Fairness and Machine Learning

Machine Learning Theory (MLT)

Edinburgh

Rik Sarkar

Fairness issues and examples

- Data is biased
  - All present employees belong to a certain group or demographic
  - Recruitment software learns bias toward the group (even when the group is not explicitly stated in applications)
- Data is insufficient
  - Data on a small group is sparse
  - Accuracy on the group is low
- Fairness is hard to define

Typical considerations

- Entities:
  - Individuals and differences across individuals
  - Groups

- ML outcome
  - A decision where the positive outcome is clear, e.g. Admission, loan approval etc
  - Accuracy of prediction of an ML algorithm, e.g. Correct recommendation, correct classification decision for the person/group

Fairness: Hard to achieve and hard to measure

- Consider the admissions case:
  - Group A has higher test scores
  - Group B has lower test scores

  - Who should get admission offers?

Fairness: Hard to achieve and hard to measure

- Consider the admissions case:
  - Group A has higher test scores
  - Group B has lower test scores

  - Who should get admission offers?

- What if:
  - A has access more resources (money), better schools, better healthcare, better test prep
  - Who should get the admission offers?

- Problems:
  - What is fair in one situation may not be in another
  - What is fair in one perspective (objective) may not be in another

- Many types definitions of fairness, often not mutually compatible

Bias/fairness concepts

- Fairness through unawarerness
- Group fairness
- Calibration
- Error rate balance
- Representational fairness
- Counterfactual fairness
- Individual fairness

Different Definitions

- Fairness through unawareness:
  - Don’t record protected/sensitive attributes. Don’t use them
  - Proxies exist: name, address, school…
- Group fairness
  - Same prediction rates for all groups
  - E.g. Same percentage of each group should be admitted
  - Different groups may have different base rates (qualifications). Can lead to high errors
- Calibration
  - Probability of positive prediction same for same score and different groups
  - Function of a score (e.g. risk score)

- Error rate balance
  - Equal false positive rates across groups
  - Incompatible with calibration
- Representational fairness
  - Learn latent representation Z to minimize group information
  - Loses information, and accuracy
- Counterfactual fairness
  - Group A should not cause prediction Y
  - Assumes we know the biases
- Individual fairness
  - Similar individuals should be treated similarly
  - Nearby points are likely to be classified similarly

An impossibility result

- Suppose we are making classifications based on a score S, and there are 2 groups A and B
- Then the following are three possible notions of fairnesss
  - Calibration
    - Probability of positive prediction same for same score and different groups
  - Balance for the positive class
    - Average of scores for positive class in A = Average of scores for positive class in B
  - Balance for the negative class
    - Average of scores for negative class in A = Average of scores for negative class in B
- Three desirable properties, but impossible to achieve all except for in special cases [Kleinberg 2016]

Fairness in binary classification

- Individual fairness: Two people with similar features are treated similarly
  - If their feature values are similar, the function f making decision outputs should produce similar values
- Idea: Similar data points should be treated similarly
  - Remember we have seen Lipschitz functions before: if $|x - y|$ is small, then $|f(x) - f(y)|$ must be small
- Plan: extend this to fairness for classifiers

Individual fairness

- Suppose, V is data points, $C = \{0,1\}$ is set of classes, $t(x)$ is the true class of any x
- $f: V \to \Delta C$ is a randomized classifier, where ΔC is the set of distributions over C
  - $f(x)$ assigns probabilities of x belonging to each class
  - E.g. $f(x) = (0.2, 0.8)$ x belongs to class 0 with prob. 0.2, to class 1 with probability 0.8
- Distance functions to define Lipschitzness:
  - Distance between points $d: V \times V \to [0,1]$
  - Distance between output distributions $D: \Delta C \times \Delta C \to \mathbb{R}$
- $\boldsymbol{f}$ **is fair if it is Lipschitz**: $\forall x, y \in V,\ D\big(f(x), f(y)\big) \leq d(x, y)$

There is a simple fair classifier!

