Fairness in Machine Learning: Part 1

Wednesday, September 19, 2018

CompSci 590: Privacy and Fairness in Data Science

Outline

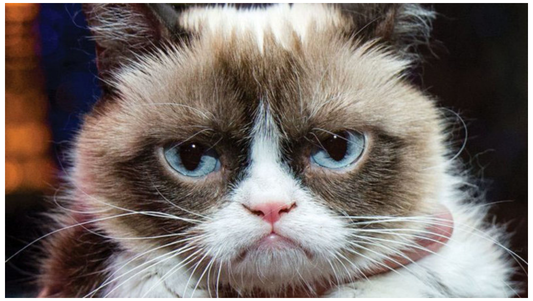
- Warmup: Fairness Through Awareness (Dwork, et. al, ITCS 2012)
  - Recap: Binary Classification
  - Linear Programming and Differential Privacy

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

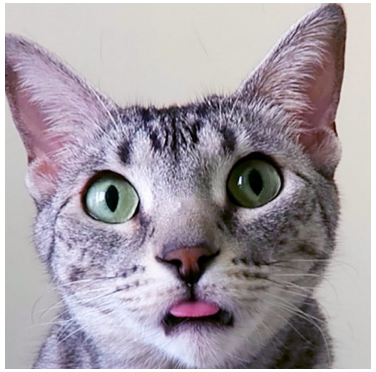
Binary Classification

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

= cat



= cat



= cat

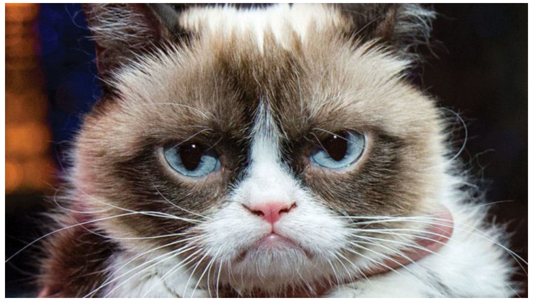
!= cat

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.

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 $D\big(f(x), f(y)\big) = 0$. □
- **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:

$$\begin{array}{ll} Min. & L(f,V) \\ s.t. & D\big(f(x),f(y)\big) \le d(x,y) \quad \forall x,y \in V \end{array}$$

- 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

- f is only fair ***ex ante***.
  - f makes a promise about your distribution; **Ex post**, you might receive an arbitrarily unfair draw from that distribution. As we saw, this is necessary.

- 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!

- 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.

Outline

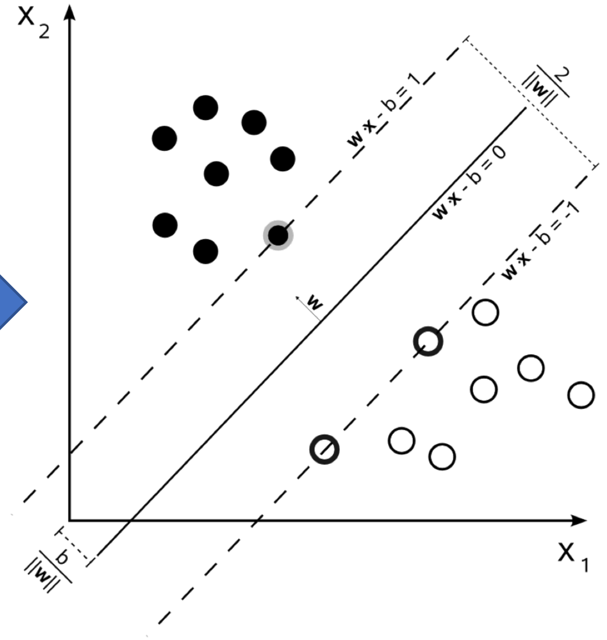
- ~~Warmup: Fairness Through Awareness (Dwork, et. al, ITCS 2012)~~
  - ~~Recap: Binary Classification~~
  - ~~Linear Programming and Differential Privacy~~

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

Recap: Disparate Impact

- Suppose we are contracted by Duke 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 a classifier *f* has **disparate impact (DI)** τ (0 < τ < 1) if:

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

that is, if the protected class is positively classified less than τ times as often as the unprotected class. (legally, τ = 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 don’t know what machine learning algorithms someone will use to build a classifier, but you get to see *X* and *Y*.
- Can we verify that a classifier on *Y* will **not** have disparate impact with respect to *X*?
- Yes! But how? Disparate impact is defined in terms of *C* (which we don’t know), so how can we search for high DI classifiers?
- **Big idea**: A classifier learned from *Y* will not have disparate impact if *X* cannot be predicted from *Y*.
- This is exciting because it means we can check a data set itself for possible problems, even without knowing the labels.

