



## PRÀCTICA

### Presentació

Pràctica sobre el desenvolupament i aplicació d'algorismes d'aprenentatge automàtic.

### 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

Seguint amb les dades utilitzades en les altres PACs, disposem d'un conjunt de dades de fallides de projectes empresarials (arxius Qualitative\_Bankruptcy\_train.csv i Qualitative\_Bankruptcy\_train\_test.csv). Els projectes poden estar en fallida (B) o funcionament (NB). Es vol desenvolupar un classificador a partir de les diferents variables (6 en total, veure la descripció en el fitxer Qualitative\_Bankruptcy\_info.txt) per veure si podem predir amb suficient precisió la fallida.

Volem construir un model capaç de predir la classe de l'exemple utilitzant l'algorisme de classificació probabilístic Naïve Bayes. L'objectiu és implementar l'algorisme i avaluar els resultats amb les dades esmentades.



La pràctica està formada per 3 exercicis amb diferent pes. La memòria de la pràctica ha de contenir les respostes a cadascun dels exercicis.

## Descripció de la PRÀCTICA

### Exercici 1. Naïve Bayes (20%)

El Naïve Bayes és un algorisme de classificació probabilístic pensat per treballar amb conjunt de dades d'atributs nominals. Cerqueu informació sobre el mateix i descriuiu-lo en detall, posant l'enfasi en com es podria implementar.

Aquest algorisme té un problema numèric: si apareix en el conjunt de test un valor d'un atribut que no ha aparegut en el conjunt d'entrenament, obtenim una probabilitat de 0, que fa que la global també sigui 0. Això es soluciona aplicant alguna tècnica de suavitzat (*smoothing* en anglès). Cerqueu informació sobre les mateixes, escolliu una i descriuiu-la en detall.

### Exercici 2. Weka (20%)

Utilitzant el Weka apliqueu **dos** algorismes diferents d'aprenentatge supervisat (un d'ells que sigui el Naïve Bayes) per predir la classe de l'arxiu de test. Doneu la precisió (*accuracy*) de la classificació obtinguda amb cada mètode. Discutiu els resultats.

### Exercici 3. Implementació del Naïve Bayes (60%)

Implementeu l'algorisme del Naïve Bayes per tal de predir si els exemples del conjunt de test són B o NB, tractant directament els atributs nominals.

Cal que tingueu en compte:

1. Cal explicar l'algorisme implementat, explicant tots aquells detalls que considereu rellevants i les decisions de disseny preses. Feu especial esment en els passos de la implementació de l'algorisme i la tècnica de suavitzat.
2. Una taula amb almenys la precisió, la matriu de confusió i el temps de càlcul de l'algorisme, comparant els resultats amb els obtinguts al Weka (exercici 2).
3. Un apèndix amb el llistat del codi font del vostre programa.
4. I, en general, una justificació de tot el que estigueu fent.

Heu de lliurar el programa que hagueu implementat per realitzar l'exercici 3 (es recomana que sigui en Python o Java, però la podeu realitzar en Delphi,



Lazarus, C, C++, Visual Basic o “similar” ). La qualitat del codi (estructura, comentaris...) és un dels criteris importants de correcció.

## Recursos

### Bàsics

Per a realitzar aquesta pràctica disposeu d'uns fitxers adjunts (Qualitative\_Bankruptcy\_info.txt, Qualitative\_Bankruptcy\_test.csv, Qualitative\_Bankruptcy\_train.csv) on trobareu les dades corresponents.

## Criteris de valoració

Els tres exercicis d'aquesta pràctica es valoraran amb 2, 2 i 6 punts respectivament, repartits de la forma següent:

**Exercici 1:** aquest exercici val 2 punts que valorarà amb 1,25 la part del Naïve Bayes i amb 0,75 la part de la tècnica de suavitzat. Es valorarà significativament la claredat, simplicitat i estructura del discurs; així com l'enfoc de cara a la implementació del mateix.

**Exercici 2:** es valoraran la inclusió de la taula de resultats amb 1,5 punts. Els resultats hauran de contenir com a mínim la precisió (*accuracy*), el nombre d'exemple erronis i les matrius de confusió de tots els mètodes especificats a l'enunciat. Els 0,5 punts restants s'adjudiquen als comentaris, valoracions i justificacions de tot l'exercici.

**Exercici 3:** la valoració per aquest exercici va fer fases:

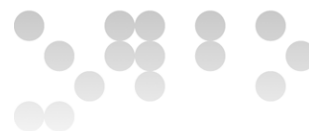
- El funcionament del programa té una valoració binària de 0 o 4 punts en funció de si funciona o no.
- La qualitat del codi i la presentació es porten un punt. Com a qualitat s'entén: l'estructura, comentaris, disseny...; i com a presentació: la sortida del programa, l'ús d'entorns de finestres, l'entrada de les dades, la possible parametrització del programa...
- El punt restant valorarà: les conclusions generals, l'anàlisi de resultats, les comparacions amb el Weka...

## Format i data de lliurament

Cal lliurar la pràctica en un únic fitxer comprimit que contingui la memòria en format *pdf* i el codi font del programa.

El nom del fitxer ha de ser CognomsNom\_AC\_Practica amb l'extensió .zip.

**Data Límit: 30/05/2014 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.