- Proof:
  - Take a constant f e.g. $f = (0.5, 0.5)$
  - Then always, $D\big(f(x), f(y)\big) = 0$

- But this classifier is not very useful!

Randomization is necessary

- The only fair deterministic classifier is constant function
- Proof sketch:
  - A deterministic classifier f outputs values in $\{0, 1\}$, therefore $D\big(f(x), f(y)\big) \in \{0,1\}$
  - Suppose x, y belong to different classes
  - If f is fair, then $D\big(f(x), f(y)\big) \leq 1$, Therefore $D\big(f(x), f(y)\big) = 0$

- A useful individually fair classifier must be randomized
- Reminiscent of differential privacy.

Utility

- Outputs $(0.5, 0.5)$ of a deterministic classifier are not useful

- A randomized classifier is better but just $\forall x, y \in V,\ D\big(f(x), f(y)\big) \leq d(x, y)$ is not sufficient.
  - E.g. $D\big(f(x), f(y)\big) = (0.5{,}0.5)$ satisfies this condition, but not useful
- We also need to maximize utility
  - E.g. If $t(x)$ is the true classification of x, then expected value of $f(x)$ should be as close to $t(x)$ as possible.
  - E.g. Maximize : $L(f, V) = \frac{1}{|V|}\sum_{x \in V} |\mathbb{E}[f(x)] - t(x)|$

Comments

- The fairness is in terms of the distribution.
  - A particular classification can be unfair due to probabilistic nature
- Does not guarantee anything about groups
  - If groups have different feature values, individual fairness can still be unfair to groups

Group fairness: Disparate impact

- Suppose we are building a classifier for college admission
- $D = (X, Y, C)$ is a labelled dataset $C = 1$ means admitted (positive class). X is a protected feature (e.g. disability) and $X = 0$ means protected. Y is all other features.
- Classifier f has disparate impact (DI) τ $(0 < \tau < 1)$ if:
  - $\frac{\Pr[f(Y)=1|X=0]}{\Pr[f(Y)=1|X=1]} \leq \tau$

- If the protected class is positively classified less than τ times as often as the unprotected class ($\tau = 0.8$ is a common threshold)

Disparate impact

- Suitable if the protected/unprotected status are uncorrelated to true value or qualification
- But the data may be biased
  - So we need to measure the disparate impact of data
- Has been legally used in US to block hiring decisions with disparate impact. Stopped using test scores that are correlated with race

Certifying disparate impact

- Suppose we have the dataset
- How can we verify that classification based on Y will not lead to DI?
- The problem is that we may not know what classifier C will be used.
- So, we need to measure DI for the data irrespective of classifier.
- Idea:
  - Any classifier C will not have disparate impact if X cannot be predicted from Y

- So, we can check a dataset for DI to see if the X labels can be predicted

Definitions

- **Balanced error rate (BER).**
- Suppose $g: Y \rightarrow X$ predicts protected group status, then
  - $BER(g(Y), X) = \frac{\Pr(g(Y)=0|X=1)+\Pr(g(Y)=1|X=0)}{2}$

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

Disparate impact characterization

- Suppose B is the fraction of data points with $X = 0$ that are classified as $C = 1$
  - I.e. fraction of data points that constitute the positive examples in the protected class
- Theorem: D is $\left(\frac{1}{2} - \frac{B}{8}\right)$-predictable if and only if it admits a classifier with DI = 0.8

- Proof. (if)
  - Suppose there is a classifier f with DI=0.8
  - We can use f as the predictor g for the protected class
  - When f gives positive classification, g predicts unprotected group, when f gives negative, g predicts protected group

- $BER(g(Y), X) = \frac{\Pr(f(Y)=0|X=1)+\Pr(f(Y)=1|\ X=0)}{2}$
- $= \frac{1-\Pr\left(f(Y)=1\middle|X=1\right)+B}{2}$
- $\leq \frac{1-\Pr\left(f(Y)=1\middle|X=0\right)/0.8+B}{2}$
- $= \frac{1}{2} - \frac{B}{8}$

