

# Project 4: Algorithm Implementation and Evaluation

Group 6: Rishav Agarwal (ra3141), Jingwei Liao (jl5983),  
Ke Liu (kl3344), Jiuru Wang (jw4150), Xinran Wang (xw2809)





01

# Introduction to Dataset





# Dataset: Bank (Used for A3 Evaluation by the paper)

**Number of variables:** 16      **Number of observations:** 45211

- **One-hot encode:** job, martial, contact, education, output
- **Numeric:** age, day, month
- **Binary:** default, housing, loan

**Sensitive Attribute:** age (also removed from feature to avoid disparate treatment)

**Criteria:** 25-60 protected group, otherwise non-protected group

Using Logistic Regression to predict whether a person subscribed to term deposit in investment.





# Dataset: COMPAS

compas-scores-two-years.csv

- **Features:** Age, Charge Degree, Gender, Prior Counts, Length Of Stay
- **Predicted Label:** Two Year Recid (whether or not the defendant recidivated within two years)
- **Sensitive Attribute:** Race (Caucasian: 1, African-American: 0)
- **Data Splitting:** training: validation: testing = 5 : 1 : 1





# Data preprocessing on CAMPAS

- **Age:** age < 25, 25 < age < 45, age > 45
- **Charge Degree:** Misdemeanor or Felony
- **Gender:** Male or Female
- **Prior Counts:** 0, 1-3, larger than 3
- **Length of Stay:** < = 1 week, < = 3 months or > 3 months

Reference: *How We Analyzed the COMPAS Recidivism Algorithm – ProPublica*  
<https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb>





02

# Introduction to Algorithms



# Algorithms: A3 & A7



Maximizing Fairness  
under Accuracy  
Constraints (Gamma and  
Fine-Gamma)



Fairness-aware Feature  
Selection



## A3: Maximizing Fairness under Accuracy Constraints (Gamma and Fine-Gamma)

---

### General Goal

Design classifiers—convex margin-based classifiers like logistic regression and support vector machines (SVMs)—that avoid both disparate treatment and disparate impact, and can additionally accommodate the “business necessity” clause of disparate impact doctrine.

### Two distinct notion

1. Disparate treatment: decision based on subject's sensitive attribute information
  2. Disparate impact: outcome hurts/benefits people with certain sensitive attribute values
-



# Way to quantify disparate impact



“80%-rule”  
or “p%-rule”



Decision  
Boundary  
Covariance



# Formulas



Logistic Loss



Decision Boundary  
Covariance

$$\begin{aligned}\text{Cov}(\mathbf{z}, d_{\boldsymbol{\theta}}(\mathbf{x})) &= \mathbb{E}[(\mathbf{z} - \bar{\mathbf{z}})d_{\boldsymbol{\theta}}(\mathbf{x})] - \mathbb{E}[(\mathbf{z} - \bar{\mathbf{z}})]\bar{d}_{\boldsymbol{\theta}}(\mathbf{x}) \\ &\approx \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x}_i),\end{aligned}\quad (2)$$



Maximizing Accuracy  
Under Fairness  
Constraints

$$\begin{aligned}\text{minimize} \quad & L(\boldsymbol{\theta}) \\ \text{subject to} \quad & \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x}_i) \leq \mathbf{c}, \\ & \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq -\mathbf{c},\end{aligned}$$



Maximizing Fairness Under  
Accuracy Constraints

$$\begin{aligned}\text{minimize} \quad & \left| \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x}_i) \right| \\ \text{subject to} \quad & L(\boldsymbol{\theta}) \leq (1 + \gamma)L(\boldsymbol{\theta}^*),\end{aligned}$$



# A7: Fairness-aware feature selection

- A framework for feature selection by computing the fairness-utility score for each feature which captures its **accuracy** and **discriminatory** impacts.
- The score depends only on the joint statistic of the data and not on the particular classifier at hand.

