Sesión Especial 22

Fair & Trustworthy Machine Learning

Organizers:

- Paula Gordaliza (Universidad Pública de Navarra)
- Jesús F. López-Fidalgo (Universidad de Navarra)

Description:
The aim of this session is to bring together researchers who are making valuable contributions to the field of Fair and Trustworthy Machine Learning from different perspectives: optimisation, counterfactual explanations, causality, optimal transport, trimming, bayesian approach, etc.

Schedule

Jueves, 25 de enero:

11:30 – 12:00	Jean-Michel Loubes (Institut de Mathématiques, Université de Toulouse III - Paul Sabatier and Artificial and Natural Intelligence Toulouse Institute (ANITI)) *Fairness for Deep Neural Network using Optimal Transport*
12:00 – 12:30	Ainhize Barrainkua (Basque Center for Applied Mathematics) *Expanding Fairness Horizons: Enhancing Classifier Fairness via Modified Uncertainty Sets in Robust Risk Minimization*
12:30 – 13:00	Elena de Diego (DATAI, Universidad de Navarra) *Enhancing Fair Learning via an extended total repair transport approach*
13:00 – 13:30	Hristo Inouzhe (Universidad Autónoma de Madrid) *Impact of nursing home status on healthcare outcomes: the Basque Country's case during the COVID-19 pandemic*
16:00 – 16:30	Álvaro Méndez Civieta (IBiDat, uc3m - Santander Big Data institute) *Fair ML strategies and Multi-Sensitive variables in practice*
16:30 – 17:00	Carlos Mougan (University of Southampton) *Beyond Demographic Parity: Redefining Equal Treatment*
17:00 – 17:30	Iris Domínguez (Universidad Pública de Navarra) *Cuando la IA sale mal, sesgos y discriminación*
17:30 – 18:00	Benoit Rottembourg (INRIA) *Contributions of surrogate models to marketplace pricing audits*

Fairness for Deep Neural Network using Optimal Transport

Jean-Michel Loubes,

IMT, ANITI

loubes@math.univ-toulouse.fr

Abstract: The increasingly common use of neural network classifiers in industrial and social applications of image analysis has allowed impressive progress these last years. Such methods are, however, sensitive to algorithmic bias, i.e., to an under- or an over-representation of positive predictions or to higher prediction errors in specific subgroups of images. We then introduce in this paper a new method to temper the algorithmic bias in Neural-Network-based classifiers. Our method is Neural-Network architecture agnostic and scales well to massive training sets of images. It indeed only overloads the loss function with a Wasserstein-2-based regularization term for which we back-propagate the impact of specific output predictions using a new model, based on the Gâteaux derivatives of the predictions distribution. This model is algorithmically reasonable and makes it possible to use our regularized loss with standard models. We apply it to bias mitigation in image and text analysis.

Expanding Fairness Horizons: Enhancing Classifier Fairness via Modified Uncertainty Sets in Robust Risk Minimization

Ainhize Barraiunkua, Paula Gordaliza, Novi Quadrianto, José Antonio Lozano

Basque Center for Applied Mathematics

abarrainkua@bcamath.org

Abstract: In recent years, fairness in machine learning has become a focal point of research and practice due to the realization that algorithms can propagate biased decision-making when trained on data reflecting societal inequalities. This study presents a novel approach to enhance fairness assurances within Minimax Risk Classifiers by adapting the uncertainty set, contrasting with conventional methods that typically redefine the loss function. This innovative formulation significantly reduces the requirement for sensitive information, necessitating access to just a few instances. Notably, our empirical observations reveal that Minimax Risk Classifiers offer robust fairness guarantees, even when sensitive information remains inaccessible. This makes them a fitting choice for scenarios where the collection of such data might be illegal or individuals opt to withhold it due to privacy concerns. These findings underscore the appeal of Minimax Risk Classifiers in situations constrained by data privacy considerations.

Acknowledgments: This research was supported by a European Research Council (ERC) Starting Grant for the project "Bayesian Models and Algorithms for Fairness and Transparency", funded under the European Union's Horizon 2020 Framework Programme (grant agreement no. 851538); by the Basque Government under grant IT1504-22 and through the BERC 2022-2025 program; by the Spanish Ministry of Science and Innovation under the grants PID2022-137442NB-I00 and PID2021-128314NB-I00, and through BCAM Severo Ochoa accreditation CEX2021-001142-S / MICIN / AEI / 10.13039/501100011033.

Enhancing Fair Learning via an extended total repair transport approach

Elena M. De-Diego, Paula Gordaliza, Jesús López-Fidalgo

Datai, Institute of Data Science and Artificial Intelligence, University of Navarra
TECNUN School of Engineering, University of Navarra, Campus Universitario

emartindedi@unav.es

Abstract: Automated decision-making systems are increasingly being employed in various domains such as healthcare, recruitment, and criminal justice. This has made the intersection of artificial intelligence (AI) and ethics a crucial issue in recent years. Fair learning has established itself as a very active area of research which tries to ensure that predictive algorithms are not discriminatory towards any individual at individual or group level, based on demographic characteristics. Total repair is a recent statistical method that mitigates bias by transforming original characteristics while maintaining conditional distributions with respect to the protected attribute (e.g., age). This adjustment is achieved by mapping the conditional distributions of each sensitive group to their respective Wasserstein barycenter using discrete optimal transport plans, which establish one-to-one correspondences. This fact imposes a limitation: the previously computed optimal transport maps cannot be applied to new samples. We study a computationally efficient method designed to extend the optimal transport plan between two probability distributions over Euclidean space (R^d) to incorporate new observations without requiring a complete re-computation of the discrete optimal transport plan. Our approach leverages the concept of the auction algorithm, enabling us to establish that the computational burden associated with the minimum cycle mean problem - a pivotal step in the creation of a consistent estimation - is considerably reduced when compared to the conventional methodology employing Karp's algorithm.

