Statistical Measures of Fairness in AI

Eric Xie, Yagnik Panguluri, Anders Gyllenhoff, Caroline Gihlstorf

Agenda

Classification

- Overview
- Distance Metrics
- Fairness Metrics

Discussion on Fairness in Classification

Implementations of Fairness

- A new loss function based on a distribution difference metric
- Learning a distance metric and using intermediate representations
- Post processing for classification fairness

Discussion on Implementations of Fairness

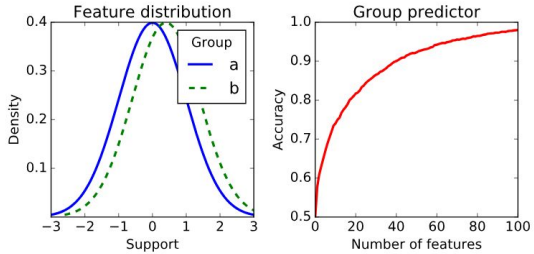
Classification: Overview

Classification is the task of assigning labels to data points based on input features

It applies to both future outcomes and unknown past events

Classification relies on patterns in data that connect observed characteristics (covariates X) to an outcome variable (Y)

A classifier function f(X) predicts $\hat{Y}$, the estimated outcome



Represent the population as a probability distribution

Use SDT to develop a classifier that makes predictions

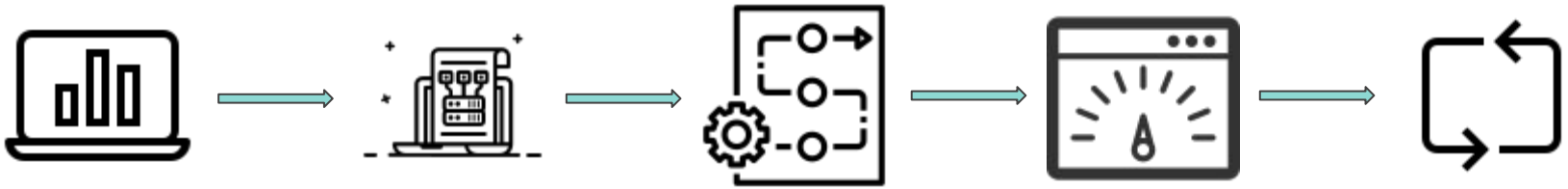
What is Fairness in Classification?

Fairness in classification ensures that predictions do not systematically disadvantage individuals based on sensitive attributes (e.g., race, gender, disability).

Discrimination is defined as disparities that are not justified by legitimate differences in the population

Key Concepts	Fairness Criteria
Many classification tasks use features X that may implicitly encode an individual's status in a protected category We define A as a sensitive attribute (e.g. race, gender). If a classifier decision depends on A, it may be discriminatory	**Independence:** Decisions should not depend on group membership **Separation:** Error rates should be equal across groups **Sufficiency:** The model's predictions should be equally accurate for all groups

Constructing Fair Classification Tasks



Choose what fairness means in the context of the classification task

Identify the bias in the data and model

Select a fairness approach (pre-processing, in-processing, post-processing)

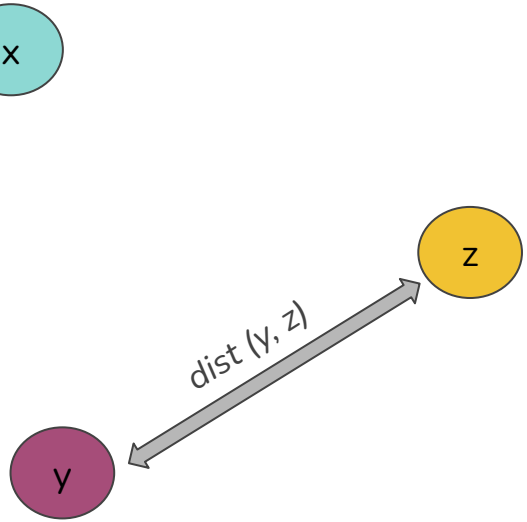
Reformulate as an optimization problem

Evaluate and iterate and assess using metrics like AUC, disparity scores, and calibration checks

Distance Metrics

Determine the “closeness” of two individuals in some space to approximate how similar they are

Sometimes they are assumed (e.g., Dwork et al., (2011)), other times they are computed explicitly (e.g., Zemel et al., (2013))



