[02/12/2025], Lecture Notes: In-Process Interventions

[Srinath Bellamkonda, Murali Krishna Prodduturi]

CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025
Abolfazl Asudeh; www.cs.uic.edu/∼ *asudeh/teaching/archive/cs516spring25/*

1 Introduction

In this lecture, we discuss in-process interventions in machine learning. The objective is, given a potentially biased training set, to train a model that satisfies certain fairness constraints. Specifically, we focus on binary classification tasks and aim to modify the training algorithm A to ensure that the output model is fair based on predefined fairness criteria.

2 Problem Formulation

Let M be a model trained by minimizing a loss function that measures the difference between the predicted outcome $h_\theta(x)$ and the true label y:

$$\min \text{loss}(h_\theta(x), y)$$

In addition to minimizing the loss, we introduce fairness constraints to ensure that some measure of unfairness is less than a small value ϵ:

$$\text{unfairness} \leq \epsilon, \text{ for some notions of fairness}$$

The challenge is that these fairness constraints are typically non-convex, making the optimization problem difficult to solve. Our goal is to either make the problem convex or find efficient ways to handle the non-convexity.

3 Paper 1: Fairness Constraints: Mechanisms for Fair Classification (Zafar et al.)

The first paper we discuss is *Fairness Constraints: Mechanisms for Fair Classification* by Zafar et al. [2]. This paper addresses the non-convexity issue by redefining fairness constraints to make the optimization problem convex. However, it has several limitations:

- It only works for **binary classification** tasks.
- It assumes a **binary sensitive attribute**.
- It supports only one notion of fairness: **demographic parity**.

3.1 Demographic Parity Example

Consider a classifier h and a test set where points are colored based on a sensitive attribute (e.g., red and blue groups). Demographic parity requires that the ratio of positive predictions for each group should be equal. For example:

- For the red group: 2 out of 6 points are classified positively.
- For the blue group: 7 out of 10 points are classified positively.

This violates demographic parity because the acceptance rates differ significantly between groups.

3.2 Relaxing Fairness Constraints

To make the problem easy to deal with, the authors relax the fairness definition from discrete (binary outcomes) to continuous. Instead of counting positive predictions, they consider the **distance of points from the decision boundary**. Points closer to the decision boundary are given less weight, while points farther away are given more weight. This is done because the model is more certain of its prediction for points that are farther away from the decision boundary compared to points that are closer to the decision boundary. This relaxation allows the fairness constraint to be expressed as a convex function.

3.3 Formulating Fairness as Covariance

Demographic parity requires that the prediction $h_\theta(x)$ is independent of the sensitive attribute S. In other words, the probability of a positive outcome should be the same across all groups defined by the sensitive attribute. Mathematically, this is expressed as:

$$P(h_\theta(x) = 1|S = s_1) = P(h_\theta(x) = 1|S = s_2) \quad \text{for all groups } s_1, s_2.$$

To enforce this, we need to ensure that the sensitive attribute S is statistically independent of the model's predictions. Independence between two random variables S and $h_\theta(x)$ implies that their covariance is zero:

$$\text{Cov}(S, h_\theta(x)) = 0.$$

However as discussed earlier, directly working with $h_\theta(x)$ (which is binary) is challenging because it leads to non-convex constraints. To address this, the authors propose using the **distance to the decision boundary** $d_\theta(x)$ as a continuous proxy for the model's predictions. The distance $d_\theta(x)$ captures how confident the model is in its prediction, with larger distances indicating higher confidence.

