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 els guanys de les persones a partir de la informació del cens. Volem predir si superen els 50k dòlars l'any o no.

Els arxius de dades "csv" tenen 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 "txt" conté la descripció d'aquests atributs.

Aquests arxius pertanyen al problema "Census Income Data Set" 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 "adult.data.csv" conté les dades del cens de 32.561 persones. 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-lo 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.

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

Cerca informació i dona raonadament la diferència entre els mètodes de "clustering": k-means, k-medians i k-medoids. Dóna també per a quin tipus de dades es recomanable escollir un mètode o un altre d'entre aquests.

Recursos

Bàsics

Per a realitzar aquesta PAC disposeu d'uns fitxers adjunts ("adult.data.csv", "adult.names.txt", "train.csv" i "test.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, repartits de la forma següent:

Exercici 1:

- Apartat 1 (0,5 punts): valoració de l'aplicació del kNN (inclou el de l'apartat 4). Es valorarà la descripció/inclusió a l'informe de la matriu de distancies, els vots, les prediccions i la precisió (*accuracy*).
- Apartat 2 (1 punt): valoració de l'aplicació del k-means supervisat. Es valorarà la descripció/inclusió a l'informe de la selecció dels centroides inicials, els passos intermitjos del procés, els centroides i grups finals de cadascuna de les classes; i la matriu de distancies, els vots, les prediccions i la precisió (*accuracy*) en aplicar l'1NN.
- Apartat 3 (1 punt): valoració de l'aplicació dels arbres de decisió. Es valorarà la descripció/inclusió a l'informe del càlcul de les bondats de tots els atributs pertinents per totes les iteracions, la representació gràfica de l'arbre, les prediccions del classificador i la precisió (*accuracy*).
- Apartat 4 (0,5 punts): valoració de l'aplicació del PCA i/o les justificacions pertinents. Es valorarà significativament la descripció/inclusió a l'informe de la diferència de tractament del PCA en aplicar-lo als conjunts d'entrenament i test.

Exercici 2:

Es valoraran la inclusió de la taula de resultats amb 2 punts. Els resultats hauran de contenir com a mínim la precisió (*accuracy*), el nombre d'exemples correctes i les matrius

de confusió de tots els mètodes especificats a l'enunciat. El punt restant s'adjudica als comentaris, valoracions i justificacions de tot l'exercici.

Exercici 3:

Aquest exercici val 2 punts que valorarà: les conclusions generals, l'anàlisi de resultats, les comparacions entre mètodes, les comparacions entre diferents conjunts...

Exercici 4:

Aquest exercici val 2 punts on es valorarà significativament la claredat i la simplicitat de la resposta.

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: 2 de 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.