- Proof (only if)
- Suppose D is (1/2 – B/8)-predictable
- Then there is a g with $BER(g(Y), X) \leq (\frac{1}{2} - \frac{B}{8})$

- We use g for classification
  - $DI(g(Y), C) = \frac{\Pr(g(Y)=1|X=0)}{\Pr(g(Y)=1|X=1)} = \frac{B}{B+1-2BER(g(Y),X)} \leq \frac{B}{\frac{5B}{4}} = 0.8$

Certifying disparate impact in practice

- We can estimate predictability of D
  - E.g. by training some standard classifiers to predict X from Y
- The fraction B can be estimated from training data.
- Thus we can get an estimate of disparate impact

Removing disparate impact

- Suppose we decide that a dataset has disparate impact
- Can we remove it? i.e. repair the data to eliminate DI?

- We need to change D so that X is no longer predictable

- We want a modified $(X, \bar{Y})$ so that $BER(g(\bar{Y}), X) > \epsilon$

Removing DI

- Suppose Y is a simple number like exam score
- Trivial solution: set $Y = 0$ for everyone!
  - Check that BER=1/2
- We need a better algorithm

Removing disparate impact

- Algorithm
- Let p_y^x be fraction of people with protected status x, with score at most y
- Take data point (x_i, y_i). Calculate $p_{y_i}^{x_i}$
- Find y_i^{-1}such that $p_{y_i^{-1}}^{1-x_i} = p_{y_i}^{x_i}$
- Repair $\overline{y_i} = median(y_i, y_i^{-1})$
- Equalising people from different groups with same rank.

- The algorithm preserves rank within groups. [Feldman et al. Certifying and removing disparate impact, 2015.]

- For multiple features, repeat the algorithm for each feature

- Limitations of DI
  - Usually does not allow perfect classifiers
  - Easy to build bad classifiers that satisfy DI criteria
  - Assumes the groups have same intrinsic merit. I.e. real qualification does not vary by group
    - May not be always true.

- Let us look at some other definitions

Demographic parity

- A classifier C satisfies demographic parity if C is independent of sensitive attribute A

- i.e. For groups a, b (distinguished by A) : $\Pr_a\{C = 1\} = \Pr_b\{C = 1\}$
- Approximate versions
  - $\frac{\Pr_a\{C=1\}}{\Pr_b\{C=1\}} \geq 1 - \epsilon$ or
  - $|\Pr_a\{C = 1\} - \Pr_b\{C = 1\}| \leq \epsilon$

- Sort of the reverse of DI
  - DI measures inequality
  - Parity measures equality

True positive parity (TPP)

- Equal opportunity
- For binary variables C and Y (labels)
- A classifier C satisfies TPP if for groups a, b:
  - $\Pr_{a}\{C = 1|Y = 1\} = \Pr_{b}\{C = 1|Y = 1\}$

False positive parity

- Similarly:
  - $\Pr_a \{C = 1|Y = 0\} = \Pr_b\{C = 1|Y = 0\}$

Equalized odds or positive rate parity

- Both TPP and FPP

Predictive Value Parity

- Positive Predictive Value Parity
  - $\Pr_a \{Y = 1 | C = 1\} = \Pr_b \{Y = 1 | C = 1\}$
- Negative Predictive Value Parity
  - $\Pr_a \{Y = 1 | C = 0\} = \Pr_b \{Y = 1 | C = 0\}$

- Classifier C satisfies predictive value parity if it satisfies both of the above.

PRP and PVP are incompatible

- With different base rates across groups, and a perfect classifier is not known, then either
  - Positive rate parity fails or
  - Predictive value parity fails

- Exercise for you: construct examples to show these

Summary

- Many types of definitions
- Demographic parity or disparate impact
  - Used in law
  - Does not allow perfect classification
  - Achieved by modifying training data
- Equal odds/opportunity
  - Perfect classification is possible
  - Different groups can get different rates of positive prediction
  - Achieved by post processing the classifer
  - Measure of the classifier being fair