Maximum Empirical Risk Minimization

Abstract

In this work, we look at the problem of unfairness in machine learning models arising out of improper representation of minority groups of people in the training data. In such cases, the ERM model of training is known to result in a model which is unfair to the minority group. We propose an extension of the ERM model which results in a fair prediction, when the training points come from two unknown groups of people. We assume that we do not have any knowledge of the sensitive attribute, i.e. we do not have any knowledge of which group a particular training point belongs to. We only assume that the sample points come from two distinct groups with one these being the majority group and the other being the minority group. We further assume we have knowledge of the size of majority and minority groups in the training data.

1 Introduction

There are a number of machine learning algorithms that focus on improving the overall accuracy by reducing the expected loss on the underlying distribution that generates the data. However, such models overlook the fact that the underlying distribution is often a composition of multiple distributions each representing a certain group of people such as people from different genders, races etc. In such cases, not all of these groups are equally well represented [1]. Since minority groups have a lower contribution in the overall loss of the model, the machine learning algorithm will often concern itself more with the majority group resulting in more erroneous predictions for the minority group [2]. In this paper, we propose a machine learning model consisting of an ensemble of classifiers that reduces this representation disparity by using a max-min optimization over all the classifiers in the hypothesis class and a set of training data that ensures a fair representation for the minority group.

Demographic Information Most online platforms which collect data do not collect data about the sensitive attributes of people like ethnicity, gender, nationality etc. Even in cases, where the platforms collect such information, they are at the most optional for user to disclose. Most users do not feel safe in disclosing such private information especially if they belong to any particular group which has suffered from some kind of socio-political bias historically. Thus, the sensitive information collected by online platforms are often incomplete and unreliable. In this work, we hence consider it reasonable to assume that we do not have any knowledge of the sensitive attributes. However, we assume that they exist in the form of a majority group and a minority group with their respective proportions in the society known to us. The respective percentages of people belonging to the minority group and majority group are often known from publicly available data like those collected census departments, independent surveys of population or from any other demographic indicators.

Fairness These days, machine learning methods are being used for decision making and making recommendations to a large extent. This includes applications like spam filtering, bank loan approvals, making hiring decisions, government administrative tasks, etc. In such cases, it becomes important for machine learning frameworks to be fair to people belonging to all racial or gender or any such protected groups. Anti-discrimination laws in many countries prevent the unfair treatment based on sensitive attributes like economic standards, race, religion, nationality etc. Hence, it becomes mandatory to devise machine learning approaches which behave in a fair way.

The sensitive attribute can be seen as some socially acquired quality like economic status(rich, poor), gender(male, female), race(caucasian, african-american) so on. We see that most of these attributes are often used implicitly or explicitly for social discrimination. Machine learning often picks up these discriminations as "safe" attributes like educational status, address, that are often highly correlated with sensitive attributes. Thus there is no way to teach the model to make decisions based on only "safe" attributes and overlook sensitive attributes. This is often because we cannot source enough data against the socially prevalent correlations.

Another considerable challenge in this area is to use a proper definition for fairness. Many such mathematical definitions of fairness are used in practice to judge how fair a particular model is [3, 4, 5, 6]. There is no agreement in the community on what is a good definition of fairness. And it has also been shown in many works that many definitions of fairness are often at loggerheads with each other and cannot be simultaneously satisfied [7, 8, 9].

2 Related Work

Earlier works have tried to solve this problem in a variety of ways. In [9], Hardt *et al.* introduce the notion of "equalized odds" for the case of supervised learning, where they consider fairness in the prediction with respect to a specific protected attribute A whose value is known for every data item being trained on. However, in many cases such as speech recognition, the demographic information may not available.

Zafar *et al.* in [3] aim to solve decision boundary unfairness in convex margin-based classifiers for the binary classification setting with only two possible values for the sensitive attribute. They formulate it as a convex optimization problem with decision boundary covariances. In [10], they introduce a new notion of fairness known as disparate mistreatment and solve the problem of different misclassification rates for people from different social groups.