The covariance between the sensitive attribute S and the distance to the decision boundary $d_\theta(x)$ is given by:

$$\text{Cov}(S, d_\theta(x)) = E[(S - \bar{S})(d_\theta(x) - E[S - \bar{S}]E[d_\theta(x)])]$$

Since $E[S - \bar{S}] = 0$ and $E[d_\theta(x) - E[d_\theta(x)]] = 0$, this simplifies to:

$$\text{Cov}(S, d_\theta(x)) = E[(S - \bar{S})d_\theta(x)]$$

Using the training set to estimate this, we get:

$$\text{Cov}(S, d_\theta(x)) \approx \frac{1}{n}\sum_{i=1}^{n}(S_i - \bar{S})d_\theta(x_i)$$

For a linear classifier, the distance to the decision boundary is given by $d_\theta(x_i) = \theta^T x_i$. Substituting this into the expression, the fairness constraint becomes:

$$\left|\frac{1}{n}\sum_{i=1}^{n}(S_i - \bar{S})\theta^T x_i\right| \leq \tau$$

Or simply,

$$|\text{Cov}(S, d_\theta(x))| \leq \tau$$

Here, τ is a small threshold that controls the allowable deviation from perfect fairness. This fairness constraint is also convex as it can be expressed as the intersection of two linear inequalities:

$$\frac{1}{n}\sum_{i=1}^{n}(S_i - \bar{S})\theta^T x_i \leq \tau \text{ and } \frac{1}{n}\sum_{i=1}^{n}(\bar{S} - S_i)\theta^T x_i \leq \tau$$

And the intersection of two linear inequalities is always convex.

4 Paper 2: A Reductions Approach to Fair Classification (Agarwal et al.)

The second paper, *A Reductions Approach to Fair Classification* by Agarwal et al. [1], addresses the limitations of the first paper. Specifically:

- It works for **any machine learning model**, not just binary classifiers.
- It supports **multiple sensitive attributes**, which can be non-binary.
- It accommodates **multiple fairness notions**, not just demographic parity.

4.1 Key Idea: Fairness Wrapper

The authors propose a **fairness wrapper** that can be applied to any black-box training algorithm. The wrapper iteratively adjusts the model to satisfy fairness constraints without requiring access to the internal workings of the training algorithm. The only requirement is that the training algorithm must support **weighted samples**.

4.2 Unified Fairness Template

The authors introduce a unified template for fairness definitions. For each group j and fairness notion F, they define a performance measure $\mu_j(\theta)$:

$$\mu_j(\theta) = E[g(X, Y, S, h_\theta)|\epsilon(X, f, S)]$$

Utilizing this template for demographic parity:

$$g_j = f_\theta(X), \epsilon_j = \{S = s_j\}$$

$$\mu_j(\theta) = E[f_\theta(X)|S = s_j]$$

$$\mu_j^*(\theta) = E[f_\theta(X)]$$

where $\mu_j^*(\theta)$ is the overall performance of the model. The fairness constraint is then expressed as:

$$|\mu_j(\theta) - \mu_j^*(\theta)| \leq \tau$$

as we want to make sure the expected value for the performance of group j vs. the overall performance of the model is less than or equal to τ. We can similarly define fairness constraints for all the groups and store them in a matrix $M_\mu(\theta)$. This gives us:

$$M_\mu(\theta) \leq \tau$$

4.3 Optimization Reformulation

The problem is reformulated as a constrained optimization:

$$\min \text{loss}(h_\theta(x), y) \quad \text{subject to} \quad M\mu(\theta) \leq \tau$$

where M is a matrix representing the fairness constraints for every group. This can be solved using **Lagrangian multipliers**.

4.4 Lagrangian Multipliers

The Lagrangian function is defined as:

$$L(\theta, \lambda) = \text{loss}(h_\theta(x), y) - \lambda(M\mu(\theta) - \tau)$$

The goal is to optimize $L(\theta, \lambda)$ with respect to both θ and λ. A standard approach to doing this would be to take the derivatives with respect to λ and θ and setting them equal to zero. However, we cannot take the derivatives of a black-box learning algorithm. Therefore this must be done using gradient descent.

4.5 Challenges with Gradient Descent

One challenge is that the optimization problem involves both minimization (with respect to θ) and maximization (with respect to λ). This creates a **saddle point** in the optimization landscape, which can make convergence impossible through gradient descent. However, the authors propose a technique to handle this issue which is discussed in the next lecture.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification, 2018.

[2] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2017.