AI & Ethics - Final Presentation

Analysis of Fair Learning without Sensitive Demographic Information

Team 2: Yuji Roh, HyungJun Yoon

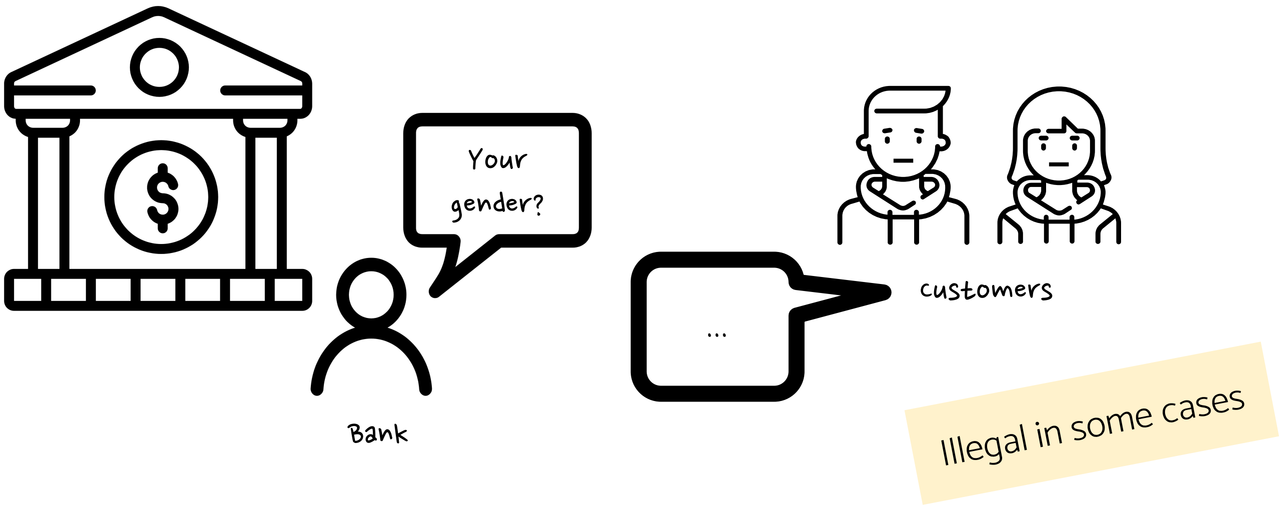
Introduction

Motivation

Related work

Motivation

- Problem
  - Most fairness studies assume that the protected demographics (e.g., gender, race) are available
  - However, this assumption may not be true in several real world applications



Main research question

How can we train a machine learning model to improve fairness
when **we do not know the protected group memberships** neither at training nor inference time?

Related work

- Fairness without demographics in repeated loss minimization [1]

- Fair learning with private demographic data [2]

- Fairness without demographics through adversarially reweighted learning [3]

[1] Hashimoto et al., ICML 2018 [2] Mozannar et al., ICML 2020 [3] Lahoti et al., NeurIPS 2020

Related work

- Fairness without demographics in repeated loss minimization [1]
  - **An initial work** to achieve fairness with assuming fully missing demographic data
  - Optimize any **worst-case distribution** using distributionally robust optimization (DRO)
- Fair learning with private demographic data [2]
- Fairness without demographics through adversarially reweighted learning [3]

[1] Hashimoto et al., ICML 2018 [2] Mozannar et al., ICML 2020 [3] Lahoti et al., NeurIPS 2020

Related work

- Fairness without demographics in repeated loss minimization [1]
- Fair learning with private demographic data [2]
  - **Privatize sensitive attributes** for fair model optimization
  - This strategy needs to access sensitive group information
- Fairness without demographics through adversarially reweighted learning [3]

[1] Hashimoto et al., ICML 2018 [2] Mozannar et al., ICML 2020 [3] Lahoti et al., NeurIPS 2020

Related work

- Fairness without demographics in repeated loss minimization [1]

- Fair learning with private demographic data [2]

- Fairness without demographics through adversarially reweighted learning [3]
    - The most recent work assuming fully missing demographic data
    - Focus on addressing **computationally-identifiable errors** on the data

[1] Hashimoto et al., ICML 2018 [2] Mozannar et al., ICML 2020 [3] Lahoti et al., NeurIPS 2020

Related work

- Fairness without demographics in repeated loss minimization [1]

- Fair learning with private demographic data [2]

Our Target Paper

- Fairness without demographics through adversarially reweighted learning [3]
  - The most recent work assuming fully missing demographic data
  - Focus on addressing **computationally-identifiable errors** on the data

[1] Hashimoto et al., ICML 2018 [2] Mozannar et al., ICML 2020 [3] Lahoti et al., NeurIPS 2020