In [11], Woodworth *et al.* use the notion of "equalized odds" to learn a fair predictor for a finite training sample with respect to the sensitive attribute. They use post-hoc corrections that require knowledge of the joint distribution ($\hat{Y}$,A,Y) where $\hat{Y}$ is the predictor, A is the protected attribute and Y is the target. Feldman *et al.* in [4] study the notion of disparate impact and devise a way to measure it by finding out how well the protected attribute can be predicted from the other attributes. This is then used to make the data unbiased. However, all these works also assume that we know the value of the protected attributes for every data point.

Dwork *et al.* in [12] introduce fairness in classification by finding out the similarity of the data points with respect to the classification problem and maximizing the objective under the fairness constraint which states that similar subjects should be treated similarly. In [13], Hébert-Johnson use a strong notion of non-discrimination where they identify multiple subpopulations and perform calibration for each of subpopulations. This multi-calibration ensures that we obtain accurate results for each of the subpopulations. However, getting a similarity measure or a multi-calibration for real-world tasks is quite difficult.

In [2], Hasimoto *et al.* study the ERM setting where the loss with respect to a user discourages the user to use the system and hence, a higher loss with respect to a particular community results in less number of people from that community using the system. This aggravates the representation disparity for the aforementioned community that becomes even more underrepresented resulting in further loss. They use Distributionally Robust Optimization (DRO) algorithm that minimizes the maximum risk observed for all groups across all iterations.

The authors assume that the distributions for each of the groups lie inside a chi-squared ball of each other. However, in the real world scenario, we cannot guarantee that the distribution of the population groups will have χ^2-divergence. Also, in many cases such as the NYPD Stop question-and-frisk program (SQF), we do not have multiple iterations to see the effect of the loss.

3 Problem

Suppose our sample points come from a distribution D_p which is a composite of two distributions D_0 and D_1 such that

$$D_p = pD_1 + (1-p)D_0$$

This can also be though of as flipping a weighted coin which lands on Heads with probability p and Tails with probability $1-p$. Whenever the coin lands on Head, we sample a point iid from D_1 and D_0 otherwise. Let us assume $p < 1/2$.

In a more practical setting, we can think of this as sampling features from two different groups of people. The smaller group gets picked with the smaller probability p while the larger group gets picked with the larger probability $1-p$. The two groups of people have features following two completely different distributions D_1 and D_0 respectively.

Going ahead, let us say we have a finite hypothesis set $\mathcal{H}$ from which we want to pick a hypothesis which achieves a small error on D_p and also on the minority group D_1, that is we want a $h \in \mathcal{H}$ with a small $\mathbb{E}_{z\sim D_p}[l(h,z)]$ and a small $\mathbb{E}_{z\sim D_1}[l(h,z)]$. Let us assume our hypothesis is realizable in a fair sense. We define this as follows:

Definition 3.1. Fair Reliability: A hypothesis set is Fair-Realizable if there exists a $h \in \mathcal{H}$ such that both of the following are satisfied:

$$\mathbb{E}_{z\sim D_p}[l(h,z)] = 0 \quad \mathbb{E}_{z\sim D_1}[l(h,z)] = 0$$

Let us consider the setting of PAC Learning for ERM algorithm for the distribution D_p with a finite Fair-Realizable $\mathcal{H}$. When we sample m points from D_p and run the ERM on it, PAC-Learning guarantees:

$$L_{D_p}(h) < \epsilon$$

with probability at least $1-\delta$, whenever,

$$m \geq \frac{1}{\epsilon}\log\frac{|\mathcal{H}|}{\delta}$$

However the PAC setting does not guarantee anything about $L_{D_1}(h)$. In the worst case, it is possible that the ERM algorithm learnt a h which does not make any errors on D_0 and makes all its errors only on D_1. In this case:

$$\epsilon > L_{D_p}(h) = pL_{D_1}(h) + (1-p)L_{D_0}(h)/geqpL_{D_1}(h)$$

$$\frac{\epsilon}{p} > L_{D_1}(h)$$

This is a poor guarantee and we wish to improve this with a modified ERM algorithm which we present in the next section. The problem of unfairness in ERM persists irrespective of how large we choose m. Choosing a larger m only reduces ϵ, but the loss on D_1 still only satisifes $L_{D_1}(h) < \frac{\epsilon}{p}$. In the next section we begin by choosing significantly larger set of samples from D_p but show that we achieve better guarantees on $L_{D_1}(h)$.

