FITXA IDENTIFICATIVA

Dades de l'Assignatura

Codi	36427
Nom	Agrupament i varietats
Cicle	Grau
Crèdits ECTS	6.0
Curs acadèmic	2021 - 2022

Titulació/titulacions

Titulació	Centre	Curs	Període
1400 - Grau Eng.Informàtica	Escola Tècnica Superior d'Enginyeria	4	Primer quadrimestre
1406 - Grau en Ciència de Dades	Escola Tècnica Superior d'Enginyeria	3	Primer quadrimestre

Matèries

Titulació	Matèria	Caràcter
1400 - Grau Eng.Informàtica	16 - Matèria Optativa	Optativa
1406 - Grau en Ciència de Dades	9 - Aprenentatge automàtic i mineria de dades	Obligatòria

Coordinació

Nom	Departament
MARTINEZ GIL, FRANCISCO	240 - Informàtica

RESUM

L'assignatura 'Agrupaments i Varietats' és el complement natural de l'assignatura 'Aprenentatge Màquina' que s'impartix també en el primer quadrimestre del tercer curs. Es revisen les tècniques més importants per a trobar estructures i patrons en conjunts de dades que no estan etiquetats. L'assignatura es dividix en dos parts. En la primera es tracten les tècniques més esteses d'aprenentatge no supervisat, conegudes també com a tècniques d'agrupament (clustering) . Es revisen l'agrupament jeràrquic, l'agrupament de per particions en les seues variants 'Hard' (K-Means) i 'Soft' (Fuzzy-C-Means) , i models d'agrupaments no compactes com l'agrupament basat en densitat (DBSCAN) i el basat en grafos (agrupament espectral) En la segona part s'amplien els coneixements sobre reducció de la dimensionalidad en dades ja introduïts en l'assignatura de segon curs 'Models Lineals'. Mentres que en esta assignatura s'introduïx PCA com el model lineal més important per a la reducció de dimensionaliad, en 'Agrupaments i Varietats' estendrem esta problemàtica a models no lineals, esdecir, a organitzacions espacials de dades que no poden ser

modelades per mitjà de hiperplans. Comença esta segona part amb un model senzill, els Mapes Autoorganitzatius. Posteriorment es revisaran algunes tècniques dins del grup conegut com 'Manifold Learning', específicament ISOMAP i Locally Esbossar Embedding. Finalment es revisarà la tècnica per a la visualització de dades coonocida com a t-SNE, la qual cosa ens connecta amb l'assignatura 'Visualització de dades' de segon curs. A banda de les diferents tècniques presentades, l'alumne adquirirà /revisará coneixements sobre conceptes essencials en aprenentatge màquina, com són els conceptes de similaritat, mètriques , espai de característiques, partició, o el problema de l'explosió combinatòria de la dimensionalidad.

A causa dels nombrosos temes que tracta d'abordar esta assignatura quadrimestral, la presentació dels temes serà necessàriament superficial, presentant les idees bàsiques de cada tècnica, els tipus de problemes que pretén resoldre i exposant els algoritmes bàsics.

Els coneixements impartits en esta assignatura són la base per a comprendre tècniques d'aprenentatge presentades en les assignatures 'Procesado del Llenguatge Natural' i les diferents optatives d'anàlisi de dades geogràfiques, d'àudio i veu, de salut i Web i xarxes socials.

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

Es recomana haver superat les assignatures: Algebra, Tractament de dades i Models lineals

1400 - Grau Eng.Informàtica

- SI3 - Capacitat per participar activament en l'especificació, el disseny, la implementació i el manteniment dels sistemes d'informació i comunicació.
- C1 - Capacitat per conèixer els fonaments, els paradigmes i les tècniques propis dels sistemes intel·ligents, i analitzar, dissenyar i construir sistemes, serveis i aplicacions informàtiques que utilitzen aquestes tècniques en qualsevol àmbit d'aplicació.
- C2 - Capacitat per adquirir, obtenir, formalitzar i representar el coneixement humà en una forma computable per a la resolució de problemes mitjançant un sistema informàtic en qualsevol àmbit d'aplicació, particularment els relacionats amb aspectes de computació, percepció i actuació en ambients o entorns intel·ligents.
- C3 - Capacitat per conèixer i desenvolupar tècniques d'aprenentatge computacional i dissenyar i implementar aplicacions i sistemes que les utilitzen, incloent-hi les dedicades a extracció automàtica d'informació i de coneixement a partir de grans volums de dades.

