PAC 2

Presentació

Segona activitat d'avaluació continuada del curs. En aquesta PAC es practicaran els algorismes bàsics de classificació.

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

L'objectiu d'aquesta prova d'avaluació és classificar les dades dels arxius adjunts relacionats amb problemes de fertilitat. Volem diagnòsticar els pacients segons si tenen l'esperma normal o alterat.

L'arxiu de dades fertility.csv té un format tipus taula, on cada fila correspon a un exemple. L'última columna és la classe i la resta corresponen als atributs de l'exemple. L'arxiu adjunt attributes.txt conté la descripció d'aquests atributs.

Aquests arxius pertanyen al problema "Fertility" del repositori d'aprenentatge de l'UCI:

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

Descripció de la PAC

Exercici 1

a) Construïu els models de classificació basats en el veí més proper per valors de k u i tres a partir de l'arxiu train.csv. És a dir, heu de construir els models: 1NN i 3NN. Un cop tingueu el model, heu de classificar els exemples de l'arxiu test.csv.

b) Construïu els models de classificació basats en el k–means per a un valor de k de dos a partir de l'arxiu train.csv. Un cop tingueu el model, heu de classificar els exemples de l'arxiu test.csv.

c) Construïu un arbre de decisió a partir de l'arxiu train.csv i classifiqueu amb ell els exemples de l'arxiu test.csv.

d) Apliqueu l'algorisme del PCA per reduir la dimensionalitat dels conjunts anteriors conservant el 95% de la variància. Utilitzeu el 1-NN i 3-NN sobre els conjunts reduïts de la mateixa forma que en el primer apartat. S'obtenen resultats comparables? Expliciteu les diferències en aplicar el PCA sobre el conjunt d'entrenament i sobre el de test.

Exercici 2

L'objectiu d'aquest segon exercici és la construcció d'un model amb un conjunt de dades proper al que utilitzaríem en una aplicació pràctica per a un cas real. Per realitzar aquest exercici, teniu a la vostra disposició una eina anomenada Weka a la Web; la seva direcció és:

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

L'arxiu adjunt fertility.csv conté 100 exemples de pacients amb problemes d'esperma amb dues classes. Aquest arxiu està en un dels formats d'entrada del Weka.

Se us demana l'estudi de cara a construir un classificador. Per a realitzar-los heu de:

- Provar els algorismes per validació creuada (cross-validation). És a dir, no utilitzar un sol arxiu de train i un de test. Aquesta opció la permet realitzar el Weka de forma automàtica.

- Provar almenys els algorismes: naïve bayes, algun tipus d'arbre de decisió (mostrant a l'informe l'arbre generat), AdaBoost, xarxes neuronals (perceptró multicapa) i algun altre algorisme.
- Provar les Support Vector Machines (SMO i/o libSVM al Weka) amb diferents tipus de kernel.
- Apliqueu el PCA per reduir la dimensionalitat conservant el 95% de la variància i torneu a aplicar els mateixos algorismes que al conjunt original.

Comenteu els resultats obtinguts i justifiqueu tot el que feu.

Exercici 3

Realitzeu una valoració global comparant els mètodes i els diferents exercicis i redacteu unes conclusions globals sobre l'aplicació dels mètodes a aquests conjunts de dades. Els criteris de correcció de la PAC invaliden una A si tots els processos no estan ben justificats i comentats.

Exercici 4

Doneu una definició intuïtiva i curta del aprenentatge incremental ("incremental learning") i digueu com creieu que s'aplica a les màquines de vectors de suport.

Recursos

Bàsics

Per a realitzar aquesta PAC disposeu d'uns fitxers adjunts (attributes.txt, train.csv, test.csv i fertility.csv) on trobareu les dades corresponents a la base de dades de la UCI en un format ja llegible directament pel SW recomanat.

Criteris de valoració

Els quatre exercicis d'aquesta PAC es valoraran amb 3, 3, 2 i 2 punts respectivament.

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

Format i data de lliurament

Cal lliurar la PAC en un pdf adjunt al registre d'activitats d'avaluació continuada.

El nom del fitxer ha de ser CognomsNom_AC_PAC2 amb l'extensió .pdf (PDF).

Data Limit: 7 Maig 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.