Fairness as a Linear Optimization Problem

- A standard method of feasibly deriving a fair model is by reformulating the problem as a linear optimization problem
  - Simple model constraints
  - Data
  -

Defining Fairness Metrics

Class-Blindness:

- Applies a single threshold across all groups, removing the protected attribute from the dataset

- This runs the risk of unfairness caused by redundant encodings
  - Predicting protected class attributes from other features

Defining Fairness Metrics

Statistical/Demographic Parity:

Percentage of the population determines expected percentage of classification outcomes

Defining Fairness Metrics

Statistical/Demographic Parity: Does it work

Group Level: Yes

Individual Level: No

Example: companies may maliciously choose to interview people from a protected group without the necessary qualifications to establish a negative track record for people in the group (equal number of interviews, unequal treatment)

Defining Fairness Metrics

Equal Opportunity:

Within a protected attribute (A), those that deserve a positive classification (Y=1) have equal correct prediction rates ($\hat{y} = 1$).

$$\Pr\left\{\widehat{Y} = 1 \mid A = 0, Y = 1\right\} = \Pr\left\{\widehat{Y} = 1 \mid A = 1, Y = 1\right\}$$

Defining Fairness Metrics

Equalized Odds:

Adds an additional constraint onto “Equal Opportunity,” requiring equal incorrect positive rates as well

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=y\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=y\right\}, \quad y \in\{0,1\}$$

Discussion (Part I) - 10 minutes

What does it mean to you to implement fairness in classification models?

Given the four aforementioned fairness metrics, can you think of scenarios where each would be more/less effective?

- Class-Blindness: Applies a single threshold across all groups, removing the protected attribute from the dataset
- Statistical/Demographic Parity: Percentage of the population determines expected percentage of classification outcomes
- Equal Opportunity: Within a protected attribute (A), those that deserve a positive classification (Y=1) have equal correct prediction rates ($\hat{y} = 1$).
- Equalized Odds: Adds an additional constraint onto “Equal Opportunity,” requiring equal incorrect positive rates as well

Implementations of Fairness

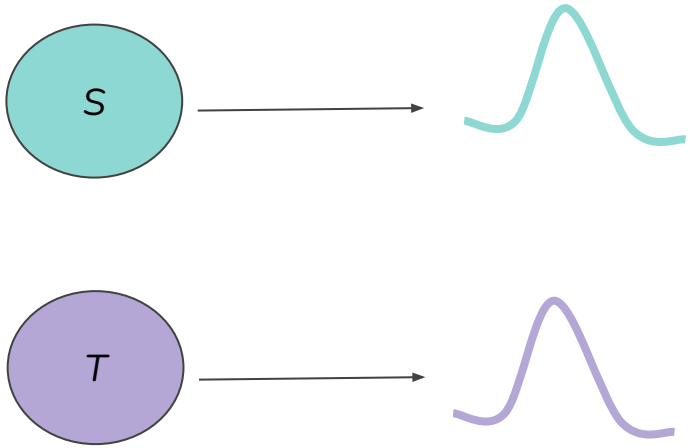
Two predominant branches for methods to achieve fairness

- Data massaging: modifying labels of data samples so that the proportions of the positive labels are equal in both protected groups
- Regularization Strategy: adding a regularization term to the classification training objects that quantify bias and discrimination, maximizing accuracy and minimizing discrimination in models.

We will be discussing the later of the two.

Implementations of Fairness

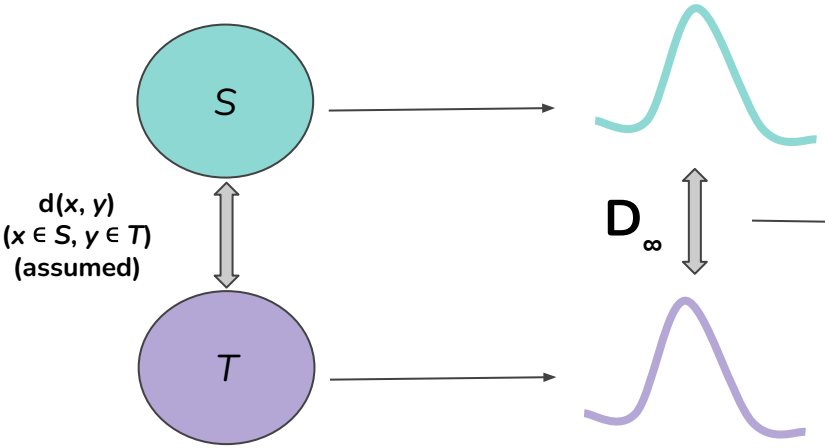
Dwork et al., (2011):



