Machine Learning Fairness

Group 10

Algorithms

- Prejudice Remover Regularizer(A5)
- Fairness-aware Feature Selection (A7)

COMPAS Dataset

Features: Gender, Age, Prior Count, Charge Degree, Length of Stay

Response: two_year_recid

Protected attribute: race (Caucasian/ African-American)

Algorithm 1: Prejudice Remover Regularizer(A5)

Framework

- Focused on classification and built regularizers into logistic regression models
- The parameters are tuned so as to maximize the log-likelihood :

$$\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) = \sum_{(y_i,\mathbf{x}_i,s_i)\in\mathcal{D}} \ln \mathcal{M}[y_i|\mathbf{x}_i,s_i;\boldsymbol{\Theta}].$$

- Adopted two types of regularizers
  (1) a standard one to avoid over-fitting, used an L2 regularizer $\|\boldsymbol{\Theta}\|_2^2$
  (2) $R(\mathcal{D},\boldsymbol{\Theta})$ to enforce fair classification

- The objective function to minimize is obtained :
  (where λ and η are positive regularization parameters)

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2,$$

Framework

- Logistic Regression
- Prejudice Remover Regularizer $\mathrm{R}_{\mathrm{PR}}(\mathcal{D}, \boldsymbol{\Theta})$ is :

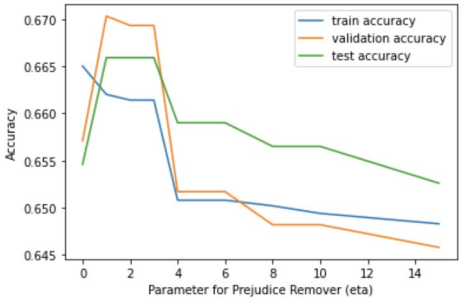
$$\sum_{(\mathbf{x}_i, s_i) \in \mathcal{D}} \sum_{y \in \{0,1\}} \mathcal{M}[y|\mathbf{x}_i, s_i; \boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]},$$

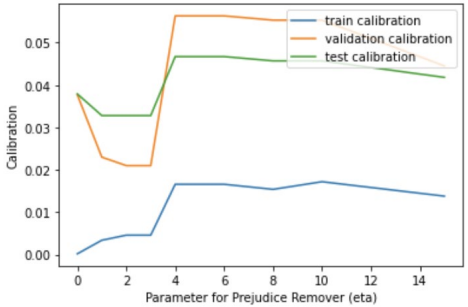
Objective Function

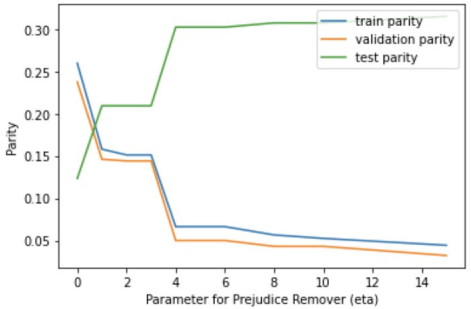
Minimize $$\sum_{(y_i, \mathbf{x}_i, s_i)} \ln \mathcal{M}[y_i|\mathbf{x}_i, s_i; \boldsymbol{\Theta}] + \eta \, \mathrm{R}_{\mathrm{PR}}(\mathcal{D}, \boldsymbol{\Theta}) + \frac{\lambda}{2} \sum_{s \in \mathcal{S}} \|\mathbf{w}_s\|_2^2,$$

Model Evaluation

- To compare with A7 paper, our first model chose 5 features same with A7.
- We use 3 evaluation metrics: **Accuracy, Calibration, Parity**







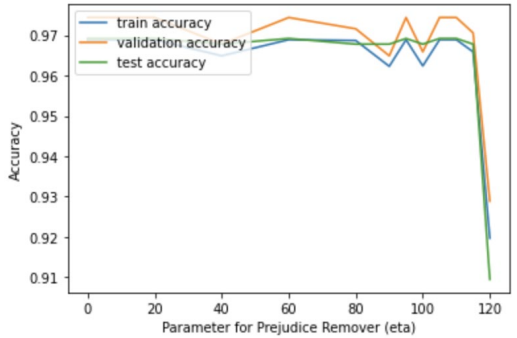
Model Performance

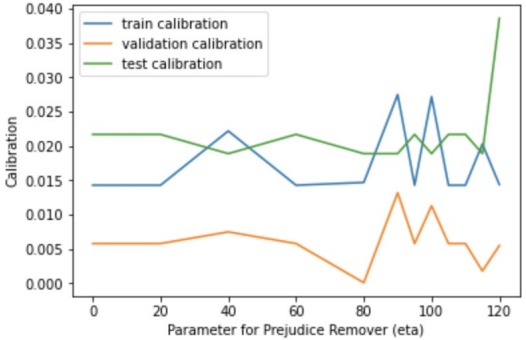
- To achieve high accuracy, low calibration and low parity, we decided to choose eta = 2.

