FITXA IDENTIFICATIVA

Dades de l'Assignatura

Codi	36426
Nom	Aprenentatge màquina
Cicle	Grau
Crèdits ECTS	6.0
Curs acadèmic	2021 - 2022

Titulació/titulacions

Titulació	Centre	Curs	Període
1406 - Grau en Ciència de Dades	Escola Tècnica Superior d'Enginyeria	3	Primer quadrimestre

Matèries

Titulació	Matèria	Caràcter
1406 - Grau en Ciència de Dades	9 - Aprenentatge automàtic i mineria de dades	Obligatòria

Coordinació

Nom	Departament
LAPARRA PEREZ-MUELAS, VALERO	242 - Enginyeria Electrònica
MARTIN GUERRERO, JOSE DAVID	242 - Enginyeria Electrònica
MUÑOZ MARI, JORDI	242 - Enginyeria Electrònica

RESUM

L’assignatura “Aprenentatge Màquina” suposa el primer contacte de l’estudiant del Grau en Ciència de Dades amb els models matemàtics no lineals i els corresponents algorismes d’aprenentatge que permeten l’extracció d’informació emmagatzemada en bases de dades per tal de resoldre fonamentalment problemes dels següents tipus:

- Classificació
- Agrupament
- Regressió
- Predicció

- Planificació

Cadascun d'estos problemes requerix una aproximació que pot ser diferent des del punt de vista de l'aprenentatge. Per tant, l'assignatura comença revisant conceptes bàsics i definicions per a establir el marc de treball que permetrà introduir els diferents tipus d'aprenentatge automàtic que s'estudiaran: aprenentatge supervisat, no supervisat, semisupervisat, actiu i per reforç. A continuació es vorà la manera d'avaluar els resultats en estos problemes, les diferents mètriques existents, la necessitat de fer subdivisions de conjunts per a garantir un millor funcionament i les possibles millores que es poden plantejar, com ara les tècniques de *boosting* o el *bagging* per a generar *ensembles*.

Una vegada establert el marc de treball, s'estudien diferents models d'aprenentatge automàtic per a resoldre els problemes descrits anteriorment. D'esta manera, "Aprenentatge Màquina" dóna un pas més respecte a assignatures de cursos anteriors on els alumnes s'han centrat majorment en una anàlisi de dades més descriptiva o en models basats en aproximacions lineals.

Respecte als models, s'estudien els mètodes majorment utilitzats i més populars hui en dia amb l'excepció de les xarxes neuronals que es descriuen a l'assignatura "Models Connexionistes", en el segon quadrimestre. En particular, es descriuran en detall les característiques, funcionament, adequació a diferents problemes i la interpretació de models basats en màquines de vector suport *(Support Vector Machines, SVM)* i arbres de decisió; en el cas dels arbres tindrà especial rellevància la generalització amb *ensembles* ja que dóna lloc als models de boscos aleatoris *(Random Forest, RF*), que són una de les aproximacions més potents que es poden utilitzar per a resoldre problemes de regressió i classificació.

Les classes de teoria s'impartiran en castellà i les classes pràctiques i de laboratori segons consta en la fitxa de l'assignatura disponible en la web del grau.

CONEIXEMENTS PREVIS

Relació amb altres assignatures de la mateixa titulació

No heu especificat les restriccions de matrícula amb altres assignatures del pla d'estudis.

Altres tipus de requisits

Lassignatura semmarca en el 3r curs del grau, on sassumix que lestudiant té coneixements mínims dels processos implicats en una anàlisi intel·ligent de dades, com ara el filtratge de dades sorolloses, outliers, i dades perdudes, així com la representació i interpretació de les dades en un espai multidimensional i la generació de visualitzacions que permeten obtindre informació útil del problema. No hi ha requisits addicionals per a poder seguir lassignatura; dels continguts estudiants en els dos prime

1406 - Grau en Ciència de Dades

- (CG02) Capacitat de resoldre problemes amb iniciativa, creativitat, i de comunicar i transmetre coneixements, habilitats i destreses, comprenent la responsabilitat ètica i professional de l'activitat del Científic de Dades.