Acknowledgments: This work was supported by the Gobierno de Navarra through the MRR Investigo 2023 program (grant no. 0011-4001- 2023-000071) and by the Spanish Ministry of Science and Innovation under the grant PID2021-128314NB-I00.

Impact of nursing home status on healthcare outcomes: the Basque Country's case during the COVID-19 pandemic

Hristo Inouzhe, Irantzu Barrio, Paula Gordaliza, María Xosé Rodríguez-Álvarez ,Itxaso Bengoechea, José M. Quintana

Universidad Autónoma de Madrid

hristo.inouze@uam.es

Abstract: We explore the impact of nursing home status on healthcare outcomes such as hospitalisation, mortality and in-hospital mortality during the COVID-19 pandemic. There have been some public claims on restrictions on access to hospitals and treatments for nursing home residents during the beginning of the pandemic in some areas of Spain, which raised a public outcry about the fairness of such measures. In this work, the case of the Basque Country is studied under a rigorous statistical approach and physician's perspective. As fairness/unfairness is hard to model mathematically and has strong real-world implications, this work concentrates on the following simplification: evaluating if nursing home status had a direct effect on healthcare outcomes once accounted for other meaningful patients' information such as age, health status and period of the pandemic, among others. The methods followed here are a combination of established statistical techniques as well as new proposals from the fields of causality and fair learning. The results suggest that as a group, people in nursing homes were significantly less likely to be hospitalised, and considerably more likely to die, even in hospitals, compared to their non-residents counterparts during most of the pandemic. Further data collection and analysis are needed to guarantee that this is solely/mainly due to nursing home status.

Fair ML strategies and Multi-Sensitive variables in practice

Álvaro Méndez Civieta,

Universidad Carlos III

alvaro.mendez@uc3m.es

Abstract: This presentation encompasses a comprehensive exploration of fairness in machine learning, encapsulating a detailed literature review of various fairness metric definitions and algorithmic implementations. Our research systematically assesses ten machine learning algorithms tailored for addressing bias in models when considering a single sensitive variable. Additionally, we propose a straightforward yet effective approach to extend these algorithms to accommodate multiple sensitive variables concurrently, facilitating the application of fairness in real-world scenarios.

Beyond Demographic Parity: Redefining Equal Treatment

Carlos Mougan, Laura State, Antonio Ferrara, Salvatore Ruggieri, Steffen Staab

University of Southampton

carmougan@gmail.com

Abstract: Liberalism-oriented political philosophy reasons that all individuals should be treated equally independently of their protected characteristics. Related work in machine learning has translated the concept of equal treatment into terms of equal outcome and measured it as demographic parity (also called statistical parity). Our analysis reveals that the two concepts of equal outcome and equal treatment diverge; therefore, demographic parity does not faithfully represent the notion of equal treatment. We propose a new formalization for equal treatment by (i) considering the influence of feature values on predictions, such as computed by Shapley values decomposing predictions across its features, (ii) defining distributions of explanations, and (iii) comparing explanation distributions between populations with different protected characteristics. We show the theoretical properties of our notion of equal treatment and devise a classifier two-sample test based on the AUC of an equal treatment inspector. We study our formalization of equal treatment on synthetic and natural data. We release explanationspace, an open-source Python package with methods and tutorials.

Acknowledgments: This work was supported by Horizon 2020 funding from the European Union through the Marie Skłodowska-Curie Actions (860630) for the ”NoBIAS” project.

Cuando la IA sale mal, sesgos y discriminación

Iris Domínguez, Daniel Paternain Dallo, Mikel Galar Idoate

Universidad Pública de Navarra

iris.dominguez@unavarra.es

Resumen: Con el continuo despliegue de nuevas aplicaciones de Inteligencia Artificial, los problemas éticos asociados se vuelven cada vez más importantes. En esta charla revisaremos los conceptos de sesgo y discriminación desde la perspectiva de la inteligencia artificial, así como las formulaciones matemáticas y metodologías que nos permiten operativizarlos. En particular, nos centramos en la caracterización de los sesgos presentes en los datasets usados para el entrenamiento en problemas de Deep Learning, y de qué formas afectan a las predicciones de estos modelos.

Contributions of surrogate models to marketplace pricing audits

Benoit Rottembourg, Jeanne Mouton

INRIA

benoit.rottembourg@inria.fr

Abstract: In a global context where competition authorities are investigating and sanctioning Amazon marketplace for practices of self-preferencing at the expense of their business users and consumers (Italian AGCM 2021, EU Commission 2022, UK CMA on-going since 2022, US FTC on-going since 2023), we observe a trend of imposing remedies on dominant players in digital markets. In addition, the Digital Market Act, shifting from an ex-post enforcement approach to ex-ante obligations on designated gatekeepers, is strengthening auditing power over these gatekeepers, which risk heavier penalties in the event of non-compliance. Therefore, competition authorities and regulators need tools to audit the compliance of these dominant players in the e-commerce sector over the obligations and remedies they are imposing on dynamic, and personalized algorithms. Most of these algorithms embed Machine-Learning components, introducing opacity and potentially biases in the decision-making process. The aim of our presentation is to explore the benefits of using black-box auditing techniques to provide insights into the behavior of these online algorithms. We anchor our research in the literature of product preeminence from vertically integrated players, of choice ranking, and of the specific literature related to Amazon search ranking, automatic pricing and Buy Box's algorithms. Through a study of the pricing and ranking of several thousand products on Amazon, from 2017 to 2023, we will illustrate the potential of surrogate models and the decision-support elements they might provide. We seek to outline the limits of such models, their role in guiding the auditor, and raise the question of their probative value in the DMA context.