PAC 1

Presentació

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

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 categoritzar les dades dels arxius adjunts relacionats amb els guanys d'una persona a partir de la informació del cens. Volem agrupar persones segons si superen o no els 50k dòlars anuals.

Els arxius de dades tenen un format tipus taula, on cada fila correspon a un exemple. L'última columna representa la classe i la resta corresponen als atributs de l'exemple. L'arxiu adjunt "txt" conté la descripció dels 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

L'objectiu d'aquest exercici és mirar si es pot categoritzar l'arxiu petit de dades (small.csv). En concret, se us demana:

1. Efectueu, si és necessari, el tractament previ de les dades. Justifiqueu totes les decisions que prengueu.
2. Utilitzeu el k-means (nítid) per categoritzar les dades del arxiu esmentat en dues categories, ignorant les columnes no pertinents. Quin és el nivell de precisió[1] del resultat?
3. Apliqueu l'algorisme del PCA per reduir la dimensionalitat del conjunt anterior conservant el 95% de la variància. Utilitzeu el k-means (nítid) sobre el conjunt reduït de la mateixa forma que en l'apartat anterior. S'obtenen resultats comparables?

Exercici 2

L'objectiu d'aquest exercici és categoritzar les dades amb el mètode aglomeratiu. En concret, se us demana la construcció de tres dendrogrames per al conjunt de dades de l'exercici anterior (small.csv) utilitzant la distància euclidea i els mètodes del lligam simple, el lligam complet i la mitja com a criteris d'aglomeració. Doneu les precisions de la mateixa forma que en l'exercici anterior.

Exercici 3

L'objectiu d'aquest exercici és utilitzar una eina per a categoritzar l'arxiu adjunt "adult.data.csv" en dues categories. Aquesta eina s'anomena Weka i la teniu a la seva plana Web:

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

En concret, se us demana l'aplicació del mètode del k-means (el trobareu com a SimpleKMeans) i un altre mètode a escollir sobre les dades de l'arxiu esmentat.

Exercici 4

1

Per calcular la precisió heu de comparar la categoria resultant per cada exemple amb la seva classe (última columna de l'arxiu).

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.

Recursos

Bàsics

Per a realitzar aquesta PAC disposeu d'uns fitxers adjunts ("adult.names.txt", "small.csv" i "adult.data.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 (0,75 punts): valoració del tractament de dades i la justificació de totes les decisions preses.
- Apartat 2 (1,5 punts): valoració de l'aplicació del k-means (inclou la de l'apartat 3). Es valorarà la descripció/inclusió a l'informe de la selecció dels centroides inicials, els passos intermitjos del procés, els centroides i grups finals i la precisió.
- Apartat 3 (0,75 punts): valoració de l'aplicació del PCA i/o justificacions pertinents.

Exercici 2:

Es valoraran amb el mateix pes (0,5 punts) la descripció/inclusió a l'informe de: la matriu de distàncies, els ordres d'agrupació, els dendrogrames resultants, les categories resultants, les precisions i els comentaris, valoracions i justificacions de tot l'exercici.

Exercici 3:

Es valorarà la inclusió de la taula de resultats amb 0,5 punts. Els resultats hauran de contenir com a mínim la precisió (*accuracy*), el nombre d'exemples erronis i les matrius de confusió. Els 0,5 punts restants s'adjudiquen als comentaris, valoracions i justificacions de tot l'exercici.

Exercici 4:

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

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_PAC1 amb l'extensió .pdf (PDF).

Data Límit: 11 Abril 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.