CY 7790, Lecture 19 Notes
Fairness in ML

Hye Sun Yun

November 29, 2021

1 Zafar et al. Fairness Beyond Disparate Treatment Disparate Impact: Learning Classification without Disparate Mistreatment. In WWW 2017

Presented by Pablo Kvitca

Problem Statement. With the rise of ML/classifiers being used in real-world applications, there have been concerns about the potential unfairness towards people with certain traits or sensitive attributes. In addition to disparate treatment and disparate impact, the authors of the paper introduce new notion of unfairness called disparate mistreatment. They show how this can be used in synthetic and real world datasets and show methods to avoid disparate mistreatment. Disparate impact is useful when ground truth is not available. When the correctness of decisions can be determined, disparate mistreatment can not only be accurately assessed, but also avoids reverse-discrimination, making it a more appealing notion of fairness.

Threat Model. There isn't a clear adversary in this case, but a potential adversary might want certain systems to exhibit disparate mistreatment that might be hard to detect with the definition of disparate impact or disparate treatment. However, there are clear objectives by those who develop ML systems as they try to develop fair classification models. In this paper, the objective is to avoid disparate mistreatment. The developer needs knowledge of predicted label class, ground-truth label class, and sensitive attribute labels in order to detect and avoid disparate mistreatment.

Methodology. The authors of the paper first formalize the three different notions of unfairness: disparate treatment, disparate impact, and disparate mistreatment. They show that unfairness definitions of disparate treatment and disparate impact are limiting and show how disparate mistreatment can be a better way to think about unfairness. They show how to train classifiers without disparate mistreatment.

The following provides definitions of the 3 notions of unfairness:

- **Disparate treatment**:
    1. arises when a system provides different outputs for groups of people with the same values for non-sensitive attributes but different values for sensitive attributes.
    2. a binary classifier does not suffer from disparate treatment if : $P(\hat{y}|x, z) = p(\hat{y}|x)$, where $\hat{y}$ is the classifier outputs, x is the feature vector, and z is the sensitive feature.
- **Disparate impact**:
    1. arises when a system provides outputs that benefit/hurt a group of people sharing the same sensitive attributes, more frequently than other groups.
    2. the notion of disparate impact is independent of the "ground truth" information about the decisions.

3. a binary classifier does not suffer from disparate impact if : $P(\hat{y} = 1|z = 0) = p(\hat{y} = 1|z = 1)$, where $\hat{y}$ is the classifier outputs and z is the sensitive feature.
4. avoiding disparate impact by requiring decision outcomes to be proportional can risk introducing reverse-discrimination.

- **Disparate mistreatment**:

1. misclassificaton rates may be different for groups of people with different sensitive attributes.
2. misclassification rate can vary: overall, false negative/positive, false omission, false discovery. Figure 1 shows the various ways of measuring misclassification rates.
3. a binary classifier does not suffer from disparate mistreatment if the misclassification rates for different groups of people having different values of the sensitive feature z are the same.

		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y \mid y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

Figure 1: This figure describes various ways of measuring misclassification rates. In addition to the overall misclassification rate, there are other ways of measuring error rates. False-negative rate and false positive rate are defined as fractions over the class distribution in the ground truth labels. False discovery rate and false omission rate are defined as fractions over the class distribution in the predicted labels.

Measure	Ground Truth	Risk Reverse Discrimination
Disparate Treatment	INDEPENDENT	NO
Disparate Impact	INDEPENDENT	YES
Disparate Mistreatment	DEPENDENT	NO

Figure 2: This figure shows the differences of the three different notions of unfairness.

Figure 2 provides a simple table of how the three different notions of differ from each other.

