



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

Aprendizaje Computacional

14 de Diciembre de 2020

Apellidos:
DNI:

Nombre:

PARTE 1

1. Se pretende encontrar el mejor clasificador para los datos del Titanic publicados en Kaggle (<https://www.kaggle.com/c/titanic>), que también podemos encontrarlos en el Campus Virtual.

La forma en la que vamos a abordar el problema es la siguiente:

a) En primer lugar, realizaremos un examen del dataset y estudiaremos los atributos y si aparecen datos faltantes. Se pide obtener un dataset en el que se hayan eliminado los atributos innecesarios y se haya solucionado los atributos faltantes. El significado de los atributos son los siguientes:

- PassengerId: 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

Justifica tu respuesta. (0.5 puntos)

*NOTA: Para eliminar un atributo: `dataframe$atributoAeliminar<- NULL`;
para eliminar atributos nulos o NA: `dataframe <- dataframe[!is.na(dataframe$atributoconNAoNULL),]`*

b) Entrenaremos los siguientes clasificadores: un árbol de decisión (paquete Rpart) , un perceptron multicapa (paquete nnet) y una maquina de soporte vectorial (paquete e1071 o Kernlab) usando validación cruzado. Obtendremos el accuracy y el área bajo la curva para cada clasificador. (0.75 puntos)

c) En tercer lugar, vamos a entrenar a cada uno de los clasificadores, usando validación cruzada y modificando sus parámetros como CP , size o el tipo de kernel, con el objetivo de encontrar el clasificador mejor de su clase. (0.75 puntos)

d) Por último, obtendremos el mejor clasificador de todos y realizaremos una predicción con el mismo usando los datos de test de Titanic. (0.5 puntos)

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



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

Aprendizaje Computacional

14 de Diciembre de 2020

Apellidos:
DNI:

Nombre:

- e) 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 PassengerId no interviene en la 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 nombres que cumpla que son lexicográficamente 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). (2.5 puntos)
- f) 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. (0.5 puntos)
- g) Clasifica los siguientes pasajeros e indica si sobrevivirán o no. (0.5 puntos)

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

2.

A partir del conjunto de datos que podemos encontrar en el campus virtual (denominado ejercicio2SVM.csv):

Determina si es separable linealmente e indica cual sería la función Kernel más adecuada (indica tipo de función y sus parámetros) (1 punto).

Calcula los siguientes parámetros de la Máquina de Soporte Vectorial que podemos obtener con el dataset anteriores y el Kernel elegido:

- Vectores Soporte. (0.25 puntos)
- Ancho del canal (0.5 puntos)
- Vector de Pesos normal al Hiperplano (W) (0.25 puntos)
- Vector B (0.25 puntos)
- La ecuación del Hiperplano y de los planos de soporte positivo y negativo. (0.75 puntos)
- Pinta el conjunto de puntos y el Hiperplano. (0.75 puntos)
- Clasifica los puntos (0.5, 0.8) y (0.6, 0.2). (0.25 puntos)