Universitat Autònoma de Barcelona

2021/2022

Matemáticas para "Big Data"

Código: 43478
Créditos ECTS: 6

Titulación	Tipo	Curso	Semestre
4313136 Modelización para la Ciencia y la Ingeniería / Modelling for Science and Engineering	OT	0	2

La metodología docente y la evaluación propuestas en la guía pueden experimentar alguna modificación en función de las restricciones a la presencialidad que impongan las autoridades sanitarias.

Contacto

Nombre: Alejandra Cabaña Nigro

Correo electrónico: AnaAlejandra.Cabana@uab.cat

Uso de idiomas

Lengua vehicular mayoritaria: inglés (eng)

Otras observaciones sobre los idiomas

Este documento es una traducción del original en inglés. En el caso de imprecisiones o discrepancias, la versión en ingléses la guía válida y oficial del curso.

Equipo docente

Pere Puig Casado

Antonio Lozano Bagen

Sundus Zafar

Prerequisitos

Los estdiantes deberían tener conocimientos básicos de álgebra lineal, inferencia estadística, y modelos lineales.

La experiencia previa con R y Python es recomendable.

Objetivos y contextualización

Hoy en día se están generando enormes cantidadesde datos en muchos campos, y el propósito de este curso es aprender a extraer información a partir de estos datos.

El objetivo de este curso es para aprender y aplicar varios métodos matemáticos y estadísticos para el descubrimiento de patrones enconjuntos de datos.

Cuando se trabaja con datasets grandes, los procedimientos matemáticos tienen que ser escalables, así que nos ocuparemos de métodos que puedan ser escalados y/o paralelizados.

Competencias

- "Aplicar el pensamiento lógico/matemático: el proceso analítico a partir de principios generales para llegar a casos particulares; y el sintético, para a partir de diversos ejemplos extraer una regla general."
- Analizar, sintetizar, organizar y planificar proyectos de su campo de estudio.
- Aplicar las técnicas de resolución de los modelos matemáticos y sus problemas reales de implementación.
- Concebir y diseñar soluciones eficientes, aplicando técnicas computacionales, que permitan resolver modelos matemáticos de sistemas complejos.
- Extraer de un problema complejo la dificultad principal, separada de otras cuestiones de índole menor.
- Formular, analizar y validar modelos matemáticos de problemas prácticos de distintos campos.
- Resolver problemas complejos aplicando los conocimientos adquiridos a ámbitos distintos de los originales

Resultados de aprendizaje

1. "Aplicar el pensamiento lógico/matemático: el proceso analítico a partir de principios generales para llegar a casos particulares; y el sintético, para a partir de diversos ejemplos extraer una regla general."
2. Analizar, sintetizar, organizar y planificar proyectos de su campo de estudio.
3. Aplicar técnicas de Estadística Bayesiana para predecir el comportamiento futuro de ciertos fenómenos.
4. Extraer de un problema complejo la dificultad principal, separada de otras cuestiones de índole menor.
5. Identificar fenómenos reales como modelos de procesos estocásticos y saber extraer de aquí información nueva para interpretar la realidad
6. Resolver problemas complejos aplicando los conocimientos adquiridos a ámbitos distintos de los originales
7. Resolver problemas reales de análisis de datos identificándolos adecuadamente desde la óptica de la Estadística Bayesiana.
8. Usar paquetes estadísticos y métodos bayesianos apropiados para solucionar problemas concretos.

Contenido

Minería de Textos

- Fundamentos de Text Mining: de texto a números
- Limpieza de datos
- Tokenization
- Stemming
- Lemmatizattion
- POS,NER
- Data chunking

Estadística

- El problema de las comparaciones múltiples (multiple testing).
- Modelos lineales y Modelos Lineales generalizados: LASSO & BigLASSO, Regresión Ridge y Elastic Nets. Selección de Variables.
- Summarising the information of large data sets: sufficient statistics. Application to linear models. The Biglm package.
- Likelihood estimation problems for large data sets. Segmentation, analysis of chunks of data, methods based on meta-analysis. Applications to Generalised linear models.

Alternativa

- Análisis de datos funcionales
- Procesos Gausianos y aprendizaje automático.

Deep Learning

- Fully Connected Neural Networks.

- Convolutional Neural Networks.
- Recurrent Neural Networks.

- Keras and Tensorflow.

Metodología

Ver la versión de la guía en inglés.

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.

Actividades

Título	Horas	ECTS	Resultados de aprendizaje
Tipo: Dirigidas			
Clases teóricas	38	1,52	2, 5
Ejercicios (problemas y programación)	36	1,44	1, 8
Tipo: Autónomas			
Estudio autónomo	20	0,8	5
Homework	44	1,76	1, 2, 4, 5, 6, 8

Evaluación

La evaluación constará de ejercicios propuestos a lo largo del curso sobre los diferentes tópicos (60% de la nota) y

Un proyecto de análisis de datos según las instrucciones publicadas en el Campus Virtual.

Actividades de evaluación

Título	Peso	Horas	ECTS	Resultados de aprendizaje
Deep Learning	0.25	3	0,12	1, 2, 4, 5, 6, 8
Homework Estadística Part B	0.25	3	0,12	1, 2, 4, 5, 6, 7, 8
Homework Estadística Parte A	0.25	3	0,12	1, 2, 3, 4, 5, 6, 7
Homework Text Mining	0.25	3	0,12	1, 2, 4, 6, 8

Bibliografía

Basic references

B. Efron, T. Hastie, *Computer Age Statistical Inference*, Cambridge University Press (2016) (5th Ed 2017) https://web.stanford.edu/~hastie/CASI/index.html

G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning (with applications in R)*. Springer, 2013.

D. Skillicorn, "Understanding Complex Data. Data Mining with Matrix Decomposition". Chapman&Hall, 2007.

Complementary references

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 Statisticswith S-PLUS", Springer, New York.

Collins FS and Varmus H, "A new initiative on precision medicine". N Engl J Med. 2015 Feb 26;372(9):793-5 .

Jensen A.B. et al, "Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients". Nat Commun 2014 Jun 24; 5:4022.

J.D. Jobson, "Applied Multivariate Analysis". Vol I i II. Springer, 1992.

R. Johnson and D.W. Wichern, "Applied Multivariate Statistical Analysis". Pearson Education International, 2007.

P.Y.Lum et al., "Extracting insights from the shape of complex data using topology". Sci. Rep. 3, 1236; DOI:10.1038/srep01236 (2013).

A. Rencher, "Methods of Multivariate Analysis". Wiley Series in Probability and Mathematical Statistics, 2002.

G. Singh, F. Mémoli, G. Carlsson, "Topological methods for the analysis of High dimensional data sets and 3D object recognition". Eurographic Symp. on Point-Based Graphics, 2007

Journal of Statistical Software, http://www.jstatsoft.org/

Dealing with Data (2011) Special Issue. Science 11 February 2011:692-789

P. Kokoszka, M. Reimherr, *Introduction to Functional Data Analysis*. CRC Press.(2017).

Ramsay, J. , B. W. Silverman,*Functional Data Analysis Springer* (2nd Ed. 2005).

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