The authors of the paper also propose a method to train decision boundary-based classifiers such as logistic regression and SVMs that do not suffer from disparate mistreatment. These classifiers generally learn the optimal decision

boundary by minimizing a convex loss $L(\theta)$. The convexity of $L(\theta)$ ensures that a global optimum can be found efficiently. However, the conditions for avoiding unfairness are non-convex in general and poses this problem to be difficult. To overcome the difficulty, a proxy is proposed. The proposed way to measure disparate mistreatment uses the covariance between the users' sensitive attributes and the signed distance between the feature vectors of misclassified users and the classifier decision boundary (Figure 3). Given the proxy and the constraint (misclassification rate) to make the classifier fair, the problem can be rewritten as Figure 4. While the constraints proposed in Figure 4 can be an effective proxy for fairness, they are still non-convex. The problem can be solved efficiently by converting the constraints into a Disciplined Convex-Concave Program (DCCP). DCCP combines the ideas of disciplined convex programming (DCP) with convex-concave programming (CCP). The DCCP version of the problem is found in Figure 5

$$\begin{aligned} \mathrm{Cov}(z, g_{\boldsymbol{\theta}}(y, \mathbf{x})) &= \mathbb{E}[(z - \bar{z})(g_{\boldsymbol{\theta}}(y, \mathbf{x}) - \bar{g}_{\boldsymbol{\theta}}(y, \mathbf{x}))] \\ &\approx \frac{1}{N} \sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}), \end{aligned}$$

Figure 3: The covariance between the users' sensitive attributes and the signed distance between the feature vectors of misclassified users and the classifier decision boundary.

$$\begin{aligned} \text{minimize} \quad & L(\boldsymbol{\theta}) \\ \text{subject to} \quad & \tfrac{1}{N} \textstyle\sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}) \leq c, \\ & \tfrac{1}{N} \textstyle\sum_{(\mathbf{x}, y, z) \in \mathcal{D}} (z - \bar{z})\, g_{\boldsymbol{\theta}}(y, \mathbf{x}) \geq -c, \end{aligned}$$

Figure 4: Simplified problem based on the proxy. The covariance threshold $c \in \mathbb{R}^+$ controls how adherent to disparate mistreatment the boundary should be.

$$\begin{aligned} \text{minimize} \quad & L(\boldsymbol{\theta}) \\ \text{subject to} \quad & \tfrac{-N_1}{N} \textstyle\sum_{(\mathbf{x}, y) \in \mathcal{D}_0} g_{\boldsymbol{\theta}}(y, \mathbf{x}) \\ & + \tfrac{N_0}{N} \textstyle\sum_{(\mathbf{x}, y) \in \mathcal{D}_1} g_{\boldsymbol{\theta}}(y, \mathbf{x}) \leq c \\ & \tfrac{-N_1}{N} \textstyle\sum_{(\mathbf{x}, y) \in \mathcal{D}_0} g_{\boldsymbol{\theta}}(y, \mathbf{x}) \\ & + \tfrac{N_0}{N} \textstyle\sum_{(\mathbf{x}, y) \in \mathcal{D}_1} g_{\boldsymbol{\theta}}(y, \mathbf{x}) \geq -c, \end{aligned}$$

Figure 5: The problem converted as a Disciplined Convex-Concave Program (DCCP)

Class Discussion.

Limitations of disparate treatment: Disparate treatment only looks for equal percentages of outcomes among different groups and does not mention anything about accuracy so not the best way of looking at fairness in some situations.

Continuous values for z: What happens if z the sensitive attribute is a continuous value? We would need to create discrete groups based on the continuous values.

Extension to more sensitive attributes: Can we extend what the paper proposes (only binary classification problem with only one sensitive attribute) to have more sensitive attributes? Yes. The paper only considers one sensitive attribute, but you can extend by adding more conditions and constraints so can be easily extensible.

Bias found in data: In general, should we first think about bias in the dataset? There are definitely biases in the dataset which can impact the fairness in models. However, the definitions are for the classifier and not for the data without the model. There are no definitions of unfairness for data. Also there are many different definitions of fairness and many impossibilities of achieving different definitions of fairness at the same time.