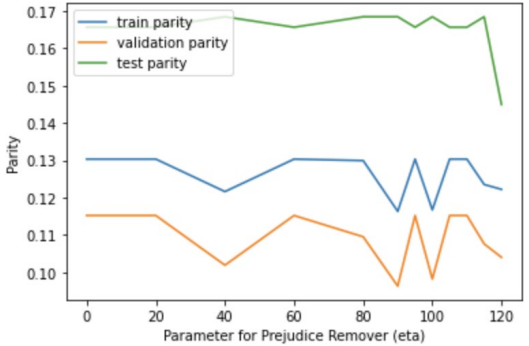
Model	Prejudice Remover Regularizer			Logistic Regression		
Evaluation	Training	Validation	**Testing**	Training	Validation	**Testing**
Accuracy	0.6614	0.6693	**0.6659**	0.665	0.6571	**0.6546**
Calibration	0.0046	0.021	**0.0328**	0.0002	0.0377	**0.0379**
Parity	0.1512	0.1441	**0.2093**	0.2596	0.2376	**0.1234**

Model Evaluation

- To get better performance, our second model chose different features based on correlation with Y.
- We use 3 evaluation metrics: **Accuracy, Calibration, Parity**







Model Performance

- To achieve high accuracy, low calibration and low parity, we decided to choose eta = 100, comparing with logistic regression with eta = 0.

Model	Prejudice Remover Regularizer			Logistic Regression		
Evaluation	Training	Validation	**Testing**	Training	Validation	**Testing**
Accuracy	0.9687	0.9716	**0.9678**	0.9689	0.9744	**0.9692**
Calibration	0.0147	0.0001	**0.0189**	0.0143	0.0058	**0.0217**
Parity	0.1299	0.1095	**0.1684**	0.1303	0.1152	**0.1656**

Summary

- As eta increases, accuracy will decrease since it sacrifices for fairness, calibration will also decrease.
- When we add prejudice remover regularizer, accuracy will decrease, calibration will also decrease.
- For this problem, the fairness looks good (calibration below 5% for all models), so the prejudice remover regularizer does not work well.

Algorithm 2: Fairness-aware Feature Selection (A7)

Fairness-aware Feature Selection

- The method is to develop a framework for fairness-aware feature selection based on correlation among features and the information theoretic measurements for accuracy and discriminatory impacts of features

- Different from some other methods. The framework depends on the joint statistics of the data rather than a particular classifier design.

Fairness-aware Feature Selection

1. We first propose information theoretic measures which quantify the impact of different subsets of features on the accuracy and discrimination.
2. We then deduce the marginal impact of each feature using Shapley value function.
3. Finally, we design a fairness utility score for each feature (for feature selection) which quantifies how this feature influences accurate as well as non-discriminatory decisions.

FFS: Step 1

- Accuracy coefficient on Subset of X:

Definition 1 (Accuracy coefficient). For a subset of features $X_S \subseteq X^n$, the accuracy coefficient of X_S is given by

$$v^{Acc}(X_S) = I(Y; X_S|\{A, X_{S^c}\}) = UI(Y; X_S\backslash\{A, X_{S^c}\}) + CI(Y; X_S, \{A, X_{S^c}\}). \quad (2.5)$$

- Discrimination coefficient on Subset of X:

Definition 2 (Discrimination coefficient). For a subset of features $X_S \subseteq X^n$, the discrimination coefficient is

$$v^D(X_S) \triangleq SI(Y; X_S, A) \times I(X_S; A) \times I(X_S; A|Y). \quad (2.6)$$

FFS: Step 1

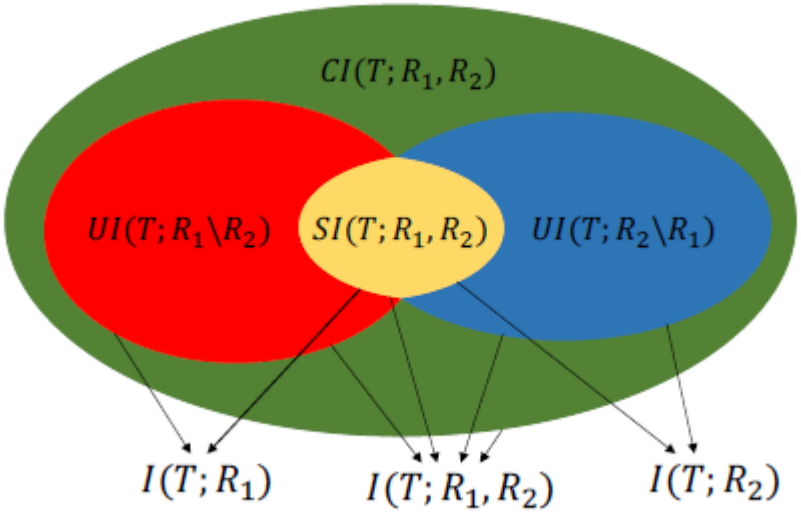


Figure 1: Decomposition of Information.

