EECS 6322 Week 13, Paper 2
Learning Fair Representations

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper proposes a new algorithm to achieve both group fairness and individual fairness in machine learning models. Group fairness is achieved when the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole, whereas individual fairness is achieved when similar individuals are treated similarly. The paper formulates a fairness optimization problem of finding a good representation of the data with two competing goals. The first is to encode the data as well as possible, and the second is to remove information about membership of a protected group. The authors also show that the intermediate representations of this method can be used for transfer learning and they propose an additional distance metric which can find important dimensions of the data for classification.

The current systems that are trained to make decisions based on historical data inherit the biases in the data which leads to bias and discrimination in decision systems that rely on statistical inference and learning. One way to improve fairness in these systems is to make the automated decision-maker blind to some attributes which may be difficult since many attributes may be correlated with the protected one. Hence, the systems need to be modified to make decisions that are not unduly biased for or against protected subgroups in the population. The important goal of this fair classification is to ensure group fairness and individual fairness. Group fairness ensures that the overall proportion of members in a protected group receiving positive (negative) classification are identical to the proportion of the population as a whole. Individual fairness ensures that any two individuals who are similar with respect to a particular task should be classified similarly.

The framework as proposed by the authors maps each individual, represented as a data point in a given input space, to a probability distribution in a new representation space. This new representation aims to lose any information that can identify whether the person belongs to the protected subgroup, while retaining as much other information as possible. Removing information about membership in a protected group can be formulated by the notion of statistical parity, which requires that the probability that a random element from X^+ maps to a particular prototype is equal to the probability that a random element from X^- maps to the same prototype. This is given by: $P(Z = k|x) = exp(-d(x, v_k))/\sum_{j=1}^{K} exp(-d(x, v_j))$. Thus, the model can be defined as a discriminative clustering model, where the prototypes act as the clusters and each input example is stochastically assigned to a prototype, which are in turn used to predict the class for that example.

The goal of the proposed model is to learn a good prototype set such that the mapping from the input to prototype set satisfies statistical parity and the mappings retain the input information, except for membership in the protected set. Thus, the learning objective can be formulated by $L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$, which the model aims to minimize. Here, the first term L_z in the objective function is given by $L_z = \sum_{k=1}^{K} |M_k^+ - M_k^-|$ where M_k^+ and M_k^- can be obtained by $M_k^+ = E_{x\epsilon X} + P(Z = k|x) = \frac{1}{|X_0^+|} \sum_{n\epsilon X_0^+} M_{n,k}$. The second term in the objective function L_x is given by $L_x = \sum_{n=1}^{N} (x_n - \hat{x}_n)^2$ where $\hat{x}_n = \sum_{k=1}^{K} M_{n,k}\boldsymbol{v}_k$. These $L_z and L_x$ terms in the objective function encourage the system to encode all information in the input attributes except for those that can lead to biased decisions. The last term L_y in the objective function ensures that the prediction of y is as accurate as possible and is given by $L_y = \sum_{n=1}^{N} -y_n log\hat{y}_n - (1 - y_n)log(1 - \hat{y}_n)$ where $\hat{y}_n = \sum_{k=1}^{K} M_{n,k}w_k$ and the values for w_k are constrained between 0 and 1.

The framework also introduces individual weight parameters for each feature dimension α_i to allow different input features to have different levels of impact. These weight parameters act as inverse precision values in the distance function and are given by $d(\boldsymbol{x}_n, \boldsymbol{v}_k, \alpha) = \sum_{i=1}^{D} \alpha_i (x_{ni} - v_{ki})^2$. Thus, as per the objective function, the first term enforces group fairness, as defined by statistical parity. Furthermore the authors also prove that even though the parity constraint does not directly address classification, under the model formulation of the framework, the two concepts are closely linked. The model also permits generalization to new examples distinct from those in the training set, since the mapping to Z is defined for any individual $x\epsilon X$. Finally, allowing the model to adapt the weights on the input dimensions ensures learning a good distance metric and the use of the same mapping function for all individuals in the group encourages individual fairness (as nearby inputs are mapped to similar representations).

