

IRAKASKUNTZA-GIDA		2011/12	
Ikastegia	226 - Informatika Fakultatea	Zikl.	Zehaztugabea
Plana	GINFOR20 - Informatikaren Ingeniaritzako Gradua	Ikastaroa	3. maila
IRAKASGAIA			
Datu Meatzaritza		ECTS kredituak:	6
GAITASUNAK / AZALPENA / HELBURUAK			
Irakasgai hau ingelesezko Data-Mining edo Machine-Learning arloetan oinarritzen da. Euskaraz, termino hauek Datu-Meatzaritza edo Ikasketa-Automatikoa moduan ezagutu ohi dira. http://en.wikipedia.org/wiki/Data_mining http://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos Adimen Artifizialean eta Estatistika klasikoan oinarritzen diren hainbat teknika biltzen ditu Datu Meatzaritzak, eta azken hamarkadan teknika horien erabilpenak gorakada handia bizi izan du, datu multzo handiak aztertu eta bertatik informazioa eta jakintza lortzeko asmoz. Gaur egungo gizartean datu multzo handiak hainbat erataraz sortzen ari dira, eta gure "arrastoa" edonon uzten dugu: erosketak supermerkatuetan, Internet-en bisitatzen ditugun orriak eta egiten ditugun erosketak, sare sozialak eta Web 2.0, gomendatze-sistemak, biologian eta bioinformatikan, banketxeekin eta telekomunikazio-konpainiekin dugun erlazioa... Horrela, datu multzo handi hauetatik informazioa eta eredu deskribatzaile eta iragarleak lortu nahian, Datu Meatzaritzak proposatzen dituen oinarritzko teknikak aurkeztuko dira kurtsoan, beti ere aplikazio errealak erakutsiz eta software desberdinen erabileran trebatuz. Horrela, Adimen Artifizialaren arlo zabalean, gaur egun Datu Meatzaritza bere atalik erabilienetarikoa bat da, eta mundu mailako ikerkuntza talde eta enpresa anitz daude bertan murgilduak. Teknika nagusien ulermen teorikoaz gain, software askeko tresnen erabileran trebatuko da ikaslea, teknika hauen aplikazioak burutu ditzan: hau, gai-zerrendako gai bakoitzerako egingo da. Disziplina honen (eta bere terminologiaren) ingelesez idatzitako testuak eta estekak ere ezinbestekoak izango dira irakasgaiaren zehar. Gaitasun espezifikokoak (titulazioarenak C1, C3, C4, C8, C9, eta konputazioa espezialitatekoak CC1, CC3, CC4, CC7) - datu meatzaritzako tekniken oinarri matematikoa ezagutu (C8, CC4) - Giza ezagutza modu konputagarrian adieraztea (CC5) - datu meatzaritzako problemak ebazti ahal izateko, erabili beharreko software-tresnak definitu eta ebaluatu (C4) - datu meatzaritzako proiektuak diseinatu eta garatzeko gaitasunean trebatu, eredu desberdinak ebaluatu eta haien artean proposena hautatu (C1, C3, CC1, CC3, CC7) - eskoletan ikasitako kontzeptuen eta egindako lanen aurkezpena (C9)			
GAI ZERRENDAA			
Lehen gaia: Datu Meatzaritzarako sarrera Sarrera gisa, aplikazio errealak aurkeztuko dira, beti ere irakasgai hau beste arlo batzurekin erlazionatuz, hala nola Adimen Artifiziala, Estatistika, eta abar. Bigarren gaia: Auzokide hurbilenaren sailkatzailea, K-NN (K-nearest neighbour) Sailkatzaile honen intuitibotasunak eta ulerterraztasunak egokia egiten du Datu Meatzaritzako lehen teknika gisa plazaratzeko. Bere oinarritzko funtzionamendua, hala nola bere parametro eta erabilerarik arruntenak plazaratuko dira. Gomendatze-sistemen oinarria dira. Hirugarren gaia: Gainbegiraturako sailkatzaileen ebaluazioa. Ezinbestekoa da, sailkatzaile eta teknika gehiago aurkeztu baino lehen, sailkatzaileen egokitasuna eta doitasuna neurtzeko dauden teknikak ezagutzea. Laugarren gaia: Sailkapen-zuhaitzak. Datu meatzaritzako teknika klasikoa dira, ulerterrazak eta intuitibotasun handikoak. Bere aldaerarik oinarritzkoenak plazaratuko dira. Ematen duten interpretaziorako aukeragatik, teknika hau oso aproposa da adituekin elkarlanean jarduteko. Medikuntzan eta banketxeetan oso erabiliak dira. Bosgarren gaia: Sailkatzaile Bayestarrak. Probabilitate teorian oinarritzen den teknika hau gaur egun ospe eta erabilera handikoa da. Teknika honek ere datuen interpretaziorako aukera handiak ematen ditu, eta oso egokia da adituekin elkarlanean erabiltzeko. Sailkatzaile honen aldaera nagusiak aztertuko dira, datubaseko aldagaien arteko erlazioen konplexutasun maila ezberdinetarako. Bioinformatikan eta industria munduan ospe handia lortu dute. Seigarren gaia: Sailkatzaileen konbinaketa. Zenbait problematan, sailkatzaile bakar bat erabili beharrean, iragarpen hobeak lortu daitezke hainbat sailkatzaile konbinatuz. Sailkatzaile ezberdinak konbinatzeko dauden erregela eta ideia nagusiak plazaratuko dira.			

