Learning Fair Representations: Supplementary Materials

Datasets

- German Credit Dataset. The dataset has 1000 instances which classify the bank account holders into credit class *Good* or *Bad*. Each person is described by 20 attributes, which include 13 categorical and 7 numerical attributes. In our experiments we consider *Age* as the sensitive attribute, following earlier papers.

- Adult Income Dataset. The dataset has 45,222 instances. The target variable indicates whether or not income is larger than 50K dollars, and the sensitive feature is *Gender*. Each data object is described by 14 attributes which include 8 categorical and 6 numerical attributes. A full description can be found at (Kohavi, 1996).

- Heritage Health Dataset. This dataset was derived from the Heritage Health Prize milestone 1 challenge. We used the same features as the winning team, Market Makers. The goal of this dataset is to predict the number of days that a person will spend in the hospital in a given year. To convert this into a binary classification task, we simply predict whether they will spend any days in the hospital that year. We split the patients into two groups based on whether they are older or younger than 65. As a pre-processing step, we binarize the data by quantizing any non-categorical variables.

For all of the datasets, in order to facilitate comparison to the other methods, notably the naive Bayes method which assumes binary variables, we transform all attributes to binary variables by using "one-hot encoding" for each categorical attribute (one variable per value, always exactly one one in the set), and quantization for numerical attributes.

After the modification, each example in the German Credit Dataset has 61 binary features while the individuals in the Adult Income Dataset are described by 103 binary features, including the sensitive feature.

For the German Credit Dataset, we optimized each method across five splits, each containing 50% of the data as a training set, 20% as validation and 30% as a test set. The Adult Income Dataset is already divided so that the training set contains two-thirds of the data with the remainder set aside for test. Here we optimized each method across five splits, each utilizing one-third of the training set as a validation set and the rest for training. For the Health dataset we split the data into five equally sized folds. We train models on three of the folds independently, and test each model using one of the remaining folds for validation and one for testing.

Tables of Results

Here we present the full quantitative results comparing the methods on the various datasets.

Table 1. Performance of the various models when optimizing discrimination

Dataset	Method	Delta	yAcc	yDiscrim
Health	LR	0.6482	0.7547	0.1064
	FNB	0.5678	0.6878	0.1200
	RKR	0.7038	0.7212	0.0174
	LFR	0.7365	0.7365	0.0000
Adult	LR	0.4895	0.6787	0.1892
	FNB	0.7711	0.7847	0.0136
	RLR	0.6494	0.6758	0.0264
	LFR	0.7018	0.7023	0.0006
German	LR	0.5517	0.6790	0.1273
	FNB	0.6314	0.6888	0.0574
	RLR	0.5842	0.5953	0.0111
	LFR	0.5867	0.5909	0.0042

Finally, an aspect of the model that bears scrutiny is its sensitivity to parameter settings and initialization. First, we found that the results of the models were consistent across splits of the data; for example, the variance of the accuracy in the validation sets were 3.12e-06, 5.80e-7 and 1.85e-04 and the variance of the discrimination was 6.38e-05, 3.43e-5 and 8.20e-04 on the Health, Adult and German datasets respectively. We also found that our model obtained fairly similar results across a range of settings for the hyperparameters, with the expected effect on the learned system. As an example, we show in Table 4 the results of our model as we vary the number of prototypes K, while maintaining the setting of the other hyperpa-

Table 2. Performance of the various models when optimizing Delta = yAcc-yDiscrim

Dataset	Method	Delta	yAcc	yDiscrim
Health	LR	0.6568	0.7656	0.1087
	FNB	0.5678	0.6878	0.1200
	RLR	0.7314	0.7553	0.0239
	LFR	0.7512	0.7521	0.0009
Adult	LR	0.5971	0.7931	0.1960
	FNB	0.7711	0.7847	0.0136
	RLR	0.6494	0.6758	0.0264
	LFR	0.7701	0.7721	0.0020
German	LR	0.5517	0.6790	0.1273
	FNB	0.6314	0.6888	0.0574
	RLR	0.6043	0.6447	0.0404
	LFR	0.6405	0.6708	0.0302

Table 3. Individual fairness

Dataset	Method	yNN
Health	LR	0.7233
	FNB	0.5893
	RLR	0.6223
	LFR	1.000
Adult	LR	0.7297
	FNB	0.5634
	RLR	0.7766
	LFR	0.8108
German	LR	0.6950
	FNB	0.6868
	RLR	0.8716
	LFR	0.9408

Table 4. Performance with varying K

Dataset	K	yAcc	yDisc
Health	10	0.7414	0.0002
	20	0.7458	0.0012
	30	0.7502	0.0025
Adult	10	0.7040	0.0012
	20	0.7548	0.0012
	30	0.7683	0.0025
German	10	0.5871	0.0039
	20	0.6611	0.0409
	30	0.6789	0.1253

rameters ($A_x = 0.01, A_y = 1, A_z = 50$). The trend is clear: adding more prototypes increases the accuracy as it allows finer classification decisions, but also leads to more discrimination as it is more difficult to remove information about membership in the protected set.