Approaches

Original approach

Limitation of the original approach

Improved approach

Original approach: Computationally-identifiable region

- Toy example

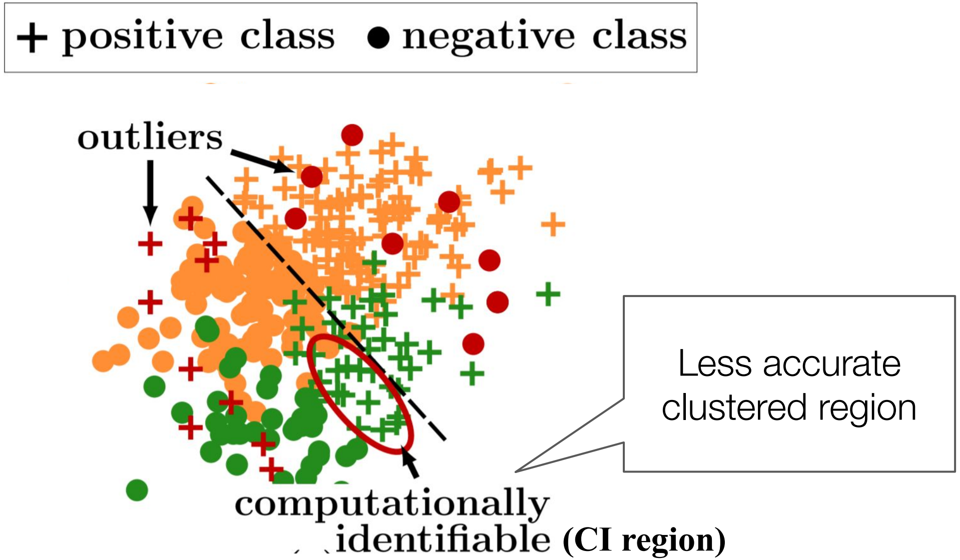


Figure 1: Computational-identifiability example

Original approach

- Adversarially reweighted learning (ARL)
  - Learner classifies the true label y
  - Adversary tries to find less accurate region and gives more weights on that region

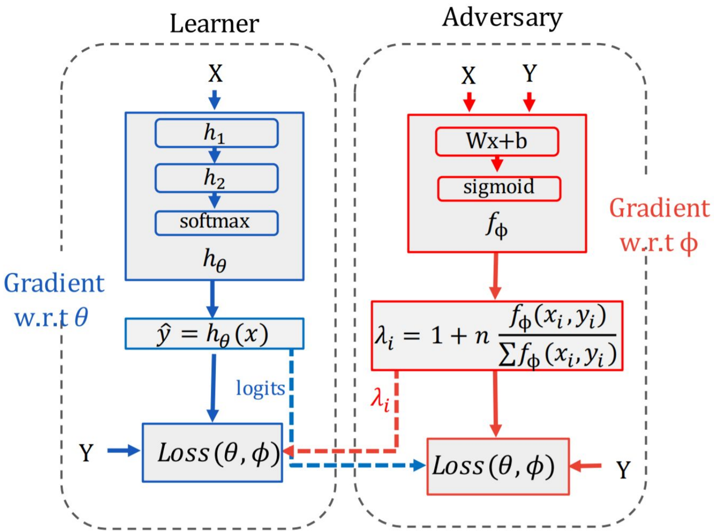


Figure 2: ARL Computational Graph

Original approach

- Adversarially reweighted learning
  - $$J(\theta, \lambda) := \min_{\theta} \max_{\lambda} L(\theta, \lambda) = \min_{\theta} \max_{\lambda} \sum_{s \in S} \lambda_s L_{\mathcal{D}_s}(h)$$
    $$= \min_{\theta} \max_{\lambda} \sum_{i=0}^{n} \lambda_{s_i} \ell(h(x_i), y_i)$$
  - The learner aims to **minimize** the objective
  - The adversary tries to **maximize** the objective

Two-player game

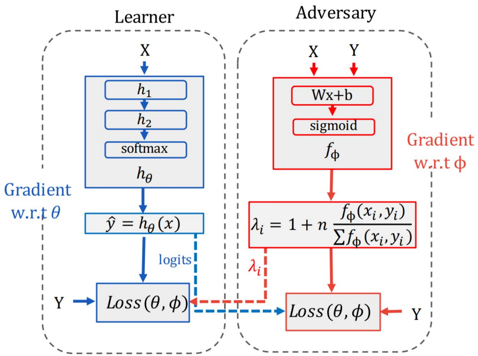


Figure 2: ARL Computational Graph

Limitation of the original algorithm

- The adversary has several limitations

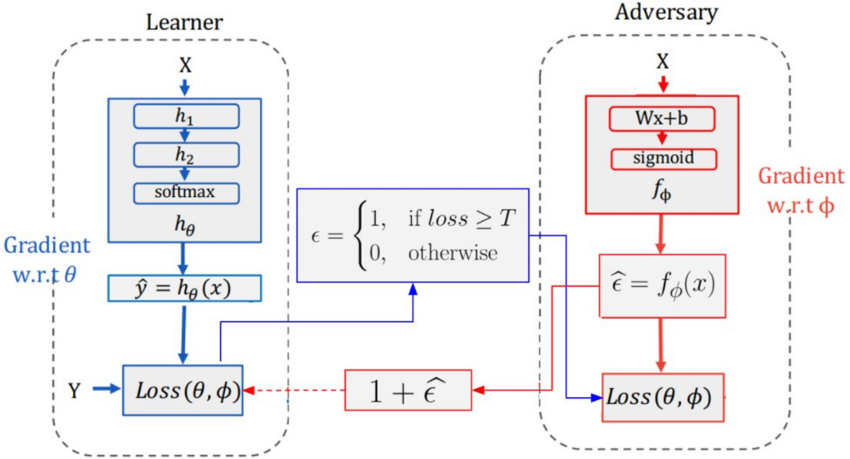
  ① Capability on identifying the less accurate region
  (i.e., computationally-identifiable region)