**Goal:** Select features that optimally satisfy accuracy and fairness requirements.

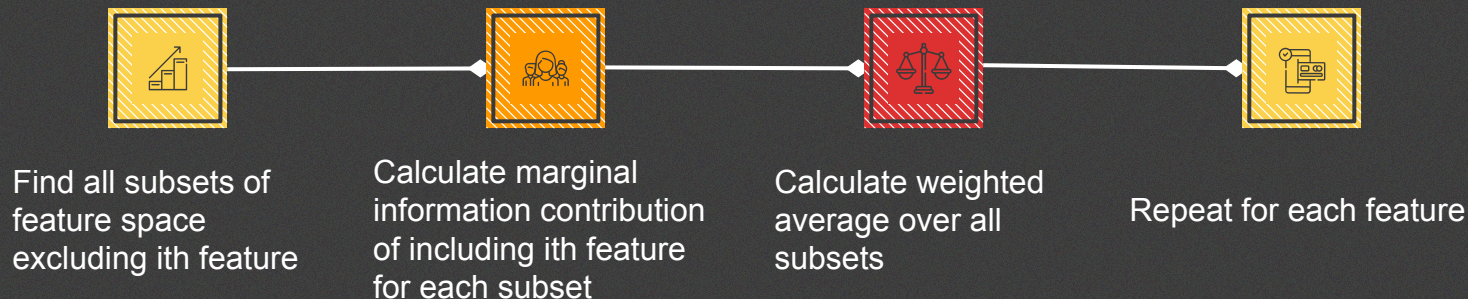
**Problem:** Tradeoff between accuracy and fairness. May remove some discriminatory features which contain important information to make prediction.



# How to measure those impacts?

- Propose accuracy coefficient and discrimination coefficient based on mutual information
- Aggregate these two coefficients by **shapley value function**

## Process







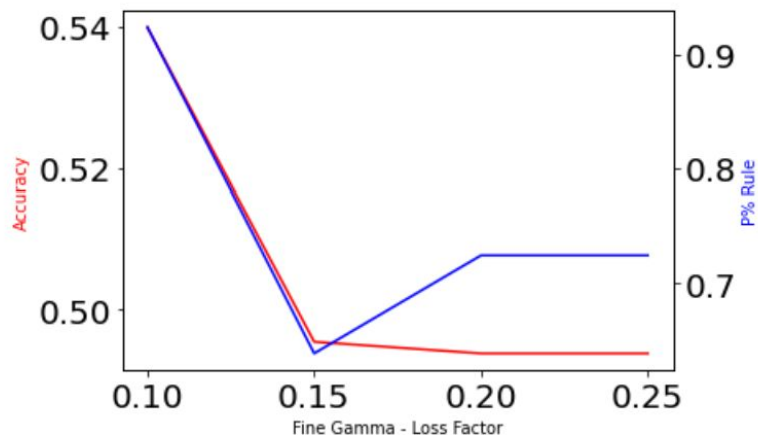
03

# Evaluation Results

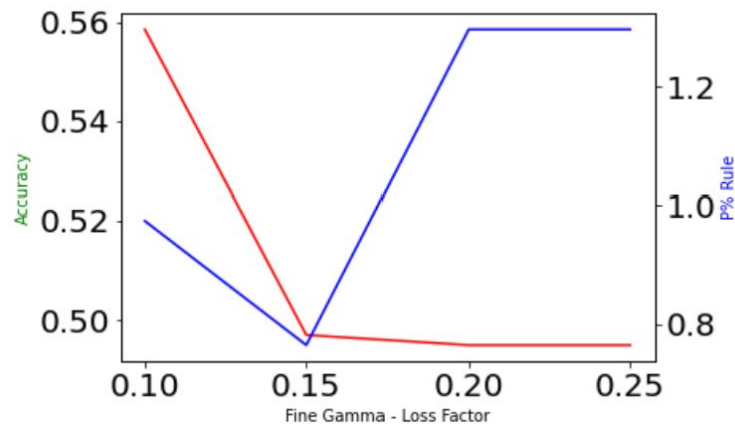


# A3: Fairness Constraints: Mechanisms for Fair Classification

Train Accuracy and P% Rule



Test Accuracy and P% Rule



We have the Train and Test Accuracy, which shows a decrease when Fine-Tuning Gamma, to give the constrained algorithm. P% also gives us the ratio between protected and unprotected class based on the test data, denoted by 1 and 0.



# Results

Initially, we used a Logistic Regression without any constraints, which gives us the accuracy of 0.67% where both the races are involved. Subsequently, we started the process of constraining the accuracy in order to ensure fairness of the model. Here, we started seeing a decrease in the accuracy but the calibration showed an accuracy, going close to 0.14, as compare to 0.07 without the constrained model.