Zazpigarren gaia: Aldagai-aukeraketarako teknikak.
Problema baterako sailkatzailea eraikitzen hasi aurretik, datubasea aztertzea ezinbestekoa gertatzen da, aldagai iragarleen artean egon daitezkeen erredundantziak eta aldagai iragarleek klasearekiko duten garrantzia maila neurtzeko. Hau burutzeko ikasketa automatikoan dauden aldagai aukeraketarako teknika nagusiak aurkeztuko dira. Interneteko eta bioinformatikako datuak aztertzeko behar-beharrezkoak dira teknika hauek.

Zortzigarren gaia: Sailkapen ez-gainbegiratua, clustering
Datubasearen ezaugarriak kontutan izanda, sailkapen ez-gainbegiratuak teknika interesgarriko multzo zabal bat ematen digu, antzekotasun handiko kasuak multzokatuz. Inkestak aztertzeko ezinbestekoa gertatzen da clustering aplikatzea.

Bederatzigarren gaia: Bilaketa heuristikoak eta algoritmo genetikoak
Adimen Artizialaren beste arlo klasiko eta ezinbesteko bati sarrera emango zaio, eta hauen artean ezagunena den algoritmo genetikoaren teknika aurkeztuko da. Logistika eta optimizazio problemak ebazteko tresna aproposa da.

IRAKASKUNTZA MOTAK

Eskola mota	M	S	GA	GL	GO	GCL	TA	TI	GCA
Ikasgelako eskola-orduak	40			20					
Ikaslearen ikasgelaz kanpoko jardueren ord.	60			30					

Legenda: M: Maistrala S: Mintegia GA: Gelako p. GL: Laborategiko p. GO: Ordenagailuko p.
GCL: P. klinikoak TA: Tailerra TI: Tailer Ind. GCA: Landa p.

Argibideak:
Irakasgai hau bi zatitan dago banatuta: %50 kontzeptu teorikoak eta ariketak egiteko eskolak izango dira, eta beste %50 laboratorio saioetan softwarea erabiliz ikasitako teknikak lantzeko. Bi zatietan erabiliko den irakaskuntza metodologia aktiboa izango da. Horrek esan nahi du, ikasleak modu aktiboan parte hartuko duela, bai klasean eta bai laboratorioan. Lan autonomoa bultzatuko da, klase presentzialetan plazaratutako lanak amaitzeko.

Irakasgaiaren amaieran, talde txikietan antolatu eta lanak egin eta aurkeztu beharko dira.

EBALUAZIOA

- Garatu beharreko azterketa idatzia
- Praktiak (ariketak, kasuak edo buruketak)
- Banakako lanak
- Taldeko lanak
- Lanen, irakurketen... aurkezpena

Argibideak:
Bi bide daude ikasgaia gainditzeko: ebaluazio globala (bukaeran), edo ebaluazio jarraitua. Ebaluazio jarraitua borondatezkoa da, eta ikaslearen parte hartze aktiboa eskatzen du; beraz, ikasleak eskoletara eta laborategietara etorri beharko du, haietan parte hartu, eta proposatutako jarduerak (ariketak, lanak, praktikak, azalpenak...) egin beharko ditu.

