Online Fair Classification for Sequential Data with Unbiased Estimators

Project Proposal for Master Thesis

Yifan Hu*

June 9, 2023

1 Introduction

In recent years, machine learning algorithms and neural networks have been widely used in various domains for classification tasks. However, there is growing concern about the potential biases and unfairness in these algorithms, particularly when dealing with sensitive attributes such as race or gender. Fairness in classification has become an important topic in machine learning research, and many fairness-aware algorithms have been proposed. In this thesis, we aim to address the problem of training a fair classifier using neural networks in an online setting, where the training data are received sequentially.

2 Problem Statement

The problem setting considers a dataset with sequential training instances, where each instance consists of a feature vector X that includes a protected attribute $A \in \{0, 1\}$ and a label Y. The goal is to develop a fair classifier $f(\theta; X)$, parameterized by θ, which can be trained incrementally as new instances arrive. The fairness requirement is to minimize the classification disparities between the estimated label distributions for different protected attributes while maintaining a good classification performance.

Existing literature in fair classification often focuses on minimizing the mean discrepancy between the estimated label distributions for different protected attributes, commonly known as minimum mean discrepancy. However, in this work, we aim to take a more general approach by considering various probability distance measures. The objective is to minimize the classification disparities under the constraint that the distance D between the estimated label distributions for $A = 0$ and $A = 1$ is small. These probability distance measures provide a flexible framework to capture different aspects of fairness and can lead to more nuanced and robust fair classifiers.

The fair classification problem can be formulated as follows:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \mathcal{L}(f(\theta; X), Y) \\ \text{s.t.} \quad & D\left(P(f(\theta; X)|A = 0), P(f(\theta; X)|A = 1)\right) \leq \epsilon, \end{aligned} \tag{1}$$

where $\mathcal{L}$ is the classification loss function, $P(f(\theta; X)|A = 0)$ and $P(f(\theta; X)|A = 1)$ are the estimated label distributions for $A = 0$ and $A = 1$, respectively, and D represents the probability distance measure. The parameter θ is learned by minimizing the classification loss function while ensuring that the disparity between the estimated label distributions is within a specified threshold ϵ.

*LAS Group, ETH Zurich Switzerland. Emails: yifan.hu@inf.ethz.ch.

However, solving the constrained fair classification problem directly can be challenging. To address this, we can introduce a penalized objective function that combines the classification error and a regularization parameter multiplied by the distance between the two distributions:

$$\underset{\theta}{\text{minimize}} \quad \mathcal{L}(f(\theta; X), Y) + \lambda \cdot D\left(P(f(\theta; X)|A = 0), P(f(\theta; X)|A = 1)\right), \tag{2}$$

where λ is the regularization parameter that controls the trade-off between classification accuracy and fairness. By adjusting the value of λ, we can emphasize either fairness or accuracy in the resulting fair classifier. The parameter θ is learned by minimizing the penalized objective function while considering the fairness constraint.

In existing literature on fair classification, the focus has primarily been on minimizing the mean discrepancy [Rychener et al., 2022] between the estimated label distributions for different protected attributes, often referred to as minimum mean discrepancy. However, in this work, we take a more comprehensive approach by considering a wider range of integral probability measures. While minimum mean discrepancy provides a useful measure of distance between distributions, it may not capture the underlying structure and shape of the distributions accurately. In contrast, we propose to leverage more general integral probability measures, such as Wasserstein distance, Sinkhorn distance, and various other distance measurements. These measures offer a more versatile framework for capturing different aspects of fairness and enable a more accurate comparison between the estimated label distributions for different protected attributes. By considering these more general integral probability measures, we aim to enhance the fairness properties of our proposed fair classification algorithm and address potential limitations of existing approaches.

3 Expected Contributions

The expected contributions of this thesis are as follows:

1. Design of unbiased gradient estimators for the penalized fair classification objective functions with various integral probability measures in offline and online settings, respectively, addressing the limitations of existing approaches. Design the online online fair classification using the unbiased gradient estimators.

2. Analysis of the complexity bounds of the proposed algorithm, including time complexity and memory requirements, providing insights into its computational efficiency and scalability.

3. Evaluation of the proposed methods on benchmark datasets for online fair classification, considering multiple integral probability measures and varying values of the regularization parameter, to assess their performance and fairness properties.

4. Analysis of the trade-offs between fairness and classification performance when using different integral probability measures, providing insights into the implications of these measures on algorithm behavior and fairness-accuracy trade-offs.

5. Recommendations for practitioners on selecting appropriate integral probability measures and regularization parameter settings based on the desired trade-offs between fairness and classification performance in online fair classification tasks, aiding decision-making in practical applications.

4 You will learn

1. State of the art fair classification techniques.

2. How to design unbiased gradient estimators in general for various applications.
3. Develop an online fair classification algorithm and analyze the complexity bounds of these algorithms.
4. Implement the proposed algorithm using appropriate machine learning frameworks and optimization techniques.

5 Timeline

The estimated timeline for completing this thesis is as follows:

- Literature review and problem understanding: **1 month**
- Algorithm design and analysis: **2 months**
- Dataset collection and preprocessing: **0.5 months**
- Experimental evaluation and performance analysis: **1.5 months**
- Results analysis and thesis write-up: **1 month**

6 Interested?

This project is primarily of theoretical nature. Before making an inquiry, please make sure that

- You are a master student.
- You have passed some optimization, machine learning, and/or statistics courses.
- You are able to read and understand (most of) the reference listed in this documents including the analysis.

If you are interested, please contact Yifan Hu (yifan.hu@inf.ethz.ch) with a CV, a transcript, and a **writing example (if possible)**. The example can be the report of a course/semester project, your bachelor thesis, or a previous publication.

References

Yves Rychener, Bahar Taskesen, and Daniel Kuhn. Metrizing fairness. *arXiv preprint arXiv:2205.15049*, 2022.