When Fairness Optimization Goes Wrong

Marco Favier[1] and Toon Calders[1]

University of Antwerp, Antwerp, Belgium
{marco.favier,toon.calders}@uantwerpen.be

Keywords: Classification, Fairness, Cake-Cutting, Impossibility Results, Cherry-Picking

Long Abstract

In our work "Cherry on the Cake: Fairness is NOT an Optimization Problem" [3], we explore how the optimization of fairness measures can have unintended consequences, leading to unfair outcomes. We achieve this by linking cake-cutting theory with multi-label classification, demonstrating the equivalence between the two problems and how findings from one field can inform the other.

In the field of fair machine learning, researchers often work with so-called *fairness metrics*: statistical measures that serve as proxies for the abstract concept of fairness. A model is typically considered fair if the values of these metrics fall within an acceptable range. However, having correct values for fairness metrics is not enough to ensure that a model is truly fair. It is often easy to demonstrate cases where a model with strong fairness metrics remains fundamentally unfair.

For instance, imagine that to meet the fairness metric's requirements, we need to hire an equal number of female and male candidates from a pool of applicants in which both genders are equally represented. If we select the top 5 male candidates and 5 random female candidates, the fairness metric may be satisfied, but the model is still far from fair. First and foremost, the most qualified female candidates are not selected, which is inherently unfair to the women who are more deserving. Second, the model risks reinforcing the stereotype that women are unsuited for the workforce, as the less qualified candidates are less likely to perform well. This could validate the beliefs of subconsciously misogynistic individuals through confirmation bias. Third, such practices undermine the value of fairness and ethics, both in machine learning and in society as a whole.

The practice of selecting individuals solely to meet fairness constraints without regard for ethical considerations is commonly referred to in the literature as *cherry-picking* [4, 2, 5]

From a practical perspective, a common way to ensure fairness metrics are met is to include them as additional objectives in the optimization process, penalizing models that do not meet the desired fairness criteria. However, these methods may not be entirely sound, as the fairness term shifts from being a descriptor of the model's fairness to becoming an objective in itself.

The well-known Goodhart's law, "When a measure becomes a target, it ceases to be a good measure," succinctly captures the issue with fairness metrics.

In our research, we demonstrate that this is not merely a speculative concern but a mathematical fact. We prove that, under specific conditions, the optimal choice for a fairness metric may involve cherry-picking, even if the model is not explicitly directed to do so. This outcome is a direct consequence of the optimization process itself rather than an intended result. This poses a risk, as unaware practitioners could inadvertently end up with a cherry-picking model without realizing it.

On the other hand, we also prove that for a few fairness metrics, the optimal solution does not exhibit these problematic behaviors, improving on previous work by Corbett-Davies et al. and Menon et al. [1, 6]. We outline the conditions under which this is possible or impossible.

The main theoretical tool used to prove these results is cake-cutting theory, a field of mathematics that studies how to fairly divide a heterogeneous resource among multiple participants. Just as each person has their own preferences regarding which part of the cake they find most enjoyable, each participant has specific preferences for which portion of the resource they consider most valuable.

This has a strong parallel with multilabel classification, where each label can be thought of as a player and the data points as the resource, or cake. The decision function partitions the dataset, assigning each data point to a label, and how well the label is predicted corresponds to how much a player values their piece of cake.

By formally establishing this connection, we successfully translated findings from cake-cutting theory into the realm of fair machine learning, equipping us with the essential tools to support our main results; tools that we believe will be significant for the advancement of (fair) machine learning in general.

In conclusion, our work shows that the common practice of optimizing models for fairness metrics can lead to unfair outcomes and that the optimization process itself can result in cherry-picking. Fairness metrics should be used as descriptors of the model, and optimizing them is a problematic misuse.

What we recommend is to first use a well-calibrated probabilistic model to accurately assign scores to each data point, and then adjusting the decision threshold to satisfy the desired fairness criteria. This approach prevents the emergence of unusual probabilistic artifacts from optimizing the fairness metric and increases the likelihood that the model will be fair.

As a final note, we want to stress that our work is not a critique of fairness metrics. The authors strongly believe that fairness metrics are a necessary tool to evaluate the fairness of a model. However, we must remain vigilant about the problems that may arise from the misuse of these metrics. In particular, we should not confuse the fairness of a model with the value of its fairness metrics.

This misconception is particularly dangerous, as it empowers individuals to exploit the system for their own benefit. When we focus solely on fairness metrics rather than fairness itself, it becomes easy to manipulate statistics and produce models that appear fair only on paper. Our work seeks to raise awareness of this issue and provide a mathematical foundation to better understand the problem.

References

1. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining. pp. 797–806 (2017)
2. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
3. Favier, M., Calders, T.: Cherry on the cake: Fairness is not an optimization problem. arXiv preprint arXiv:2406.16606 (2024)
4. Fleisher, W.: What's fair about individual fairness? In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 480–490 (2021)
5. Goethals, S., Martens, D., Calders, T.: Precof: counterfactual explanations for fairness. Machine Learning pp. 1–32 (2023)
6. Menon, A.K., Williamson, R.C.: The cost of fairness in binary classification. In: Conference on Fairness, accountability and transparency. pp. 107–118. PMLR (2018)