  ② Unstability in the model training

Improved approach

- We suggest a modified adversary
  - Give loss information directly to the adversary
  - Adversary learns to capture the CI regions via the loss-based labels

    Label: 1 if an example has a high loss,
    0 otherwise
- Benefits
  - Intuitively learning how to capture the CI region
  - Without the two-player game => More stable



Datasets

Real-world data

Synthetic data

Target datasets

AdultCensus

<Target label (Y) attribute>
Whether one's annual income > $50k

Kohavi, 1996, Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.

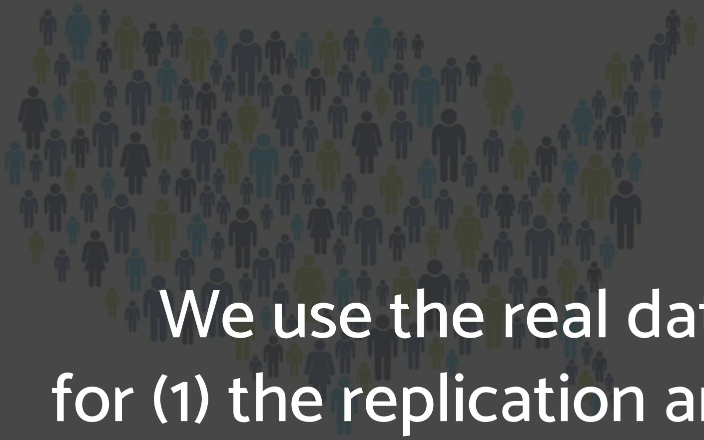
COMPAS

<Target label (Y) attribute>
Whether each criminal makes recidivism

Angwin et al., 2016, There's software used across the country to predict future criminals. And its biased against blacks.

Target datasets

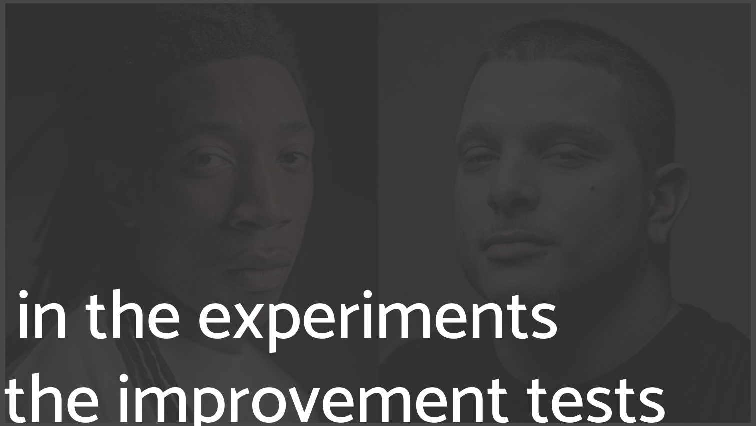
We use the real datasets in the experiments for (1) the replication and (2) the improvement tests



AdultCensus

<Target label (Y) attribute>
Whether one's annual income > $50k

Kohavi, 1996, Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.

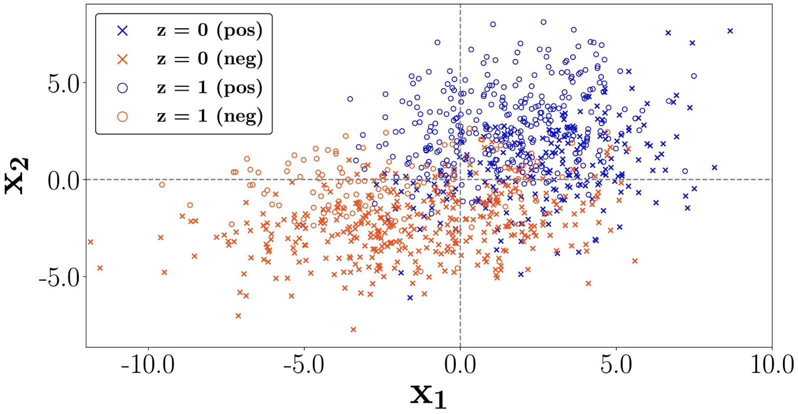


COMPAS

<Target label (Y) attribute>
Whether each criminal makes recidivism

Angwin et al., 2016, There's software used across the country to predict future criminals. And its biased against blacks.