- (CG03) Capacitat per a la realització de models, càlculs, informes, planificació de tasques i altres treballs anàlegs en l'àmbit específic de la Ciència de Dades.
- (CT03) Habilitat per defensar el seu treball amb rigor i arguments, exposant-ho de forma adequada i precisa, recolzant-se en els mitjans necessaris.
- (CT05) Capacitat per avaluar els avantatges i inconvenients de diferents alternatives metodològiques i/o tecnològiques en diferents àmbits d'aplicació.
- (CE03) Capacitat per resoldre problemes de classificació, modelització, segmentació i predicció a partir d'un conjunt de dades.
- (CE07) Capacitat per modelar la dependència entre una variable resposta i diverses variables explicatives, en conjunts de dades complexes, mitjançant tècniques d'aprenentatge màquina, interpretant els resultats obtinguts.
- (CE13) Saber dissenyar, aplicar i avaluar algorismes de Ciència de Dades per a la resolució de problemes complexos.
- (CB3) Que els estudiants tinguen la capacitat d'arreplegar i interpretar dades rellevants (normalment dins de la seua àrea d'estudi) per emetre judicis que incloguen una reflexió sobre temes rellevants d'índole social, científica o ètica.
- (CB4) Que els estudiants puguen transmetre informació, idees, problemes i solucions a un públic tant especialitzat com no especialitzat.

Els resultats d'aprenentatge més importants d'esta assignatura són:

- Conèixer i implementar arbres basats en dades.
- Conèixer la base dels mètodes *kernel* i els diferents *kernels* que es poden plantejar.
- Obtindre regles d'associació a partir de bases de dades *(basket analysis)*.
- Conèixer les diferents formes d'associar-se els sistemes experts.

Cadascun d'estos quatre resultats permet en major o menor mesura adquirir totes les competències descrites anteriorment. No obstant això, particularitzant els resultats associats a les diferents competències, respecte a les competències bàsiques i generals, els resultats d'aprenentatge més rellevants seran:

- Realització de treballs i informes amb diferent nivell de guiat per tal que l'estudiant assolisca un alt grau d'autonomia a la fi del quadrimestre, tant en quant a la decisió de les aproximacions més adequades per a resoldre un problema com a l'obtenció dels millors resultats possibles i la seua interpretació en un entorn multidisciplinar (CG02, CB3, CB4).
- Tractament de conjunts de dades en diferents formats conferint a l'estudiant de la capacitat de fer un tractament de dades independentment del format d'estes (CG02, CG03).

Les competències transversals tindran associats els següents resultats:

- Realització de presentacions formals per a una audiència experta i no experta (CT03, CT05).
- Defensa dels arguments exposats en la presentació davant dels dubtes que puguen plantejar-se (CT03, CT05).
- Capacitat d'avaluació crítica del treball propi i del realitzat per altres companys i col·legues (CT05).

Finalment, les competències específiques es voran reflectides en els següents resultats d'aprenentatge:

- Desenvolupament de models de classificació, modelat, agrupament, predicció, regressió i planificació, tot entenent el processament intern realitzat per models i algorismes i tenint la capacitat de proposar variants que puguen millorar els resultats abastats (CE03, CE07).
- Capacitat per a adaptar els models a les característiques peculiars de cada problema (CE03, CE07).
- Capacitat per a triar els models més adequats a cada problema i avaluar-los amb mètriques objectives (CE13).
- Interpretació dels resultats del modelat i utilitat de les solucions aportades en un entorn multidisciplinar (CE07).
- Entendre els diferents tipus d'aprenentatge dels algorismes i esprémer la seua capacitat tant de manera individual com combinant-los, si escau (CE13).

DESCRIPCIÓ DE CONTINGUTS

1. Conceptes preliminars

Esta primera unitat temàtica introduix alguns conceptes necessaris per a poder preparar els conjunts de dades de manera que puguen ser analitzats per mètodes daprenentatge automàtic. En particular es voran els següents continguts:

1. Definicions: mostra, patró, característica, algorisme, dimensionalitat,...
2. Normalització i codificació
3. Selecció de característiques

3.1. Mètodes filter

3.2. Mètodes wrapper

4. Extracció de característiques(*): Descomposició en valors singulars, Anàlisi de Components Principals, Anàlisi de Components Independents, ...

