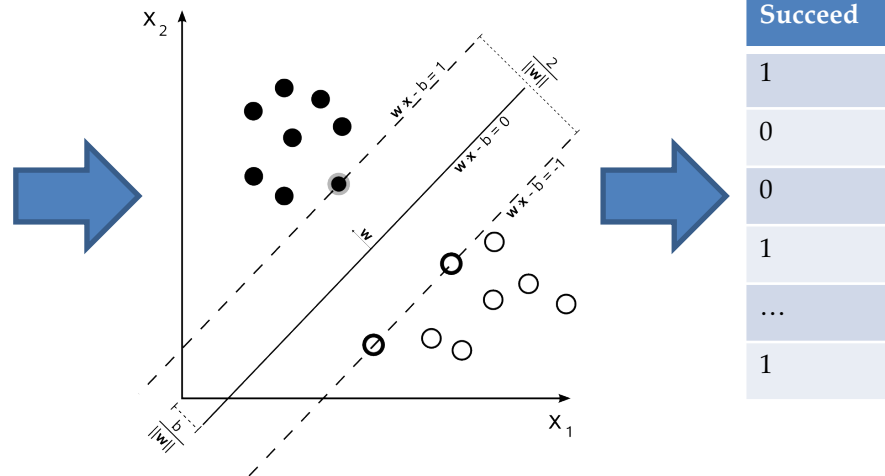
Recap: Disparate Impact

- Let $D=(X, Y, C)$ be a labeled data set, where $X = 0$ means protected, $C = 1$ is the positive class (e.g., admitted), and Y is everything else.

- We say that a classifier f has **disparate impact (DI) of τ** ($0 < \tau < 1$) if:

$$\frac{\Pr(f(Y) = 1 \mid X = 0)}{\Pr(f(Y) = 1 \mid X = 1)} \leq \tau$$

that is, if the protected class is positively classified less than τ times as often as the unprotected class. (legally, $\tau = 0.8$ is common).

Recap: Disparate Impact

- Why this measure?

- Arguably the only good measure if you think the **data** are biased **and** you have a strong prior belief protected status is uncorrelated with outcomes.

  - E.g., if you think that the police *target* minorities, and thus they have artificially higher crime rates because your data set isn't a random sample.

- "In Griggs v. Duke Power Co. [20], the US Supreme Court ruled a business hiring decision illegal if it resulted in disparate impact by race even if the decision was not explicitly determined based on race. The Duke Power Co. was forced to stop using intelligence test scores and high school diplomas, qualifications largely correlated with race, to make hiring decisions. The Griggs decision gave birth to the legal doctrine of *disparate impact*…" (Feldman et. al, KDD 2015).

Certifying Disparate Impact

- Suppose you are given $D = (X, Y, C)$.

- Can we verify that a new classifier learned on Y aiming to predict C will **not** have disparate impact with respect to X?

- **Big idea**: A classifier learned from Y will not have disparate impact if X cannot be predicted from Y.

- Therefore, we can check a data set itself for possible problems, even without knowing what algorithm will be used.

Certifying Disparate Impact – Definitions

- **Balanced Error Rate**: Let $g: Y \rightarrow X$ be a predictor of the protected class. Then the balanced error rate is defined as

$$BER(g(Y), X) = \frac{\Pr(g(Y) = 0 \mid X = 1) + \Pr(g(Y) = 1 \mid X = 0)}{2}$$

- **Predictability**: D is ϵ-predictable if there exists $g: Y \rightarrow X$ with $BER(g(Y), X) \leq \epsilon$.

Certifying Disparate Impact – Characterization

- **Theorem (simplified).** If $D = (X, Y, C)$ admits a classifier f with disparate impact 0.8, then D is is $(1/2 - B/8)$-predictable, where $B = \Pr(f(Y) = 1 \mid X = 0)$.

- **Proof sketch.** → Suppose D admits a classifier $f: Y \to C$ with disparate impact 0.8.

- Use f to predict X.

- If f positively classifies an individual, predict they are not in the protected class, otherwise predict that they are in the protected class.

