

Fairness-aware Relational Learning and Inference

Golnoosh Farnadi^{1*}, Behrouz Babaki^{2*}, Lise Getoor¹

¹University of California, Santa Cruz, USA,

²KU Leuven, Belgium

1 Introduction

AI and machine learning methods have become essential tools in decision-making processes that affect people’s lives, such as loan approval, advertising, employment, education, criminal risk assessment, policing, etc. Concerns of bias and algorithmic discrimination (Boyd, Levy, and Marwick 2014) have raised attention in recent years and there is starting to be works on methods for ensuring fairness. These include dealing with biases through unawareness/blindness, awareness/Lipschitz property (Dwork et al. 2012), demographic parity/disparate impact (Feldman et al. 2015), preference-based (Zafar et al. 2017), and equality of opportunity (Hardt, Price, and Srebro 2016).

While there is starting to be fairness in ML, there is no work that considers relational information. Since many forms of bias occur in social context, leveraging relational information is essential. The goal of fairness in classic setting is to design algorithms that make fair predictions across both *protected* and *unprotected* groups where protected group is recognized based on a protected attribute such as age, gender, race, religion, etc.

In this paper, our main objective is to develop fairness in relational machine learning models while the protected group is based on protected attributes such as gender, race, etc, but we expand the definitions to include the relational context in which discrimination may occur. This extension leads to a richer notion of fairness, that both allows for more complex statistical modeling of dependencies, and richer definitions of fairness. We believe these richer models are needed to realistic settings such as employment discrimination in which biases in producing reports from employees influence on their promotion, or discrimination in reviewing process of papers based on the rank of the universities.

In this paper, we first introduce and formulate fairness in relational setting, then we propose 1) *fairness-aware constrained conditional inference* subject to common data-oriented fairness measures and 2) *fairness-aware learning* by incorporating decision-oriented fairness measures.

2 Fairness in relational setting

In this section, we formalize relational fairness using first-order logic: An atom is an expression of the form $p(a_1, a_2, \dots, a_n)$ where p is a *predicate symbol*, and each argument a_1, a_2, \dots, a_n is either a constant or a variable. The finite set of all possible substitutions of a variable to a constant for a particular variable a is called its *domain* D_a . If all variables in $p(a_1, a_2, \dots, a_n)$ are substituted by some constant from their respective domain, then we call the resulting atom a *ground atom*. A formula is defined by induction: every atom is a formula. If α and β are formulae, then $\alpha \vee \beta$, $\alpha \wedge \beta$, $\neg\alpha$, $\exists x \alpha$, $\forall x \alpha$ are formulae. An interpretation I is a mapping that associates a truth value $I(P)$ to each ground atom P .

We denote formula F which has only one free variable v (i.e. other variables in F are quantified) by $F[v]$. The *population* defined by $F[v]$ is the set of substitutions of v for which $F[v]$ holds. A *discriminative pattern* is a pair $DP[v] \equiv (F_1[v], F_2[v])$, where $F_1[v]$ and $F_2[v]$ are formulae.

Example 1. *The two formulae in the discrimination pattern $DP[v] := ((\forall u, \neg\text{Affiliated}(v, u) \vee \neg\text{TopRank}(u)), \text{Student}(v))$ specify two populations, namely the individuals who are not affiliated with a top-rank institute and the student authors.*

Given an interpretation I , the *protected group*

$$PG \equiv \{v : F_1[v] \wedge F_2[v]\}$$

is defined as the set of all instances hold for variable v for which $F_1[v] \wedge F_2[v]$ is true under interpretation I , that is, $I(F_1[v] \wedge F_2[v]) = 1$. Similarly, the *unprotected group*

$$UG \equiv \{v : \neg F_1[v] \wedge F_2[v]\}$$

is defined as the set of all instances holds for variable v for which $I(\neg F_1[v] \wedge F_2[v]) = 1$. A *decision atom* $d(v)$ is an atom containing exactly one variable v that specifies a decision affecting the protected group which is defined either by law or end-user.

Example 2. *The protected group of the discrimination pattern specified in Example 1 is*

$$PG := \{v : (\forall u, \neg\text{Affiliated}(v, u) \vee \neg\text{TopRank}(u)) \wedge \text{Student}(v)\}$$

*Equal contributors.

and the unprotected group is

$$UG := \{v : (\exists u, (\text{Affiliated}(v, u) \wedge \text{TopRank}(u)) \wedge \text{Student}(v))\}.$$

The decision atom $d[v] := \text{Presents}(v)$ indicates the benefit of presenting a paper at a conference.

