PRÀCTICA

Presentació

Pràctica sobre el desenvolupament i aplicació d'algoritmes d'aprenentatge automàtic.

Competències

Competències de grau

- Capacitat per utilitzar els fonaments matemàtics, estadístics i físics i comprendre els sistemes TIC.

- Capacitat per analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.

- Capacitat per conèixer les tecnologies de comunicacions actuals i emergents i saber-les aplicar, convenientment, per dissenyar i desenvolupar solucions basades en sistemes i tecnologies de la informació

- Capacitat per proposar i avaluar diferents alternatives tecnològiques i resoldre un problema concret

Competències específiques

- Capacitat per utilitzar la tecnologia d'aprenentatge automàtic més adequada per a un determinat problema.

- Capacitat per avaluar el rendiment dels diferents algorismes de resolució de problemes mitjançant tècniques de validació creuada.

Objectius

Seguint amb les dades utilitzades en les altres PACs, disposem d'un conjunt de dades de 100 casos d'anàlisis de la qualitat de l'esperma (fitxers fertility_diagnosis_train.csv i fertility_diagnosis_test.csv). Els resultats dels anàlisis poden ser Normal (N) o Alterat (O). Es vol estudiar la correlació de diferents variables (9 en total, veure la descripció en el fitxer attributes.txt) per veure si alguna d'aquestes es pot utilitzar per predir si l'esperma seria N o O.

Volem construir un model capaç de predir la classe de l'exemple, normal o alterat, utilitzant un arbre de decisió amb una lleugera modificació (veure exercici 3). L'objectiu és implementar l'algoritme i avaluar els resultats amb les dades esmentades.

La pràctica està formada per 3 exercicis amb diferent pes. La memòria de la pràctica ha de contenir les respostes a cadascun dels exercicis.

Descripció de la PRÀCTICA

Exercici 1. Bondat (20%)

Desenvolupeu un programa que, donades unes dades d'entrenament, sigui capaç de calcular la bondat de cada atribut. Esbrineu i discutiu quins són els millors atributs del fitxer fertility_diagnosis_train.csv.

Exercici 2. Weka (20%)

Utilitzant el Weka apliqueu **dos** algoritmes diferents d'aprenentatge supervisat (un d'ells que sigui un arbre de decisió) per predir la classe del fitxer de test. Doneu la precisió de la classificació obtinguda amb cada mètode. Discutiu els resultats.

Exercici 3. Implementació d'un arbre de decisió amb poda (60%)

Implementeu l'algorisme d'arbre de decisió per tal de predir si un cas del conjunt de test és N o O. Cal fer una petita modificació en l'algoritme original de l'arbre: realitzar una **poda**. Seguint l'analogia de la poda d'arbres real, la poda d'un arbre de decisió es realitza per a disminuir la mida del arbre (per evitar casos de *overfitting* de les dades) sense disminuir-ne la seva precisió. La manera més senzilla és realitzar la poda basada en el criteri de l'error de precisió: a partir de les fulles, cada node es substitueix amb la seva classe més popular. Si la precisió de la predicció després de la poda no es veu afectada, es manté la poda. S'aniran realitzant diferents podes successivament mentre es compleixi el criteri de l'error.

Cal que tingueu en compte:

1. Cal explicar l'algorisme implementat, explicant tots aquells detalls que considereu rellevants i les decisions de disseny preses. Feu especial esment en els passos de la implementació de l'arbre de decisió i la poda.
2. Una taula amb almenys la precisió, la matriu de confusió i el temps de càlcul de l'algorisme, comparant els resultats amb els obtinguts al Weka (exercici 2).
4. Un apèndix amb el llistat del codi font del vostre programa.
5. I, en general, una justificació de tot el que estigueu fent.

Heu de lliurar el programa que hagueu implementat per realitzar els exercicis 1 i 3 (es recomana que sigui en Java, però la podeu realitzar en Pascal, Delphi, Python, C, C++, Visual Basic o "similar"). La qualitat del codi (estructura, comentaris...) és un dels criteris importants de correcció.

Recursos

Bàsics

Per a realitzar aquesta PRÀCTICA disposeu d'uns fitxers adjunts (attributes.txt, fertility_Diagnosis_test.csv, fertility_Diagnosis_train.csv) on trobareu les dades corresponents.

Criteris de valoració

Els tres exercicis d'aquesta PRÀCTICA es valoraran amb 2, 2 i 6 punts respectivament.

Raoneu la resposta en tots els exercicis. Les respostes sense justificació no rebran puntuació.

Format i data de lliurament

Cal lliurar la PRÀCTICA en un únic fitxer comprimit (ex.zip) que contingui la memòria i els codi font dels diferents exercicis al registre d'activitats d'avaluació continuada.

El nom del fitxer ha de ser CognomsNom_AC_PAC2 amb l'extensió .zip (ZIP).

Data Límit: 31/05/2013 a les 24 hores.

Per a dubtes i aclariments sobre l'enunciat, adreceu-vos al consultor responsable de la vostra aula.

Nota: **Propietat intel·lectual**

Sovint és inevitable, en produir una obra multimèdia, fer ús de recursos creats per terceres persones. És per tant comprensible fer-ho en el marc d'una pràctica dels estudis d'Enginyeria Informàtica, sempre i això es documenti clarament i no suposi plagi en la pràctica.

Per tant, en presentar una pràctica que faci ús de recursos aliens, s'ha de presentar juntament amb ella un document en què es detallin tots ells, especificant el nom de cada recurs, el seu autor, el lloc on es va obtenir i el seu estatus legal: si l'obra està protegida pel copyright o s'acull a alguna altra llicència d'ús (Creative Commons, llicència GNU, GPL ...). L'estudiant haurà d'assegurar-se que la llicència que sigui no impedeix específicament seu ús en el marc de la pràctica. En cas de no trobar la informació corresponent haurà d'assumir que l'obra està protegida pel copyright.

Hauran, a més, adjuntar els fitxers originals quan les obres utilitzades siguin digitals, i el seu codi font si correspon.