	Models	Gammas	Accuracy	Accuracy_AA	Accuracy_CA	Calibration
0	Original Model without Constraints	-	0.679671	0.705128	0.634286	0.070842
1	Model with Constraints	0.1	0.588571	0.541667	0.541667	0.046905
2	Model with Constraints	0.15	0.577143	0.451923	0.451923	0.125220
3	Model with Constraints	0.2	0.582857	0.445513	0.445513	0.137344



# A7: Feature selection & Model Evaluation

- Hyperparameter( $\alpha$ ) tuning:

$$\mathcal{F}_i = \phi_i^{Acc} - \alpha \phi_i^D$$

- Calculate fairness-utility score for each feature under a choice of  $\alpha$
- Include features with high score or exclude features with low score(similar to Best Subsets Algorithm)
- Build ML model using features; evaluation metrics: accuracy and calibration

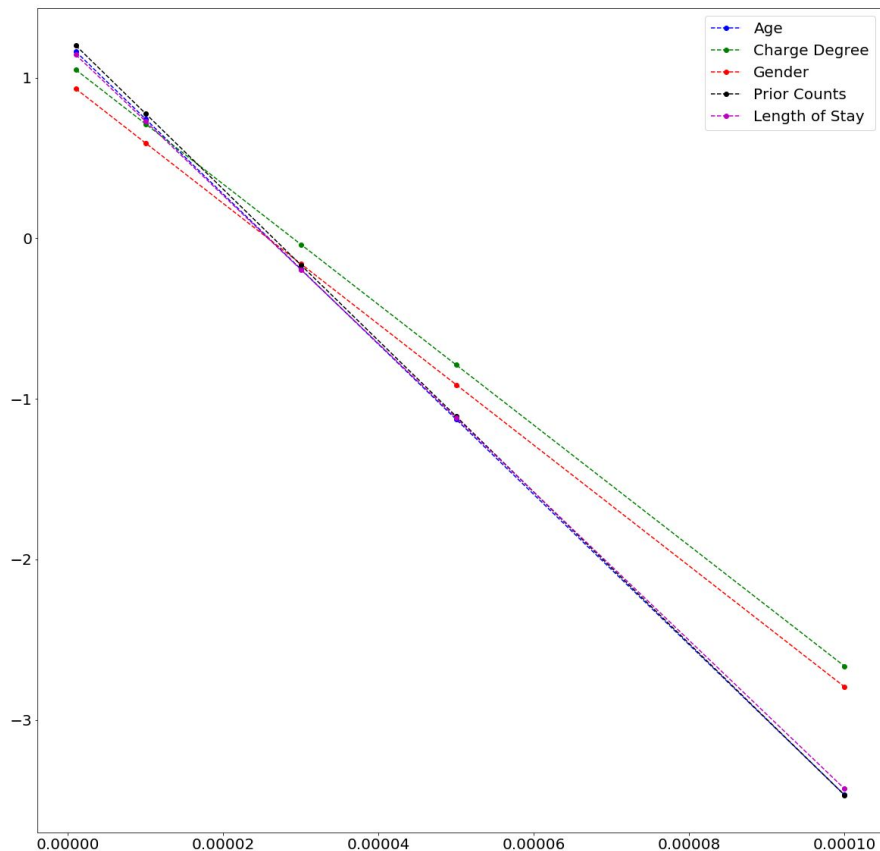
Example:  $\alpha = 0.00001$

Remove Gender to get a subset of size 4

	Features	Accuracy	Discrimination	fairness-utility score
0	Age	1.210114	46772.615106	0.742387
1	Charge Degree	1.087360	37525.470930	0.712105
2	Gender	0.969253	37630.375782	0.592949
3	Prior Counts	1.248108	47161.541173	0.776493
4	Length of Stay	1.190087	46187.281878	0.728214



# Tuning $\alpha$



To get a subset of size 4:

1) When  $\alpha=0.000001$  and  $\alpha=0.00001$ :  
remove **Gender**;

2) When  $\alpha=0.00003$ :  
remove **Length of Stay**;

3) When  $\alpha=0.00005$ :  
remove **Age**;

4) When  $\alpha=0.0001$ :  
remove **Prior Counts**.

To get a subset of size 3:

Remove the feature with the second  
Lowest score under each chosen  $\alpha$

...



# Feature selection results

- To maintain a considerable accuracy and complexity of the model, we compared the complete model and models with feature size of 4 under different choice of  $\alpha$

	Model	Alpha	Accuracy	Calibration
0	complete model	/	0.679671	0.070842
1	without gender	0.000001, 0.00001	0.673511	0.043388
2	without length of stay	0.00003	0.655031	0.085897
3	without age	0.00005	0.599589	0.097473
4	without prior counts	0.0001	0.562628	0.066538

- The model without **gender** is probably the best model considering the accuracy and discrimination effect at the same time.
- As expected, when  $\alpha$  increases, accuracy decreases. Calibration should also decrease but there is some variation. (complexity of the model? No outlying accuracy/discrimination coefficient? )



# References

Algorithm 3: Fairness Constraints: Mechanisms for Fair Classification Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi

Algorithm 7: Information Theoretic Measures for Fairness-aware Feature Selection (Sajad Khodadadian, Mohamed Nafea, AmirEmad Ghassami, Negar Kiyavash)





---

# Thanks!

Do you have any questions?