3 Fairness-aware Inference

To formulate fairness in the relational setting, in this section we propose fairness-aware constrained conditional inference subject to common data-oriented fairness measures. We first introduce these fairness measures and then re-define them using the notation introduced in Section 2. Let a and c denote the counts of denial (i.e., negative decisions) for protected and unprotected groups, and n_1 and n_2 denote their sizes, respectively. Let $p_1 = a/n_1$ be the proportion of benefit denied for the protected group, and $p_2 = c/n_2$ be the proportion of benefit denied for the unprotected group. Using p_1 and p_2 we can define three well-known fairness measures as follows:

1. **Risk difference:** $RD = p_1 - p_2$, also known as absolute risk reduction. The UK uses RD as it's legal definition of fairness measure.
2. **Risk Ratio:** $RR = p_1/p_2$, also known as relative risk. The EU court of justice has given more emphasis on the RR as a measure of fairness.
3. **Relative Chance:** $RC = 1 - p_1/1 - p_2$ also known as selection rate. The US laws and courts mainly refer to the RC as a measure of fairness. For further information we refer to (Pedreschi, Ruggieri, and Turini 2012).

Notice that RR is the ratio of benefit denial between the protected and unprotected groups, while RC is the ratio of benefit granting.

Now, we can formulate these measures using the formalism defined in Section 2. Given the decision atom $d(v)$ and discriminative pattern $DP(F_1[v], F_2[v])$, the counts of denial for both protected and unprotected groups are computed by the following equations:

$$\begin{aligned} a &\equiv \sum_{v \in D_v} I(\neg d(v) \wedge F_1[v] \wedge F_2[v]) \\ c &\equiv \sum_{v \in D_v} I(\neg d(v) \wedge \neg F_1[v] \wedge F_2[v]) \\ n_1 &\equiv \sum_{v \in D_v} I(F_1[v] \wedge F_2[v]) \\ n_2 &\equiv \sum_{v \in D_v} I(\neg F_1[v] \wedge F_2[v]) \end{aligned}$$

Using these counts, the fairness measures can be computed as: $RD \equiv a/n_1 - c/n_2$, $RR \equiv \frac{a/n_1}{c/n_2}$, and $RC \equiv \frac{1-a/n_1}{1-c/n_2}$. Finally, we introduce the notion of δ -fairness.

Definition 1 (δ -fairness). *If a fairness measure for a decision making process falls within some δ -window, then the*

process is δ -fair. Given $0 \leq \delta \leq 1$, the δ -windows for measures $RD/RR/RC$ are defined as:

$$\begin{aligned} -\delta &\leq RD \leq \delta \\ 1 - \delta &\leq RR \leq 1 + \delta \\ 1 - \delta &\leq RC \leq 1 + \delta \end{aligned}$$

The standard MAP inference aims at finding values that maximize the conditional probability of unknowns. Once a decision is made according to these values, one can use the fairness measure to quantify the degree of discrimination. To develop fairness-aware inference, we propose to incorporate fairness in MAP inference by adding the δ -fairness constraints to the underlying optimization problem of MAP inference.

Consider risk difference, RD , where $RD \equiv \frac{a}{n_1} - \frac{c}{n_2}$. The δ -fairness constraint $-\delta \leq RD \leq \delta$ can be encoded as the following constraints:

$$\begin{aligned} n_2 a - n_1 c - n_1 n_2 \delta &\leq 0 \\ n_2 a - n_1 c + n_1 n_2 \delta &\geq 0 \end{aligned}$$

Similarly, from $RR \equiv \frac{a/n_1}{c/n_2}$ and the δ -fairness constraint $1 - \delta \leq RR \leq 1 + \delta$ we obtain:

$$\begin{aligned} n_2 a - (1 + \delta) n_1 c &\leq 0 \\ n_2 a - (1 - \delta) n_1 c &\geq 0 \end{aligned}$$

And finally, $RC \equiv \frac{1-a/n_1}{1-c/n_2}$ and the δ -fairness constraint $1 - \delta \leq RC \leq 1 + \delta$ gives:

$$\begin{aligned} -n_2 a + (1 + \delta) n_1 c - \delta n_1 n_2 &\leq 0 \\ -n_2 a + (1 - \delta) n_1 c + \delta n_1 n_2 &\geq 0 \end{aligned}$$