1) Learn a mapping

Implementations of Fairness

Dwork et al., (2011):



1) Learn a mapping
2) Define optimization for fairness

$$D_\infty(P, Q) = \sup_{a \in A} \log \left(\max \left\{ \frac{P(a)}{Q(a)}, \frac{Q(a)}{P(a)} \right\} \right).$$

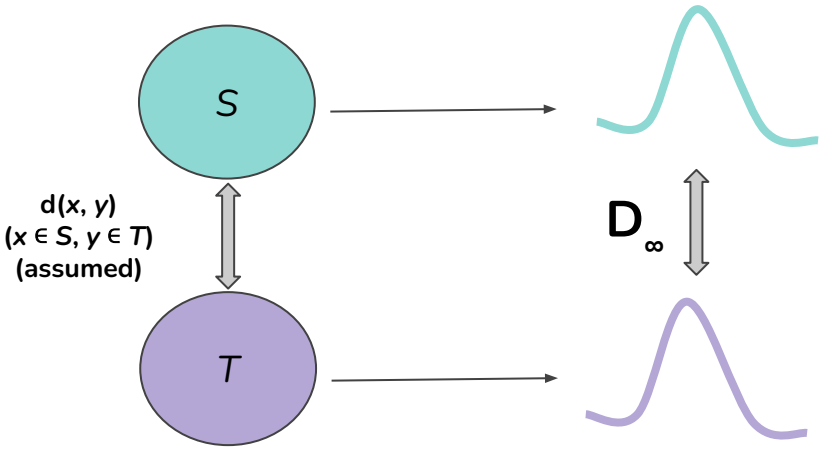
(Dwork et al., (2011), Equation 6, page 5)

Individuals

Distributions over outputs

Implementations of Fairness

Dwork et al., (2011):



Individuals

Distributions over outputs

1) Learn a mapping
2) Define optimization for fairness
3) Minimize loss

Lipschitz constraint on fairness

$$\text{opt}(\mathcal{I}) \stackrel{\text{def}}{=} \min_{\{\mu_x\}_{x \in V}} \mathop{\mathbb{E}}_{x \sim V} \mathop{\mathbb{E}}_{a \sim \mu_x} L(x, a) \quad (2)$$

$$\text{subject to} \quad \forall x, y \in V, : \; D(\mu_x, \mu_y) \leq d(x, y) \quad (3)$$

$$\forall x \in V : \quad \mu_x \in \Delta(A) \quad (4)$$

Figure 1: The Fairness LP: Loss minimization subject to fairness constraint

(Dwork et al., (2011), Figure 1, page 5)

Implementations of Fairness

Dwork et al., (2011):

Bias:

$$\text{bias}_{D,d}(S,T) \stackrel{\text{def}}{=} \max \mu_S(0) - \mu_T(0)$$ (Dwork et al., (2011), Equation 8, page 8)

Implementations of Fairness

Dwork et al., (2011):

Bias:

$$\text{bias}_{D,d}(S,T) \stackrel{\text{def}}{=} \max \mu_S(0) - \mu_T(0)$$ (Dwork et al., (2011), Equation 8, page 8)

↓

Statistical Parity & Lipschitz Condition:

- A *(D, d)*-Lipschitz mapping entails statistical parity up to the bias value.
- Less stringent Lipschitz condition for dissimilar distributions often still entails statistical parity

Implementations of Fairness - Critical Analysis

Dwork et al., (2011):

