Guia docent
340458 - MIDA-I7P23 - Mineria de Dades

Última modificació: 03/04/2024

Unitat responsable: Escola Politècnica Superior d'Enginyeria de Vilanova i la Geltrú
Unitat que imparteix: 723 - CS - Departament de Ciències de la Computació.

Titulació: GRAU EN ENGINYERIA INFORMÀTICA (Pla 2018). (Assignatura optativa).

Curs: 2024 **Crèdits ECTS:** 6.0 **Idiomes:** Català

PROFESSORAT

Professorat responsable: MARIO MARTÍN MUÑOZ

Altres: MARIO MARTÍN MUÑOZ

CAPACITATS PRÈVIES

És necessari que l'estudiant conegui tècniques avançades de programació (no és necessari cap llenguatge de programació en concret). És convenient que es tinguin coneixements bàsics de probabilitat i d'estadística.

REQUISITS

Dessitjable que hagi cursat o cursi PPROP.

COMPETÈNCIES DE LA TITULACIÓ A LES QUALS CONTRIBUEIX L'ASSIGNATURA

Específiques:
1. CECO7. Capacitat per a conèixer i desenvolupar tècniques d'aprenentatge computacional i dissenyar i implementar aplicacions i sistemes que les utilitzin, incloent les dedicades a extracció automàtica d'informació i coneixement a partir de grans volums de dades.
2. CEIS4. Capacitat d'identificar i analitzar problemes i dissenyar, desenvolupar, implementar, verificar i documentar solucions programari sobre la base d'un coneixement adequat de les teories, models i tècniques actuals.

Transversals:
3. EMPRENEDORIA I INNOVACIÓ: Conèixer i comprendre l'organització d'una empresa i les ciències que regeixen la seva activitat; capacitat per comprendre les regles laborals i les relacions entre la planificació, les estratègies industrials i comercials, la qualitat i el benefici.

METODOLOGIES DOCENTS

L'assignatura té classes de teoria on s'aprendran els conceptes i les tècniques principals de mineria de dades. En aquestes classes s'adquiriran els coneixements científics i tecnològics adequats per tractar problemes reals de mineria de dades.
També hi haurà classes de laboratori on l'estudiant aprendrà l'ús de eines modernes per la mineria de dades en casos reals. Aquestes eines s'utilitzaran en casos reals per mineria de dades en casos de corpus de texts i de xarxes socials.

OBJECTIUS D'APRENENTATGE DE L'ASSIGNATURA

La mineria de dades tracta sobre l'anàlisi de grans volums de dades de manera que en resulti coneixement útil i interpretable. Aquest tipus de tecnologia és especialment important quan en tots els dominis comencem a estar inundats de informació en estat brut que ens sobrepassa i no ens permet assimilar-la. Pensem per exemple en empreses que tracten volums de milions de clients com Amazon o empreses que han de tractar milions de pagines webs com google. Aquetes tècniques permeten treballar amb conjunt de dades d'aquesta magnitud, de manera que puguem per exemple ordenar o catalogar pagines web segons el seu contingut o descobrir els tipus de lectors de llibres segons les preferències de compres que fan i per tant fer-hi recomanacions, etc.

Els objectius de l'assignatura són doncs conèixer els fonaments científics i tecnològics per fer aquests anàlisis així com l'adquisició de les habilitats tècniques i coneixement pràctics necessaris per fer l'anàlisi d'un cas real amb èxit.

HORES TOTALS DE DEDICACIÓ DE L'ESTUDIANTAT

Tipus	Hores	Percentatge
Hores grup petit	30,0	20.00
Hores aprenentatge autònom	90,0	60.00
Hores grup gran	30,0	20.00

Dedicació total: 150 h

CONTINGUTS

1. Introducció a la Mineria de Dades.

Descripció:
Característiques comunes dels problemes de mineria de dades. Descripció de casos reals.

Objectius específics:
Identificar en quins casos és útil aplicar les tècniques de mineria de dades.

Activitats vinculades:
Primer parcial
Examen final

2. Caracterització i preparació de dades

Descripció:
Descripció del tipus de dades que podem trobar en un cas real. Concepte d'espai d'estats. Descripció complexa d'objectes. Dades i soroll. Neteja de les dades. Selecció de variables. Reducció de la dimensió de les dades. Emmagatzematge de dades eficient.

