CS 226 Final Project Report: Learning Comprehensively Fair Representations

Jonathan Chu, Yi Lin Wang

May 7, 2020

1 Introduction

For our final project, we worked on improving upon current techniques for learning fair data representations. Given a dataset X, a representation of this data is an alternate space Z which encodes the important features of X. As such, we are often interested in different mappings $X \to Z$ depending on how they encode properties of X. The general goal within the research area of fair representations is to learn mappings $X \to Z$ such that Z preserves essential features of X for important classification tasks while obscuring features which may be exploited unfairly. One way of formulating this condition is that Z should encode all essential properties of X except for certain *sensitive attributes* A, a framework which is very convenient for stipulating conditions which respect group fairness. This way, classification can be performed on Z in a *fair* way; perhaps respecting statistical (demographic) parity, in that probabilities of assignment are the same in expectation for groups in X that do or do not possess sensitive attributes, or equalized odds, in which the false positive and false negative rates across groups with different sensitive attributes are the same.

While current research has produced techniques which are able to learn representations representing various definitions of group fairness, none are able to guarantee individual fairness. Individual fairness presents a more microscopic notion of group fairness, in which similar individuals should face similar outcomes. In this project, we extend methods of learning fair representations to include individual fairness. Building off of the LAFTR framework [3] for learning fair representations using adversarial training, we propose a modification to the objective function which imposes a constraint that individuals close by in X are mapped to points in the latent space Z which are also not too far part. We then describe experiments and challenges using this method. This approach not only provides the group fairness guarantees and modular representation advantage of Z achieved by LAFTR, but further respects individual fairness - thus producing "comprehensively" fair representations.

2 Background and Previous Work

2.1 Notions of Fairness

First, we provide formal definitions of group fairness and individual fairness. Consider a dataset of individuals $X \in \mathbb{R}^n$, which posses a sensitive attribute $A \in \{0, 1\}$, and labels $Y \in \{0, 1\}$. A predictor outputting predictions $\hat{Y} \in \{0, 1\}$ respects *demographic parity* if

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1).$$

This is not always an appropriate criterion, however, in particular if the natural base rates disagree:

$$P(Y = 1|A = 0) \neq P(Y = 1|A = 1).$$

A different criterion then is *equalized odds*, in which it is instead the false positive and false negative rates in each group that agree:

$$P(\hat{Y} \neq Y|a = 0, Y = y) = P(\hat{Y} \neq Y|A = 1, Y = y) \ \forall y \in \{0, 1\}.$$

A variant of the above condition in which we only have $y = 0$ is known as *equal opportunity*. Now consider a slightly different classification scenario where individuals $x \in X$ are (randomly) mapped to distributions of outcomes $x \to M(x)$. Given a metric d on X a notion of distance of distributions D, in the classical definition by Dwork et al. (2012), we say the classifier is *individually fair* if

$$D(M(x), M(x')) \leq d(x, x').$$

2.2 Fair Representations

We build on existing work in learning fair representations of data, taking inspiration from two works in particular. Madras et al. (2018) proposed "Learning adversarially fair and transferable representations"(LAFTR) which used an adversarial learning method that attempts to satisfy the various notions of group fairness defined previously. In their framework, there is an input set X, label set Y, and sensitive information set A. The approach constructs a representation Z of X, and the goal is to mazimize accuracy of the classifier $g : Z \to Y$ while minimizing accuracy of an adversary $h : Z \to A$. They achieve this by constructing Z using an encoder-decoder framework as pictured below, and use the loss function

$$L(f, g, h, k) = \alpha L_C(g(f(X, A)), Y) + \beta L_{Dec}(k(f(X, A), A), x) + \gamma L_{Adv}(h(f(X, A), A)).$$

The advantage here is that the representation space Z is modular, so that it can be used for other classification tasks, despite not having been constructed while training for other tasks. This work did not mention individual fairness, so we wanted to obtain similar modularity advantages while accounting for individual fairness in some way.

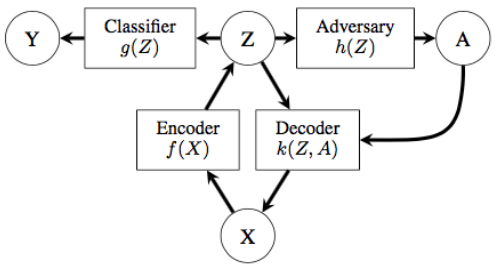

Figure 1: Madras et al. (2018)

How can we fit individual fairness into this framework? Recall the classical definition of individual fairness by Dwork et al. (2012) under classification: we have a set of individuals whom we

wish to classify, and a probabilistic classifier C that maps each individual to a probability distribution over classifications such that similar individuals are mapped to similar probability distributions, where similarity is defined by some distance on distributions, for example the total variation distance.