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

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

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

Certifying Disparate Impact – Characterization

- **Theorem.** *D* is (1/2 – B/8)-predictable if and only if it admits a classifier with disparate impact 0.8, where B is the fraction of fraction of data points with X=0 that are classified as C=1.
- **Proof.** → Suppose *D* has disparate impact 0.8.
- Then *D* itself gives a function $f\colon Y \to C$ that has disparate impact 0.8.
- The predictor of *X* from *Y* is simple: just use *f*!
- 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 – Characterization

- ← Suppose *D* is (1/2 – B/8)-predictable.
- Then we have a predictor $g\colon Y \to X$ with $BER(g(Y), X) \leq$ (1/2 – B/8).
- We will classify using f.

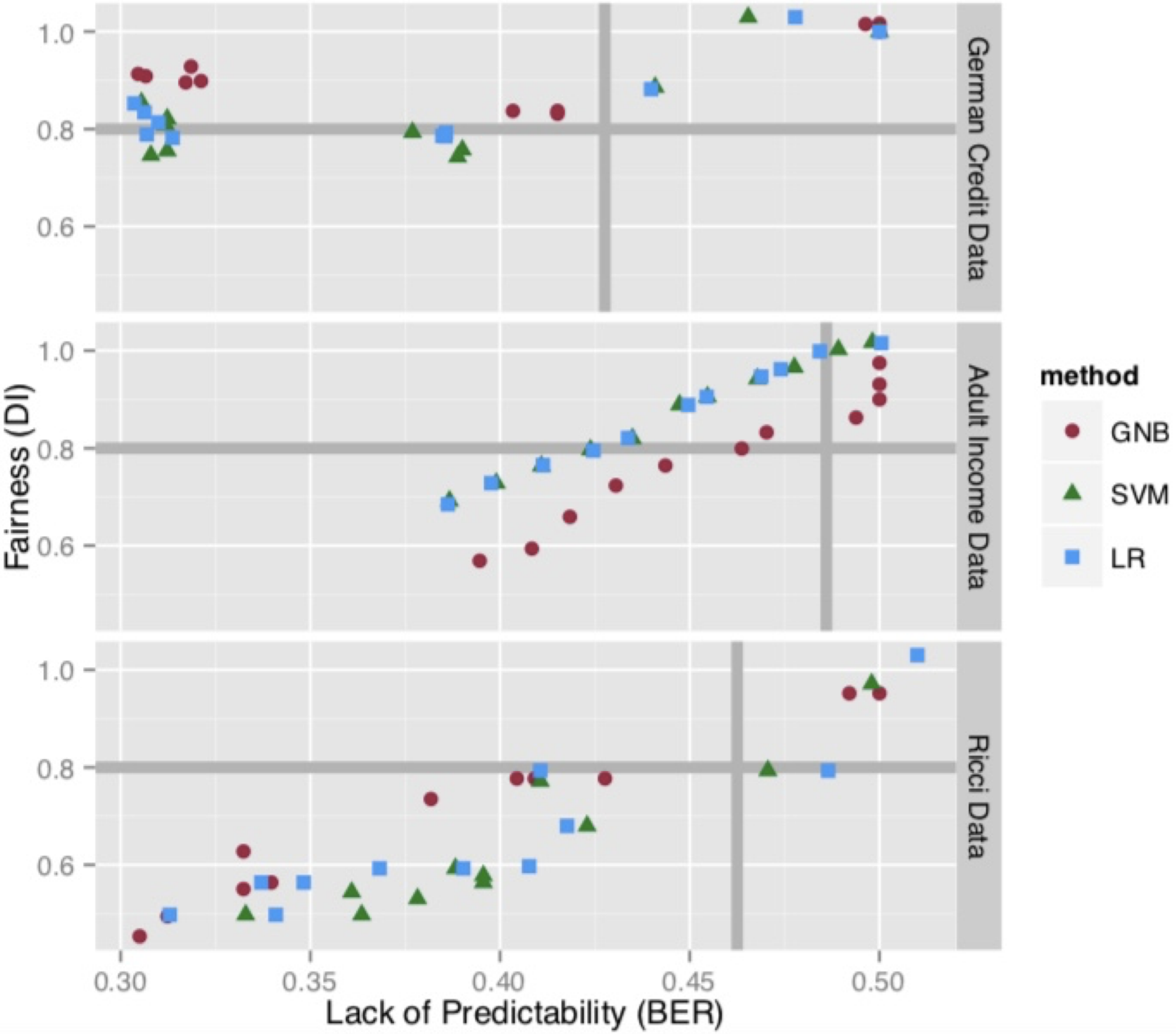
$$DI(\mathrm{g(Y), C}) = \frac{\Pr(g(Y) = 1 \mid X = 0)}{\Pr(g(Y) = 1 \mid X = 1)}$$

$$= \frac{B}{B + 1\ - 2BER(g(Y), X)}$$

$$\leq \frac{B}{\frac{5B}{4}} = 0.8.\ \square$$

Certifying Disparate Impact

- Disparate impact related to predictability. So what?
- Given *D*, we estimate:
    1. The predictability (call it ϵ) of *D*.
    2. B, that is, the fraction of the protected class with C=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?