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.
- (CE06) Capacitat per representar i visualitzar conjunts de dades per a l'extracció de coneixement.
- (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.

Conéixer el concepte de clustering i algoritmes principals (Jeràrquics, de particions) (CB3, CB4, CG2, CG3,CT03, CT05, CE03, CE13)

Conéixer els principals algoritmes de varietats (manifolds) que existixen. (CB3, CB4, CG2, CG3,CT03, CT05, CE03, CE13)

Conéixer els problemes dels algoritmes plantejats i possibles solucions. (CB3, CB4, CG2, CG3,CT03, CT05, CE03, CE13)

DESCRIPCIÓ DE CONTINGUTS

1. Introducció

1.1 Aprenentatge no supervisat
1.2 Noció de similaridad. Concepte de mètrica. Tipus de mètriques
1.3Conceptos bàsics: espai de característiques, vector de característiques, partició, matriu de proximitat.

1.4 Concepte d'agrupament. Taxonomia

2. Agrupament jeràrquic

2.1 Idees bàsiques. Agrupaments aglomerativo i divisiu.
2.2 Tipus de Linkage
2.3 Algoritmes bàsics
2.4 Dendogramas. Interpretació. Coeficient de correlació

3. Agrupament per particions

3.1 Idees bàsiques. Agrupaments Soft i Hard
3.2 Agrupament 'hard'. Algoritme de K-Means
3.3 Problemes associats. Inicialització. Elecció del número de clusters.
3.4 Agrupament 'soft'. Algoritme Fuzzy-C-Means
3.5 Problemes associats.

4. Agrupament basat en grafos

4.1 Agrupament basat en grafos. Agrupament espectral
4.2 Idees bàsiques. Representació de dades en grafos.
4.3 Laplaciana del grafo. Descomposició espectral.
4.4 Algoritmes . Implementacions amb llibreries.

5. Agrupament basat en densitat

5.1 Idees bàsiques. Conceptes de densitat
5.2 Algoritme DBSCAN
5.3 Implementacions en llibreries i exemples

6. Evaluació i validació d'agrupaments

6.1 Tècniques per a l'avaluació d'agrupaments
6.1.1 Puresa. Índex aleatori (Rand Index) . Mesura F
6.2 Tècniques per a la validació de clusters
6.2.1 Regularitat. Compacitat i aïllament. Estadístics de Hopkins. SSE

7. Introducció ales tècniques de reducció de dimensionalitat

7.1 El problema de la maledicció de la dimensionalidad
7.2 Concepte de Manifold i problemàtica dels agrupaments no lineals.
7.3 Concepte de dimensionalidad intrínseca

8. Mapes auto-organitzatius

8.1 Concepte d'aprenentatge competitiu. Concepte de SOM.
8.2 Algoritme. Punts forts i dèbils de l'aproximació. Implementacions en llibreries.
8.3 Exemples

9. Aprenentatje en varietats

9.1 Introducció. Idees bàsiques sobre les tècniques de Manifold Learning
9.2 Algoritme ISOMAP. Punts forts i dèbils. Paràmetres.
9.3 Algoritme Locally Linear Embedding (LLE) . Punts forts i dèbils. Paràmetres.
9.4 Examples

10. Reducció de dimensionalidad per a visualització

10.1 Introducció al problema.
10.2 Algoritme t-SNE. Punts forts i dèbils. Comportament
10.3 Implementacions en llibreries.
10.4 Exemples.