(*)Estos mètodes i daltres més sofisticats sestudiaran amb molt més de detall en lassignatura Agrupament i Varietats

2. Esquemes daprenentatge i problemes associats

La segona unitat temàtica descriu els diferents tipus daprenentatge que definixen els algorismes que fan la tasca dextracció de informació en els models daprenentatge automàtic. Depenent de lesquema daprenentatge utilitzat es poden abordar diferents problemes sempre sobre la base de què estiguen definits mitjançant un conjunt de dades. Els diferents apartats que sestudiaran en la unitat temàtica II són:

1. Aprenentatge supervisat
1.1. Problemes de classificació
1.2. Problemes de regressió i predicció
2. Aprenentatge no supervisat: Problemes dagrupament i segmentació(*)
3. Aproximacions semisupervisades
3.1. Aprenentatge semisupervisat
3.2. Aprenentatge actiu
4. Aprenentatge per reforç: problemes doptimització

(*)Estos mètodes sestudiaran amb profunditat en lassignatura Agrupament i Varietats

3. Avaluació de models

La tercera unitat temàtica se centra en les diferents mètriques que es poden utilitzar per tal davaluar el rendiment de models daprenentatge automàtic. Dependent del tipus de problema abordat i per tant de lesquema daprenentatge, les mètriques a considerar són diferents. A més, sanalitzarà la conveniència duna o altra mètrica en funció de lobjectiu final que es perseguisca de manera que els models puguen esbiaixar-se per tal de minimitzar o maximitzar certs criteris. També es descriuran algunes problemàtiques que cal tindre en compte i diferents tècniques que permeten millorar el rendiment dels models i fer-los més útils de cara a la seua utilització en un problema real. La taula de continguts de la unitat temàtica III és la següent:

1. Sobreentrenament i sobreajust
2. Divisió del conjunt de dades
2.1. Hold-out
2.2. V-fold
2.3 Leave-one-out
3. Avaluació del rendiment de models daprenentatge automàtic
3.1. Problemes de classificació
3.2. Problemes de regressió
4. Millora de models
4.1. Boosting
4.2. Comitès dexperts: ensembles
4.3 Bagging

4. Màquines de Vectors Suport

Les Màquines de Vectors Suport (SVMs, pel seu nom en anglès) són un model daprenentatge inicialment proposat per a tasques de classificació però que amb modificacions pot aplicar-se també per a regressió. Són especialment indicades en espais dispersos amb poca densitat de dades, empitjorant prou el seu funcionament a mesura que augmenta el nombre de mostres del conjunt de dades. En esta unitat temàtica es revisarà el seu funcionament i característiques, desglossat de la següent manera:

1. Introducció
2. Hiperplà de separació òptima
3. El truc del kernel
4. Regressor basat en vectors suport (SVR)

5. Arbres de decisió

Esta última unitat temàtica teòrica descriu els models daprenentatge automàtic que estan basats en estructures darbres, començant amb models darbre senzill i finalitzant amb els models de boscos aleatoris (Random Forests, RFs) que fent ús de bagging generen un conjunt darbres, suposant un model que representa lestat de lart en problemes de classificació i regressió. La taula de continguts de la unitat temàtica és:

1. Representació
2. Entropia i guany dinformació
3. Poda
4. Algorismes principals
4.1. ID3
4.2. C4.5
4.3. CART: Classification and Regression Trees
4.4. CHAID: Chi-square Automatic Interaction Detection
5. Ensembles darbres
5.1. Bosc aleatori (Random Forest, RF)
5.2. Arbres extremadament aleatoris (Extremely Randomized Trees, ERTs)

6. Temes actuals en Aprenentatge Màquina

Després dhaver estudiat els mètodes més coneguts daprenentatge màquina, lobjectiu desta unitat és finalitzar lassignatura amb la descripció dalguns dels temes que més recentment sestan proposant i investigant en el camp de laprenentatge automàtic per tal que lalumne conega els darrers desenvolupaments. La següent taula de continguts és per tant flexible i adaptable a laparició de noves propostes interessants:

1. Aprenentatge profund.
2. Aprenentatge automàtic quàntic.
3. Noves aplicacions de laprenentatge màquina.

7. Pràctiques de laboratori

Finalment, per la seua importància en lassignatura sha considerat convenient incloure com a una unitat temàtica independent les pràctiques a realitzar al laboratori (aula informàtica), on lestudiant aprendrà a implementar els models descrits en les classes de teoria. Molts dels mètodes analitzats en lassignatura només cobren sentit quan es desenvolupen en entorn de laboratori en el que es pot observar la seua potencialitat, ja que pot resultar relativament complicant entendre totes les seues característiques de funcionament atenent només a lestudi teòric i a la realització dexercicis i problemes senzills. Es plantegen sis pràctiques de laboratori que corresponen amb els continguts teòrics prèviament descrits en les anteriors unitats temàtiques:

1. Preprocessament de conjunts de dades
1.1. Normalització
1.2. Codificació
1.3. Selecció de característiques:
1.3.1 Mètodes filter
1.3.2 Mètodes wrapper
2. Extracció de característiques
2.1. Separació de conjunts: Hold-out, V-fold, Leave-one-out
2.2. Anàlisi de Components Principals (PCA)
3. SVMs en problemes de classificació
3.1. Exemples
3.2. Aprenentatge actiu amb SVMs
4. SVRs en problemes de regressió
5. Models basats en arbre de decisió per a problemes de classificació i regressió
6. Models basats en ensembles darbres
6.1. Bagging
6.2. Random Forest

VOLUM DE TREBALL

ACTIVITAT	Hores	% Presencial
Classes de teoria	32,00	100
Pràctiques en laboratori	20,00	100
Pràctiques en aula	8,00	100
Assistència a esdeveniments i activitats externes	2,00	0
Elaboració de treballs en grup	6,00	0
Elaboració de treballs individuals	5,00	0
Estudi i treball autònom	35,00	0
Lectures de material complementari	4,00	0
Preparació d'activitats d'avaluació	20,00	0
Preparació de classes de teoria	4,00	0
Preparació de classes pràctiques i de problemes	4,00	0

Resolució de casos pràctics	7,00	0
Resolució de qüestionaris on-line	3,00	0
TOTAL	**150,00**	

METODOLOGIA DOCENT

Les metodologies docents utilitzades en esta assignatura són:

MD1 - Activitats teòriques (CG03, CB4, CT03, CT05, CE03, CE07, CE13): Desenvolupament expositiu de la matèria amb la participació de l'estudiant en la resolució de qüestions puntuals. Realització de qüestionaris individuals d'avaluació.

MD2 - Activitats pràctiques (CG02, CB3, CT05, CE03, CE07, CE13): Aprenentatge mitjançant resolució de problemes, exercicis i casos d'estudi pels quals s'adquirixen competències sobre els diferents aspectes de la matèria.

MD4 - Treballs en laboratori i/o aula d'ordinador (CG02, CG03, CB3, CB4, CT03, CT05, CE03, CE07, CE13): Aprenentatge mitjançant la realització d'activitats desenvolupades de manera individual o en grups reduïts i dutes a terme en laboratoris i/o aules d'ordinador.

A continuació, es descriu amb més detall el mètode d'ensenyament-aprenentatge a utilitzar en l'assignatura. La metodologia docent tindrà dos enfocs diferents, un per a les classes teòriques i de problemes i un altre per a les classes pràctiques de laboratori. Es farà servir l'Aula Virtual i les seues utilitats, especialment en el que pertoca a la disposició de material, a les avaluacions automàtiques i a les classes remotes, si escau.