Ebaluazio jarraitua ikastaroaren hasieran aukeratu ahal izango da, eta behin betiko bihurtuko da adieraziko diren epeetan (ikastaroaren %60 - %80 igarota), ikasleak hala eskatuta eta irakasleak ikaslearen errendimendua egiaztatu ondoren.

Irakasgaiaren alde teorikoa bukaeran egingo den azterketa baten bidez ebaluatuko da, eta azken notaren %50 izango da.

Alde praktikoaren ebaluazioari dagokionez, kurtsoan zehar hainbat lan praktiko indibidual burutuko dira, gai teorikoen erritmoari egokituta. Lan motzak izango dira, eta entregatze-epeak gai teorikoak bukatu adinakoak. Kurtsoaren bukaera aldera, behin datu meatzaritzako teknika gehienak "jaso" eta gero, taldeko lan praktiko bat burutuko da datu meatzaritzako problema erreal baten inguruan, eta egindako lanaren emaitzak klasean aurkeztu beharko dira. Alde praktikoaren pisua azken notan %50ekoa izango da.

Ebaluazio globala: %100, azterketa teorikoa eta derrigorrezko praktiken entrega.

NAHITAEZ ERABILI BEHARREKO MATERIALAK

BIBLIOGRAFIA

Oinarrizko bibliografia
- I. H. Witten, E. Frank (2005). Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann. (Second edition)



- B. Sierra (ed.) (2006). Aprendizaje Automático: conceptos básicos y avanzados. Prentice Hall.
- E. Alpaydin (2004). Introduction to Machine Learning. MIT Press.
- T. Mitchell (1997). Machine Learning. McGraw Hill.
- J. Han, M. Kamber (2006). Data Mining: concepts and techniques. Morgan Kaufmann. (Second edition)
- J. Hernández, M. J. Ramírez, C. Ferri (2004). Introducción a la Minería de Datos. Pearson Prentice Hall.

Gehiago sakontzeko bibliografia

- O. Pourret, P. Naïm, B. Marcot (2008). Bayesian networks: a practical guide to applications. Wiley.
- L.I. Kuncheva (2004). Combining Pattern Classifiers. Wiley.
- H. Liu, H. Motoda (ed.) (2008). Computational Methods of Feature Selection. Chapman & Hall/CRC.
- C.M. Bishop (2006). Pattern Recognition and Machine Learning. Springer.
- S. Brunak, P. Baldi (2001). Bioinformatics: the machine learning approach. MIT Press. (Second edition).
- B. Liu (2006). Web Data Mining: exploring hyperlink, contents and usage data. Springer.

Aldizkariak

- Machine Learning Journal. Springer.
- Journal of Machine Learning Research. Electronic publication.
- Data Mining and Knowledge Discovery. Springer.
- Bioinformatics. Oxford University Press.

Interneteko helbide interesgarriak

<http://www.sc.ehu.es/ccwbayes/docencia/dm/main1112.htm>

Bertan materiala eta irakasgaiako interesgarri den informazioa topatuko duzue. Irakasgaiaren egunerokotasuna eramateko ezinbesteko lotura.

<http://www.cs.iastate.edu/~honavar/Courses/cs673/machine-learning-courses.html>

Ikasketa automatikoaren inguruko Unibertsitate mailako kurtsoak (materiala, ariketak, gardenkiak eta abar).

<http://videolectures.net/Top/Computer Science/Machine Learning/>

Videolectures.net: ikasketa automatikoaren inguruko on-line hitzaldiak, mundu mailako ikertzaile onenen eskutik.

<http://archive.ics.uci.edu/ml/>

Ikasketa automatikoko teknikak testatzeko datubaseen bilduma (University California Irvine).

<http://www.kdnuggets.com/>

Ikasketa automatikoaren eta datu meatzaritzaren erreferentzi webgunea (aplikazio industrialak, berriak, inkestak, eta abar).

<http://www.cs.waikato.ac.nz/ml/weka/>

WEKA ikasketa automatikorako software askea (Waikato University, New Zealand).