VOLUM DE TREBALL

ACTIVITAT	Hores	% Presencial
Classes de teoria	32,00	100
Pràctiques en laboratori	20,00	100
Pràctiques en aula	8,00	100
Elaboració de treballs en grup	10,00	0
Elaboració de treballs individuals	10,00	0
Estudi i treball autònom	20,00	0
Lectures de material complementari	10,00	0
Preparació d'activitats d'avaluació	10,00	0
Preparació de classes de teoria	10,00	0
Preparació de classes pràctiques i de problemes	10,00	0
Resolució de casos pràctics	5,00	0
Resolució de qüestionaris on-line	5,00	0

TOTAL	**150,00**

METODOLOGIA DOCENT

En les activitats teòriques es desenrotllaran els temes exposant-los el professor utilitzant mitjans audiovisuals, proporcionant una visió global i integradora dels continguts. Es fomentarà la participació de l'estudiant en classe a través de preguntes durant l'exposició i la realització de qüestions simples per a fixar els conceptes que es presenten. (CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)

En les classe de resolució de problemes es resoldran preferentment per estudiants els problemes plantejats amb una setmana d'antelació , de tal manera que l'alumno/a tinga temps suficient per a treballar-ho a casa. S'incentivarà la discussió dels problemes en l'aula. (CB3, CB4, CG2, CG3, CT03, CT05, CE03, CE13)

En els laboratoris es plantejaran tasques de major complexitat i envergadura que les propostes en les activitats de l'aula. Es fomentarà el treball en grup, per parelles, de les pràctiques propostes. També l'alumno/a es familiaritzarà amb les llibreries de càlcul científic de Python així com amb ferramentes de presentació de codi com és Jupyter Notebook. (CB5, CT03, CT05, CE03, CE13)

AVALUACIÓ

Primera convocatòria:

Es realitzarà almenys una prova objectiva parcial durant el quadrimestre d'impartició del curso.

Es realitzarà un examen final en la primera convocatoria.

El valor de les proves parcials podrà arribar fins al 50% de la nota de teoria (SE1) .

La resta del percentatge se li assignarà a la prova final.

El percentatge sobre la nota final d'esta part (SE1) serà 45%.. (CB5, CT03, CT05, CE03, CE13)

El valor de les pràctiques de laboratori (SE2) representarà un 35% de la nota total de l'assignatura (CB5, CT03, CT05, CE03, CE13)

El valor de l'avaluació contínua (SE3) representarà el 20% de la nota total (CB5, CT03, CT05, CE03, CE13)

És necessari traure una nota mínima de 4.5 en cada una de les parts anteriors (SE1, SE2, SE3) per a poder superar l'assignatura.

Segona convocatòria

La nota d'avaluació contínua (SE3), per ser activitats presencials ,es considera no recuperable en la segona convocatòria. Encara que no s'imposa la restricció de la nota mínima de la primera convocatòria.

Per a la nota de pràctiques de laboratori , es plantejará la defensa d'una pràctica o pràctiques resum dels continguts o bè es realitzarà un exam pràctic al terminar l'exam de teoria que suposarà el 100% de la nota (SE2) en segona convocatoria.

La restricció de nota mínima es manté per a la teoria.

La nota de teoria serà únicament la nota de l'examen final de segona convocatòria (sense considerar els parcials realitzats en el curs).

La nota final s'obtindrà com 50% SE1, 30% SE2, 20% SE3

En qualsevol cas, el sistema d’avaluació es regirà pel que s’estableix en el Reglament d’Avaluació i Qualificació de la Universitat de València per a Graus i Màsters (https://webges.uv.es/uvTaeWeb/MuestraInformacionEdictoPublicoFrontAction.do?accion=inicio&idEdictoSeleccionado=5639)

REFERÈNCIES

Bàsiques

- Data Mining. Concepts and Techniques. J. Han, M. Kamber , J. Pei. Morgan-Kauffman. 3ª edición. 2012
- Introduction to Data Mining. Pang-Ning Tan, M. Steinbach, A. Karpatne, V. Kumar. Pearson. 2ª edición. 2018
- An introduction to Statistical Learning . Gureth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. Springer (2013)
- Pattern Recognition. Sergios Theodoridis. AP (2009) Hay versión electrónica

Complementàries

- Scikit-Learn Users Guide (Hay versión electrónica)
- Python Data science handbook. Jacob Vanderplas. O'Reilly. (2016)

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.