UAB
Universitat Autònoma de Barcelona

Matemáticas para Big Data

Código: 45562
Créditos ECTS: 6

2025/2026

Titulación	Tipo	Curso
Modelling for Science and Engineering	OP	1

Contacto

Nombre: Amanda Fernandez Fontelo

Correo electrónico: amanda.fernandez@uab.cat

Equipo docente

Sundus Zafar

Idiomas de los grupos

Puede consultar esta información al final del documento.

Prerrequisitos

Los estudiantes deben tener conocimientos básicos de álgebra lineal, probabilidad, inferencia estadística y modelos lineales. La experiencia previa con los programas R y Python es muy recomendable.

Objetivos y contextualización

Actualmente, se generan grandes volúmenes de datos en ámbitos muy diversos como el ámbito de la salud, la ingeniería, las ciencias sociales, la economía, etc. Este crecimiento exponencial de los datos representa, al mismo tiempo, un reto y una gran oportunidad para extraer información relevante que permita tomar decisiones fundamentadas, optimizar procesos o generar nuevo conocimiento. El objetivo principal de este curso es proporcionar al estudiante los conocimientos matemáticos, estadísticos y computacionales, así como las herramientas necesarias para procesar, analizar y modelizar grandes conjuntos de datos. Además, se hace énfasis en la interpretación y uso de la información obtenida, con el fin de transformar los datos en conocimiento útil que permita extraer conclusiones más precisas y tomar decisiones más acertadas. El curso se centra especialmente en el aprendizaje y la aplicación de algunos métodos matemáticos, estadísticos y computacionales para identificar patrones, tendencias y relaciones en conjuntos de datos masivos y complejos.

Resultados de aprendizaje

1. CA27 (Competencia) Aplicar las herramientas matemáticas del análisis de las grandes bases de datos a la resolución de problemas en el ámbito empresarial o de la investigación
2. CA28 (Competencia) Integrar las herramientas matemáticas del análisis de las grandes bases de datos a otras herramientas en entornos de trabajo multidisciplinares
3. CA29 (Competencia) Comunicar a un público especializado y no especializado los resultados obtenidos al aplicar los métodos estadísticos de análisis de grandes bases de datos en distintos ámbitos
4. CA30 (Competencia) Trabajar en equipos multidisciplinares en el desarrollo de proyectos donde se apliquen técnicas de deep learning
5. KA21 (Conocimiento) Describir las técnicas de aprendizaje automático profundo (Deep Learning) usadas en el análisis de grandes bases de datos
6. KA22 (Conocimiento) Describir las herramientas matemáticas empleadas en el análisis de grandes bases y volúmenes de datos
7. SA27 (Habilidad) Aplicar técnicas matemáticas de tratamiento de grandes bases de datos para analizar fenómenos particulares, como patrones de comportamiento del consumidor, tendencias de mercado o análisis de redes sociales.
8. SA27 (Habilidad) Aplicar técnicas matemáticas de tratamiento de grandes bases de datos para analizar fenómenos particulares, como patrones de comportamiento del consumidor, tendencias de mercado o análisis de redes sociales.
9. SA27 (Habilidad) Aplicar técnicas matemáticas de tratamiento de grandes bases de datos para analizar fenómenos particulares, como patrones de comportamiento del consumidor, tendencias de mercado o análisis de redes sociales.
10. SA28 (Habilidad) Interpretar los resultados obtenidos después de analizar una base de datos de gran tamaño

Contenido

Block 1. Text Mining (10 h):

- Fundamentals of Text Mining - From text to numbers.
- Data cleaning.
- Tokenization.
- Stemming.
- Lemmatization.
- POS, NER.
- Data chunking.

Block 2. Statistics for Big Data (18 h):

- Topic 1. The principle of sufficiency: Summarising the information from large, complex datasets.
  Sufficient statistics and the factorisation theorem.
  Classic examples of sufficient statistics.
- Topic 2. Classical linear models with large, complex datasets.
  A quick review of classical linear models and the ordinary least squares estimator. Examples.
  Use of sufficient statistics for the estimation of classical linear models. Examples.
- Topic 3. Generalised linear models with large, complex datasets: Logit and Poisson models.
  A quick review of logit and Poisson models, as well as the associated likelihood-based estimators.
  Introduction to the idea of segmentation: The "Divide and Recombine" method and meta-analysis-based methods.
  The estimation of logit and Poisson models using the idea of segmentation. Examples.
  Some recent advances in the idea of "Divide and Recombine".
- Topic 4. The problem of multiple testing and the false discovery rate.
  Introduction to knockoff variables.

Block 3. Deep Learning (10 h):

- Fully Connected Neural Networks.
- Convolutional Neural Networks.
- Recurrent Neural Networks.
- Keras and Tensorflow.

