

Titulació	Tipus	Curs
Modelització per a la Ciència i l'Enginyeria / Modelling for Science and Engineering	OP	1

Professor/a de contacte

Nom: Amanda Fernandez Fontelo

Correu electrònic: amanda.fernandez@uab.cat

Equip docent

Sundus Zafar

Idiomes dels grups

Podeu consultar aquesta informació al [final](#) del document.

Prerequisits

Els estudiants han de tenir coneixements bàsics d'àlgebra lineal, probabilitat, inferència estadística i models lineals. L'experiència prèvia amb els programaris R i Python és molt recomanable.

Objectius

Actualment, es generen grans volums de dades en àmbits molt diversos com l'àmbit de la salut, l'enginyeria, les ciències socials, l'economia, etc. Aquest creixement exponencial de les dades representa, alhora, un repte i una gran oportunitat per extreure informació rellevant que permeti prendre decisions fonamentades, optimitzar processos o generar nou coneixement. L'objectiu principal d'aquest curs és dotar l'estudiant dels coneixements matemàtics, estadístics i computacionals, així com de les eines necessàries per processar, analitzar i modelitzar grans conjunts de dades. A més, es posa èmfasi en la interpretació i aprofitament de la informació obtinguda, amb l'objectiu de transformar les dades en coneixement útil que permeti extreure conclusions més precises i prendre decisions més acurades. El curs se centra especialment en l'aprenentatge i aplicació d'alguns mètodes matemàtics, estadístics i computacionals per identificar patrons, tendències i relacions en conjunts de dades massius i complexos.

Resultats d'aprenentatge

1. CA27 (Competència) Aplicar les eines matemàtiques de l'anàlisi de les grans bases de dades a la resolució de problemes en l'àmbit empresarial o de la recerca.

2. CA28 (Competència) Integrar les eines matemàtiques de l'anàlisi de les grans bases de dades a altres eines en entorns de treball multidisciplinaris.
3. CA29 (Competència) Comunicar a un públic especialitzat i no especialitzat els resultats obtinguts d'aplicar els mètodes estadístics d'anàlisi de grans bases de dades en diferents àmbits.
4. CA30 (Competència) Treballar en equips multidisciplinaris en el desenvolupament de projectes on s'apliquin tècniques de deep learning.
5. KA21 (Coneixement) Descriure les tècniques d'aprenentatge automàtic profund (deep learning) usades en l'anàlisi de grans bases de dades.
6. KA22 (Coneixement) Descriure les eines matemàtiques emprades en l'anàlisi de grans bases i volums de dades.
7. SA27 (Habilitat) Aplicar tècniques matemàtiques de tractament de grans bases de dades per analitzar fenòmens particulars, com patrons de comportament del consumidor, tendències de mercat o anàlisi de xarxes socials.
8. SA27 (Habilitat) Aplicar tècniques matemàtiques de tractament de grans bases de dades per analitzar fenòmens particulars, com patrons de comportament del consumidor, tendències de mercat o anàlisi de xarxes socials.
9. SA27 (Habilitat) Aplicar tècniques matemàtiques de tractament de grans bases de dades per analitzar fenòmens particulars, com patrons de comportament del consumidor, tendències de mercat o anàlisi de xarxes socials.
10. SA28 (Habilitat) Interpretar els resultats obtinguts després d'analitzar una base de dades de mida molt gran.

Continguts

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.