Respecte a les classes teòriques, l'aprenentatge es farà en els dos sentits, des del professor fins l'alumne, i des de l'alumne fins el professor. En la part que naix del docent, hi haurà dues fonts de generació de coneixement. Per una banda, la classe magistral en la que el professor introduirà els conceptes nous que van apareixent tot relacionant-los amb els coneixements previs dels estudiants per tal de facilitar la seua comprensió. Per una altra banda, l'alumne disposarà amb anterioritat a la seua explicació en classe de material per a poder preparar mínimament les classes teòriques i així agilitzar el procés d'ensenyament-aprenentatge; este mateix material contindrà també la informació necessària per a que l'estudiant puga complementar i repassar la informació rebuda en classe.

En el procés que fluix des de l'estudiant hi haurà també dues aproximacions que es corresponen amb les fonts de generació de coneixement anteriorment comentades. La classe magistral del professor es vorà reforçada amb la resolució d'exercicis pràctics i problemes de creixent complexitat a mesura que avance la unitat temàtica; si bé el docent realitzarà algun exercici com a exemple, la majoria d'estos problemes hauran de ser resolts pròpiament pels estudiants per a garantir una comprensió total dels continguts de l'assignatura, que malgrat tindre una forta component teòrica, a la fi el seu punt fort és l'aplicació pràctica a diferents problemes i estos exercicis i problemes fan que l'estudiant puga esbrinar les particularitats dels diferents algorismes en la seua aplicació a problemes de diversa índole. A banda, l'estudiant realitzarà treballs amb una component de recerca individual en continguts que podrien considerar-se com una versió sofisticada dels descrits en l'assignatura; en particular, es planteja que l'estudiant puga realitzar, presentar i defensar treballs sobre mètodes actuals d'aprenentatge automàtic, que al mateix temps, faran que millore el seus fonaments dels continguts més bàsics que es descriuen en detall en l'assignatura. Es contempla que alguns d'estos treballs puguen ser voluntaris i es realitzen de manera individual o per parelles.

Respecte a les classes pràctiques, poden distingir-se tres metodologies docents. Primerament, i amb anterioritat a la realització de les pràctiques, els estudiants hauran de preparar-se de manera autònoma els exercicis pràctics a realitzar, consultat els dubtes que els puguen sorgir amb els professors preferentment abans de la realització de la pràctica. Este aspecte es vorà recolzat amb la realització de un qüestionari curt i senzill al començament de la pràctica per a comprovar que la preparació s'ha realitzat correctament.

La segona metodologia docent que apareix en les pràctiques és el propi treball a realitzar durant la sessió de pràctiques, que bàsicament consistirà en la programació en Python que permeta trobar la solució correcta als problemes plantejats en la pràctica. Este treball es realitzarà de manera individual o per parelles i tindrà en tot moment la supervisió del professor; en primer lloc, perquè els exercicis seran prèviament explicats i en segon perquè l'alumne podrà en tot moment consultar els seus dubtes per tal de poder avançar correctament en la realització de la pràctica.

La tercera aproximació docent en pràctiques torna a donar el protagonisme a l'alumne, que a la finalització de la pràctica haurà de ser capaç de fer una discussió crítica dels resultats assolits i contestar correctament a les preguntes formulades pel professor i a realitzar els exercicis plantejats en la sessió. Depenent de les circumstàncies (ocupació de l'aula, classe presencial o no, etc.) esta discussió podria realitzar-se de manera automàtica mitjançant les eines disponibles a l'Aula Virtual.

AVALUACIÓ

La qualificació final de l'assignatura s'obtindrà com a resultat de la mitjana pesada entre les parts de teoria i de prà
D'acord amb els crèdits assignats a cada part, la teoria tindrà una representació de 2/3 en la nota final i la pràctica e

La nota de teoria corresponent a la primera convocatòria eixirà com a resultat de:

- SE1 (60%; CG02, CG03, CB3, CB4, CT05, CE03, CE07, CE13): Proves objectives, consistents en un o més qüestions teòriques, problemes sintètics i problemes pràctics reals. Per tal de superar l'assignatura, se exigirà una q mínima de 4 (sobre 10) en esta part.

- SE2 (30%; CG02, CG03, CB3, CB4, CT03, CT05, CE03, CE07, CE13): Treballs, memòries i exposicions or

- SE3 (10%; CG02, CB4, CT03, CE03, CE07, CE13): Avaluació contínua de cada alumne, basada en la partici d'implicació de l'alumne en el procés d'ensenyament-aprenentatge, tenint en compte l'assistència regular a les acti presencials previstes i la resolució de qüestions i problemes proposats periòdicament.

