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 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

Conjuntament amb aquesta PAC se us proporcionen dos programes en Python (kmeans_sklearn.py i pca_basic.py) per si us poden ser d'ajuda. Aquests programes no tenen per què funcionar amb les dades subministrades i el tractament previ de les dades que duen a terme no té per què ser l'adequat en aquesta PAC. Els programes es proporcionen només com a referència.

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

Per tal de realitzar aquest exercici no heu d'utilitzar un programa que el resolgui, sinó que s'espera que mostreu i justifiqueu cada una de les passes i els càlculs realitzats.

En aquest exercici haureu de mirar si es possible categoritzar l'arxiu petit de dades (small.csv). La darrera columna en aquest arxiu denota la classe. En particular, se us demana:

1.- Efectueu, si és necessari, el tractament previ de les dades. Justifiqueu totes les decisions que prengueu. En posteriors apartats, quan es parli de "small.csv" es sobreentendrà que treballareu amb el resultat del tractament de dades que faceu aquí.

2.- Utilitzeu el k-means nítid per a categoritzar les dades de l'esmentat arxiu en dues categories, ignorant les columnes no pertinents. Quin és el nivell de precisió[1] del resultat?

3.- Apliqueu l'algorisme PCA per a 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. Comparau els resultats.

Exercici 2

En aquest exercici haureu de categoritzar les dades amb el mètode aglomeratiu. En particular, heu de construir tres dendrogrames per al mateix conjunt de dades que a l'exercici anterior (small.csv) utilitzant la distància euclidiana. Com a criteris d'aglomeració emprau el lligam simple per al primer dendrograma, el lligam complet per al segon i la mitja per al tercer. Calculeu les precisions de la mateixa forma que a l'exercici anterior.

Per tal de realitzar aquest exercici no heu d'utilitzar un programa que el resolgui, sinó que s'espera que mostreu i justifiqueu cada una de les passes i els càlculs realitzats.

Exercici 3

En aquest exercici treballareu amb l'arxiu de dades gran (car.csv). El primer que heu de fer és aplicar a aquestes dades el mateix tractament previ que heu aplicat al primer apartat de l'Exercici 1. Després haureu d'emprar una biblioteca de Python per tal de categoritzar les dades de car.csv tractades tal i com s'indicarà.

---

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

La biblioteca s'anomena scikit-learn i la podeu descarregar i llegir-ne la documentació a la seva plana Web:

http://scikit-learn.org

En particular, se us demana que apliqueu, sobre les dades esmentades, el mètode k-means (anomenat KMeans) amb dos criteris distints per a triar els centroides inicials i un altre mètode a escollir distint a KMeans. Per a cada categorització heu de mostrar, com a mínim, precisió (accuracy), nombre d'exemples categoritzats erròniament i matriu de confusió.

No heu d'adjuntar el codi, però sí heu de mostrar per a cada un dels tres casos (k-means amb el primer criteri, k-means amb el segon i mètode a escollir) la línia o línies que defineixen i parametritzen el categoritzador.

Exercici 4

Realitzau una valoració global comparant els mètodes dels exercicis anteriors i els resultats que n'heu obtingut. Redacteu unes conclusions globals. Els criteris de correcció de la PAC invaliden una A si tots els processos no estan ben justificats.

Recursos

Bàsics

Per a realitzar aquesta PAC disposeu d'uns fitxers adjunts ("specs.txt", "small.csv" i "cars.csv") on trobareu les dades corresponents a la base de dades de la UCI en un format ja llegible directament pel SW recomanat. També us adjuntem dos exemples en Python ('kmeans_sklearn.py' i 'pca_basic.py'). Noteu que aquests exemples no funcionen directament amb les dades proporcionades. Si els utilitzeu per a resoldre la PAC, els haureu de revisar amb cura i parametritzar adequadament. En particular, el tractament de dades que es du a terme en aquests arxius és purament a mode d'exemple, i això no significa que sigui el tractament que heu de fer vosaltres. Si ho creieu convenient, podeu preprocessar les dades manualment o amb un full de càlcul.

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 intermedis 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 i de la línia o línies de codi sol·licitades amb 0,5 punts per cada un dels tres mètodes demanats. 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: 9 d'abril de 2019 a les 24h

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.