Certifying Disparate Impact – Characterization

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

$$= \frac{1 - \Pr(f(Y) = 1 \,|\, X = 1) + B}{2}$$

$$\leq \frac{1 - \Pr(f(Y) = 1 \,|\, X = 0)/0.8 + B}{2}$$

$$= \frac{1}{2} - \frac{B}{8}$$

Certifying Disparate Impact

- Disparate impact is related to predictability. So what?
- Given D, we estimate:
    1. The predictability (call it ϵ) of D.
    2. B, the fraction of class X=0 predicted to have outcome 1.
- This yields an estimate on the *possible* disparate impact of *any* classifier built on D.
- How do we get these estimates?
    1. Use an SVM to predict X from Y while minimizing BER.
    2. The empirical estimate from D.
- That's a lot of estimation! How does it work in practice?
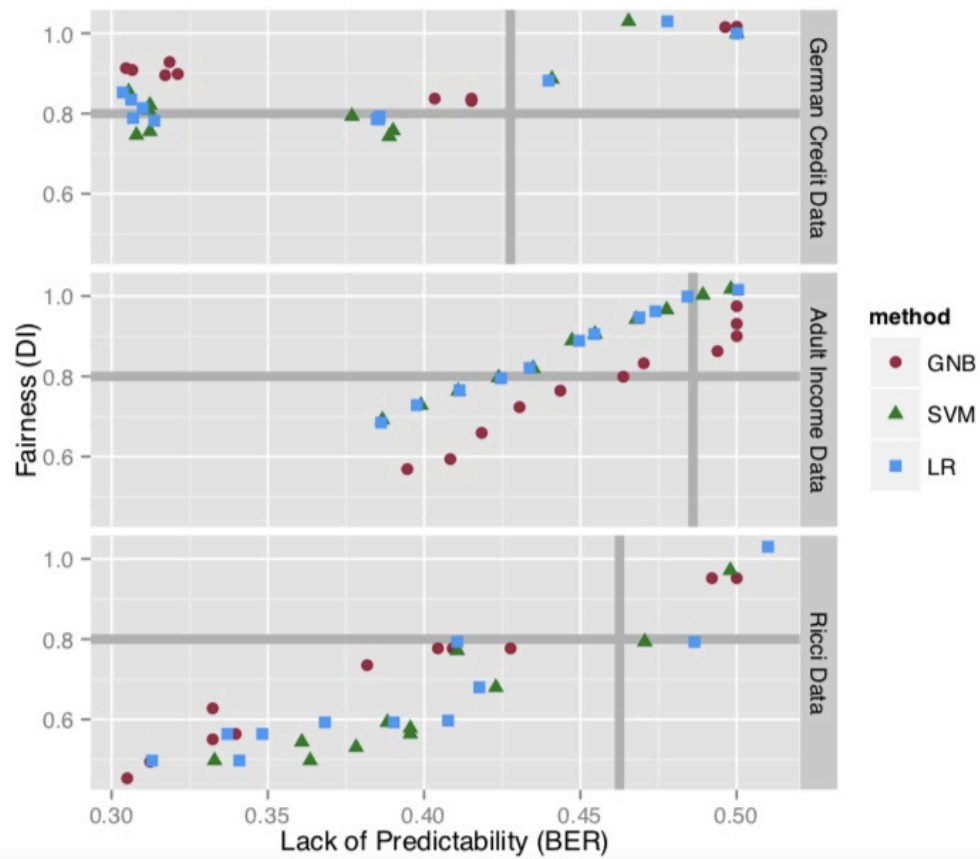
Removing Disparate Impact

- Suppose we find that X and Y **do** admit disparate impact. What do we do?

- Can we define a "repair" protocol that works the same way, on a data itself, without even needing to know the labels?

- We want to change D so that it is no longer predictable. How can we do this?

- Formally, given (X, Y), we want to construct a repaired data set $(X, \bar{Y})$ such that for all $g : Y \rightarrow X$, $BER(g(Y), X) > \epsilon$, where ϵ depends on the strength of guarantee we want.

Removing Disparate Impact

- For simplicity, suppose that Y is a single well ordered numerical attribute like SAT score.

- **Claim.** Perfect Repair is always possible.

- **Proof.** Just set Y to 0 for every individual.