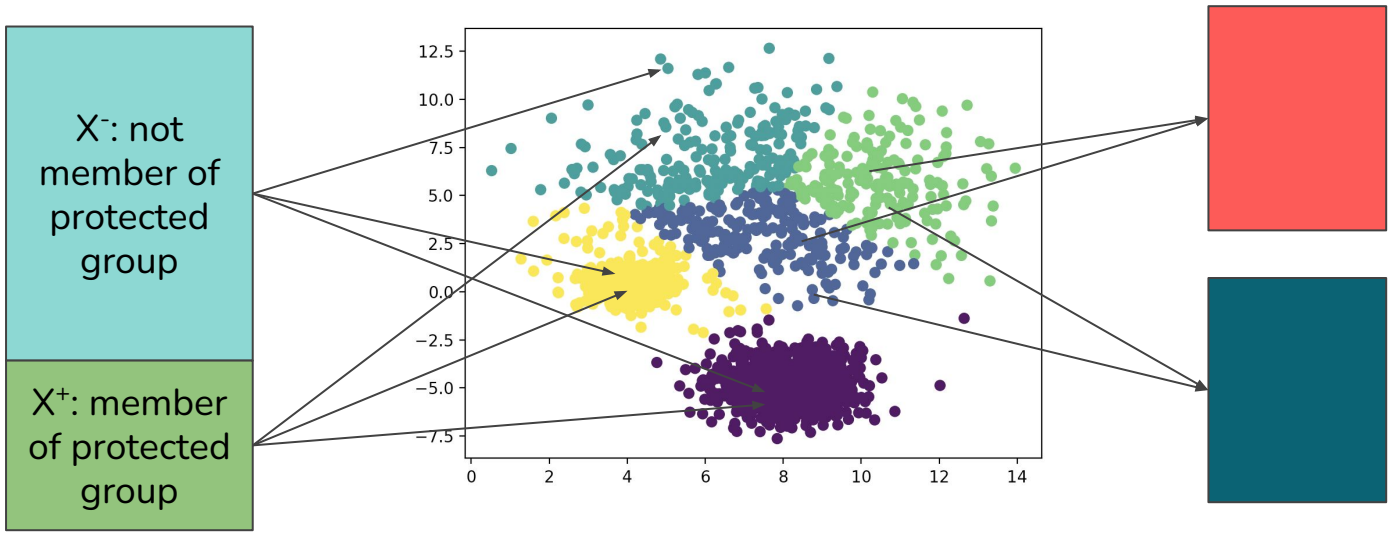
- Assume non-overlapping groups of individuals (unrealistic - people can hold multiple protected identities at once)
- Distance metric is theoretical
  - May need to account for bias in distance metrics
- Work centers around a targeted advertising scenario - to what extent is this ethical?

Implementations of Fairness - LFR

How can we derive our own distance metric?

Continuing with a regularization methodology, Zemel et al. establish the Learning Fair Representations (LFR) framework

Implementations of Fairness - LFR



Input Data $x \in X$

- Z, a multinomial RV, where each K value is a intermediate “prototype”.
- v_k is a vector in the same space as x

Classification

Distance

$$d(\mathbf{x}_n, \mathbf{v}_k) = ||\mathbf{x}_n - \mathbf{v}_k||_2$$

Implementations of Fairness - LFR

LFR Framework Overview, a discriminative clustering model:

1. Mapping the training data (X_0) to some Z to satisfy statistical parity

$$P(Z = k|\mathbf{x}^+ \in X^+) = P(Z = k|\mathbf{x}^- \in X^-), \forall k \quad (1)$$

2. The mapping to the Z space retains information in X, except about membership to the protected group

$$P(Z = k|\mathbf{x}) = \exp(-d(\mathbf{x}, \mathbf{v}_k))/\sum_{j=1}^{K} \exp(-d(\mathbf{x}, \mathbf{v}_j)) \quad (2)$$

3. The new mapping from X to Y is close to f: X → Y

Implementations of Fairness - LFR

Where $M_{n,k}$ denotes the probability that data point x_n maps to v_k

$$M_{n,k} = P(Z = k|\mathbf{x}_n) \quad \forall n, k \qquad (3)$$

Now have a learning system minimizing the objective

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y \qquad (4)$$

where A_z, A_x, and A_y are hyper parameters tuned to balance the tradeoffs between statistical parity (L_z), mapping Z to be a good description of X (L_x), and to maximize the accuracy of prediction y (L_y).

Implementations of Fairness - LFR

An important note:

$$L_y = \sum_{n=1}^{N} -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n) \qquad (10)$$

$$\hat{y}_n = \sum_{k=1}^{K} M_{n,k} w_k \qquad (11)$$

$yhat_n$ is the prediction for y_n=1, based on marginalizing over each prototype's prediction for Y and weighted by their respective probabilities.

$$d(\mathbf{x}_n, \mathbf{v}_k, \alpha) = \sum_{i=1}^{D} \alpha_i (x_{ni} - v_{ki})^2 \qquad (12)$$

Case Study - LFR

Metrics:

