# Learning Optimal Fair Scoring Systems for Multi-Class Classification

Julien Rouzot, Julien Ferry, Marie-José Huguet

LAAS-CNRS, Université de Toulouse, CNRS, INSA, Toulouse, France
{jrouzot,jferry,huguet}@laas.fr

**Keywords** : *Mixed-integer linear programming, supervised learning, multi-class classification, fairness, interpretability.*

## 1 Introduction

Machine Learning (ML) models are increasingly used for decision making, in particular in high-stakes applications such as credit scoring, medicine or recidivism prediction. However, there are growing concerns about these models with respect to their lack of interpretability and the undesirable biases they can generate or reproduce [2, 4]. While the concepts of interpretability and fairness have been extensively studied by the scientific community in recent years, few works have tackled the general multi-class classification problem under fairness constraints, and none of them proposes to generate fair and interpretable models for multi-class classification. In ML, a classification task refers to a predictive problem in which, given a set of samples characterized by some input features, a *model* aims to predict the label associated to this input. Supervised learning methods process a labelled dataset and exploit the correlations learnt from the data to produce a model. We will focus on the more general multi-class classification, in which the output of the classification algorithm can take only one value in a set $\mathcal{K}$ of possible classes with $|\mathcal{K}| > 2$. In our work [3], we used Mixed-Integer Linear Programming techniques to produce inherently interpretable optimal scoring systems under sparsity and fairness constraints, for the multi-class classification setup.

## 2 FAIRScoringSystems : A framework to generate fair and interpretable models for multi-class classification

With the SLIM framework [5], C. Rudin and B. Ustun proposed to generate interpretable and sparse optimal scoring systems for binary classification using Mixed-Integer Linear Programming. A binary scoring system can be represented as a table in which each row associates a Boolean condition over the dataset's features to a number of points. If a new sample satisfies the given condition, the associated number of points is added to the score of the sample (which is later compared to a threshold to compute the binary prediction). We extend binary scoring systems to multi-class classification using the *one-vs-all* paradigm [1]. More precisely, one scoring system is generated for each label $k$ of the dataset, whose purpose is to detect examples belonging to class $k$. To classify a new sample, each scoring system is applied and the class corresponding to the scoring system with the highest score is predicted.

In the literature, statistical fairness for multi-class classification is measured by applying binary statistical fairness metrics on each label. We propose a more generic setup in which the set of labels is partitioned into a subset of *sensitive labels* and a subset of *unsensitive labels*. We then apply the binary fairness metrics on the *sensitive labels* only.

FAIRScoringSystems [3] generates multi-class scoring systems, maximizing accuracy or balanced accuracy, given multi-class fairness and sparsity constraints. We use Mixed Integer

SCORE FOR quality == LOW

| | |
|---|---|
| type == white ? | -4 PTS |
| volatile acidity == HIGH ? | 3 PTS |
| chlorides == LOW ? | -3 PTS |
| total sulfur dioxide == LOW ? | -4 PTS |
| alcohol == LOW ? | 4 PTS |

SCORE FOR quality == MED

| | |
|---|---|
| fixed acidity == LOW ? | -2 PTS |
| volatile acidity == HIGH ? | -2 PTS |
| pH == HIGH ? | -2 PTS |
| alcohol == LOW ? | -2 PTS |
| alcohol == HIGH ? | 3 PTS |

SCORE FOR quality == HIGH

| | |
|---|---|
| fixed acidity == LOW ? | -4 PTS |
| fixed acidity == HIGH ? | -1 PTS |
| pH == HIGH ? | -4 PTS |
| alcohol == LOW ? | -4 PTS |
| alcohol == HIGH ? | 4 PTS |

FIG. 1: Multi-class scoring system for wine classification

Linear Programming to learn the scoring systems' coefficients with optimality guarantees (or a bounded optimality gap). Due to the declarative nature of the approach, such constraints can be tuned by the user and additional operational constraints can easily be handled. More precisely, we limit the number of non-zeros coefficients (i.e. number of lines) in each scoring system and the statistical fairness violation on the chosen metric(s), on the training set.

We empirically evaluate FAIRScoringSystems using one synthetic and two real-world datasets. We compare the performances of FAIRScoringSystems with two baseline methods: `FAIR` (fair baseline, majority class constant classifier) and `SVM` (accurate baseline, support vector machine black-box). Figure 2 displays the Pareto fronts for accuracy/fairness trade-offs for different sparsity parameters.

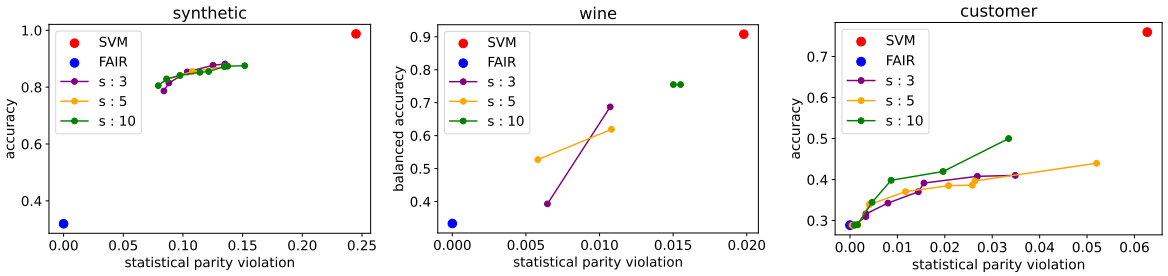


FIG. 2: Test set *accuracy* or *balanced accuracy* for different *sparsity* constraints and *fairness* values (*statistical parity* metric), for the three datasets (left: *synthetic*, middle: *wine*, right: *customer*)

Our experimental results demonstrates that FAIRScoringSystems is able to generate interesting trade-offs between accuracy, fairness and sparsity on both synthetic and real-world multi-class classification datasets of various shapes. While reaching and proving optimality for difficult datasets (*i.e.*, non-linearly separable, with high numbers of samples and features) is computationally challenging, our method can still be used to produce well-performing models. Moreover, to the best of our knowledge, this is the first work tackling both interpretability and fairness, in the context of multi-class classification.

# References


[1] Mohamed Aly. Survey on multiclass classification methods. *Neural Netw*, 19(1-9):2, 2005.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

[3] Julien Rouzot, Julien Ferry, and Marie-José Huguet. Learning optimal fair scoring systems for multi-class classification. In *ICTAI 2022-The 34th IEEE International Conference on Tools with Artificial Intelligence*, 2022.

[4] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

[5] Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Mach. Learn.*, 102(3):349–391, 2016.