- Recall that $BER(g(Y), X) = \dfrac{\Pr(g(Y)=0 \mid X=1) + \Pr(g(Y)=1 \mid X=0)}{2}$

- Then on the repaired data, the balanced error rate of **any** classifier is ½, which is the maximum possible balanced error rate. □
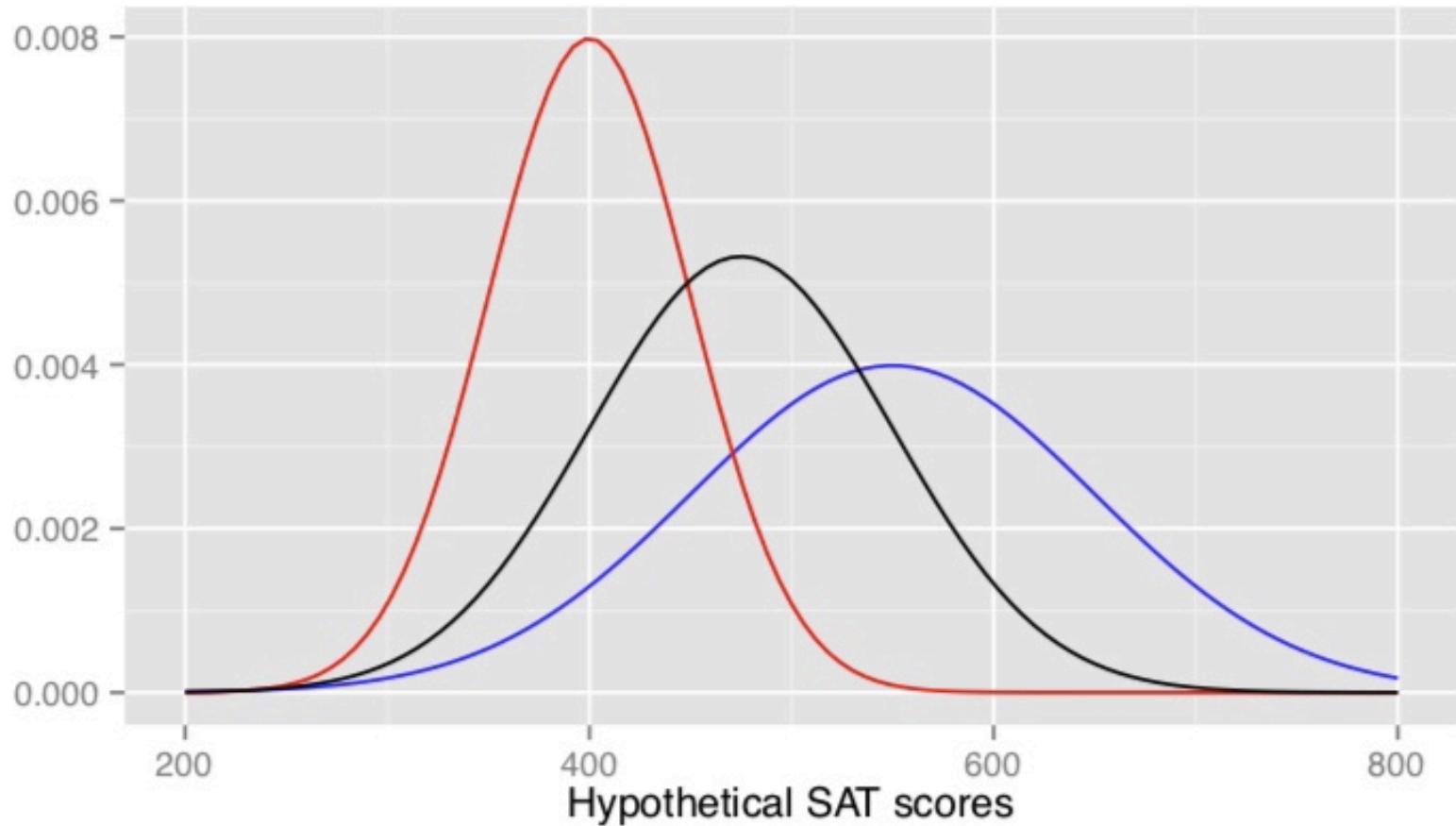
Removing Disparate Impact

- We would like a smarter way, that preserves the ability to classify accurately.

- More specifically, we want to transform Y in a way that preserves rankings within the protected group and within the nonprotected group (but not necessarily across).

- Ideally, this leads to a smooth transformation that still allows us to perform reasonably accurate classification. How?

Removing Disparate Impact

- Assume we have a single well ordered numerical attribute and that the protected and unprotected groups have equal size.

- **Algorithm.**

  - Let p_y^x be percentage of agents with protected status x whose numerical score is at most y.

  - Take a data point (x_i, y_i). Calculate $p_{y_i}^{x_i}$.

  - Find y_i^{-1} such that $\left(p_{y_i^{-1}}^{1-x_i} \right) = p_{y_i}^{x_i}$.

  - Repair $\bar{y}_i = median(y_i, y_i^{-1})$.