One possible interpretation of individual fairness under this representation space approach is to consider a situation similar to the one described above, where we construct a representation Z of X and want to maximize the accuracy of a classifier $g : Z \to Y$, but this time we desire that nearby points in X are mapped to nearby points in Z.

This approach was partially inspired by that of Ruoss et al. (2020), who proposed "Learning certifiable individually fair representations" (LCIFR), which is a framework for learning representations that respect individual fairness, which can then be used without knowing the fairness metric, which allows for modular use.

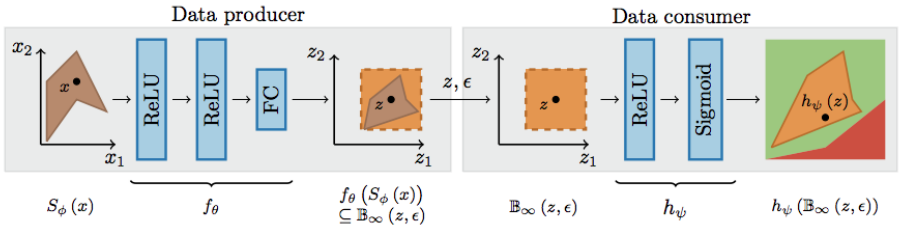

Figure 2: Ruoss et al. (2020)

The idea is that for any z in the representation space, individuals which are similar up to a certain bounded region by the metric should all receive the same classification. Below, we formalize these ideas in the definition of a loss function term as well as evaluation metrics to incorporate this sense of individual fairness.

3 Contribution

We propose a modification to the objective loss in LAFTR so that we retain the classification loss that comes from supervised learning with the "real" labels of a dataset, but also consider a loss term that is determined by the number of pairs of points that are close on X but too far apart when mapped to Z. The representation space should encode the data "accurately" but subject to constraints: here the constraints are specifically that we don't have too many "close" points that are mapped too far away from each other in the representation space.

Let d_X be a nonnegative metric on X and let d_Z be a nonnegative metric on Z. We use the loss function

$$L(f, g, h, k) = \alpha L_C(g(f(X, A)), Y) + \beta L_{Dec}(k(f(X, A), A), x) + \gamma L_{Adv}(h(f(X, A), A)) + \delta L_{Ind},$$

where

$$L_{Ind} = \frac{1}{|(X, A) \times (X, A)|} \sum_{(x,x') \in (X,A) \times (X,A)} ReLU\left(\frac{d_Z(f(x), f(x'))}{d_X(x, x')} - M\right)$$

for some ratio M. In our implementation, we set $M = 1$.

This "individual fairness loss" term tells us that for pairs of data points in X at a distance δ away from each other, when they are mapped to Z and have distance ϵ from each other, we use a threshold that is activated when the ratio ϵ/δ is too large. Intuitively, we want this to be the case because the closer the points in X, the more strict we want our requirements for closeness on Z, and the farther the points in X, the looser we can be about how far their mapped representatives are in Z.

4 Experiments

Our experimental approach was to implement our loss function, train models using the LAFTR framework, and evaluate the individual fairness of the learned representations using metrics. Our implementation and modifications, based on LAFTR (Madras et al., 2018), can be found at `https://github.com/jonathanchu33/laftr`.

4.1 Dataset

We used the UCI adult dataset, one of the datasets experimented on by Madres et al. (2018). The first question was how to define a suitable metric on this dataset for individual fairness metric evaluation; the original dataset consists of 14 categorical attributes, which are discrete and some non-numeric. Instead, the authors use a post-processed version of the dataset, in which the original data is encoded using one hot representation which assigns binary vectors with a single 1 component to categorical data, eventually leading to 112 attributes. The data is then normalized by subtracting the mean and dividing by the standard deviation along every dimension. We defined our metric distance on the dataset as simply the ℓ^2 norm on this postprocessed representation.

4.2 Implementation

We implemented our proposed loss function from section 3. Due to the computational constraints required for a dataset of n points to check all n^2 pairs of points, we only calculated our loss function over the pairs of points in batches of size 64. We faced challenges integrating our loss function to the existing optimization framework of LAFTR, however. We hypothesized that this was because the natural pairwise dependency of our loss function on the datapoints presented a difficulty for the autodifferentiation. Our approach, for instance, requires that batche sizes be ≥ 2. Because of this difficulty, we trained our models on the original loss function and evaluated the values of our individual fairness loss term in training. Then, we assess the efficacy of this loss term by proxy, by comparing it to our individual fairness metrics below.

4.3 Metrics

We consider 3 metrics to asses the individual fairness of the learned representation. Given two individuals $x, x' \in X$, two of the metrics are based on the ratio

$$\frac{d_Z(f(x), f(x'))}{d_X(x, x')}.$$

This is very clearly inspired by the definition of individual fairness; an individually fair representation should provide a reasonable bound on this ratio for all pairs of individuals. We thus compute

the average ratio across the entire dataset as well as the maximum such ratio.