4 Fair version of ERM

We first present our algorithm Maximum Empirical Risk Minimization (MERM) :

Definition 4.1. MERM: We sample mn points from D_p in an iid fashion and call it S. We partition S into $S_1, S_2, \ldots S_n$ each with m sample points.

Let us pick a h such that

$$L_{S_i}(h) = \frac{1}{m}\sum_{z\in S_i} l(h,z)$$

$$L_S(h) = \max_{1\leq i\leq n} L_{S_i}(h)$$

$$h^* = \text{argmin}_{h\in\mathcal{H}} L_S(h)$$

We say $MERM(\mathcal{H}) = h^*$.

We note that MERM differs from ERM in the way it defines $L_S(h)$. In the case of the ERM setting, we would have defined

$$L_S(h) = \frac{1}{n}\sum_{i=1}^{n} L_{S_i}(h)$$

In the case of MERM we replace the average with the maximum and hence the name Max-ERM:

$$L_S(h) = \max_{1s\leq i\leq n} L_{S_i}(h)$$

We now discuss how large we should choose n to be. We want n to be large enough to be able to guarantee that at least one of our sample sets is Minor-Centric which we define as follows:

Definition 4.2. Minor-Centric: A sample set S_i is called Minor-Centric if it contains more samples from the minority group D_1 than the majority group D_0

Let us denote the probability of a sample set being minor-centric as α.

The probability of a sample set to be minor-centric is at least as big as the probability of containing equal number of samples from both groups in the sample set:

$$P(S_i \text{ is Minor-Centric}) = \alpha \geq \binom{m}{m/2} p^{m/2}(1-p)^{m/2}$$

By Stirling's Approximation, we bound the binomial coefficient as:

$$\binom{m}{m/2} \geq \frac{2^m}{\sqrt{2m}}$$

Using this, we can bound the probability of a sample set to be minor-centric as:

$$\alpha \geq \frac{\left(2\sqrt{p(1-p)}\right)^m}{\sqrt{2m}}$$

Of the n sample sets, the probability that none of the sample sets are minor-centric is

$$(1-\alpha)^n \leq e^{-n\alpha} \leq \exp\Big(-n\frac{\left(2\sqrt{p(1-p)}\right)^m}{\sqrt{2m}}\Big)$$

Let us define δ_f as $\delta_f = \exp\Big(-n\frac{\left(2\sqrt{p(1-p)}\right)^m}{\sqrt{2m}}\Big)$

We know that $\mathcal{H}$ is fair-realizable. This means that for our MERM setting, there exists a h which achieves 0 loss on all samples and hence might get chosen by MERM. This further means for any h returned by MERM, it would achieve 0 loss on all the sample points.

We claim that the h returned by MERM would satisfy $P(L_{D_1}(h) < 2\epsilon) > (1-\delta_f)(1-\delta)$

Proof: With probability of $1-\delta_f$, we have a minor-centric sample set amon our n sample sets. On this sample set, our h must give 0 loss. But still, if it were to give more than 2ϵ loss on D_1, then we must have got unlucky to choose "bad samples" from D_1. This happens with probability $(1-2\epsilon)$ samples. And we must have picked atleast $m/2$ of them since our sample set is minor centric. Together, this would happen with probability

$$(1-2\epsilon)^{m/2} \leq e^{-(m/2)(2\epsilon)} = e^{-m\epsilon} = \delta$$

Hence with probability $(1-\delta_f)(1-\delta)$, we have

$$P(L_{D_1}(h) < 2\epsilon)$$

5 Conclusion

It is important that we guarantee fairness for machine learning algorithms as they are used in various critical applications that require it to be non-discriminatory. Empirical risk minimization tends to be unfair due to its inclination to the losses of the majority group in the population. By using MERM, we can guarantee that we will be able to obtain highly accurate models that also generalize well to the minority group with low error.

References

[1] Patrick J Grother, George W Quinn, and P Jonathon Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST interagency report*, 7709:106, 2010.

[2] Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*, 2018.

[3] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

[4] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[5] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

[6] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[7] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[8] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[9] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[10] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[11] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[13] Ursula Hébert-Johnson, Michael P Kim, Omer Reingold, and Guy N Rothblum. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.