COURSE DATA

Data Subject	
Code	46579
Name	Causal inference and machine learning
Cycle	Master's degree
ECTS Credits	3.0
Academic year	2023 - 2024

Study (s)

Degree	Center	Acad. year	Period
2262 - M.U. en Ciencia de Datos	School of Engineering	1	Second term

Subject-matter

Degree	Subject-matter	Character
2262 - M.U. en Ciencia de Datos	11 - Causal inference and machine learning	Obligatory

Coordination

Name	Department
MONTORO PONS, JUAN DE DIOS	110 - Applied Economics

SUMMARY

Causal Inference and Machine Learning aims to provide a solid understanding of the fundamental concepts and techniques used to estimate causal relationships as well as to explore recent advances with the incorporation of methods based on machine learning. All this is motivated through applications using data from different fields of study, mainly economics but also other areas such as sociology, medicine, epidemiology, biology or ecology.

The course gradually addresses the main aspects of causal inference. It starts from a research design aimed at answering a causal question (effect of a treatment X on the response Y) and the introduction of the differences between the use of experimental and observational data. Specifically, we explore the distinction between data obtained through controlled experiments and data in which the researcher lacks control over the generation process. Empirical research using quantitative methods and machine learning relies mostly on observational data, such as data collected from an organization’s internal records, from instruments designed for specific purposes (surveys), or generated through online activity to name just a few. In this context, the strengths and weaknesses of each data type are discussed and the main challenge associated with the use of observational data to test causal relationships, i.e. the problem iof identification, is examined. This arises from the difficulty of estimating precise and unbiased causal relationships using non-experimental data that, in the absence of a theoretical model and an adequate empirical strategy, only provides correlations.

From a theoretical perspective, complementary approaches for the modeling of causal relationships are proposed. We introduce DAGs (Directed Acyclic Graphs) as a tool to represent and visualize the causal structure of a model and help in its identification strategy. Additionally, the potential outcomes model explores the idea of the contrast between observed outcomes and those that would have been observed under different treatment conditions (counterfactuals). The latter allows us to identify the problem of causal inference as a missing value problem and connects it with recent developments in machine learning.

Next, we introduce different techniques for the estimation of causal effects under the assumption of selection on observables. Here we deal with those methods that seek to mitigate bias in observational data stemming from the existence of confounders (observed variables that influence both the response and the treatment). Specifically, parametric methods such as regression, and semi-parametric/non-parametric methods such as matching or propensity score weighting are discussed.

The catalog of techniques for selection on observables is extended with the introduction of methods that arise from recent developments in machine learning. Thus, double machine learning, causal trees and causal forests are incorporated. These methods allow to retrieve not only average causal effects, but also their distribution, introducing the possibility of exploring heterogeneity in causal effects and the factors with which it is associated.

The last block of the subject deals with selection techniques on unobservables, a situation in which there are latent traits affecting treatment and outcome (or, to put it differently, there are confounders that cannot be measured). This is a complex methodological challenge that requires the introduction of additional hypotheses for the estimation of causal effects. The validity of the results is based on the plausibility of these restrictions and on an appropriate validation strategy. This part includes designs whose origin is in econometrics: instrumental variables, selection models, fixed effects, differences in differences, and regression discontinuity designs. Likewise, the literature on machine learning offers flexible alternatives for solving the problem, such as instrumental forests.

PREVIOUS KNOWLEDGE

Relationship to other subjects of the same degree

There are no specified enrollment restrictions with other subjects of the curriculum.

Other requirements

There are no prerequisites.

OUTCOMES

2262 - M.U. en Ciencia de Datos

- Students should be able to integrate knowledge and address the complexity of making informed judgments based on incomplete or limited information, including reflections on the social and ethical responsibilities associated with the application of their knowledge and judgments.
- Students should communicate conclusions and underlying knowledge clearly and unambiguously to both specialized and non-specialized audiences.
- Students should demonstrate self-directed learning skills for continued academic growth.
- Students should possess and understand foundational knowledge that enables original thinking and research in the field.
- Be able to assess the need to complete their technical, scientific, language, computer, literary, ethical, social and human education, and to organise their own learning with a high degree of autonomy.
- Capacidad de organización y planificación de actividades de investigación, desarrollo y consultoría en el área de ciencia de datos.
- Ser capaces de acceder a herramientas de información (bibliográficas y de empleo) y utilizarlas apropiadamente.
- Ser capaces de asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio, aplicando los conocimientos adquiridos en la identificación de salidas profesionales y yacimientos de empleo.
- Extraer conocimiento de conjuntos de datos en diferentes formatos.
- Modelar la dependencia entre una variable respuesta y varias variables explicativas, en conjuntos de datos complejos, mediante técnicas de aprendizaje máquina, interpretando los resultados obtenidos.
- Diseñar y poner en marcha soluciones basadas en análisis de datos teniendo en cuenta los requisitos específicos para cada aplicación.

