FICHA IDENTIFICATIVA

Datos de la Asignatura

Código	46579
Nombre	Inferencia causal y aprendizaje máquina
Ciclo	Máster
Créditos ECTS	3.0
Curso académico	2023 - 2024

Titulación(es)

Titulación	Centro	Curso	Periodo
2262 - M.U. en Ciencia de Datos	Escuela Técnica Superior de Ingeniería	1	Segundo cuatrimestre

Materias

Titulación	Materia	Caracter
2262 - M.U. en Ciencia de Datos	11 - Inferencia causal y aprendizaje máquina	Obligatoria

Coordinación

Nombre	Departamento
MONTORO PONS, JUAN DE DIOS	110 - Economía Aplicada

RESUMEN

La asignatura **inferencia causal y aprendizaje máquina** tiene como objetivo proporcionar una comprensión sólida de los conceptos fundamentales y las técnicas utilizadas para la estimación de relaciones causales así como explorar avances recientes con la incorporación de métodos basados en el aprendizaje máquina. Todo esto se motiva con aplicaciones para datos procedentes de diversos campos de estudio, principalmente economía aunque también de otras áreas como sociología, medicina, epidemiología, biología o ecología.

El curso aborda de manera gradual los principales aspectos de la inferencia causal. Partiendo del diseño de una investigación que busca dar respuesta a una cuestión causal (efecto de un tratamiento X sobre la respuesta Y) se introducen las diferencias entre el uso de datos **experimentales y observacionales**. En concreto se explora la distinción entre datos obtenidos a través de experimentos controlados y datos en los que el investigador carece de control sobre su proceso de generación. La investigación empírica en métodos cuantitativos y aprendizaje máquina se apoya mayoritariamente en datos observacionales, como son los datos recopilados a partir de registros de organizaciones, provenientes de instrumentos diseñados con fines específicos (encuestas), o generados a través de la actividad en redes por citar tres ejemplos. En este contexto, se discuten las fortalezas y debilidades de cada tipo de datos y se examina el principal

desafío asociado al uso de datos observacionales para establecer relaciones causales: **el problema de la identificación**. Este surge por la dificultad de estimar relaciones causales precisas e insesgadas utilizando datos no experimentales que, en ausencia de un modelo teórico y una estrategia empírica adecuada, solo proporcionan correlaciones.

Desde una perspectiva teórica se proponen seguidamente enfoques complementarios para la modelización de relaciones causales. Introducimos los **DAGs** (*Directed Acyclic Graphs* o *Grafos Acíclicos Dirigidos*) como una herramienta para representarlas y visualizarlas que ayudan a comprender la estructura causal de un modelo y a su identificación. Adicionalmente, el **modelo de resultados potenciales** explora la idea de contraste entre los resultados observados y aquellos que se habrían observado bajo diferentes condiciones de tratamiento (**contrafactuales**). Este último permite identificar el problema de la inferencia causal como un problema de valores faltantes y lo conecta con los desarrollos recientes en aprendizaje máquina.

A continuación, se desarrollan diferentes técnicas para la estimación de efectos causales bajo el supuesto de **selección en observables.** Aquí se agrupan aquellos métodos que buscan mitigar el problema del sesgo en datos observacionales por la existencia de confusión en variables observables (variables que influyen tanto en la respuesta como en el tratamiento). En concreto se discuten métodos paramétricos como la regresión, y semiparamétricos/no-paramétricos como el emparejamiento (*matching*) o la ponderación por puntaje de propensión (*propensity score weighting*).

El catálogo de técnicas para la selección en observables se amplía con la introducción de métodos que surgen de **recientes desarrollos en aprendizaje máquina**. Así, se incorporan *double machine learning*, árboles causales (c*ausal trees*) y bosques causales (c*ausal forests*). Estos métodos permiten no solo recuperar efectos causales medios, sino también la distribución de los mismos, introduciendo la posibilidad de explorar la heterogeneidad en los efectos causales y los factores con la que se asocia.

El último bloque de la materia aborda las técnicas de **selección en no observables,** una situación en la que existen variables no observadas que afectan tanto a la respuesta como al tratamiento. Es este un desafío metodológico complejo que dificulta la identificación del efecto causal y exige la introducción de hipótesis adicionales para su estimación. La validez de los resultados se apoya en la plausibilidad de estas restricciones y en una adecuada estrategia de validación de los mismos. Se incluyen en esta parte diseños cuyo origen está en la **econometría**: variables instrumentales, modelos de selección, efectos fijos, diferencias en diferencias, y regresión en discontinuidad. Asimismo, la literatura en aprendizaje máquina ofrece alternativas flexibles para la solución del problema como son los bosques instrumentales (*instrumental forests*).