Target datasets

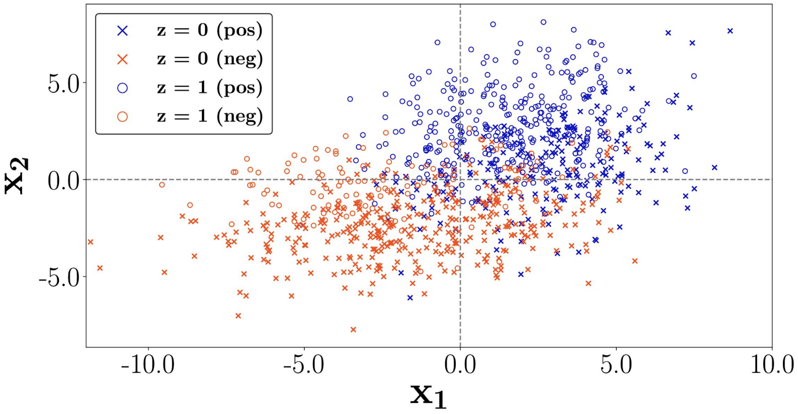


Synthetic data

Zafar et al., AISTATS 2017,
Fairness constraints: Mechanisms for fair classification

- Generate 2,000 examples
  - Two non-sensitive attributes **x1 and x2**
  - A sensitive attribute **z**
  - A label **y**
- More details are in the report :)

Target datasets



Synthetic data

Zafar et al., AISTATS 2017,
Fairness constraints: Mechanisms for fair classification

- Generate 2,000 examples
  - A sensitive attribute **z**
  - A label **y**
- More details are in the report :)

Synthetic data is utilized for analyzing
(1) the limitations of the original design
and
(2) the validity of the improved design

Experiments

Three experiments

Replication

1. Performance evaluation of the original ARL
    - Show the limitation of the previous approach
    - Check the validity of our replicated algorithm

Analysis

2. Evaluation on CI region identification
    - Demonstrate the limitation in capturing CI region with synthetic data
    - Show improved CI region identification of the new ARL

Improvement

3. Performance evaluation of the improved ARL
    - Compare performances of original ARL and our improved ARL

Three experiments

Replication

1. Performance evaluation of the original ARL
   - Show the limitation of the previous approach
   - Check the validity of our replicated algorithm

Analysis

2. Evaluation on CI region identification
   - Demonstrate the limitation in capturing CI region with synthetic data
   - Show improved CI region identification of the new ARL

Improvement

3. Performance evaluation of the improved ARL
   - Compare performances of original ARL and our improved ARL

Our Main Contribution

Evaluation Metrics

We set **AUC** (area under the ROC curve) as our main metric

- To consider robustness to class imbalance in our data

1) AUC avg : Average AUC of all "samples"
2) AUC macro-avg : Average AUC of all "groups"
3) AUC min : Minimum AUC value among groups
4) AUC minority : AUC value of minority group

→ Overall performance

→ Target for improvement

Experimental Settings

Data: AdultCensus / COMPAS / Synthetic data

Model network design (based on basic feed-forward network)

- Learner : linear classifier
- Adversary : linear classifier for AdultCensus / COMPAS,

  1 additional hidden layer (32 units) for synthetic data

Experimental Settings

Hyperparameter selection for ***T*** value

- If the output loss from the learner is larger than threshold *T*, the output works as a positive label for adversary to be classified as a sample in CI region.
- **Larger T → wider CI region**

We manually set T values by finding the best working parameter for each dataset

Results: Evaluation of Original ARL

Evaluation from replication paper

Table 1: Main results: ARL vs DRO

dataset	method	AUC avg	AUC macro-avg	AUC min	AUC minority
Adult	Baseline	0.898	0.891	0.867	0.875
Adult	ARL	**0.907**	**0.915**	**0.881**	**0.942**
COMPAS	Baseline	**0.748**	**0.730**	**0.674**	0.774
COMPAS	ARL	0.743	0.727	0.658	**0.785**

→ ARL improves AUC compared to baselines

Results: Evaluation of Original ARL

Evaluation from replication paper

Table 1: Main results: ARL vs DRO

dataset	method	AUC avg	AUC macro-avg	AUC min	AUC minority
Adult	Baseline	0.898	0.891	0.867	0.875
Adult	ARL	**0.907**	**0.915**	**0.881**	**0.942**
COMPAS	Baseline	**0.748**	**0.730**	**0.674**	0.774
COMPAS	ARL	0.743	0.727	0.658	**0.785**