LEARNING OUTCOMES

Distinguish between pure prediction problems and causal inference problems. Interpret the identification error as associated with observational data and big data. Know the main approaches to modeling causal effects: the Rubin model (potential results) and graph-based models. Know and apply estimation methods under selection in observables (regression, matching, reweighting) and non-observables (instrumental variables, selection models, fixed effects, differences in differences, and discontinuity regression). Know and apply double ML and causal forests.

DESCRIPTION OF CONTENTS

1. Introduction to causal inference

Pure prediction problems vs. causal inference problems
Experimental data and observational data
The problem of identification
Directed Acyclic Graphs (DAGs)
The model of potential results (Rubin's model)
Types of causal effects
Identification strategies

2. Selection on observables

Selection on observables and identification: parametric and non-parametric methods
Matching
Propensity score matching
Diagnosis
Doubly robust estimation
Inverse probability weighting

3. Causality and machine learning

Causal inference and prediction revisited
Causal trees and causal forests
Estimation of heterogeneous effects (HTE), average by groups (GATE), and conditional effects (CATE)
Treatment prioritization rules: rank-weighted average treatment effect (RATE)
Double machine learning (DML)
Other machine learning developments applied to causal inference

4. Selection on unobservables

Selection models
Instrumental variables
Panel data: fixed effects and random effects.
Panel data: differences in differences
Discontinuity regression
Machine learning developments in selection on unobservables: instrumental forests

WORKLOAD

ACTIVITY	Hours	% To be attended
Theory classes	19,00	100
Laboratory practices	9,00	100
Theoretical and practical classes	2,00	100
Development of group work	15,00	0
Study and independent work	5,00	0
Readings supplementary material	5,00	0
Preparation of evaluation activities	5,00	0
Preparing lectures	5,00	0
Preparation of practical classes and problem	5,00	0
Resolution of online questionnaires	5,00	0
TOTAL	75,00	

TEACHING METHODOLOGY

Theoretical activities. Expository development of the subject with the participation of the student in the resolution of specific questions. Carrying out individual evaluation questionnaires.

Practical activities. Learning through problem solving, exercises and case studies through which skills are acquired on different aspects of the subject.

Work in laboratory and/or computer classroom. Learning by carrying out activities carried out individually or in small groups and carried out in computer classrooms.

EVALUATION

1. One or several tests that will consist of both theoretical-practical questions and problems (30%)

2. Evaluation of the practical activities from the elaboration of works/reports, oral presentations and e-learning tools of the University (60%).

3. Evaluation based on the participation and degree of involvement of the student in the teaching-learning process, taking into account regular attendance at the planned face-to-face activities and the resolution of questions and problems proposed periodically (10%).

REFERENCES

Basic

- Cunningham, S. (2021). Causal inference. Yale University Press. (disponible online en https://mixtape.scunning.com/
- Huntington-Klein, N. (2021). The effect: An introduction to research design and causality. Chapman and Hall/CRC. (disponible en https://theeffectbook.net/

Additional

- Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. Annual Review of Economics, 10, 465-503. (https://www.annualreviews.org/doi/10.1146/annurev-economics-080217-053402)
- Angrist, J. D., & Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton university press.
- Athey, Susan (2017). Beyond prediction: Using big data for policy problems. Science, 355(6324), 483-485. (https://www.science.org/doi/10.1126/science.aal4321)
- Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (2021). DoubleML--An Object-Oriented Implementation of Double Machine Learning in R. arXiv preprint arXiv:2103.09603. (https://arxiv.org/pdf/2103.09603.pdf)
- Cameron, A.C. & Trivedi, P.K. (2005). Microeconometrics. Methods and Applications. Cambridge University Press.
- Cerulli, G. (2015). Econometric evaluation of socio-economic programs Theory and applications. Springer.
- Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." Journal of Economic Perspectives, 31 (2): 87-106. (https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. Advances in methods and practices in psychological science, 1(1), 27-42 (https://journals.sagepub.com/doi/full/10.1177/2515245917745629).
- Varian, H. R. (2016). Causal inference in economics and marketing. Proceedings of the National Academy of Sciences, 113(27), 7310-7315. (https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1319839
- Wager S. & Athey, S. (2018) "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests", Journal of the American Statistical Association. https://doi.org/10.1080/01621459.2017.1319839