Exchange programme Vrije Universiteit Amsterdam

Vrije Universiteit Amsterdam - Exchange programme Vrije Universiteit Amsterdam - 2024-2025

Exchange

Vrije Universiteit Amsterdam offers many English-taught courses in a variety of subjects, ranging from arts & culture and social sciences, neurosciences and computer science, to economics and business administration.

The International Office is responsible for course approval and course registration for exchange students. For details about course registration, requirements, credits, semesters and so on, please visit the exchange programmes webpages.

Fair, Transparent and Interpretable Machine Learning

Course Code	E_EOR3_FML
Credits	6
Period	P5
Course Level	300
Language Of Tuition	English
Faculty	School of Business and Economics
Course Coordinator	prof. dr. J. Schaumburg
Examiner	prof. dr. J. Schaumburg
Teaching Staff	prof. dr. D. Wozabal, S.H. Kooiker, prof. dr. J. Schaumburg
Teaching method(s)	Study Group, Lecture

Course Objective

Students know about, and are able to apply basic statistical concepts of fairness in algorithmic decision making. They also understand the challenges and limitations of these concepts in real-world settings. Furthermore, students know and are able to apply different approaches to interpret the model outcomes of supervised machine learning methods, including both intrinsically interpretable models and model-agnostic methods.

Course Content

Machine learning algorithms are increasingly used to make or improve predictions, which then serve as a basis for decision making. Examples include bank lending, college admissions, and bail decisions in criminal proceedings.

Though the use of algorithmic decision making is often justified as being "more objective" than human decision making, there are many instances demonstrating that it can produce biased or discriminatory predictions or decisions that unfairly disadvantage certain individuals or groups. Awareness of this issue and knowledge about approaches to address it are of high importance for data scientists and policy makers.

Another highly relevant aspect for decision making based on data is the interpretability of the estimation outcomes and decisions obtained using a machine learning method. One possibility is to restrict the class of applied algorithms to interpretable models (e.g. decision trees, linear regression, logistic regression). However, "black-box" methods (e.g. random forests, deep neural networks) have proven to be highly effective in many more complex settings, while not providing a means to understand the sources of a particular prediction or decision. Model-agnostic methods such as partial dependence plots, Local Surrogate models (LIME), and Shapley Values are important concepts enhancing interpretability and allowing for comparisons of any set of machine learning outcomes.

This course is divided into two parts. The first part of the course (weeks 1-3) addresses the topic of interpretability in supervised machine learning settings. We will study local and global methods to interpret the outcomes of black-box models and apply them to a range of real-world examples. The second part (weeks 4-6) is concerned with fairness in machine learning. We will introduce formal definitions of fairness, analyze real-world data sets, and discuss what algorithmic decision making can and cannot achieve.

Additional Information Teaching Methods

4 hours per week of lectures, 2 hours per week of tutorials.

Method of Assessment

Written exam, group assignment.

Entry Requirements

Knowledge of machine learning methods, programming knowledge in Python.

Literature

Selected scientific papers and book chapters.

Explanation Canvas

All course materials and additional information will be disseminated via Canvas.