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 dades d'interès sobre cotxes a l'hora de decidir si comprar-los. Volem agrupar els cotxes en funció de si la seva compra seria acceptable o no.

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 "specs.txt" conté la descripció dels atributs.

Aquest conjunt de dades (*dataset*) és una versió modificada del *Car Evaluation Database* que es troba al següent enllaç:

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

Tanmateix, vosaltres heu de treballar amb els arxius proporcionats, no amb les dades disponibles a l'enllaç anterior.

Descripció de la PAC

Aquesta PAC consta de cinc exercicis. Es recomana que llegiu l'apartat "Criteris de Valoració" per a saber quina de la informació que proporcioneu a les vostres respostes tindrà més pes en l'avaluació de la PAC.

Exercici 1

Els arxius adjunts contenen dades que es faran servir en posteriors exercicis. En particular, 'car.csv' conté un conjunt gran de dades i tant 'train.csv' com 'test.csv' són uns pocs exemples pertanyents a 'car.csv'.

En aquest exercici heu d'analitzar aquestes dades, decidir si és necessari dur a terme algun tractament previ i si aquest tractament hauria d'esser el mateix per tots tres arxius o no. Abans de prendre aquestes decisions és convenient que llegiu els enunciats de la resta d'exercicis per a saber com s'emprarà cada un dels arxius.

Si és necessari fer el tractament, feu-lo explicant cada una de les decisions preses i mostrau a la documentació les dades de 'train.csv' i 'test.csv' després del tractament. Si no és necessari fer el tractament, justificau molt clarament els motius.

En posteriors exercicis, si es necessita un tractament previ de les dades de 'train.csv', 'test.csv' o 'car.csv' emprareu les dades resultants d'aquest exercici. Si no es necessita tractament previ, heu d'emprar les dades originals.

Exercici 2

a) Construïu els models de classificació basats en el veí més proper emprant k=1 i k=3 a partir de l'arxiu "train.csv". És a dir, heu de construir els models 1NN i 3NN. Amb cada un d'aquests dos models heu de classificar els exemples de l'arxiu "test.csv"

b) Construïu el model de classificació basat en k-means per a k=2 a partir de l'arxiu "train.csv". És a dir, heu de construir el model per a 2-means. 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". Un cop tingueu l'arbre, classifiqueu amb ell els exemples de l'arxiu "test.csv".

d) Apliqueu PCA per a reduir la dimensionalitat dels conjunts anteriors conservant el 95% de la variància. Utilitzeu ara 1NN i 3NN sobre els conjunts reduïts de la mateixa forma que en el primer apartat. Compareu els resultats. Indiqueu clarament les diferències en aplicar PCA sobre el conjunt d'entrenament i sobre el de test.

Exercici 3

L'objectiu d'aquest exercici és la construcció d'un model amb un conjunt de dades realista. El conjunt de dades realista és el proporcionat per 'car.csv', el qual conté 1728 exemples de cotxes. Per a la construcció del model haureu de programar en Python emprant la biblioteca sklearn (http://scikit-learn.org)

La biblioteca sklearn proporciona iteradors com *KFold*, *RepeatedKFold*, *GroupKFold o StratifiedKFold* especificament dissenyats per a dur a terme validacions creuades (*cross-validation*). Llegiu-ne la documentació i concentrau-vos en un d'ells. També proporciona implementacions d'algorismes com kNN, SVM, xarxes neuronals, etcètera.

Emprant l'iterador de validació creuada escollit i les dades de 'car.csv', feu una validació creuada de, almenys, kNN, SVM i algun altre classificador inclòs en sklearn de la vostra elecció. Justifiqueu cada decisió que prengueu.

Mostrau els principals resultats obtinguts i indiqueu les conclusions que se'n poden extreure.

Exercici 4

Realitzeu una valoració global comparant els mètodes dels 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 5

L'algorisme kNN es basa en determinar els exemples del conjunt d'entrenament més propers a l'exemple que volem classificar. Aquest criteri de proximitat requereix de l'ús d'una distància la qual, habitualment, és la distància euclidiana. Ara bé, existeixen moltes distàncies distintes a l'euclidiana, com ara la distància Manhattan, la distància Mahalanobis o la distància Levenshtein.

Cerqueu informació sobre aquestes distàncies i analitzeu el seu possible ús en kNN. Es podrien emprar? Quins avantatges i inconvenients presentarien respecte de la distància euclidiana? Alguna d'elles seria particularment útil en algun tipus d'aplicació?

No s'acceptaran respostes que continguin text copiat. S'espera que reflexioneu sobre el que es demana i que us expresseu amb les vostres paraules, fins i tot per a definir termes.

Recursos

Bàsics

Per a realitzar aquesta PAC disposeu d'uns fitxers adjunts ("car.csv", "specs.txt", "train.csv" i "test.csv") on trobareu les dades corresponents a la base de dades de la UCI en un format fàcilment llegible pel programari recomanat.

Criteris de valoració

Els cinc exercicis d'aquesta PAC es valoraran amb 1, 3, 3, 1.5 i 1.5 punts respectivament, repartits de la forma següent:

Exercici 1 (1 punt)

Es valorarà la justificació de les decisions preses (0.5 punts) i la correcció de les dades tractades que es sol·liciten (0.5 punts)

Exercici 2 (3 punts)

- a) (0.5 punts). Valoració de l'aplicació del kNN, incloent també el de l'apartat (d). Es valorarà la descripció i inclusió a l'informe de la matriu de distàncies, els vots, les prediccions i la precisió (*accuracy*).
- b) (1 punt). Valoració de l'aplicació del k-means supervisat. Es valorarà la descripció i inclusió a l'informe de la selecció dels centroides inicials, les passes intermèdies dels procés, els centroides i grups finals de cada classe, així com la matriu de distàncies, els vots, les prediccions i la precisió (*accuracy*) en aplicar el 1NN.
- c) (1 punt) Valoració de l'aplicació dels arbres de decisió. Es valorarà la descripció i inclusió a l'informe del càlcul de les bondats de tots els atributs pertinents per a totes les iteracions, la representació gràfica de l'arbre, les prediccions del classificador i la precisió (*accuracy*).
- d) (0.5 punts) Valoració de l'aplicació del PCA i les justificacions pertinents. Es valorarà particularment la descripció i inclusió a l'informe de la diferència de tractament del PCA en aplicar-lo als conjunts d'entrenament i de test.

Exercici 3 (3 punts):

Es valorarà la inclusió de la taula de resultats amb 2 punts. Els resultats per a cada classificador hauran de contenir com a mínim: i) el promig i la desviació típica del temps de construcció del model, ii) el promig i la desviació típica del temps de classificació, iii) les *accuracy*, *precision* i recall, iv) el número de classificacions correctes i incorrectes i la matriu de confusió. El punt restant s'adjudica als comentaris, valoracions i justificacions de tot l'exercici. No és precís que adjunteu el codi que hageu realitzat.

Exercici 4 (1.5 punts):

Es valoraran les conclusions generals, la comparació entre mètodes, la comparació entre distints conjunts de dades, l'anàlisi dels resultats, ...

Exercici 5 (1.5 punts):

Es valorarà la descripció de les distàncies esmentades (0.75 punts) i l'anàlisi del seu possible ús en kNN (0.75 punts)

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: 3 de maig de 2019 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.