CD03

Fairness and Interpretability in Machine Learning

Equidad e Interpretabilidad en Aprendizaje Automático

Organizers **Organizadores** **Antolatzaileak**

Pablo Morala Miguélez
(Universidad Carlos III de Madrid)

Paula Gordaliza Pastor
(Universidad Pública de Navarra)

Description **Descripción** **Deskribapena**

The adoption of machine learning and artificial intelligence techniques across various fields requires careful oversight to ensure their use is fair, non-discriminatory, and understandable to those affected. Two key areas have emerged to address these concerns: fairness and interpretability. Fairness focuses on detecting and preventing biases, especially those impacting vulnerable groups, both in data and models. Interpretability, on the other hand, aims to make model decisions understandable, allowing the assignment of importance to original variables and thus facilitating fairness analysis.

La adopción de técnicas de machine learning e inteligencia artificial en diversos campos requiere un control cuidadoso para asegurar un uso justo, no discriminatorio y comprensible para las personas afectadas. Así, han surgido dos áreas clave para abordar estas preocupaciones: el fairness y la interpretabilidad. El fairness se enfoca en detectar y prevenir sesgos, especialmente contra colectivos vulnerables, tanto en los datos como en los modelos. La interpretabilidad, por su parte, busca hacer entendibles las decisiones de los modelos, permitiendo asignar importancia a las variables originales y facilitando así el análisis de fairness.

MSC Codes **Códigos MSC** **MSC Kodeak**

62-XX
(primary)

Slots **Bloques** **Blokeak**

2.B (Aula 0.16); 2.C (Aula 0.16)

QR Code **Código QR** **QR Kodea**

Session Schedule **Horario de la Sesión** **Saioaren Ordutegia**

J16 | 16:30-16:50 | 0.16
Functional relevance based on the Shapley value
Cristian Pachón García (Universitat Politècnica de Catalunya)

J16 | 17:00-17:20 | 0.16
Fourier Analysis in CNN
Isabel María Moreno Cuadrado (University Complutense)

J16 | 17:30-17:50 | 0.16
Sensitivity Analysis of NSUM Estimators in the context of Social-Networks
Sergio Díaz-Aranda (IMDEA Networks)

J16 | 18:00-18:20 | 0.16
Feature interactions in XAI: a comparison study of SHAP extensions and NN2Poly
Pablo Morala (IBiDat - Universidad Carlos III de Madrid)

V17 | 9:30-9:50 | 0.16
Fair Partial Least Squares
Adrián Pérez-Suay (Universitat de València)

V17 | 10:00-10:20 | 0.16
Integrating bias mitigation techniques for fair and accurate machine learning with multiple sensitive variables
Sandra Benítez-Peña (Universidad Carlos III de Madrid)

V17 | 10:30-10:50 | 0.16

Estimating Average Treatment Effects through Generalized Trimming: Applications to Decision Auditing

Hristo Inouzhe (Universidad Autónoma de Madrid)

Thursday 16
16:30-16:50
[Room 0.16]

Jueves 16
16:30-16:50
[Aula 0.16]

Osteguna 16
16:30-16:50
[Gela 0.16]

Functional relevance based on the Shapley value

Cristian Pachón García
(Universitat Politècnica de Catalunya)

Consider a scalar-on-function regression problem, where the goal is to predict a scalar response from a functional predictor. Several predictive models have been proposed in the Functional Data Analysis literature, but many of them are difficult to interpret since it is hard to identify the relevance of the functional predictors. In this work, we extend relevance measures based on the Shapley value from multivariate to functional predictors by adapting concepts from the continuous games theory.

Joint work with Pedro Delicado.

Thursday 16
17:00-17:20
[Room 0.16]

Jueves 16
17:00-17:20
[Aula 0.16]

Osteguna 16
17:00-17:20
[Gela 0.16]

Fourier Analysis in CNN

Isabel María Moreno Cuadrado
(University Complutense)

In an era where explainability in Deep Learning (DL) is crucial, this talk demonstrates how mathematics supports advances in DL, particularly in Convolutional Neural Networks (CNN). The presentation is divided into two parts: a mathematical exploration of Fourier analysis and its application to image filtering, followed by a computational analysis using the Fast Fourier Transform (FFT) to optimize CNN training. This approach aims to improve efficiency, aligning with the goals of "Green AI".