- Accuracy: Classification accuracy
- Discrimination: Bias with respect to the sensitive feature in classification
- Consistency: Model classification prediction for kNN(x)

LFR consistently achieves lowest levels of discrimination, but maintains high accuracy

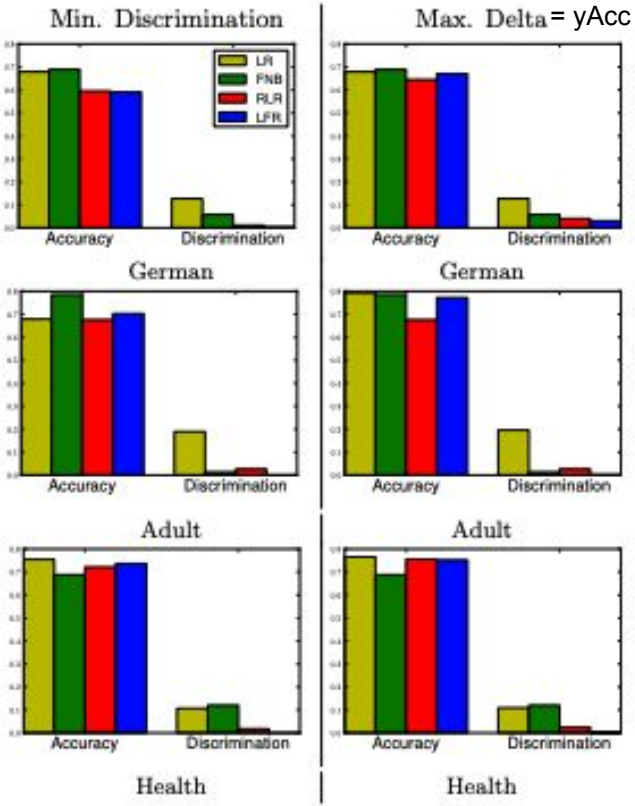


Figure 1. Results on test sets for the three datasets (German, Adult, and Health), for two different model selection criteria: minimizing discrimination and maximizing the difference between accuracy and discrimination.

Case Study - LFR

LFR achieves better individual fairness on each dataset, rewarding Z's preservation of information about the features in X.

Models selected based on discrimination.

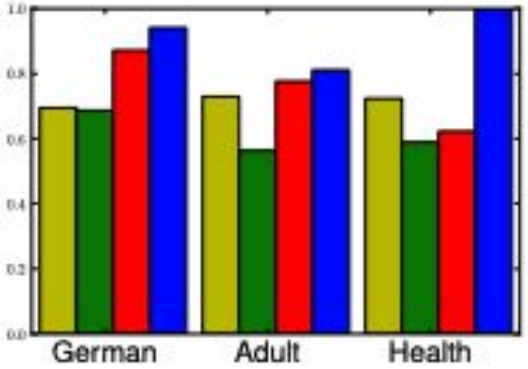


Figure 2. Individual fairness: The plot shows the consistency of each model's classification decisions, based on the yNN measure. Legend as in Figure 1.

Case Study - LFR

Measure protected group (S) information in the model by building a predictor to predict S from Z.

Optimize predictor to minimize difference with actual s_n and test prediction for S = sAcc score.

sAcc is shift towards the lower bound in all dataset.

Models selected based on delta.

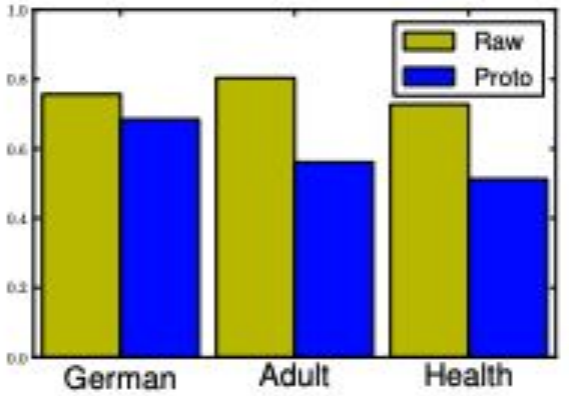


Figure 3. The plot shows the accuracy of predicting the sensitive variable (*sAcc*) for the different datasets. Raw involves predictions directly from all input dimensions except for S, while Proto involves predictions from the learned fair representations.