Activitats formatives i Metodologia

Títol	Hores	ECTS	Resultats d'aprenentatge
Tipus: Dirigides			
Sessions de problemes i pràctiques	11	0,44	CA27, CA28, CA29, CA30, SA27, SA28
Sessions de teoria	19	0,76	CA28, CA29, KA21, KA22
Tipus: Supervisades			
Sessions de problemes i pràctiques	8	0,32	CA27, CA28, CA29, CA30, SA27, SA28
Tipus: Autònomes			
Ampliació de conceptes introduïts a les sessions de teoria	43	1,72	CA28, KA21, KA22, SA28
Tasques per treballar els conceptes introduïts a les sessions presencials	50	2	CA27, CA28, CA29, CA30, KA21, KA22, SA27

El curs s'organitza en tres blocs temàtics independents, cadascun impartit per un professor diferent. Mentre que el primer bloc del curs (10 h) introdueix conceptes de mineria (generalment grans conjunts de) dades, el segon bloc (18 h) es centra en els mètodes i coneixements estadístics per la modelització d'aquests grans conjunts de dades i, particularment, en com ajustar i fer inferència de models lineals clàssics i models lineals generalitzats (model logístic i de Poisson) quan tenim grans quantitats de dades. Finalment, el tercer bloc del curs (10 h) introdueix l'alumnat en alguns dels mètodes més rellevants de l'aprenentatge profund, amb especial atenció a les xarxes neuronals i les seves aplicacions.

En general, cada bloc combina sessions de teoria i introducció de conceptes per part del professor amb sessions pràctiques que poden ser sessions dirigides o bé sessions de treball autònom. Les sessions de teoria es poden acompanyar amb diapositives les quals es compartiran al Moodle del curs, mentre que les sessions de problemes i pràctiques dirigides poden acompanyar-se per exemples pràctics amb R i/o Python, els quals generalment també es podran compartir al Moodle del curs. Els professors, si així ho consideren, podran compartir, mitjançant el Moodle del curs, material addicional el qual s'haurà de treballar per part dels alumnes per aprofundir en els conceptes introduïts a classe.

Nota: es reservaran 15 minuts d'una classe, dins del calendari establert pel centre/titulació, perquè els alumnes completin les enquestes d'avaluació de l'actuació del professorat i d'avaluació de l'assignatura.

Avaluació

Activitats d'avaluació continuada

Títol	Pes	Hores	ECTS	Resultats d'aprenentatge
-------	-----	-------	------	--------------------------

Projectes Bloc 1	26	5	0,2	CA27, CA28, CA29, KA22, SA27, SA28
Projectes Bloc 2	48	9	0,36	CA27, CA28, CA29, KA22, SA27, SA28
Projectes Bloc 3	26	5	0,2	CA27, CA28, CA29, CA30, KA21, KA22, SA27, SA28

L'avaluació del curs es fa independentment en cadascun dels blocs descrits. El pes de cada bloc coincideix amb el número d'hores del bloc en relació amb el total d'hores del curs. En general, en cada bloc es plantejaran un conjunt de projectes que es podran treballar de manera individual o en grups, i que en alguns casos, una petita part del projecte requerirà la comunicació de continguts i resultats oralment. Els projectes dels blocs 1 i 3 tenen un pes de l'assignatura d'un 26% cadascun mentre que els projectes del bloc 2 tenen un pes de l'assignatura d'un 48%.

Per a cada projecte es penjarà al Moodle del curs l'enunciat amb la descripció del que es demana, tot el material necessari per a desenvolupar el projecte (conjunt de dades, fonts addicionals d'informació, etc.), la data i el mecanisme d'entrega del projecte, així com altres detalls que el professor consideri rellevants. Les qualificacions de cada projecte així com les qualificacions finals de cada bloc i del curs es publicaran també al Moodle del curs.

Bibliografia

Referències bàsiques:

- 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

Referències 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 Statistics with S-PLUS. Springer, New York.

Els professors podran proporcionar altres referències d'interès per a cada bloc, les quals estaran disponibles a través de la pàgina de Moodle.

Programari

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

Grups i idiomes de l'assignatura

La informació proporcionada és provisional fins al 30 de novembre de 2025. A partir d'aquesta data, podreu consultar l'idioma de cada grup a través daquest [enllaç](#). Per accedir a la informació, caldrà introduir el CODI de l'assignatura

Nom	Grup	Idioma	Semestre	Torn
(TEm) Teoria (màster)	1	Anglès	segon quadrimestre	tarda