- The algorithm is easier to draw than to write. Once you understand it, the proof that it preserves rank and is not predictable is obvious.
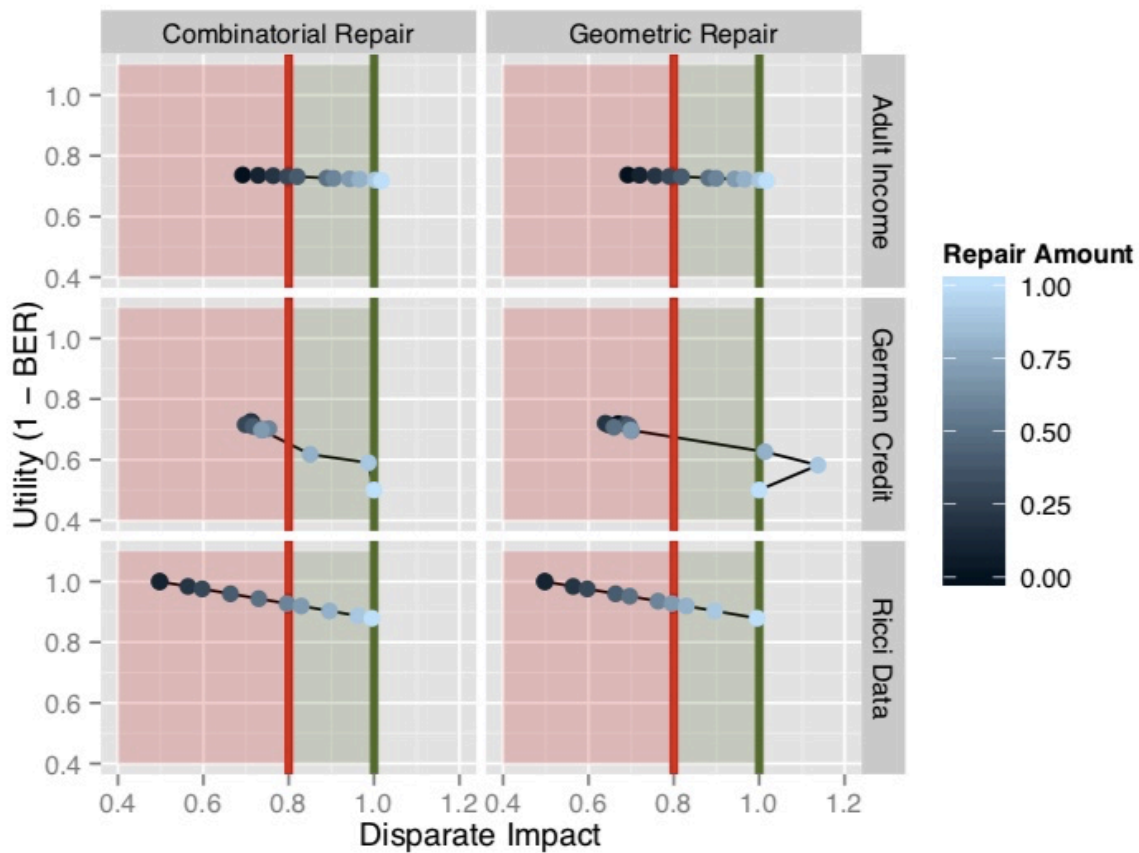
Removing Disparate Impact

Removing Disparate Impact

- If Y is more than just one attribute, Feldman et. al repair each attribute individually.

- The same basic idea can be extended for a partial repair algorithm, that still allows some disparate impact, but modifies the data less.

- Of course, preserving rank doesn't guarantee that the resulting dataset can still be used to train good classifiers. Here's what Feldman et. al observe in practice on their experiments.

Disparate Impact – Limitations

- Typically forbids the "perfect" classifier.
- Allows "laziness." For example, here is a disparate impact free classifier:
  - Accept the top 50% (by SAT score) of men who apply
  - Accept a random sample of 50% of the women who apply.
- Arguably this is a biased classifier, but it doesn't have disparate impact.
- It also assumes that there is not a fundamental difference between the two groups. If that assumption isn't true, disparate impact might not make sense, and could be viewed as "anti-meritocratic."

Conclusion

- We saw an approach based on differential privacy for providing optimal utility subject to *individual* fairness.

  - But this had limitations: in particular, it's not clear where the distance metric on individuals comes from.

- We saw an approach based on the predictability of the sensitive attribute for certifying and removing disparate impact - a measure of equality of outcomes.

- Next section, we will consider consider a different approach: equality of opportunity, rather than outcomes.