Implementations of Fairness

Post-processing approaches can avoid altering the model training process

- Hardt et al. explore post-hoc model constraints satisfied by modifying the Bayes threshold
- Select an optimal model that satisfies $TPR_{A=1} = TPR_{A=0}$, and $FPR_{A=1} = FPR_{A=0}$

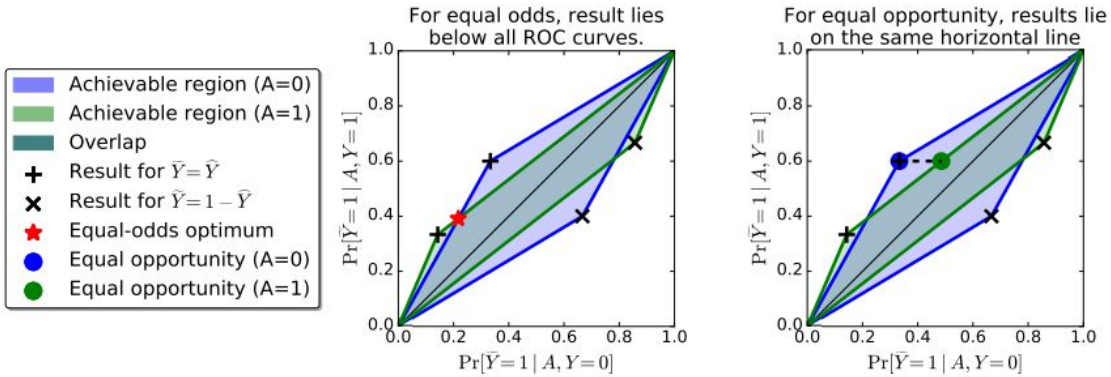


Figure 1: Finding the optimal equalized odds predictor (left), and equal opportunity predictor (right).

Implementations of Fairness

- The cost of implementing this fairness method scales linearly with the Kolmogorov distance *d*
- The distance between the constrained ROC curve and the optimum is bounded by d√(2) per metric (2d√(2) for equalized odds)

$$\mathbb{E}\ell(\widehat{Y}, Y) \leqslant \mathbb{E}\ell(Y^*, Y) + 2\sqrt{2} \cdot d_{\mathrm{K}}(\widehat{R}, R^*)$$

Case Study (FICO Credit Scores)

These metrics are applied in a case study regarding loan profitability, with race as the protected attribute

- Maximum Profit: Maximizes profits regardless of fairness
- Race-blind: Applies a single threshold across the entire dataset, removing the race attribute.
- Demographic Parity: Each racial group receives loans at equal rates
- Equal Opportunity: $TPR_{A=1} = TPR_{A=0}$
- Equalized Odds: $TPR_{A=1} = TPR_{A=0}$ and $FPR_{A=1} = FPR_{A=0}$

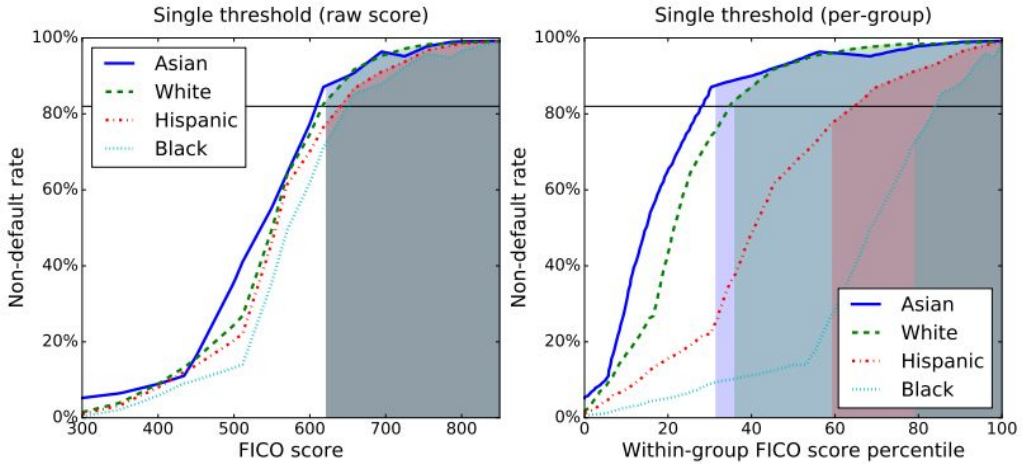
Case Study (FICO Credit Scores)

Maximum Profit: 82% non-default rate overall, representing the optimal Bayes classifier performance

- Race-blind: 99.3% of the Maximum
- Equal Opportunity: 92.8%
- Equalized Odds: 80.2%
- Demographic Parity: 69.8%

Case Study (FICO Credit Scores)

- Under the race-blind approach, Black non-defaulters are significantly less likely to qualify for loans compared to White or Asian applicants.



