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

Aprendizaje Computacional
15 de Noviembre de 2016

Apellidos: **Nombre:**
DNI:

PARTE 2

1. A continuación se muestra un subconjunto de los pasajeros del Titanic:

PassengerId	Survived	Pclass	Name	Sex	Age
1	0	3	Braund, Mr. Owen Harris	male	22.00
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00
3	1	3	Heikkinen, Miss. Laina	female	26.00
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00
5	0	3	Allen, Mr. William Henry	male	35.00
6	0	3	Moran, Mr. James	male	NA
7	0	1	McCarthy, Mr. Timothy J	male	54.00
8	0	3	Palsson, Master. Gosta Leonard	male	2.00
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00

Los atributos que se muestran son:

- PassangerId: identificador del pasajero.
- Survived: Atributo de clasificación. 1, sobrevive; 2, no sobrevive.
- Pclass: La clase del pasajero: 1 primera, 2 segunda o 3 tercera.
- Name: Nombre del pasajero:
- Sex: Sexo del pasajero
- Age: Edad del pasajero

a) Realiza un árbol de decisión (tipo ID3). Itera hasta el nivel 1. Dibuja el árbol, describe las operaciones y muestra los datos usados en cada una de las iteraciones. El atributo PassangerId no interviene en le construcción del árbol. El atributo Survived se modifica con las etiquetas Si (1) y No (0). El atributo Pclass con el literal de la clase. El atributo Name se etiqueta con "Menor o igual que Futrelle" para aquellos nombre que cumpla que son lexicograficamente menores a la palabra Futrelle y "Mayor que Futrelle" para aquellos nombres mayores a Futrelle. El atributo Age, se etiqueta como "No adulto" a los pasajeros con edades menor o igual a 22 años, "Adulto" para los pasajeros con edad mayor a 22 años y "Desconocido" para los pasajeros que no se conozca su edad (NA).

b) Para el siguiente nivel considera un algoritmo de poda tal que se etiqueta el nodo con la clase mayoritaria del atributo objetivo (**Survived**). Completa el árbol.

c) Clasifica los siguientes pasajeros e indica si sobrevivirán o no.

Survived	Pclass	Name	Sex	Age
1	1	Sloper, Mr. William Thompson	male	28.00
0	3	Palsson, Miss. Torborg Danira	female	8.00

d) Entrena un árbol Rpart con los datos anteriores y compara la precisión (accuracy) de ambos clasificadores.

e) Entrena Rpart con el conjunto completo de datos del Titanic (que lo puedes encontrar en el campus virtual).

f) Dibuja el árbol, muestra la tabla ROC (los datos de Test lo puedes encontrar en el campus virtual) y calcula el área bajo la curva.

g) Muestra la matriz de confusión. Calcula el accuracy.

h) Poda el árbol. Elige el CP que consideres mejor. Justifica la elección.

i) Dibuja el árbol podado y muestra la tabla ROC y la matriz de confusión.

j) Entrena un Maquina de Vectores Soporte. Muestra la curva ROC y el área bajo la curva (usa el conjunto completo de datos del Titanic y los datos que se encuentran en el campus virtual).

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

Aprendizaje Computacional
15 de Noviembre de 2016

Apellidos: **Nombre:**
DNI:

2. Una señal (signal) tiene los siguientes valores en función de la entrada (In):

In	0	10	20	30	40	50
Signal	4	22	44	60	82	89

Entrena un clasificador de regresión lineal (lm) con los datos anteriores (usa el modelo Signal = $\beta_0 + \beta_1$In cuya fórmula es Signal ~ In). Predice los siguientes valores:

In	5	15	25	35	45

Encuentra los parámetros β_0, β_1. Dibuja en una gráfica los puntos predichos frente a los valores de entrada In (Sugerencia: En R si quisiéramos pintar la gráfica de los valores de la señal de entrada en función de In seria plot(In , Signal,)).
Repite los pasos anteriores con el modelo Signal = $\beta_0 + \beta_1 A + \beta_2 A^2$ cuya fórmula es Signal ~ In + **I(**In^2).
¿Cuál de los dos modelos se ajusta mejor al conjunto de datos de entrenamiento?

3. El conjunto de datos Kyphosis (disponible con el paquete Rpart) nos indica si, tras una operación en la columna vertebral, una muestra de 81 niños presentan o no deformaciones en la columna vertebral (columna ***Kyphosis***). El resto de columnas son:

Age : Año en meses.
Number: Numero de la vértebra involucrada.
Start : Numero de la primera vértebra involucrada

Compara la precisión (accuracy) del conjunto de datos anterior usando validación cruzada (10% para test el resto para entrenamiento). Obtén los conjuntos aleatoriamente) mediante árboles de decisión (rpart) y perceptrón multicapas (usa 5 capas). Compara la precisión (accuracy).
Entrena un árbol de decisión (rpart) y un perceptrón multicapas (5 capas) usando validación cruzada (10% para test el resto para entrenamiento; obtén los conjuntos aleatoriamente).
¿Es posible encontrar un árbol de decisión Rpart y un perceptrón muticapa tal que sus accuracy sean próximos modificando las parámetros CP y numero de neuronas, respectivamente?