2 Hardt et al. Equality of Opportunity in Supervised Learning. In NeurIPS 2016

Presented by Zohair Shafi

Problem Statement. This paper aims to propose a criterion for discrimination against a specified sensitive attribute in supervised learning. They show how to optimally adjust any learned predictor so as to remove discrimination according to their definitions of equalized odds and equal opportunity. Simple demographic parity is not enough since it doesn't ensure fairness and often reduces the utility of the model.

Threat Model. There isn't a clear adversary in this case, but a potential adversary might want certain systems to exhibit unfairness. Similar to the previous paper, the objective of the developers of the classifiers is to avoid disparate mistreatment.

Methodology.

The authors of the paper propose a simple, interpretable, and easily checkable notion of nondiscrimination with respect to a specified protected attributes. They argue that, unlike demographic parity, their notion of equalized odds and equal opportunity provide meaningful measures of discrimination, while allowing for higher utility. Demographic parity does not ensure fairness and can accept random individuals in a demographic so long as percentages of acceptance match.

The following provides the definitions of equalized odds and equal opportunity:

- **Equalized odds**:
  1. has both true positive parity $Pr\hat{Y} = 1|A = 1, Y = 1 = Pr\hat{Y} = 1|A = 0, Y = 1$ and false positive parity $Pr\hat{Y} = 1|A = 0, Y = 0 = Pr\hat{Y} = 1|A = 1, Y = 0$ where $\hat{Y}$ is the classifier output, Y true labels, and A the protected attribute.
  2. provides equal bias and equal accuracy in all demographics
- **Equal opportunity**:
  1. True positive parity $Pr\hat{Y} = 1|A = 1, Y = 1 = Pr\hat{Y} = 1|A = 0, Y = 1$ where $\hat{Y}$ is the classifier output, Y true labels, and A the protected attribute.
  2. a weaker, though still interesting, notion of non-discrimination and can allow for better utility

Achieving non-discrimination:

- **Binary classification - discrete**:
  1. To achieve equalized odds we need to verify that both false positive rate and true positive rate are equal. $\gamma_a(\hat{Y}) \stackrel{\text{def}}{=} (Pr(\hat{Y} = 1|A = a, Y = 0), Pr(\hat{Y} = 1|A = a, Y = 1))$. $\hat{Y}$ satisfies equalized odds if and only if $\gamma_0(\hat{Y}) = \gamma_1(\hat{Y})$.
  2. The optimal derived predictor can be obtained as a solution to the linear program in Figure 6. Finding equalized opportunity requires removing one of the equality constraints from the linear program and just focusing on TPR.

3. Figure 7 shows the geometric solution - intersecting ROC curves.

- **Real-value classification - continuous**:

1. Use a search method for this case
2. Choose threshold t such that: $\hat{Y} = IR > t$. Optimal threshold should be chosen to balance FPR and TPR to minimize loss. Also multiple thresholds can be used to protect different attributes.
3. The ROC curves have to intersect for different sensitive attribute groups.
   $C_a(t) \stackrel{\text{def}}{=} (Pr(\hat{R} > t|A = a, Y = 0), Pr(\hat{R} > t|A = a, Y = 1))$
   However, the curves might not intersect expect for trivial points at (0,0) and (1,1). Randomization can be used to fill the span of possible derived predictors and allow for more intersection.
4. To find the optimal threshold: 1) can have single threshold or mixture of thresholds and 2) use a search method to find the optimal threshold. Figure 8 provides the intuition behind finding the threshold.
5. Bayes optimal classifier can be a threshold predictor of R where the threshold depends on the loss function. Given random variables (X, A) and target variable Y, the Bayes Optimal Regressor is given as $R = \text{argmin}_{r(x,a)} \mathbb{E}[Y - r(X, A))^2]$ where $r(x, a) = \mathbb{E}[Y|X = x, A = a]$.

$$\begin{aligned} \min_{p} \quad & \mathbb{E}\ell(\widetilde{Y}_p, Y) \\ \text{s.t.} \quad & \gamma_0(\widetilde{Y}_p) = \gamma_1(\widetilde{Y}_p) \\ & \forall_{y,a} 0 \leqslant p_{ya} \leqslant 1 \end{aligned}$$

Figure 6: Linear program for optimal derived predictor for equalized odds.