Objectius específics:
Conèixer com tractar, transformar, netejar i simplificar les dades per preparar-les pels algorismes de data mining.

Activitats vinculades:
Primer parcial
Examen final
Primera activitat
Segona activitat
Pràctica

3. Introducció a la classificació

Descripció:
Classificació com a generador de coneixement. Classificació supervisada versus classificació no supervisada. Classificadors com capses negres. Tècniques d'avaluació de les classificacions.

Objectius específics:
Coneixer els conceptes basics de la classificació com a generador de coneixement.
Coneixer quan aplicar mètodes supervisats o no supervisats
Coneixer com avaluar els resultats d'un classificador (mesures i tècniques).

Activitats vinculades:
Primer parcial
Examen final
Activitat 1
Activitat 2
Pràctica

4. Naive Bayes i Veins propers

Descripció:
Mètodes basics de classificació supervisada. Mètode de Naive Bayes. Mètode dels Veïns Propers. Cas pràctic detecció de SPAM.

Objectius específics:
Aprendre dues tècniques bàsiques molt efectives en la classificació supervisada
Identificar els punts forts i els febles de les dues tècniques per identificar quan usar-les i quan no.

Activitats vinculades:
Primer parcial
Examen final
Activitat 1
Pràctica

5. Arbres de decisió.

Descripció:
Concepte d'arbres de decisió. Generació d'arbres petits i concepte de navalla d'Ockam. Mesures d'informació. Atribut maximament discriminant. Decision Stumps. Obtenció de regles de decisió. Arbres de regressió.

Objectius específics:
Conèixer la tècnica dels arbres de decisió, com es generen i com obtenir-ne sense patir de sobregeneralització.
Conèixer les limitacions de la tècnica i reconèixer de forma efectiva quan fer-la servir.

Activitats vinculades:
Primer parcial
Examen final
Primera activitat
Pràctica

6. Xarxes Neuronals

Descripció:
Models de xarxes neuronals simples: Perceptró. Algorísme d'aprenentatge de la delta rule. Model de xarxa neural feedforward. Cas pràctic de reconeixement de dígits. Aprenentatge per backpropagation. Xarxes neurals profundes.

Objectius específics:
Coneixer els fonaments científics de les xarxes neuronals.
Coneixer les virtuts i els problemes de les xarxes neuronals.
Aprendre a fer servir xarxes neuronals per un problema concret.
Aprendre a evaluar en quines situacions és adequat fer servir una xarxa neuronal i quan no.

Activitats vinculades:
Primer parcial
Examen final
Activitat 1
Practica final

7. Maquines de support vectorial y boosting

Descripció:
Algorísmes avançats d'aprenentatge supervisat. Maquines de suport vectorial: Model bàsic i model amb slacks. Concepte de kernel i adequació a les SVM. Transducció.
Fonaments dels mètodes d'agregació de classificadors diferents. Algorismes basats en Mostrejos. Algorísme de Boosting.

Objectius específics:
Coneixer els fonaments científics dels métodes avançats de classificació de les màquines de suport vectorial.
Coneixer com aplicar el mètode de les SVM
Coneixer els fonaments científics del consens i la agregació de classificadors diferents.
Aprendr a fer servir les tècniques d'agregació i boosting per casos concrets quan sigui adequat.

Activitats vinculades:
Segon parcial
Examen final
Practica final

8. Clustering

Descripció:
Classificació no supervisada: Clustering. Agrupació d'observacions o objectes en classes segons la seva similitud. Algorísme de k-means, classificació jeràrquica i basats en grafs.

Objectius específics:
Coneixer la potencia de la classificació no supervisada per obtenir coneixement.
Coneixer diferents métodes de clustering i aprendre a fer-los servir segons el cas.

Activitats vinculades:
Segon Parcial
Examen Final
Activitat 2

9. Regles d'associació

Descripció:
Descobriment de regularitats en conjunts enormes de dades mitjançant les regles d'associació.

Objectius específics:
Aprende a fer servir els algorísme de regles d'associació

Activitats vinculades:
Segon parcial
Examen final
Activitat 2

ACTIVITATS

PRIMER PARCIAL

Descripció:
Examen teòric de la primera part de l'assignatura, fins al tema 5 d'arbres de decisió inclòs.

Objectius específics:
Testar els coneixements científics de l'estudiant en la primera part de l'assignatura.