→ ARL improves AUC compared to baselines

→ But poor result on COMPAS

Results: Evaluation of Original ARL

Evaluation from replication paper

Table 1: Main results: ARL vs DRO

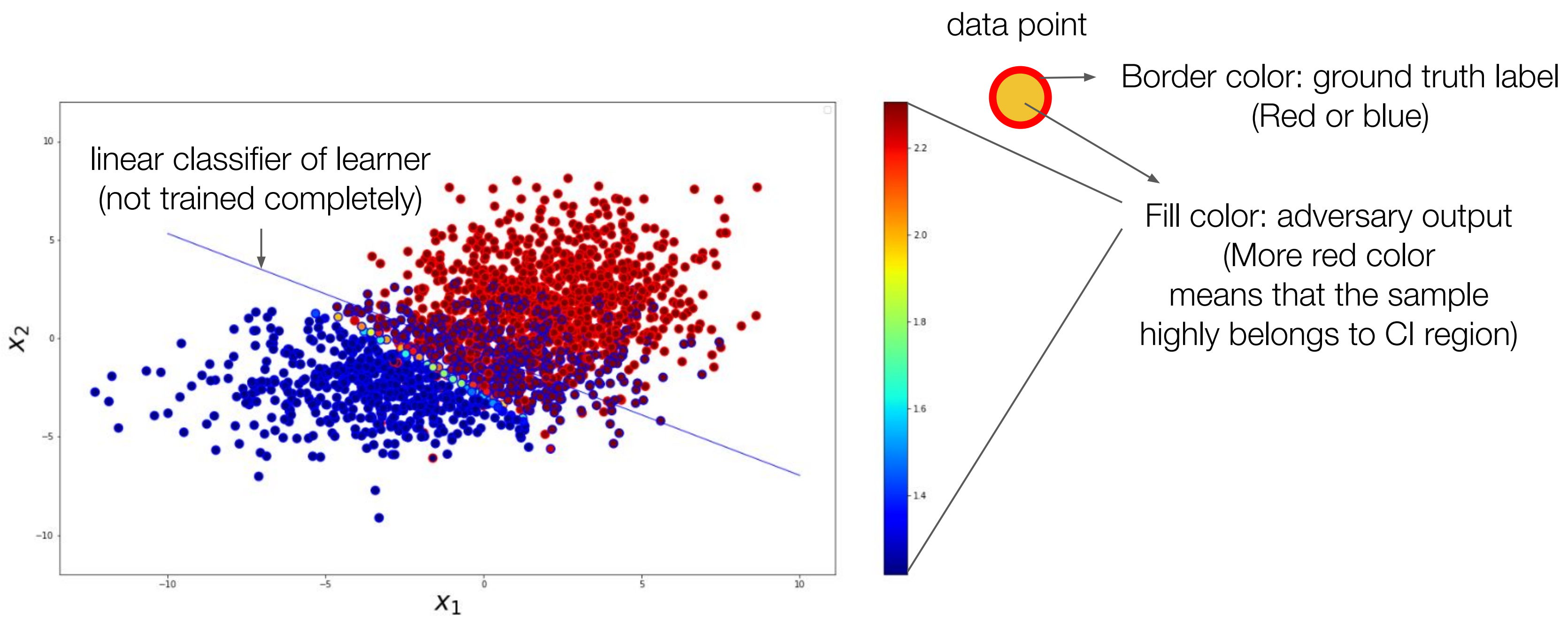
dataset	method	AUC avg	AUC macro-avg	AUC min	AUC minority
Adult	Baseline	0.898	0.891	0.867	0.875
Adult	ARL	**0.907**	**0.915**	**0.881**	**0.942**
COMPAS	Baseline	**0.748**	**0.730**	**0.674**	0.774
COMPAS	ARL	0.743	0.727	0.658	**0.785**

Replicated ARL performance

Dataset	Method	AUC avg	AUC macro-avg	AUC min	AUC minority
	LR	0.698	0.695	0.688	0.688
AdultCensus	ARL	0.703	0.703	0.694	0.710
	LR	0.677	0.639	0.602	0.623
COMPAS	ARL	0.663	0.630	0.601	0.601