Respecte a la qualificació de pràctiques, el 40% de la nota correspondrà amb SE2 (CG02, CG03, CB3, CB4, CT03, CE03, CE07, CE13) i el 60% amb la qualificació obtinguda en la pràctica final, que tindrà lloc en la darrera sessió (SE1; CG02, CG03, CB3, CB4, CT05, CE03, CE07, CE13). La pràctica final serà una prova objectiva que s'avalua individualment i que consistirà en la realització de diferents exercicis relacionats amb una o diverses pràctiques ant de superar l'assignatura, s'exigirà una qualificació mínima de 4 (sobre 10) en la pràctica final. Del 40% correspon l'avaluació contínua, el 70% correspondrà amb la realització dels exercicis proposats en la sessió de pràctiques, que avaluats pel professor a la finalització de la pràctica. El 30% restant provindrà de la preparació prèvia a la sessió de que s'avaluarà ràpidament al començament de cada sessió de pràctiques. Les pràctiques poden realitzar-se de mane o per parelles, amb l'excepció de la pràctica final, que obligatòriament serà individual. A més, el professor pot opta de manera individual les sessions regulars de pràctiques encara que estes s'hagen desenvolupat per grups de dos es

La segona convocatòria s'avaluarà com la primera amb l'excepció de què en la part de teoria, SE1 tindrà un peso d del 0%; en la part de pràctiques el 100% correspondrà a SE1, i serà necessari obtindré un mínim de 4 (sobre 10) pe amitjanar amb la part de teoria.

En qualsevol cas, el sistema d'avaluació es regirà pel que s'estableix en el Reglament d'Avaluació i Qualificació d Universitat de València per a Graus i Màsters (https://webges.uv.es/uvTaeWeb/MuestraInformacionEdictoPublicoFrontAction.do?accion=inicio&idEdictoSelecc

REFERÈNCIES

Bàsiques

- E. Alpayidin, F. Bach (2014). Introduction to Machine Learning, Third Edition, The MIT Press (disponible com a eBook per a la Universitat de València)
- S. Theodoridis (2015). Machine Learning: a Bayesian and Optimization perspective, Elsevier (disponible com a eBook per a la Universitat de València)
- D. Haroon (2017). Python Machine Learning Case Studies: Five Case Studies for the Data Scientist, Apress (disponible com a eBook per a la Universitat de València)

Complementàries

- C. M. Bishop (2016). Pattern Recognition and Machine Learning, Springer
- K. P. Murphy (2020). Machine Learning: a probabilistic perspective, Second Edition, The MIT Press
- R. O. Duda, P. E. Hart, D. G. Stark (2016). Pattern classification, Third Edition, John Wiley & Sons Inc.
- T. Hastie, R. Tibshirani, J. Friedman (2011) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer (Series in Statistics)

- S. Raschka, V. Mirjalili (2019). Python Machine Learning. Packt Publishing
- David V. (2017). Machine Learning with Python: The Basics. CreateSpace Independent Publishing Platform

ADDENDA COVID-19

Aquesta addenda només s'activarà si la situació sanitària ho requereix i previ acord del Consell de Govern

La metodologia docent de l'assignatura seguirà el Model Docent aprovat per la Comissió Acadèmica de Títol de Ciència de Dades (https://go.uv.es/cienciadatos/ModelDocentGCD1Q). En cas que es produïsca un tancament de les instal·lacions per causes sanitàries que afecte totalment o parcialment les classes de l'assignatura, aquestes seran substituïdes per sessions no presencials seguint els horaris establits. Si el tancament afectara alguna prova d'avaluació presencial de l'assignatura, aquesta serà substituïda per una prova de naturalesa similar que es realitzarà en modalitat virtual a través de les eines informàtiques suportades per la Universitat de València. Els percentatges de cada prova d'avaluació romandran invariables, segons el que s'estableix per aquesta guia.