FFS: Step 2

- As we calculated the accuracy and discrimination on subsets of X, we don't know the correlation among the features. In order to account for this correlation, we need to factor in the aggregate effect of all subsets of features that include a certain feature. It leads us to the Shapley value function:

Definition 5. Let $\mathcal{P}$ denote the power set. Given a *characteristic function* $v(\cdot) : \mathcal{P}([n]) \to \mathbb{R}$, the Shapley value function $\phi_{(\cdot)} : [n] \to \mathbb{R}$ is defined as:

$$\phi_i = \sum_{T \subseteq [n] \setminus i} \frac{|T|!(n - |T| - 1)!}{n!} (v(T \cup \{i\}) - v(T)), \ \forall i \in [n].$$

FFS: Step 3

- Finally, we can get the accuracy and discrimination coefficients on each feature, and compare them to do the feature selection.
- Fairness-utility score for a feature Xi is defined as F = acc - α*disc. Set α = 0.00125 in this case.

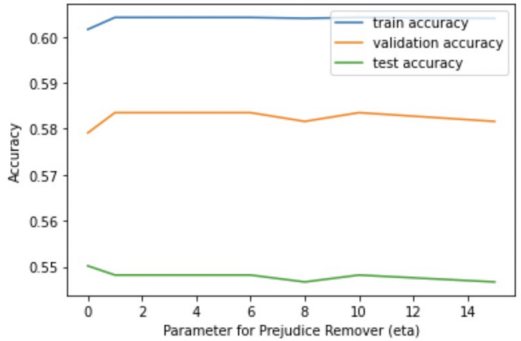
Results

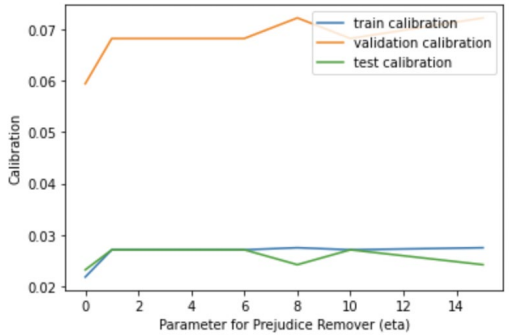
	Feature	Accuracy	Discrimination	F
1	Charge Degree	1.046473	765.7377	0.089301
2	Gender	0.973917	729.6456	0.061860
3	Age	1.181441	939.7405	0.006765
4	Prior Count	1.229856	982.4314	0.001817
5	Length of Stay	1.028396	908.0171	-0.10662

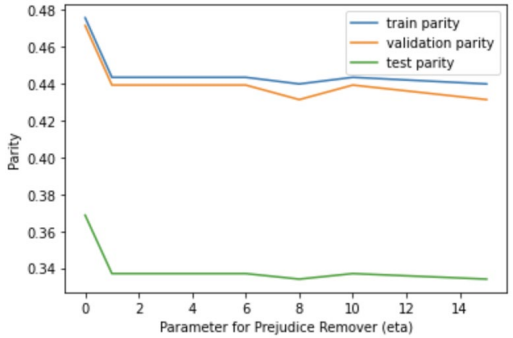
Same conclusion as paper shows.

Compare A5 and A7

- From the result of A7, we eliminate 2 most discriminative features: prior count and age and use the remaining 3 features.
- We use 3 evaluation metrics: **Accuracy, Calibration, Parity**







Compare A5 and A7

- To achieve high accuracy, low calibration and low parity, we decided to choose eta = 2.

Model	Prejudice Remover Regularizer			Logistic Regression		
Evaluation	Training	Validation	**Testing**	Training	Validation	**Testing**
Accuracy	0.6042	0.5853	**0.5482**	0.6016	0.5791	**0.5502**
Calibration	0.0271	0.0682	**0.0271**	0.0218	0.0594	**0.0232**
Parity	0.4435	0.4393	**0.337**	0.4758	0.4761	**0.3687**

References

Khodadadian, S., Nafea, M., Ghassami, A., & Kiyavash, N. (2021). Information Theoretic Measures for Fairness-aware Feature Selection. arXiv preprint arXiv:2106.00772.