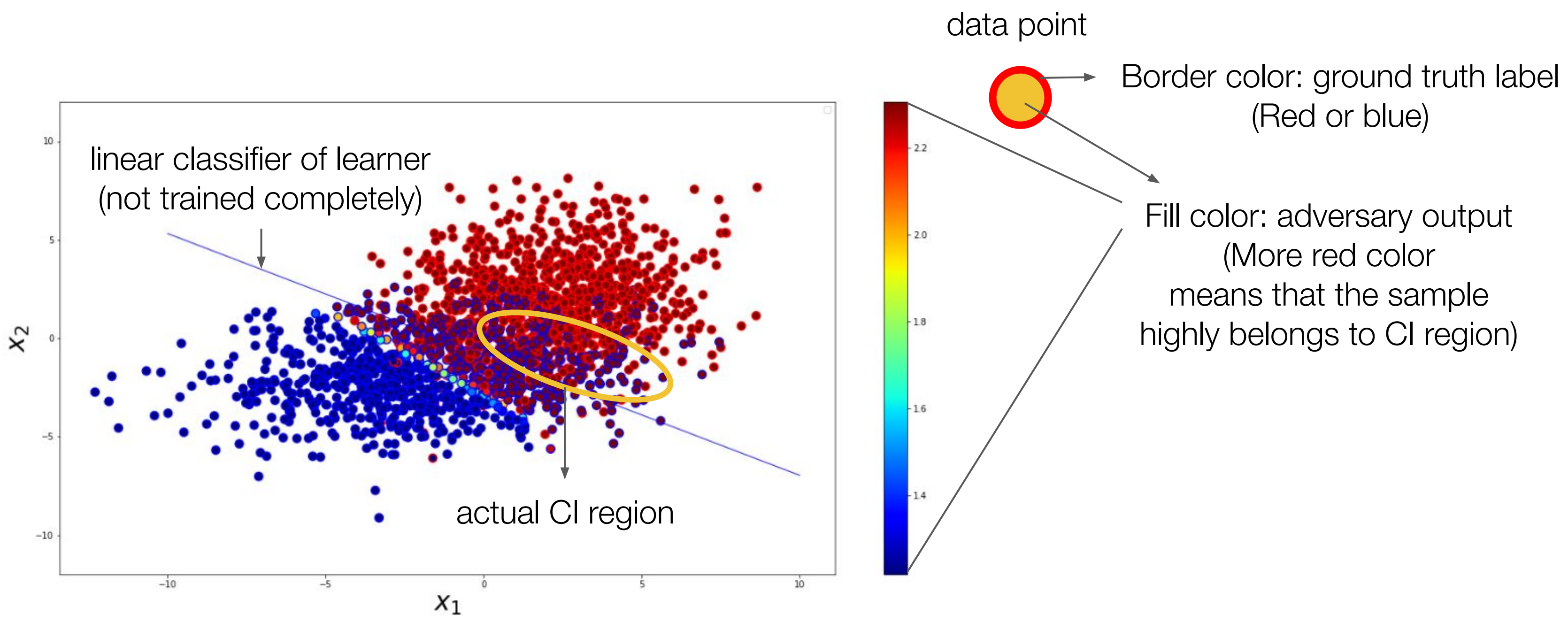
→ Consistent result in our implementation

Results: Original CI Region Identification



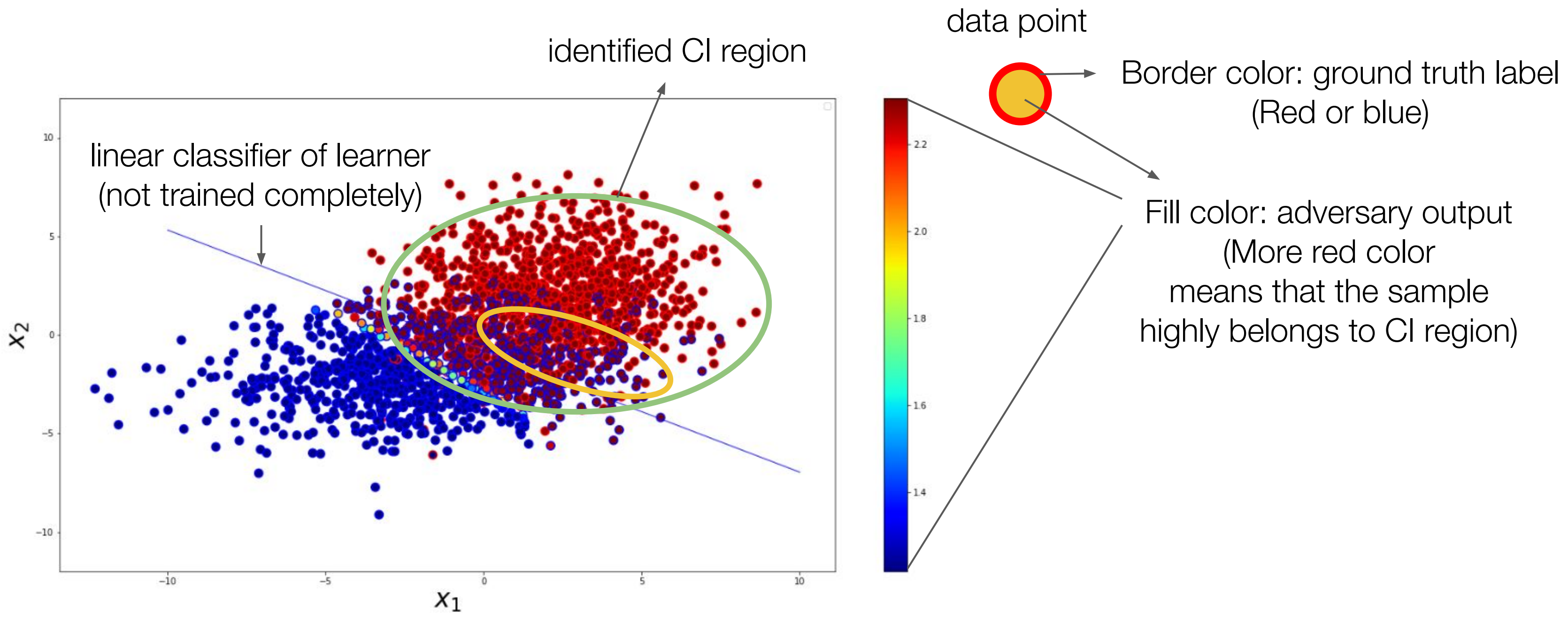
(a) Original CI region identification.