Discussion (Part II) - 20 minutes

If you were in charge of selecting the classification model, which method would you choose?

- If you were an investor selecting between companies that implemented each type of model, would your decision change?

Disregarding performance, which metric seems the most fair to you?

Is it fair to have differing thresholds based on protected class membership?

- Is it more fair to have equivalent thresholds, even if it may be more difficult for some groups to reach these thresholds given systemic issues

Do you think it is possible to fully quantify fairness in classification? Why or why not?

References

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. “Fairness and Machine Learning: Limitations and Opportunities.” MIT Press, 2023.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. “Fairness Through Awareness,” 2011. https://arxiv.org/abs/1104.3913.

Hardt, Moritz, Eric Price, and Nathan Srebro. “Equality of Opportunity in Supervised Learning,” 2016. https://arxiv.org/abs/1610.02413.

Zemel, Richard, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. “Learning Fair Representations.” In Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, III-325-III–333. ICML’13. Atlanta, GA, USA: JMLR.org, 2013.

Appendix

In order to achieve statistical parity, we want to ensure Eqn. 1, which can be estimated using the training data as:

$$M_k^+ = M_k^-, \forall k \tag{5}$$

$$M_k^+ = \mathbb{E}_{\mathbf{x} \in X^+} P(Z = k|\mathbf{x}) = \frac{1}{|X_0^+|} \sum_{n \in X_0^+} M_{n,k} \tag{6}$$

and M_k^- is defined similarly.

Hence the first term in the objective is:

$$L_z = \sum_{k=1}^{K} \left| M_k^+ - M_k^- \right| \tag{7}$$

The second term constrains the mapping to Z to be a good description of X. We quantify the amount of information lost in the new representation using a simple squared-error measure:

$$L_x = \sum_{n=1}^{N} (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2 \tag{8}$$

where $\hat{\mathbf{x}}_n$ are the reconstructions of $\mathbf{x}_n$ from Z:

$$\hat{\mathbf{x}}_n = \sum_{k=1}^{K} M_{n,k} \mathbf{v}_k \tag{9}$$

These first two terms encourage the system to encode all information in the input attributes except for those that can lead to biased decisions.

The final term requires that the prediction of y is as accurate as possible:

$$L_y = \sum_{n=1}^{N} -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n) \tag{10}$$

Here $\hat{y}_n$ is the prediction for y_n, based on marginalizing over each prototype's prediction for Y, weighted by their respective probabilities $P(Z = k|\mathbf{x}_n)$:

$$\hat{y}_n = \sum_{k=1}^{K} M_{n,k} w_k \tag{11}$$

Appendix

Tested framework with four models (each hyperparameter tuned).

1. Unregularized Logistic Regression (LR) - baseline.
2. Regularized Logistic Regression (RLR) - (Kamishima et al., 2011)
3. The Fair Naive Bayes (FNB) - four variants (Kamiran & Calders, 2009) of FNB for ranking and classification phases.
4. LFR - L-BFGS used to minimize the fairness optimization.

- **Accuracy**: measures the accuracy of the model classification prediction:

$$yAcc = 1 - \frac{1}{N}\sum_{n=1}^{N} |y_n - \hat{y}_n| \qquad (13)$$

- **Discrimination**: measures the bias with respect to the sensitive feature S in the classification:

$$yDiscrim = \left|\frac{\sum_{n:s_n=1} \hat{y}_n}{\sum_{n:s_n=1} 1} - \frac{\sum_{n:s_n=0} \hat{y}_n}{\sum_{n:s_n=0} 1}\right| \qquad (14)$$

This is a form of statistical parity, applied to the classification decisions, measuring the difference in the proportion of positive classifications of individuals in the protected and unprotected groups.

- **Consistency**: compares a model's classification prediction of a given data item $\mathbf{x}$ to its k-nearest neighbors, $kNN(\mathbf{x})$:

$$yNN = 1 - \frac{1}{Nk}\sum_{n} \left|\hat{y}_n - \sum_{j \in kNN(\mathbf{x}_n)} \hat{y}_j\right| \qquad (15)$$