CONOCIMIENTOS PREVIOS

Relación con otras asignaturas de la misma titulación

No se han especificado restricciones de matrícula con otras asignaturas del plan de estudios.

Otros tipos de requisitos

No hay requisitos previos

COMPETENCIAS

2262 - M.U. en Ciencia de Datos

- Que los/las estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios.
- Que los/las estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades.
- Que los/las estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo
- Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.
- Ser capaces de valorar la necesidad de completar su formación técnica, científica, en lenguas, en informática, en literatura, en ética, social y humana en general, y de organizar su propio autoaprendizaje con un alto grado de autonomía
- 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.

RESULTADOS DE APRENDIZAJE

Aprender a distinguir entre problemas de predicción puros y problemas de inferencia causal. Interpretar el error de identificación como asociado a datos observacionales y big data. Conocer los principales enfoques de modelización de efectos causales: el modelo de Rubin (resultados potenciales) y los modelos basados en grafos. Conocer y aplicar métodos de estimación bajo selección en observables (regresión, pareamiento, reponderación) y no observables (variables instrumentales, modelos de selección, efectos fijos, diferencias en diferencias y regresión en discontinuidad). Conocer y aplicar double ML y bosques causales

DESCRIPCIÓN DE CONTENIDOS

1. Introducción a la inferencia causal

Problemas de predicción puros vs. problemas de inferencia causal
Datos experimentales y datos observacionales
El problema de la identificación
Grafos acíclicos dirigidos (DAGs)
El modelo de resultados potenciales (modelo de Rubin)
Tipos de efectos causales
Estrategias de identificación

2. Selección en observables

Selección en observables e identificación: métodos paramétricos y no-paramétricos
Emparejamiento (matching)
Emparejamiento por puntaje de propensión (propensity score matching)
Diagnóstico
Estimación doblemente robusta
Ponderación de probabilidad inversa

3. Causalidad y aprendizaje máquina

Inferencia causal y predicción revisited
Árboles causales (causal trees) y bosques causales (causal forests)
Estimación de efectos heterogeneos (HTE), promedio por grupos (GATE), y condicionales (CATE)
Reglas de priorización de tratamiento: rank-weighted average treatment effect (RATE)
Double machine learning (DML)
Otros desarrollos de aprendizaje máquina aplicados a la inferencia causal

4. Selección de no observables

Modelos de selección
Variables instrumentales
Datos de panel: efectos fijos y efectos aleatorios
Datos de panel: diferencias en diferencias
Regresión en discontinuidad
Desarrollos de aprendizaje máquina en problemas de selección en no observables: bosques instrumentales (instrumental forests)

VOLUMEN DE TRABAJO

ACTIVIDAD	Horas	% Presencial
Clases de teoría	19,00	100
Prácticas en laboratorio	9,00	100
Clases teórico-prácticas	2,00	100
Elaboración de trabajos en grupo	15,00	0
Estudio y trabajo autónomo	5,00	0
Lecturas de material complementario	5,00	0
Preparación de actividades de evaluación	5,00	0
Preparación de clases de teoría	5,00	0
Preparación de clases prácticas y de problemas	5,00	0
Resolución de cuestionarios on-line	5,00	0
TOTAL	**75,00**	

METODOLOGÍA DOCENTE

Actividades teóricas. Desarrollo expositivo de la materia con la participación del estudiante en la resolución de cuestiones puntuales. Realización de cuestionarios individuales de evaluación.

Actividades prácticas. Aprendizaje mediante resolución de problemas, ejercicios y casos de estudio a través de los cuales se adquieren competencias sobre los diferentes aspectos de la materia.

Trabajos en laboratorio y/o aula ordenador. Aprendizaje mediante la realización de actividades desarrolladas de forma individual o en grupos reducidos y llevadas a cabo en aulas de ordenador.

EVALUACIÓN

1. Prueba objetiva, consistente en uno o varios exámenes que constarán tanto de cuestiones teórico-prácticas como de problemas (30%)
2. Evaluación de las actividades prácticas a partir de la elaboración de trabajos/memorias, exposiciones orales y herramientas de e-learning de la Universitat (60%).
3. Evaluación basada en la participación y grado de implicación del alumno en el proceso de enseñanza-aprendizaje, teniendo en cuenta la asistencia regular a las actividades presenciales previstas y la resolución de cuestiones y problemas propuestos periódicamente (10%).

Las calificaciones obtenidas en los apartados 2 y 3 se conservarán en las dos convocatorias del curso académico en que hayan sido realizadas, dado que su evaluación sólo es posible en el periodo de docencia.

REFERENCIAS

Básicas

- 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/

Complementarias

- 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