A6 and A4

Group 8

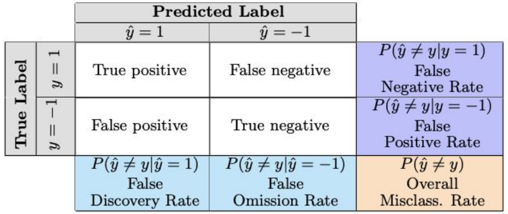
A4 Classifying without Disparate Mistreatment

- Goal: Use disparate mistreatment, a notion of unfairness which is defined in terms of misclassification rates, to build a fair classifier.
- Steps:
  - Formalizing notions of fairness
  - Convert constraints into convex form
  - Train decision boundary-based classifiers that do not suffer from disparate mistreatment by minimizing a convex loss.

Junhan Huang, Chenbohan Zhang, Xueyi Zhang

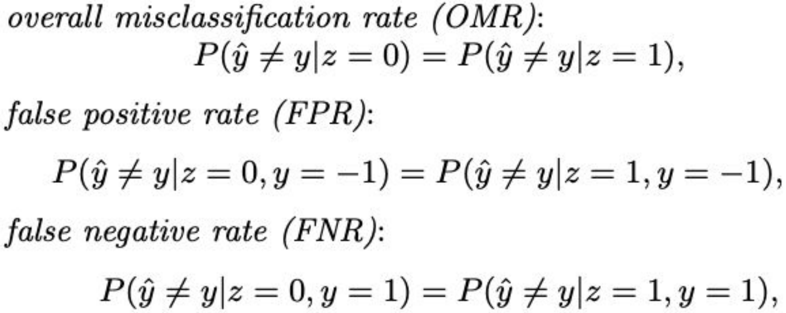
A4 Formalizing notions of fairness

A binary classifier does not suffer from disparate mistreatment if the misclassification rates for different groups of people having different values of the sensitive feature z are the same.



		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y \mid y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

Misclassification rates can be measured as false positive and false negative rates.



overall misclassification rate (OMR):

$$P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1),$$

false positive rate (FPR):

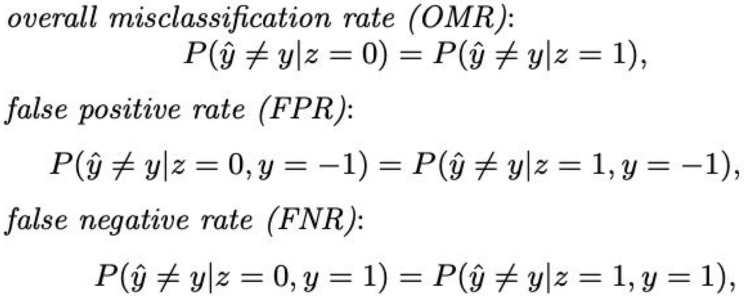
$$P(\hat{y} \neq y | z = 0, y = -1) = P(\hat{y} \neq y | z = 1, y = -1),$$

false negative rate (FNR):

$$P(\hat{y} \neq y | z = 0, y = 1) = P(\hat{y} \neq y | z = 1, y = 1),$$

A4 Convert constraints into convex form

- Original constraints.



overall misclassification rate (OMR):

$$P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1),$$

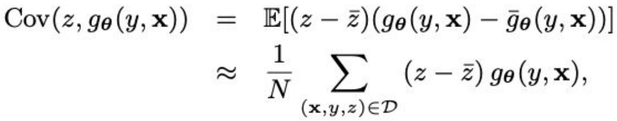
false positive rate (FPR):

$$P(\hat{y} \neq y | z = 0, y = -1) = P(\hat{y} \neq y | z = 1, y = -1),$$

false negative rate (FNR):

$$P(\hat{y} \neq y | z = 0, y = 1) = P(\hat{y} \neq y | z = 1, y = 1),$$

- Propose an effective proxy for fairness: covariance between the users' sensitive attributes and the signed distance between the feature vectors of misclassified users and the classifier decision boundary.



$$\begin{aligned} \mathrm{Cov}(z, g_{\boldsymbol{\theta}}(y, \mathbf{x})) &= \mathbb{E}[(z - \bar{z})(g_{\boldsymbol{\theta}}(y, \mathbf{x}) - \bar{g}_{\boldsymbol{\theta}}(y, \mathbf{x}))] \\ &\approx \frac{1}{N} \sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z}) \, g_{\boldsymbol{\theta}}(y, \mathbf{x}), \end{aligned}$$

- Convert constraints into a Disciplined Convex-Concave Program (DCCP).

A4 Train decision boundary-based classifiers

- Train an unconstrained logistic regression classifier.

- Train classifiers on data leads to disparate mistreatment in terms of only the false positive rate (fpr) or false negative rate (fnr).

- Train classifiers on data leads to disparate mistreatment in terms of both fpr and fnr.

A4 Data

- Input:
  - X: Features (DM: without sensitive feature; DM-sen: with sensitive feature)
  - Y: Label/result
  - Z: Sensitive feature
- Data Preprocessing:
  - Remove columns from X if it contains significant amount of Null and Missing Data
  - Transform Y (labels) into {1,-1}
  - Transform and apply standard scaling to normalize values

A6 Handling Conditional Discrimination

- Goal: train classifiers on a data set, so that they are discrimination free with respect to a given sensitive attribute, like RACE in this case.
- Steps:
    - subroutine PARTITION (X, e)
    - subroutine DELTA (s)
    - perform algorithm: Local massaging (LM) & Local preferential sampling (LPS)

Liang Hu, Wenchang Zhu

A6 Handling Conditional Discrimination

- Input:
  - X – an instance in p dimensional space
  - s – sensitive attribute
  - e – explanatory attribute
  - y – label/result
- How to select features?
  - We first dropped missing values;
  - We then used a heatmap to filter out the most popular features among all variables

A6: algorithm- delta Function

It performs a fairness or bias mitigation operation based on an individual's race and their predicted probability of belonging to a certain class, by updating their predicted label if their predicted probability falls within a certain threshold based on their race's two-year recidivism rate.

Also, the way to construct the DELTA function is the main difference between algorithms of LM and LPS, in which they employ different ways to process data (relabel or delete/duplicate data) to reduce discrimination.

A6: algorithm- Local Massaging

The Local Messaging algorithm is a message-passing algorithm used for inference on factor graphs, where messages are passed between nodes and factors in a local manner based on a belief propagation update rule.

The local massaging for every partition in the training data induced by the explanatory attribute will modify the values of labels until both $P'(+|m, ei)$ and $P'(+|f, ei)$ become equal to $P\star(+|ei)$.

A6: algorithm- Local preferential sampling

The Local Preferential Sampling algorithm is a network growth model that generates synthetic networks by selecting new nodes based on their degree and the degree of their neighbors, with a preference for nodes that have higher degrees than their neighbors.

The preferential sampling technique does not modify the training instances or labels, instead it modifies the composition of the training set. It deletes and duplicates training instances so that the labels of new training set contain no discrimination and satisfy the criteria $P'(+|m, ei) = P'(+|f, ei) = P_\star(+|ei)$.

Comparison

What do they mean?

	Baseline	LM-A6	LPS-A6	DM- A4	DM-sen- A4
Calibration (African American and Caucasian)	0.0199	0.01059	0.03958	0.014240	0.031734
Accuracy	0.66612	0.96098	0.96531	0.9040	0.6627
Distribution of two race		(0.52, 0.45)	(0.52, 0.5)		
F1-score	0.62446	0.96461	0.96733	0.88	0.69207
Recall-score	0.57816	0.97363	0.95658	0.84	0.65