# 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 l'origen de distints vins a partir de la seva composició i color. Volem agrupar els vins en funció del seu origen.

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 "Wine 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 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 2

L'objectiu d'aquest exercici és la construcció d'un model amb un conjunt de dades realista. Per a realitzar aquest exercici teniu a la vostra disposició la biblioteca sklearn per a Python:

http://scikit-learn.org/

L'arxiu adjunt "wine.csv" conté més de 100 exemples de vins en un format fàcilment llegible des de Python. Se us demana l'estudi de cara a construir un classificador. En particular, heu de:

- Provar els algorismes de validació creuada (*cross-validation*). És a dir, no heu d'utilitzar un sol conjunt d'entrenament i un sol conjunt de test. La biblioteca sklearn proporciona iteradors específicament dissenyats per a aquesta tasca, com ara *KFold*, *RepeatedKFold*, *GroupKFold*, *StratifiedKFold*. Llegiu-ne la documentació i concentrau-vos en un d'ells.
- Provar almenys els algorismes: KNN, algun tipus de SVM, xarxes neuronals (perceptró multicapa) i algun altre classificador de la vostra elecció.

Comenteu els resultats obtinguts i justifiqueu tot el que feu.

## Exercici 3

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 4**

El *K-Fold cross validation* presenta dos principals inconvenients. D'una banda, pot requerir un elevat temps d'execució. D'altra banda, només explora un subconjunt de les possibles particions de dades entre entrenament i prova. Existeixen estratègies que pretenen alleugerir aquests problemes. Una d'elles es l'anomenada *Monte-Carlo Cross Validation*. Cerqueu informació sobre aquesta estratègia, descriviu-la i analitzeu els avantatges i inconvenients que presenta en relació al K-Fold.

## Recursos

### Bàsics

Per a realitzar aquesta PAC disposeu d'uns fitxers adjunts ("wine.csv", "wine.names.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 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:**

- 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 2:**

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 3:**

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

**Exercici 4:**

Aquest exercici val dos punts. Es valorarà la descripció del *Monte-Carlo cross validation* i la comparativa amb *K-fold cross validation*.

## 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: 4 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.