The model was evaluated using different algorithms such as Fair Naive-Bayes, Regularized Logistic Regression, and Un-regularized Logistic Regression. The performance of the model was calculated using different metrics that assesses the consistency of the model classifications locally in input space, where values close to one indicate that similar inputs are treated similarly. The accuracy metric is given by $yAcc = 1 - \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n|$ *which measures the accuracy of the model classification prediction. The discrimination metric is given by* $yDiscrim = |\frac{\sum_{n:s_n=1}\hat{y}_n}{\sum_{n:s_n=1}1} - \frac{\sum_{n:s_n=0}\hat{y}_n}{\sum_{n:s_n=0}1}|$ *which measures the bias with respect to the sensitive feature S in the classification. The consistency metric is given by* $yNN = 1 - \frac{1}{Nk}\sum_{n}|\hat{y}_n - \sum_{j \epsilon kNN(x_n)}\hat{y}_j|$ *which compares a model's classification prediction of a given data item* x *to its* k*-nearest neighbors,* $kNN(x)$*. The model results are evaluated on two datasets from the UCI ML-repository, namely the German credit dataset and the Adult income dataset.*

The results that evaluate the model's performance on the validation set as compared against baselines such as Logistic Regression, Fair Naive Bayes, and Regularized Logistic Regression show that the proposed model (Learned Fair Representations) is capable of pushing the discrimination to very low values, while maintaining fairly high accuracy. The results are consistent in all three datasets, and across the validation criteria. In addition it is also seen that the model is trying to preserve information about the data. Further, comparing the models with respect to individual fairness, show that for each dataset, the proposed model obtained better individual fairness, which is likely due to the optimization criteria rewarding Z's preservation of information about X. Finally, with respect to transfer learning, the results showed that although the transferred representations suffered a small loss in accuracy, it significantly removed bias in the classification.

II. Strengths

The proposed model formulates fairness as an optimization problem of finding an intermediate representation of the data that best encodes the data while simultaneously removing any information about membership with respect to the protected group. The model was implemented on three data sets and showed positive results compared to other known techniques, which proved the model's efficiency on fairness.

III. Weaknesses

In transfer learning, although the model reduces bias in classification, the accuracy of the model decreases significantly. Hence, there is a tradeoff between the model accuracy and the reduced bias in classification.

IV. Correctness

The claims and empirical methodology as proposed by the paper are correct and are backed by theoretical grounding and comparisons against the current state of the art models. The paper also proposes evaluation metrics to measure the fairness of the framework on the validation set which prove the model to be efficient against its counterparts in terms of fairness.

V. Clarity

The paper is well written in general with the architecture, aim, and metrics clearly defined throughout the paper. The notations and parameters as used for the model architecture are also elaborated upon which supplements the understanding of the paper.

VI. Relation to prior work

Prior work in the field of fairness includes approaches that aim to achieve the first goal, group fairness, by adapting standard learning approaches through a form of fairness regularizer, or by re-labeling the training data to achieve statistical parity. Other works also include frameworks which attempt to achieve both group and individual fairness, where the goal is to define a probabilistic mapping from individuals to an intermediate representation such that the mapping achieves both. These approaches differ from the algorithm as proposed in the paper since their framework is not formulated as a learning problem and the distance metric as defined by the approaches may be unrealistic in certain settings.

On the contrary, the framework as proposed by the authors develops a learning approach to solving the fairness problem and learns a restricted form of a distance function as well as the intermediate representation. Additionally, the framework formulates fairness as an optimization problem of finding an intermediate representation of the data that best encodes the data, while simultaneously removing any information about membership with respect to the protected subgroup. The authors prove that this intermediate representation provides fairness in classification, since it is composable and not ad hoc.

VII. Reproducibility

Sufficient details in terms of the model architecture, code implementation, notations, and parameters are provided in the paper to reproduce the major results of the paper. Additionally, the baselines models, datasets, and evaluation metrics have also been clearly defined which would aid in reproducing the proposed work.

VIII. Additional comments

A further study on other forms of intermediate representations beyond prototypes, which utilize multi-dimensional distributed representations would supplement the work proposed by the authors.