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

Aprendizaje Computacional
23 de Enero de 2018

Apellidos: **Nombre:**
DNI:

PARTE ENSAMBLES

1. La URL maliciosa, también conocido como sitio web malicioso, es una amenaza común y grave para la ciberseguridad. La URL maliciosa atraaen a los usuarios confiados a ser víctimas de estafas (monetarias pérdida, robo de información privada e instalación de malware), y causan pérdidas de miles de millones de euros cada año. Por tanto, es necesario detectar y actuar sobre tales amenazas de manera oportuna. Tradicionalmente, esta detección se realiza principalmente mediante el uso de listas negras.. Sin embargo, las listas negras no pueden ser exhaustivas, y carecen de la capacidad para detectar nuevas URL maliciosas generadas. Otra técnica que se esta usando es machine lerarning para clasificar las url entre maliciosas o no. Se pide:

a) Crear un ramdom forest usando el dataset *detect-malicious-URL .csv* (que podeis encontrar en el campus virtual). Usa validación cruzada para entrenamiento y predicción. Calcula el accuracy.
b) Crea un stacking (por votación y sin pesos) con los clasificadores pobres que considere que mejore o al menos iguale el accuracy del ramdom forest.

2.
a) Teniendo en cuenta el dataset que se muestra en la Figura 1 (puntos negros se clasifican positivamente; rojos negativamente), calcula los pesos α según el algoritmo de boosting teniendo en cuenta que se han separado los puntos tal y como muestran las figuras 2 (valores positivos en el lado izquierdo de la imagen) y 3 (valores positivos en la parte superior de la imagen) .

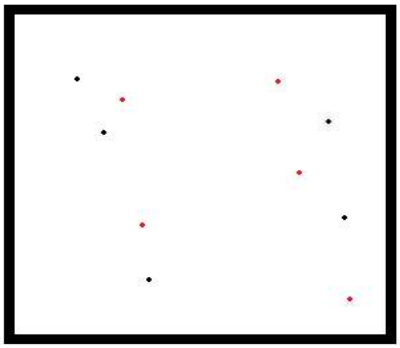

Figura 1: Dataset

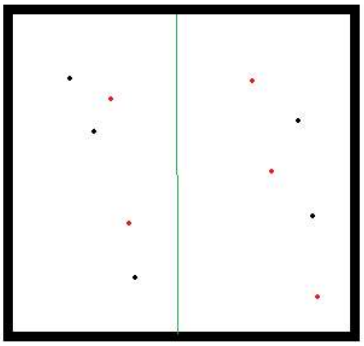

Figura 2: Separación 1

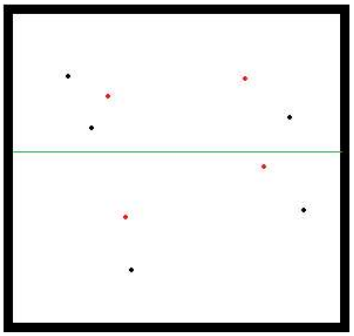

Figura 3: Separación 2

b) Los datos de la figura 1 se encuentran en un rectángulo se 100x100. Dale valores (x,y) a cada uno de los puntos (de la forma mas aproximada posible) . Crea el dataset y usalo para entrenar un boosting.
c) Predice los siguientes puntos usando los dos métodos: (0,0), (0,100), (100,0) , (100,100) y (50,50).

3.
a) Explica el siguiente programa:

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

Aprendizaje Computacional
23 de Enero de 2018

Apellidos: **Nombre:**
DNI:

```
x1=rnorm(1000)
x2=rnorm(1000)
y=2*x1+.7*x2+rnorm(1000)
df=data.frame(y,x1,x2,x3=rnorm(1000),x4=rnorm(1000),x5=rnorm(1000)
library(randomForest)
rf1 <- randomForest(y~., data=df, mtry=2,ntree=50,importance=TRUE)
importance(rf1,type=1)
```

b) Modifica el programa de forma que la importancia de las variables sea equiprobable (es decir, que cambie dependiendo de la ejecución).