4 Fairness-aware parameter learning

In this section, we first review five measures of fairness from literature. The difference between these measures and the ones introduced earlier is that the latter are based on the *decision* made by an algorithm. To explain these measures, assume that symbols tp , fn , fp , and tn denote true positive, false negative, false positive, and true negative rate, respectively. Each of the following measures assume that a decision is fair if the values of some quantity among the protected and unprotected group are the same:

1. **Overall accuracy equality:** equal values for $(tp + tn)/(tp + fn + fp + tn)$. This measure is not commonly used because it does not distinguish between the accuracy for success and failure.
2. **Demographic parity:** equal marginal distributions of the predicted classes $(tp + fp)/(tp + fn + fp + tn)$ or $(fn + tn)/(tp + fn + fp + tn)$ in both groups. This measure has been criticized as it can lead to highly undesirable decisions (Dwork et al. 2012).

3. **Equality of opportunity:** equal values for $tp/(tp + fn)$ or $tn/(fp + tn)$.
4. **Conditional use accuracy equality:** equal values for $tp/(tp + fp)$ or $tn/(fn + tn)$.
5. **Treatment equality:** equal ratio of false negatives and false positives (i.e., fp/fn or fn/fp) in both groups.

To incorporate these measures in learning the parameters of a discriminative relational model with joint probability distribution $P(y|x)$, we first introduce their logical counterparts. Let \hat{y}_j and y_j denote the actual and predicted truth values for n atoms of interest. We extend the definitions of tp , fn , fp and tn as:

$$\begin{aligned}
 tp &= \sum_{j=1}^n I(\hat{y}_j \wedge y_j) \\
 fn &= \sum_{j=1}^n I(\hat{y}_j \wedge \neg y_j) \\
 fp &= \sum_{j=1}^n I(\neg \hat{y}_j \wedge y_j) \\
 tn &= \sum_{j=1}^n I(\neg \hat{y}_j \wedge \neg y_j)
 \end{aligned}$$

Fair parameter learning is an optimization problem with two possibly conflicting goals: 1) to achieve high prediction power according to the data, and 2) ensuring fair predictions. The first goal can be for example translated into a high likelihood of the data.

In order to achieve the second goal, we add a term to the objective function that reflects the degree of fairness of predictions according to known fair truth values for a subset of variables. More specifically, let \mathbf{y}_{PG} and \mathbf{y}_{UG} denote the fair truth values of the target predicate for subsets of protected and unprotected groups, respectively. Given a fairness measure M (which can be one of the measures defined above), we aim at decreasing the value of $|M(\mathbf{y}_{PG}) - M(\mathbf{y}_{UG})|$. Combining these two goals leads to the following objective function for fairness-aware parameter learning:

$$\max_W \left\{ \log P_W(\mathbf{y}|\mathbf{x}) - \gamma \cdot \mathbb{E}_W \left[|M(\mathbf{y}_p) - M(\mathbf{y}_u)| \right] \right\}$$

where γ is a positive constant that determines the relative importance of the two components of the objective function.

5 Conclusion

In this paper, we introduce the notion of fairness in relational setting. We extend MAP inference with fairness-aware constrained conditional inference subject to common data-oriented fairness measures. In addition, we propose a fairness-aware learning algorithm that incorporates decision-oriented fairness measures to ensure fairness in learning. We believe that extending fairness to the relational setting facilitates defining complex discrimination patterns.

Many applications in social network analysis, personalized advertising, education science, and computational social science can benefit from this extension.

References

- Boyd, D.; Levy, K.; and Marwick, A. 2014. The networked nature of algorithmic discrimination. In Gangadharan, S. P., ed., *Data and discrimination: Collected essays*. Washington, DC: New America. 53–57.
- Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. S. 2012. Fairness through awareness. In Goldwasser, S., ed., *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, 214–226. ACM.
- Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In Cao, L.; Zhang, C.; Joachims, T.; Webb, G. I.; Margineantu, D. D.; and Williams, G., eds., *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, 259–268. ACM.
- Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 3315–3323.
- Pedreschi, D.; Ruggieri, S.; and Turini, F. 2012. A study of top-k measures for discrimination discovery. In Ossowski, S., and Lecca, P., eds., *Proceedings of the ACM Symposium on Applied Computing, SAC 2012, Riva, Trento, Italy, March 26-30, 2012*, 126–131. ACM.
- Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; Gummadi, K. P.; and Weller, A. 2017. From parity to preference-based notions of fairness in classification. *CoRR* abs/1707.00010.