Results: Original CI Region Identification



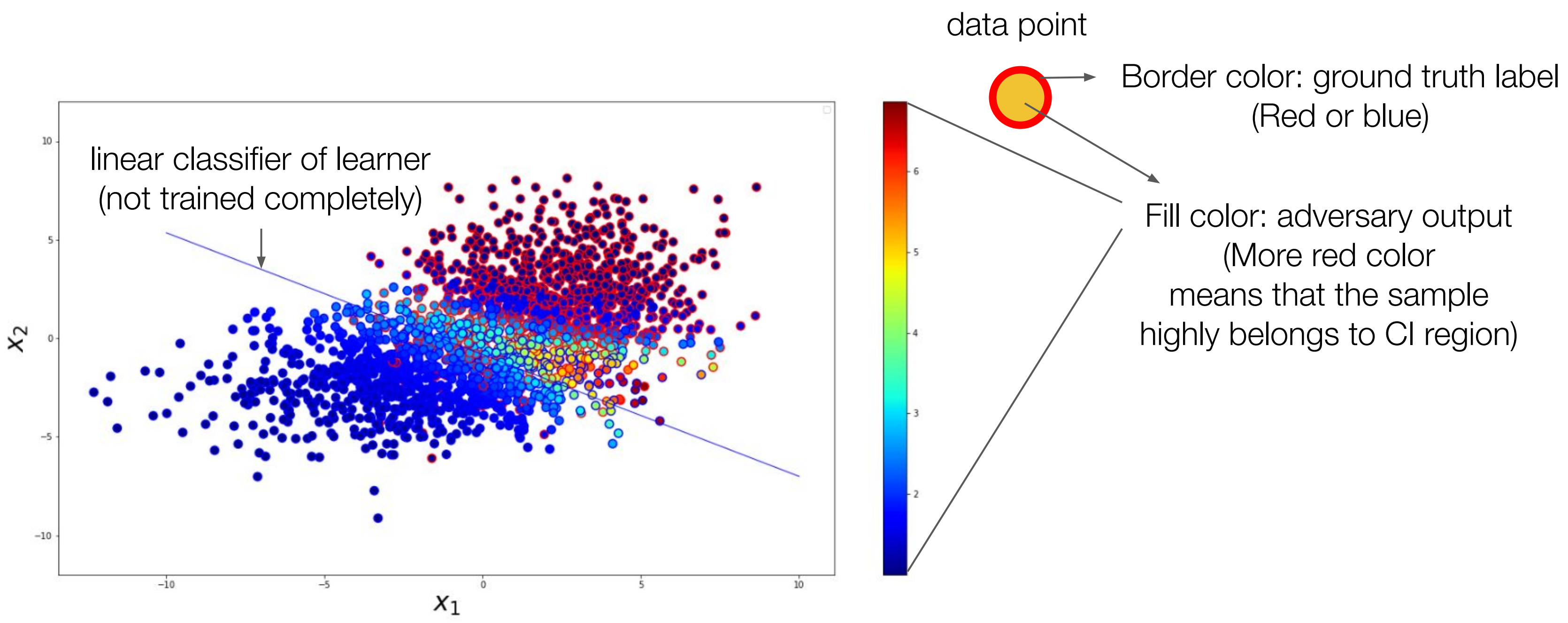
(a) Original CI region identification.

