Note on Computational Tractability of Non-Trivial Fairness Assignment

Bobby (Zelin) Lv

UW-Madison, zlv7@wisc.edu

December 2, 2019

1 Introduction

In this note, we summarize the framework of Non-Trivial Fairness Assignment given by **[JMR17]**. **[JMR17]** proved that the fairness conditions of well-calibrated, balance for negative class and balance for positive class can't be all satisfied at the same time unless the instance given can have perfect prediction or the two groups of instance have equal base rate. Furthermore, they mention that the Computational Tractability of Non-Trivial Fairness Assignment of equal base rate is unsolved. So here we note some thoughts on the Tractability of Non-Trivial Fairness Assignment of equal base rate.

2 Problem Setting

Here we briefly describe the problem setting. Most of details and intuitive definition of this problem can be found in **[JMR17]**.

2.1 Instance

An instance is given as a tuple, (G_1, G_2, σ, p). G_t is group of t and each member of each group has a feature vector from σ and a label. And each entry in $n_t \in \mathbf{R}^{|\sigma|}$ denotes the number of member in group t has feature σ. Here in this problem, we just focus on the case where the label is binary. We use $|\sigma|$ to denote the number of feature vectors. Each entry in the vector $p \in \mathbf{R}^{|\sigma|}$ in position σ denotes the probability that a member with σ belong to the positive class. And we have $P \in \mathbf{R}^{|\sigma|*|\sigma|}$ as the diagonal matrix version of p. We denote the number of members of G_t as $N_t = |G_t|$ and μ_t as the number of members of G_t belong to positive class. Since we know two groups have equal base rate, so we have $\frac{\mu_1}{N_1} = \frac{\mu_2}{N_2}$.

2.2 Fairness Assignment

Given a model, we want to have assignment of these two groups and similar to the instance, we define the assignment as a tuple (X, v, B). B is the number of bins, here we can view it as the variable of dimension of assignment matrix X and score vector v. Since we want a non-trivial assignment, $B \geq 2$. The assignment matrix $X \in \mathbf{R}^{|\sigma|*B}$ has its entry $x_{\sigma b}$ specifies the fraction of people with feature vector σ who get mapped to bin b and score vector $v \in \mathbf{R}^B$ has its entry v_b specifies the score given to member in b. Similarly, we use $V \in \mathbf{R}^{B*B}$ to represent the diagonal matrix of v.

2.3 Fairness Condition

Given an instance (G_1, G_2, σ, p), we want to find a non-trivial assignment (X, v, B) that satisfies the following three fairness conditions:

$$\begin{cases} {n_t}^T P X = {n_t}^T X V \\ \frac{{n_1}^T X V v}{\mu_1} = \frac{{n_2}^T X V v}{\mu_2} \\ \frac{\mu_1 - {n_1}^T X V v}{N_1 - \mu_1} = \frac{\mu_2 - {n_2}^T X V v}{N_2 - \mu_2} \end{cases}$$

The first condition is the well-calibrated condition and the following two are the balance for positive class and the balance for positive class, respectively. More details can be found in **[JMR17]**.

We denote the fraction in we write γ_t for the average of the expected scores assigned to members of the positive class in group t, where $\gamma_t = \frac{{n_t}^T X V v}{\mu_t}$.

Definition 2.1. *(fairness difference)*. We denote the fairness difference to be $d = \gamma_1 - \gamma_2$

If $d \geq 0$, we say this risk assignment favors group 1 and if $d \leq 0$, we say this risk assignment favors group 2.

2.4 Example

In this section, we provide a toy example to help understand the problem setting. $\sigma = \{A, B, C, D\}$ as four discrete features. $G_1 = \{(A,+),(A,+),(A,+),(B,+),(B,+),(B,-),(C,+),(C,-),(C,-)\}$ $G_2 = \{(A,+),(A,+),(A,-),(B,+),(B,-),(B,-),(C,-),(C,-),(C,-),$ $(D,+),(D,+),(D,+),(D,+),(D,+),(D,+),(D,+),(D,+),(D,+)\}$ as two groups with equal base rate and $p = \{\frac{4}{5}, \frac{2}{5}, \frac{1}{5}, \frac{5}{6}\}$.

So we found the corresponding solution is X $s.t.$ $\sum X_{1,i} + \sum X_{2,i} + \sum X_{3,i} = 3 \times \sum X_{4,i}$ and $V = I$

3 Determination

It is an open question whether there is a polynomial-time algorithm to find a fair assignment of minimum loss, or even to determine whether a non-trivial fair solution exists **[JMR17]**. Clearly, the determination problem is easier than the optimization problem of finding the minimum loss. So our focus is on the determination version of this problem.

From my intuition, we can pre-set the value of B as $|\sigma| - 1$ or 2 or $\frac{|\sigma|}{2}$ and find an assignment that satisfies the first condition. Then we can use this value to see if it favors which group and *find a assignment that will have huge difference on the favor of group*. Therefore, by the following lemma, we know there exits a non-trivial well-calibrated assignment.

Lemma 3.1. *There exists a non-trivial fair assignment if and only if there exist non-trivial well-calibrated assignments X_1 and X_2 such that X_1 weakly favors group 1 and X_2 weakly favors group 2.*

In order to find the tractability of this problem, we can relax the domain of this problem from $\mathbf{R}$ to $\mathbf{Z}$.

References

[JMR17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan.*Inherent Trade-Offs in the Fair Determination of Risk Scores*. Proceedings of the 8th Conference on Innovation in Theoretical Computer Science: 43:1–43:23.