German Credit Data
Adult Income Data
Ricci Data
Fairness (DI)
1.0
0.8
0.6
Lack of Predictability (BER)
0.30
0.35
0.40
0.45
0.50
method
● GNB
▲ SVM
■ LR

Removing Disparate Impact

- Suppose we find that X and Y **do** admit of disparate impact. What do we do?
- Can we define a "repair" protocol that works the same way, on the training 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 $g: Y \to X$, $BER(g(Y), X) > \epsilon$, 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 $BER(g(Y), X) = \frac{\Pr(g(Y)=0 \mid X=1) + \Pr(g(Y)=1 \mid X=0)}{2}$
- 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).

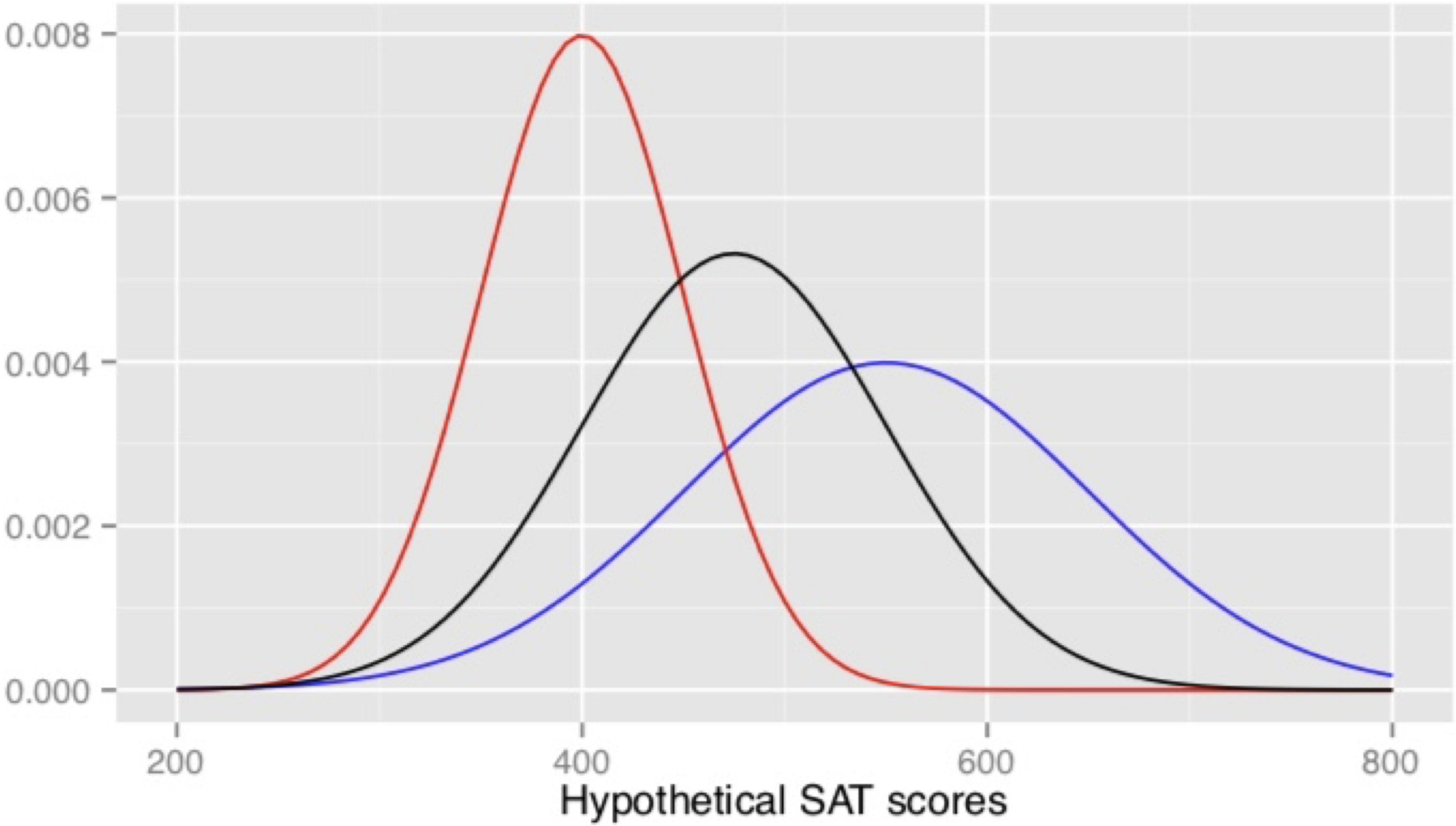
- Ideally, this leads to a smooth transformation that still allows us to perform reasonably accurate classification. How?