Thursday 16	**Jueves 16**	**Osteguna 16**
17:30-17:50	**17:30-17:50**	**17:30-17:50**
[Room 0.16]	**[Aula 0.16]**	**[Gela 0.16]**

Sensitivity Analysis of NSUM Estimators in the context of Social-Networks

Sergio Díaz-Aranda
(IMDEA Networks)

The Network Scale-up Methods (NSUM) estimate hidden populations through indirect surveys using participants' aggregated data about acquaintances. This study compares nine NSUM through simulations, examining factors like network structure, subpopulation distribution, sample size, and biases. Findings show that some lesser-used estimators excel in specific cases of network configuration and biases, while the most common estimator is less sensitive to subpopulation configuration and recall error.

Joint work with Jose Aguilar, Juan Marcos Ramírez, David Rabanedo, Antonio Fernández Anta, and Rosa E. Lillo.

doi:10.1080/00031305.2024.2421361

Thursday 16	**Jueves 16**	**Osteguna 16**
18:00-18:20	**18:00-18:20**	**18:00-18:20**
[Room 0.16]	**[Aula 0.16]**	**[Gela 0.16]**

Feature interactions in XAI: a comparison study of SHAP extensions and NN2Poly

Pablo Morala
(IBiDat - Universidad Carlos III de Madrid)

Explaining feature importance in model predictions is key in Explainable AI (XAI). However, most methods focus on single variables, overlooking interactions common in real-world problems. This work compares extensions of SHAP values, a popular interpretability method, to account for interactions, alongside NN2Poly, a neural network specific interpretability method. Simulations under various settings compare local and global explanations and propose metrics for computing importance order.

Joint work with J. Alexandra Cifuentes, Rosa E. Lillo, and Iñaki Úcar.

Friday 17 | **Viernes 17** | **Ostirala 17**
9:30-9:50 | **9:30-9:50** | **9:30-9:50**
[Room 0.16] | **[Aula 0.16]** | **[Gela 0.16]**

Fair Partial Least Squares

Adrián Pérez-Suay
(Universitat de València)

We address fair representation learning using fair Partial Least Squares (PLS) components, a technique commonly used in statistics for efficient data dimensionality reduction tailored for prediction. We introduce a novel method that integrates fairness constraints into the construction of PLS components, applicable in both linear and nonlinear cases using kernel embeddings. Our algorithm's effectiveness is demonstrated across various datasets.

Joint work with Elena M. De Diego, Paula Gordaliza and Jean-Michel Loubes.

Friday 17 | **Viernes 17** | **Ostirala 17**
10:00-10:20 | **10:00-10:20** | **10:00-10:20**
[Room 0.16] | **[Aula 0.16]** | **[Gela 0.16]**

Integrating bias mitigation techniques for fair and accurate machine learning with multiple sensitive variables

Sandra Benítez-Peña
(Universidad Carlos III de Madrid)

As machine learning's role in decision-making grows, concerns about fairness and bias have risen. Various fairness techniques address discrimination based on sensitive variables like race or gender, but little research combines these methods or considers multiple sensitive attributes simultaneously. This project explores the impact of combining fairness algorithms to enhance equity, offering insights for real-world applications like hiring and credit approval.

Joint work with Rosa Lillo, Arturo Pérez and Fabio Scielzo.

Friday 17	**Viernes 17**	**Ostirala 17**
10:30-10:50	**10:30-10:50**	**10:30-10:50**
[Room 0.16]	**[Aula 0.16]**	**[Gela 0.16]**

Estimating Average Treatment Effects through Generalized Trimming: Applications to Decision Auditing

Hristo Inouzhe
(Universidad Autónoma de Madrid)

Causal inference estimates the impact of a treatment on an outcome under strong assumptions. A key measure, the Average Treatment Effect, reflects the difference in outcome likelihood between the same population when fully treated or untreated. This applies to auditing decisions involving sensitive attributes that shouldn't influence outcomes. We introduce a generalized trimming method using Maximum Mean Discrepancies, offering a flexible alternative to existing causal inference techniques.

Joint work with Eustasio del Barrio, Paula Gordaliza and Jean Michel Loubes.