AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Name: Usman Abdul-Ganiy B. Assignment Number: 3
Course: Fairness/Privacy Date: June 3, 2019

Solution 2

1 10 features most correlated with Y and A:

Y	A
marital-status_Married-civ-spouse	sex_Female
relationship_Husband	sex_Male
education_num	relationship_Husband
marital-status_Never-married	marital-status_Married-civ-spouse
age_u30	relationship_Unmarried
hours-per-week	relationship_Wife
relationship_Own-child	occupation_Adm-clerical
capital-gain	hours-per-week
sex_Female	marital-status_Divorced
sex_Male	occupation_Craft-repair

Table 1: Correlated features

2 Classification accuracy and Δ_{DP} on test set;

Prediction = Y	Test set	
	All features	without Corr features with A
Accuracy	0.838769	0.834654
Δ_{DP}	0.155376	0.131491

Table 2: Prediction metrics for $\hat{Y}$

Table 2 reveals that the accuracy of the classifier reduces when features correlated with the sensitive attribute are removed. However, fairness is gained, this is observed from the reduction in value of the demographic parity metric.

The sensitive group with higher value on $\hat{Y}$, on average, is when $A = 1$

3 Features correlated with $\hat{Y}$:

$\hat{Y}$	$\hat{Y}/A=0$	$\hat{Y}/A=1$
capital-loss	capital-loss	education_num
education_num	relationship_Wife	capital-loss
marital-status_Married-civ-spouse	marital-status_Married-civ-spouse	education_Bachelors

Table 3: Correlated features with $\hat{Y}$

4 With attributes sex_Male and sex_Female removed;

Prediction = A	Test set	
	without {sex_Male, sex_Female }	without Corr features with A
Accuracy	0.462195	0.457896
Reweighted accuracy	0.574188	0.565746

Table 4: Prediction metrics for $\hat{A}$

Solution 3

1 Using the preprocessed data;

		Test set
$\hat{Y}=g(X)$	Accuracy	0.850071
	Δ_{DP}	0.122558
$\hat{A}=h(X)$	Accuracy	0.452306
	Reweighted accuracy	0.561279

Table 5: Prediction metrics

2 preprocessed data with MMD regularizer;

		Test set
$\hat{Y}=g(X)$	Accuracy	0.850623
	Δ_{DP}	0.137124
$\hat{A}=h(X)$	Accuracy	0.458510
	Reweighted accuracy	0.568562

Table 6: Prediction metrics

From tables 5 and 6 the preprocessed data only, without the MMD regularizer, gives a better result. This might probably be because the solution is learned end-to-end.

3 Range of α hyperparameters

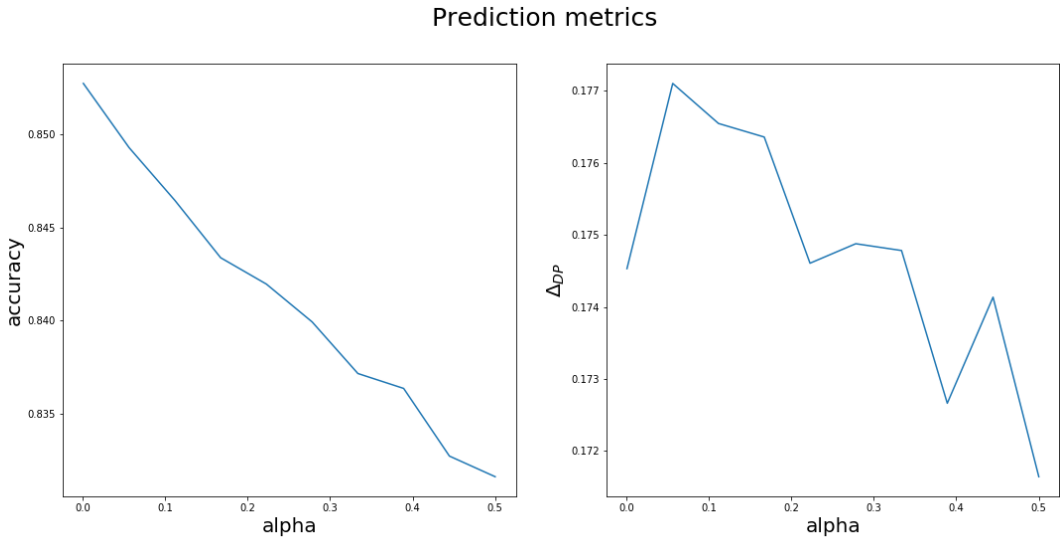

Figure 1: prediction plots for α

4 The Demographic parity seem to be the best metric to compare the methods, in the case where Y is predicted. The closer the measure to zero, 0, the fairer the classifier.

5 Another way to remove information about A is to build Generative Adversarial Networks, GANs. A fair classifier of Y can be trained by leveraging adversarial networks to build an outcome distribution of the model which doesn't depends on the sensitive attribute A. Predictions that are independent of the sensitive attribute can then be enforced.