Removing Disparate Impact

- **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 $\overline{y_i} = median(y_i, y_i^{-1})$.

- The algorithm is easier to draw than to explain, and once you understand it, the proof that it preserves rank and is not predictable is obvious.

Removing Disparate Impact

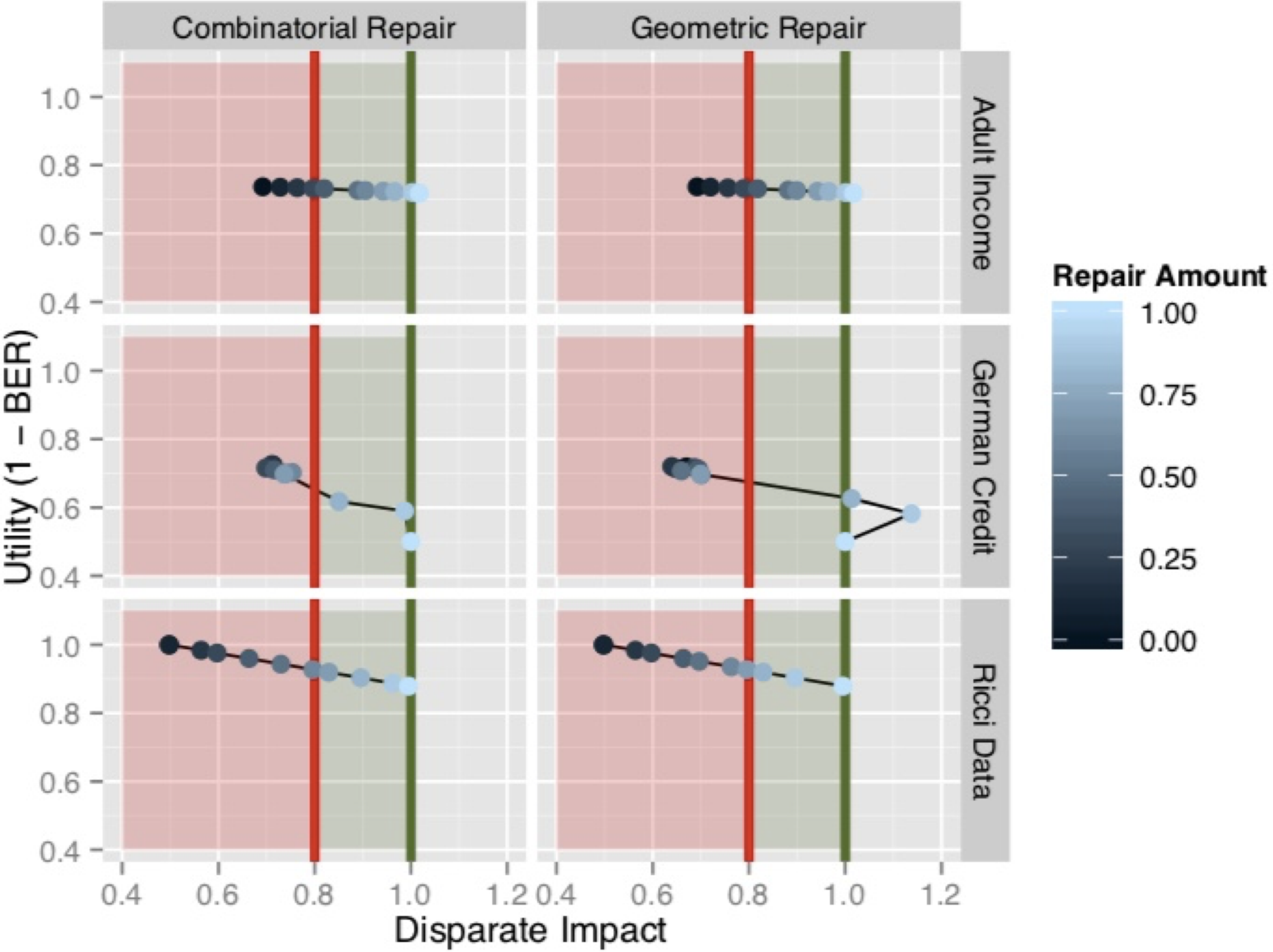


Removing Disparate Impact

- 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.


Combinatorial Repair
Geometric Repair
Adult Income
German Credit
Ricci Data
Utility (1 − BER)
1.0
0.8
0.6
0.4
Disparate Impact
0.4
0.6
0.8
1.0
1.2
Repair Amount
1.00
0.75
0.50
0.25
0.00

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 week, we will consider consider a different approach: equality of opportunity, rather than outcomes.