Material:
Transparencies de l'assignatura al campus digital

SEGON PARCIAL

Descripció:
Examen de la segona part de l'assignatura. Des de Xarxes Neuronals fins al final.

Objectius específics:
Testar els coneixements científics de l'estudiant en la segona part de l'assignatura.

Material:
Transparències de l'assignatura.

ACTIVITAT 1

Descripció:
Donat un conjunt de dades reals, tractar les dades d'entrada iobtenir un métode discrimant de la classe positiva.

ACTIVITAT 2

Descripció:
Obtenció de una classificació no supervisada d'un conjunt de dades i extreure'n conclusions.

PRÀCTICA

Descripció:
Fer tot el processament d'un conjunt de dades complexe i real.

SISTEMA DE QUALIFICACIÓ

La nota de l'assignatura s'obtindrà per una part avaluant els coneixements teòrics assolits i per altre l'aplicació d'aquests coneixements a casos reals. La part teòrica s'avaluarà com el màxim de la mitjana de dos exàmens parcials o un de final.

Nota Teoria =0,5 Prova1 + 0,5 Prova2

La nota dels laboratoris es calcularà a partir de dues activitats i d'una pràctica d'un cas real.

Nota Laboratori = 0,2 Activitat1 + 0,2 Activitat2 + 0.6 Pràctica

La nota final de l'assignatura serà la [0.7*Nota Laboratori+0.3*Nota teoria]

NORMES PER A LA REALITZACIÓ DE LES PROVES.

Les activitats es fan en grups de dues persones. La pràctica en grups de fins a 4 persones.

BIBLIOGRAFIA

Bàsica:

- Kantardzic, Mehmed. Data mining : concepts, models, methods, and algorithms [en línia]. 2a ed. New Jersey: IEEE Press, 2011 [Consulta: 15/02/2024]. Disponible a: https://onlinelibrary-wiley-com.recursos.biblioteca.upc.edu/doi/book/10.1002/9781118029145. ISBN 9781118029145.
- Han, Jiawei ; Kamber, Micheline ; Pei, Jian. Data mining : concepts and techniques [en línia]. 3rd ed. Burlington: Morgan Kaufmann, 2012 [Consulta: 14/02/2024]. Disponible a: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=729031. ISBN 9780123814791.
- Bramer, Max. Principles of data mining. 2nd ed. London: Springer, cop. 2013. ISBN 9781447148838.
- Flach, Peter A. Machine learning : the art and science of algorithms that make sense of data. Cambridge [etc.]: Cambridge University Press, 2012. ISBN 9781107096394.
- Witten, Ian H.; Frank, Eibe; Hall, Mark A. Data mining : practical machine learning tools and techniques [en línia]. 3rd ed. Burlington, MA: Morgan Kaufmann Publishers/Elsevier, 2011 [Consulta: 20/02/2024]. Disponible a: https://www-sciencedirect-com.recursos.biblioteca.upc.edu/book/9780128042915/data-mining. ISBN 9780080890364.

Complementària:

- The Top ten algorithms in data mining [en línia]. Boca Raton: CRC Press, 2009 [Consulta: 20/02/2024]. Disponible a: https://www-taylorfrancis-com.recursos.biblioteca.upc.edu/books/edit/10.1201/9781420089653/top-ten-algorithms-data-mining-xindong-wu-vipin-kumar. ISBN 9781420089646.
- Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. The elements of statistical learning : data mining, inference, and prediction [en línia]. 2nd ed. New York [etc.]: Springer, 2009 [Consulta: 02/05/2022]. Disponible a: https://link.springer.com/book/10.1007/978-0-387-84858-7. ISBN 0387952845.
- Segaran, Toby. Programming collective intelligence : building smart web 2.0 applications [en línia]. Beijing ; Sebastapol [CA]: O'Reilly, 2007 [Consulta: 18/03/2024]. Disponible a: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=443469. ISBN 9780596529321.
- Conway, Drew; White, John Myles. Machine learning for hackers [en línia]. Sebastopol, CA: O'Reilly, 2012 [Consulta: 14/02/2024]. Disponible a: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=862166. ISBN 9781449303716.
- Harrington, Peter. Machine learning in action [en línia]. Shelter Island, N.Y: Manning Publications Co., 2012 [Consulta: 16/11/2022]. Disponible a: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=nlebk&AN=2948840&site=ehost-live&ebv=EK&ppid=Page-__-1. ISBN 9781617290183.