Results: Original CI Region Identification



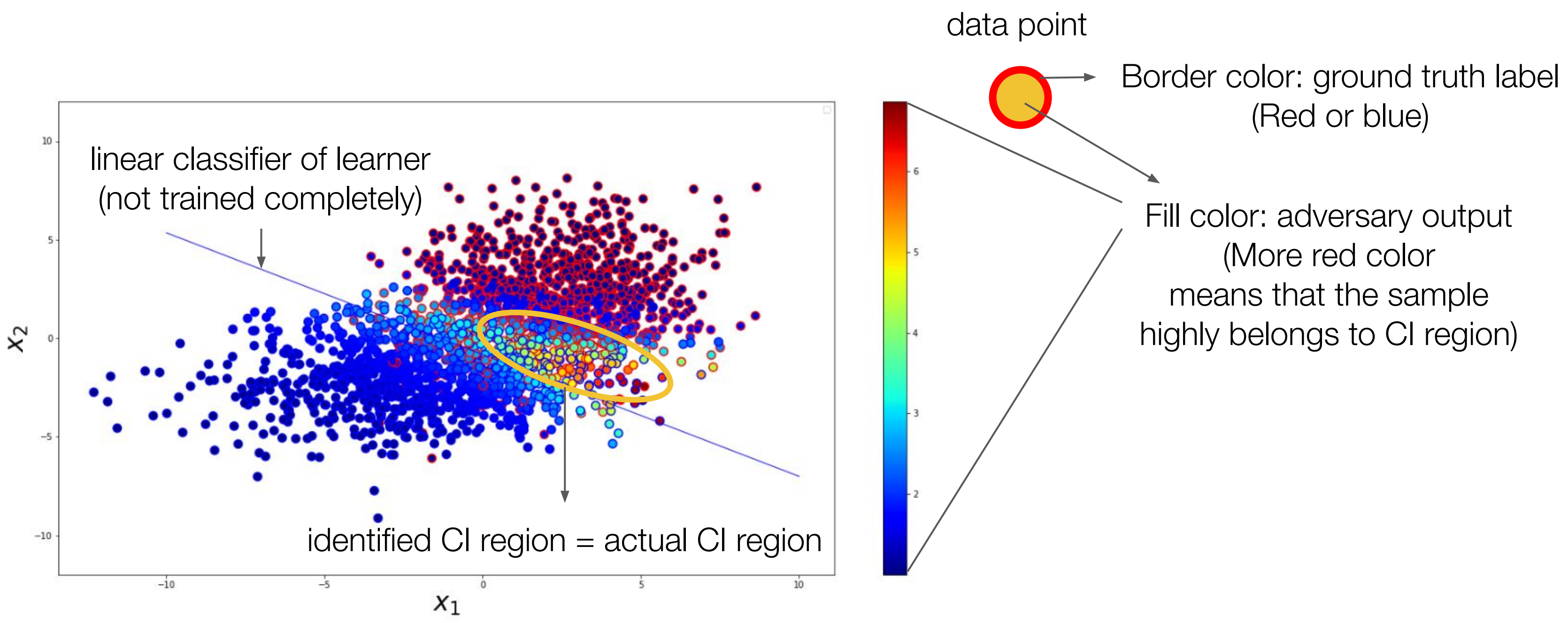
(a) Original CI region identification. → Fails to capture CI region on synthetic data

Results: **Improved** CI Region Identification



(b) Improved CI region identification.

Results: **Improved** CI Region Identification



(b) Improved CI region identification.

→ We can visually check that it captures CI region

Results: Evaluation of Improved ARL

Table 1. Algorithm performances on the AdultCensus and COMPAS datasets.

Dataset	Method	AUC avg	AUC macro-avg	AUC min	AUC minority
AdultCensus	LR	0.698	0.695	0.688	0.688
	ARL	0.703	0.703	0.694	0.710
	Improved ARL	0.747	0.753	0.735	0.779
COMPAS	LR	0.677	0.639	0.602	0.623
	ARL	0.663	0.630	0.601	0.601
	Improved ARL	0.677	0.639	0.602	0.623

→ 5.9% improvement in AUC min compared to original ARL

Results: Evaluation of Improved ARL

Table 1. Algorithm performances on the AdultCensus and COMPAS datasets.

Dataset	Method	AUC avg	AUC macro-avg	AUC min	AUC minority
AdultCensus	LR	0.698	0.695	0.688	0.688
	ARL	0.703	0.703	0.694	0.710
	Improved ARL	0.747	0.753	0.735	0.779
COMPAS	LR	0.677	0.639	0.602	0.623
	ARL	0.663	0.630	0.601	0.601
	Improved ARL	0.677	0.639	0.602	0.623

→ 5.9% improvement in AUC min compared to original ARL

→ No improvement in AUC min compared to LR, but at least does not degrade performance. Better than original ARL.

Why it does not work on COMPAS?

→ Possible assumption: *Invalid CI region in COMPAS*

Discussion & Limitation

- Practicality of the main assumption of ARL: Does CI region exist?
  - No discussion on the validity in real-world datasets
  - As dataset has high complexity it becomes hard to be analyzed
  - Could be a reason for the difficulty in improving COMPAS before
- No golden standard in selecting hyperparameter T
  - T strongly affects CI region identification task
  - If there is a systemic way to decide T, performance of our ARL would increase

Project in summary

- Big goal: Building a fair learning system without demography information
  - Replicated ARL which adaptively gives more weights to computationally-identifiable regions, by leveraging minmax game between learner and adversary sharing same objective
- Improvement approach:
  Achieving improved CI region identification without minmax game
- Our design outperforms both in CI region identification and achieving fair classification task without demography information compared to the base
  - Demonstrated improvement in CI region identification through visualization on synthetic data
  - Showed increase in 5.9% AUC-min in AdultCensus, eliminated AUC degradation in COMPAS

Project in summary

- Big goal: Building a fair learning system without demography information
  - Replicated ARL which adaptively gives more weights to computationally-identifiable regions, by leveraging minmax game between learner and adversary sharing same objective
- Improvement approach:

  Achieving improved CI region identification without minmax game
- Our design outperforms both in CI region identification and achieving fair classification task without demography information compared to the base
  - Demonstrated improvement in CI region identification through visualization on synthetic data
  - Showed increase in 5.9% AUC-min in AdultCensus, eliminated AUC degradation in COMPAS

Thank You :)