Next, we also consider classification similarity between pairs of individuals within explicitly bounded distances in the latent space. Figure 2 illustrates this intuition: for some $z \in Z$, we expect that all individuals within some fixed distance (the brown rectangle in the diagram, an l_∞ norm) to be classified the same. Thus for various distance thresholds b, we define a subset of pairs

$$S_b = \{(x, x') \in X \times X | d_Z(f(x), f(x')) \leq b\}$$

and calculate the percentage of pairs for which the desired classification similarity is true:

$$\frac{|\{(x, x') \in S_b | f(x) = f(x')\}|}{|S_b|}. \tag{1}$$

5 Results

We trained four models used by the authors of LAFTR, each of which are different variants of GANs. DP stands for demographic parity, while EqO for equal odds. The adversarial loss term varies slightly for each of these variants, but we leave the details to the original paper.

Table 1 shows the value of our individual fairness loss term compared with our first two evaluation metrics:

Model	Individual Fairness Loss	Average ratio	Maximum ratio
DP GAN	.168	.660	10976.541
DP WassGAN	.143	.650	23923.812
Weighted DP WassGAN	.027	.420	9523.564
EqO GAN	.042	.430	13239.420

Table 1: Individual Fairness Loss by model.

It is no surprise that the higher fairness losses correspond to higher average ratios. While the maximum ratios are enormous, the average ratio indicates that the learned representations are reasonable and that there are few outlying pairs at the extreme end of the worst case scenarios. The next table documents (1) as a function of b:

↓ Model / Dist. Threshold →	16	8	4	2	1	.5	.25	.1	.05	.025	.01
DP GAN	0.659	0.692	0.722	0.744	0.757	0.764	0.767	0.769	0.770	0.770	0.771
DP WassGAN	0.731	0.757	0.774	0.783	0.788	0.791	0.792	0.793	0.793	0.793	0.793
Weighted DP WassGAN	0.678	0.704	0.725	0.737	0.743	0.746	0.747	0.748	0.749	0.749	0.749
EqO GAN	0.672	0.683	0.689	0.693	0.694	0.695	0.696	0.696	0.696	0.696	0.696

Table 2: Percentage of pairs within a bounded distance (in representation space) which are both given the same classification outcome. As the bounded distance between pairs decreases, the percentage of equal classifications increases.

The trend is that the percentage of similarly classified pairs increases as the bounded region of consideration becomes smaller, as expected. Now, one might look for a direct relationship between the individual fairness losses and the percentages in Table 2 for some fixed b. If we fix b, there

does not appear to be a relationship, but the average ratios in table 1 indicates that fixing b across different models is not the correct standard for comparison, as the typical region for the same similar individuals in Z should be larger when using the models which yield a larger average ratio; this simply corresponds to the scaling of the metric.

6 Discussion and Further Work

Our results verify that the proposed individual fairness loss function is well motivated. However, one main obstacle remains the dependence on pairs of datapoints. One interesting area for future work is the development of an individual fairness constraint which can be implemented using more standard methods, as Madras et al. (2018) have done for group fairness using the standard applications of GANs.

For further exploration, one possible variant on our method is to try using other distance metrics, like various ℓ^p norms, and train on different datasets to obtain a more complete understanding of the merits of this approach. Due to the computational constraints described above on calculating distances with respect to all pairs of points in a dataset, we implicitly considered only dealing with pairs of points that are at most some fixed constant distance δ away from each other in X by calculating the loss on separate minibatches of the data.

Another possible modification is the addition of a cycle consistency loss term: $L_{cyc}(f(k(Z, A), A), Z)$. In addition to the reconstruction loss, we should expect the fair representation to be reconstructable as well, independent of sensitive attribute A. In adversarial learning literature, Zhu and Park et al. (2017) refer to constraints that $F(G(X)) \sim X$ and $G(F(Y)) \sim Y$ as cycle consistency because the conditions encourage rough bijection. This method would require thorough experimentation, as the original approach is taken from a different context of style transfer, though also using a GAN framework.

Overall, the learning of a representation which balances both individual and group aspects of fairness is most desirable. While our work makes progress towards this goal, much room remains for improvement towards more simplistic, accurate, and elegant methods.

7 References

1. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R., 2012, January. Fairness Through Awareness. In ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214-226).

2. Jun-Yan Zhu*, Taesung Park*, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, in IEEE International Conference on Computer Vision (ICCV), 2017.

3. Madras, D., Creager, E., Pitassi, T. and Zemel, R., 2018. Learning adversarially fair and transferable representations. arXiv preprint arXiv:1802.06309.

4. Ruoss, A., Balunovic, M., Fischer, M., Vechev, M., 2020, February, Learning certified individually fair representations. arXiv preprint arXiv:2002.10312v1

5. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. and Dwork, C., 2013, February. Learning fair representations. In International Conference on Machine Learning (pp. 325-333).