Dpto. Lenguajes y Ciencias de la Computación
E.T.S.I. en Informática, Universidad de Málaga

Aprendizaje Computacional
22 de Enero de 2015

Apellidos: **Nombre:**
DNI:

PARTE 2

1. En la tabla que se muestra a continuación se describen características de varios personajes de la serie Los Simpson:

Personaje	Longitud Pelo	Peso	Edad	Género
Homer	0″	250	36	H
Marge	10″	150	34	M
Bart	2″	90	10	H
Lisa	6″	78	8	M
Maggie	4″	20	1	M
Abe	1″	170	70	H
Selma	8″	160	41	M
Otto	10″	180	38	H
Krusty	6″	200	45	H

   a) Realiza un árbol de decisión (tipo ID3) tendiendo en cuenta que la **Longitud del Pelo** es dividida en dos clases con etiquetas *Menor que 5* y *Mayor que 5*; el **Peso** es dividido en *Mayor que 160* y *Menor que 160* y la **Edad** se divide en *Mayor que 40* y *Menor que 40*. Dibuja el árbol, describe las operaciones y muestra los datos usados en cada una de las iteraciones.
   b) ¿cuales son las reglas de clasificación que se deducen del árbol?
   c) Clasifica:

   e indica el genero.
   d) Poda el árbol con el siguiente criterio: si uno de los géneros de los datos a clasificar alcanza el 75% o mas, el nodo no se explota y se etiqueta con dicho genero (Por ejemplo, supongamos que en un nodo tengo 5 hombres y 1 mujer. El genero hombres supone mas de un 75% de los datos y el nodo no se expande y se etiqueta con H (Hombre). Dibuja dicho árbol y describe los atributos de un personaje que con el árbol podado este mal clasificado.

2. Realiza un programa en R que entrene un árbol tipo Rpart con los datos de la tabla siguiente:

Dpto. Lenguajes y Ciencias de la Computación
E.T.S.I. en Informática, Universidad de Málaga

Aprendizaje Computacional
22 de Enero de 2015

Apellidos: **Nombre:**
DNI:

	x	y	z
1	1	a	5
2	2	c	4
3	3	b	4
4	3	a	2
5	3	b	5
6	4	c	2
7	1	b	4

a) Crea el fichero csv con los datos anteriores.
b) Entrena el árbol teniendo en cuenta que el atributo ***y*** clasifica las diferentes entradas en tres clases.
c) Dibuja el árbol y muestra las reglas de clasificación.
d) Predice la clasificación de los siguientes datos:

	x	y	z
1	1	a	4
2	2	c	6

e) Imprime la tabla CP (Parámetro de complejidad)
f) Determina usando R el valor apropiado para el parámetro CP y úsalo para podar el árbol.
g) Dibuja el árbol podado y predice los datos del apartado ***d)*** con el árbol podado.
h) Comenta los resultados.

3. En una aplicación cliente servidor TCP/IP, los clientes realizan clasificaciones usando un perceptrón multicapa. El desarrollo se realiza en Java y el perceptrón multicapa esta desarrollado en R ¿Como se puede realizar la conexión entre Java y R? Dibuja un esquema de la arquitectura del cliente y del servidor ¿Como debemos configurar el compilador java (piensa por ejemplo en Eclipse) para ofrecer la solución planteada?
4. A continuación se muestra una tabla con los diferentes parámetros de complejidad de un conjunto de datos:

	CP	nsplit	rel error	xerror	xstd
1	0.161992664	0	1.0000000	1.0002790	0.01853630
2	0.043985638	1	0.8380073	0.8385070	0.01749290
3	0.030278222	2	0.7940217	0.7963870	0.01709283
4	0.013881619	3	0.7637435	0.7695997	0.01653832
5	0.010181164	4	0.7498619	0.7560406	0.01606136
6	0.008004043	5	0.7396807	0.7466449	0.01600352
7	0.007026176	6	0.7316767	0.7356289	0.01549501
8	0.006614587	8	0.7176243	0.7388091	0.01559568
9	0.005312278	10	0.7043951	0.7254237	0.01522645
10	0.004883811	11	0.6990828	0.7248227	0.01526605

a) Describe el método para determinar el mejor árbol para realizar la poda.
b) Determina el CP de la tabla anterior ¿Qué contienen las columnas xerror, xsts y nsplit?
c) Explica cómo se calcula la tabla anterior.

Dpto. Lenguajes y Ciencias de la Computación
E.T.S.I. en Informática, Universidad de Málaga

Aprendizaje Computacional
22 de Enero de 2015

Apellidos: **Nombre:**
DNI:

PARTE 3

5. Un método de Boosting consiste en entrenar una serie de clasificadores pobres (por ejemplo, arboles de decisión) con un subconjunto escogido aleatoriamente del dataset. Se realiza una predicción con cada uno de los clasificadores y se obtiene como salida del boosting la clase más votada (es decir, la clase con mayor número de ocurrencias). Realiza un programa en R que implemente el algoritmo descrito. Usa el dataset que consideres conveniente (ejemplo, el de 1x2 o el de la recidiva…).

   Sugerencia:

   Para poder manjar una lista de clasificadores puedes usar el código siguiente:

```
library(rpart)
df<-data.frame(x=c(1,2,3,3,3), y=factor(c("a", "a", "b", "a",
"b")),z=c(5,4,4,2,5))

mytree1<-rpart(y ~ x+z  , data = df, minbucket = 1, minsplit=1)
mytree2<-rpart(y ~ x+z  , data = df, minbucket = 1, minsplit=2)
mytree3<-rpart(y ~ x+z  , data = df, minbucket = 1, minsplit=3)

lista1<- list(mytree1, mytree2)
lista2<- list(mytree3)
w <- c(lista1,lista2)
w[[1]]
plot(w[[3]])

Dicho código crea tres clasificadores y dos listas que concatena
en una única lista. w[[1]] hace referencia a mytree1 y
plot(w[[3]]) dibuja el árbol mytree3.
```

6. Usa el comando *importance* de ***Ramdon Forest***. para determinar la importancia de los atributos en un dataset (puedes usar cualquiera que prefieras o crear uno nuevo), Dibuja también el grafico. Explica que significa que una variable tenga mayor importancia que otra.