Actividades formativas y Metodología

Título	Horas	ECTS	Resultados de aprendizaje
Tipo: Dirigidas			
Sesiones de problemas y prácticas	11	0,44	CA27, CA28, CA29, CA30, SA27, SA28, CA27
Sesiones de teoría	19	0,76	CA28, CA29, KA21, KA22, CA28
Tipo: Supervisadas			
Sesiones de problemas y prácticas	8	0,32	CA27, CA28, CA29, CA30, SA27, SA28, CA27
Tipo: Autónomas			
Ampliación de conceptos introducidos en las sesiones de teoria	43	1,72	CA28, KA21, KA22, SA28, CA28
Tareas para trabajar los conceptos introducidos en las sesiones presenciales	50	2	CA27, CA28, CA29, CA30, KA21, KA22, SA27, CA27

El curso se organiza en tres bloques temáticos independientes, cada uno de ellos impartido por un profesor diferente. Mientras que el primer bloque del curso (10 h) introduce conceptos de minería de datos, generalmente aplicada a grandes conjuntos de datos, el segundo bloque (18 h) se centra en los métodos y conocimientos estadísticos necesarios para la modelización de estos grandes volúmenes de datos, y en particular, en cómo ajustar y realizar inferencia en modelos lineales clásicos y modelos lineales generalizados (modelos logísticos y de Poisson) cuando trabajamos con grandes cantidades de información. Finalmente, el tercer bloque del curso (10 h) introduce al alumnado en algunos de los métodos más relevantes del aprendizaje profundo, prestando especial atención a las redes neuronales y sus aplicaciones.

En general, cada bloque combina sesiones teóricas y de introducción de conceptos por parte del profesorado con sesiones prácticas, que pueden ser sesiones dirigidas o bien sesiones de trabajo autónomo. Las sesiones teóricas podrán ir acompañadas de diapositivas, las que se compartirán a través del Moodle del curso. Las sesiones de problemas y las prácticas dirigidas podrán incluir ejemplos prácticos con R y/o Python, que, en general, también estarán disponibles en el Moodle del curso. El profesorado, si lo considera oportuno, podrá compartir a través de Moodle material adicional que deberá ser trabajado de forma autónoma por el alumnado con el fin de profundizar en los conceptos introducidos en clase.

Nota: se reservarán 15 minutos de una clase dentro del calendario establecido por el centro o por la titulación para que el alumnado rellene las encuestas de evaluación de la actuación del profesorado y de evaluación de la asignatura o módulo.

Evaluación

Actividades de evaluación continuada

Título	Peso	Horas	ECTS	Resultados de aprendizaje
Proyectos Bloque 1	26	5	0,2	CA27, CA28, CA29, KA22, SA27, SA28
Proyectos Bloque 2	48	9	0,36	CA27, CA28, CA29, KA22, SA27, SA28
Proyectos Bloque 3	26	5	0,2	CA27, CA28, CA29, CA30, KA21, KA22, SA27, SA28

La evaluación del curso se realiza de forma independiente en cada uno de los bloques descritos. El peso de cada bloque en la calificación final coincide con el número de horas de ese bloque en relación con el total de horas del curso. En general, en cada bloque se propondrá un conjunto de proyectos que podrán desarrollarse de forma individual o en grupo, y que, en algunos casos, incluirán una pequeña parte en la que se requerirá la presentación oral de contenidos y resultados. Los proyectos de los bloques 1 y 3 tienen un peso del 26% cada uno en la nota final de la asignatura, mientras que los proyectos del bloque 2 tienen un peso del 48%.

Para cada proyecto se publicará en el Moodle del curso un documento con el enunciado y la descripción de lo que se solicita, así como todo el material necesario para desarrollar el proyecto (conjuntos de datos, fuentes de información adicionales, etc.), la fecha de entrega, el mecanismo de entrega, y otros detalles que el profesorado considere relevantes. Las calificaciones de cada proyecto, así como las calificaciones finales de cada bloque y del curso, se publicarán también en el Moodle del curso.

Bibliografía

Referencias básicas:

- B. Efron, T. Hastie. Computer Age Statistical Inference, Cambridge University Press, 2018. https://bibcercador.uab.cat/permalink/34CSUC_UAB/1eqfv2p/alma991010753063206709
- G. James, D. Witten, T. Hastie and R. Tibshirani. An Introduction to Statistical Learning (with applications in R). Springer, 2013. https://bibcercador.uab.cat/permalink/34CSUC_UAB/1c3utr0/cdi_globaltitleindex_catalog_2960062
- D. Skillicorn. Understanding Complex Data. Data Mining with Matrix Decomposition. Chapman & Hall, 2007. https://bibcercador.uab.cat/permalink/34CSUC_UAB/1eqfv2p/alma991004136809706709

Referencias complementarias:

- B. Everitt and T. Hothorn. An introduction to Applied Multivariate Analysis with R. Springer, 2011.
- B. Everitt. An R and S+ Companion to Multivariate Analysis. Springer, 2005.
- J. Faraway. Extending de Linear Model with R. Chapman & Hall, Miami, 2006.
- J. Faraway. Linear Models with R. Chapman & Hall, Boca Raton, 2005.
- W. Härdle and L. Simar. Applied Multivariate Statistical Analysis. Springer. 2007.
- B. Ripley. Pattern Recognition and Neural Networks. Cambridge University Press, 2002.
- L. Torgo. Data Mining with R. Learning with Case Studies. Chapman & Hall, Miami. 2010
- W Venables, B Ripley. Modern Applied Statistics with S-PLUS. Springer, New York.

Los profesores podrán proporcionar otras referencias de interés para cada bloque, las cuales estarán disponibles a través de la página de Moodle.

Software

R Core Team (2021). R: A language and environment for statistical computing.
R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/.

Python

Grupos e idiomas de la asignatura

La información proporcionada es provisional hasta el 30 de noviembre de 2025. A partir de esta fecha, podrá consultar el idioma de cada grupo a través de este enlace. Para acceder a la información, será necesario introducir el CÓDIGO de la asignatura

Nombre	Grupo	Idioma	Semestre	Turno
(TEm) Teoría (máster)	1	Inglés	segundo cuatrimestre	tarde