Universitat Autònoma de Barcelona

2019/2020

"Matemàtiques per a ""Big Data"""

Codi: 43478
Crèdits: 6

Titulació	Tipus	Curs	Semestre
4313136 Modelització per a la Ciència i l'Enginyeria / Modelling for Science and Engineering	OT	0	2

Professor/a de contacte

Nom: Alejandra Cabaña Nigro

Correu electrònic: AnaAlejandra.Cabana@uab.cat

Utilització d'idiomes a l'assignatura

Llengua vehicular majoritària: anglès (eng)

Altres indicacions sobre les llengües

Aquest document és una traducció de l'original en anglès. En el cas d'imprecisions o discrepàncies, la versió en ingléses la guia vàlida i oficial del curs.

Equip docent

Albert Ruíz Cirera

Joan Gasull Jolis

Equip docent extern a la UAB

Isabel Serra

Prerequisits

Els estudiants haurien de tenir coneixements bàsics d'àlgebra lineal, inferència estadística, i models lineals.

L'experiència prèvia amb R i Python és recomanable.

Objectius

Avui dia, quantitats enormes de dades estan sent generadess a molts camps, i el propòsitl d'aquest curs és per aprendre com extreure informació a partir d' aquestes dades.

L'objectiu d'aquest curs és per aprendre i aplicar diversos mètodes matemàtics i estadístics per a la descoberta de patrons pertinents a conjunts de dades.

Quan es treballa amb datasets grans, els procediments matemàtics han de ser escalables, així que serem concernits amb mètodes que puguin ser escalata i/o paralelitzats.

Competències

- "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."
- Analitzar, sintetitzar, organitzar i planificar projectes del seu camp d'estudi.
- Aplicar les tècniques de resolució dels models matemàtics i els seus problemes reals d'implementació.
- Concebre i dissenyar solucions eficients, aplicant tècniques computacionals, que permetin resoldre models matemàtics de sistemes complexos.
- Extreure d'un problema complex la dificultat principal, separada d'altres qüestions d'índole menor.
- Formular, analitzar i validar models matemàtics de problemes pràctics de diferents camps.
- Resoldre problemes complexos aplicant els coneixements adquirits a àmbits diferents dels originals

Resultats d'aprenentatge

1. "Aplicar el pensament lògic/matemàtic: el procés analític a partir de principis generals per arribar a casos particulars; i el sintètic, para a partir de diversos exemples extreure una regla general."
2. Analitzar, sintetitzar, organitzar i planificar projectes del seu camp d'estudi.
3. Aplicar tècniques d'Estadística Bayesiana per predir el comportament futur de certs fenòmens.
4. Extreure d'un problema complex la dificultat principal, separada d'altres qüestions d'índole menor.
5. Identificar fenòmens reals com a models de processos estocàstics i saber extreure d'aquí informació nova per interpretar la realitat
6. Resoldre problemes complexos aplicant els coneixements adquirits a àmbits diferents dels originals
7. Resoldre problemes reals d'anàlisis de dades identificant-los adequadament des de l'òptica de l'Estadística *Bayesiana.
8. Usar paquets estadístics i mètodes bayesians apropiats per solucionar problemes concrets.

Continguts

Mineria de Textos

Python Crash Course

Fundamentals of Text Mining - From text to numbers

Data cleaning

N-grams, Lemmatization, Translation

Topic Modelling

Estadística

The problem of multiple testing.

Linear and Generalized linear methods: LASSO, Ridge Regression and Elastic Nets. Feature Selection.

Gaussian Processes for Machine Learning.

Alternatively,
Functional Data Analysis: Observed functional data and its computational representation,descriptive statistics and dimensionality reduction, depth measures for FD, two-sample problem for FD, Functional linear models, class cation techniques.

Anàlisi Topológica de Dades

Topology and data, quick review of linear algebra, from points to polyhedra, combinatorial topology, persistence Diagrams and software.

Aprenentatge Estadístic

Review of basic concepts and the state-of-th- art in statistical learning techniques.

Metodologia

Veure la versió de la guia en anglés.

Activitats formatives

Títol	Hores	ECTS	Resultats d'aprenentatge
Tipus: Dirigides			
Clases de Teoria	38	1,52	2, 5
Excercicis (problemes i ordinador)	36	1,44	1, 8
Tipus: Autònomes			
Estudi autònom	20	0,8	5
Projecte	44	1,76	1, 2, 4, 5, 6, 8

Avaluació

Homework: Presentació escrita dels exercicis proposats.

Projecte final: cada part del curs inclourà una sèrie de papers. Els estudiants haurán de triar-ne un i preparar una xerrada, que pot incloure exemples de dades. Aquesta tasca pot ser feta en grup.

Les dates previstes seran anunciadesd durant el curs i seràn estrictes.

Activitats d'avaluació

Títol	Pes	Hores	ECTS	Resultats d'aprenentatge
Homework	0,6	6	0,24	1, 2, 3, 4, 5, 6, 7
Projecte Final	0.4	6	0,24	1, 2, 4, 5, 6, 7, 8

Bibliografia

Referencies básiques

B. Efron, T. Hastie, *Computer Age Statistical Inference*, Cambridge University Press (2016) (5th Ed 2017)

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

Gunnar Carlsson, "Topology and data". Bull. AMS 46,2 (2009), 255-308.

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

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

Referéncies Complementàries

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.

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

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