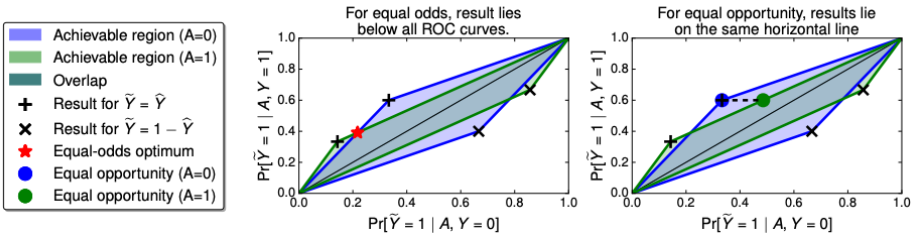

Figure 7: Finding the optimal equalized odds predictor and equal opportunity predictor

Class Discussion.

Similarities to disparate mistreatment: The definition of disparate mistreatment is the same as equalized odds when both false positive rate and false negative rate are used. The objectives in the previous paper are similar, and the definitions were similar.

Not relying on features: The proposed method in this paper does not rely on features but just on the predictor and sensitive attribute.

Graphs - ROC for post-processed classifier: For equalized odds, we need to look at points of intersection since we are looking at both TPR and FPR. However, for equal opportunity, we just look at TPR so it is just a horizontal line.

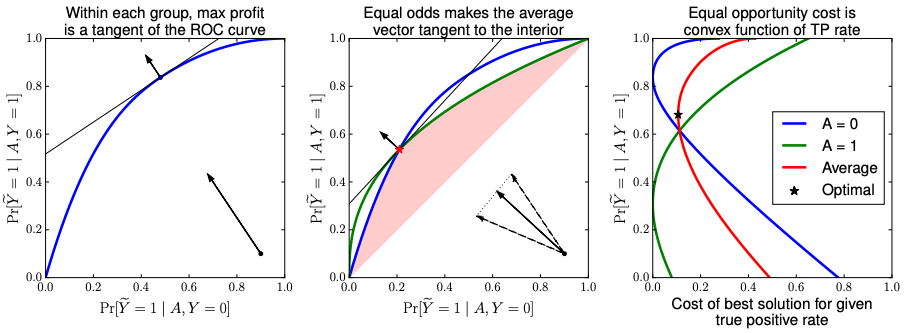

Figure 8: Finding the optimal equalized odds threshold predictor (middle), and equal opportunity threshold predictor (right). For the equal opportunity predictor, within each group the cost for a given true positive rate is proportional to the horizontal gap between the ROC curve and the profit-maximizing tangent line (i.e., the two curves on the left plot), so it is a convex function of the true positive rate (right). This lets us optimize it efficiently with ternary search.

Randomization for real-valued classifiers: Not sure what randomization means for real-valued classifiers. Maybe picking some points on the curve and use the threshold to lower the curve and therefore, achieve lower accuracy?

More than one sensitive attribute: How does finding optimal threshold work for more than one sensitive attributes? Not sure since we are not 100% how it works with one sensitive attribute. Maybe you find thresholds for each sensitive attribute group and find the intersection for all groups?

Figures in extended paper: Figures 10 and 11 in the extended paper were difficult to understand. Takeaways were that we can always lower the curve to match the equalized odds but not increase it since we want to intersect the curves. Thus, our accuracy will reduce to this process.

Comparison with previous paper approach: This paper's approach is different from the previous paper since it tries to make an unfair classifier fair after training but previous paper tries to train an unfair classifier. It is not very clear which one of these two methods are better based on evaluation metrics. One advantage of the post-processing step is that you don't necessarily have to know the training algorithm.

Selling fairness approaches to banks: It is really hard to sell these fairness approaches to banks. Loss of risk and loss of reputation and competition all influence this.