Ricardo Aler Mur

PRÁCTICA ANÁLISIS DATOS EN R

Puntuación: 2 puntos

PROGRAMAR EN R UN ALGORITMO QUE GENERE UNA VERSIÓN **SIMPLIFICADA** DE ÁRBOLES DE MODELOS

1. OBJETIVO

Se trata de, a partir de una tabla de datos de un problema de regresión, construir un árbol de modelos (para regresión) mediante un algoritmo **simplificado**. Supongamos que tenemos una tabla de datos con n atributos discretos D1 a Dn y m atributos continuos C1 a Cm, además de la clase R (continua). Es decir, un dato está formado por el siguiente vector: (D1, …, Dn, C1, …, Cm, R). Por ejemplo, en el problema de la predicción de la concentración de algas que vimos en clase, había tres atributos discretos ("season", "size", "speed") con valores (autumn, spring, summer, winter), (large, médium, small) y (high, low, medium) respectivamente. Los atributos restantes ("mxPH" "mnO2" "Cl" "NO3" "NH4" "oPO4" "PO4" "Chla") eran continuos. La clase continua a predecir es “a1”.

Se trata de construir un árbol de modelos (regresión) **simplificado** donde el primer nivel del árbol usa el mejor atributo discreto, y en las hojas se construyen modelos de regresión lineal utilizando solamente los atributos contínuos. Por ejemplo, la función que tenéis que programar, aplicada a los datos de las algas:

```
miArbolModelos(algae[,1:12])
```

debería producir la siguiente salida, con un árbol que en la raíz tiene el atributo *size*, puesto que es el que mas disminuye la desviación típica media, y en las hojas tiene tres modelos lineales:

```
  size
  *** size == large

Call:
lm(formula = salida ~ ., data = cbind(entradasContinuas, salida))

Coefficients:
(Intercept)         mxPH         mnO2           Cl          NO3
  54.266167    -3.945670     0.909984    -0.047155    -1.699866
        NH4         oPO4          PO4         Chla
   0.001846    -0.005471    -0.045740    -0.091701

  *** size == medium

Call:
lm(formula = salida ~ ., data = cbind(entradasContinuas, salida))

Coefficients:
(Intercept)         mxPH         mnO2           Cl          NO3
  54.266167    -3.945670     0.909984    -0.047155    -1.699866
        NH4         oPO4          PO4         Chla
   0.001846    -0.005471    -0.045740    -0.091701

  *** size == small

Call:
lm(formula = salida ~ ., data = cbind(entradasContinuas, salida))

Coefficients:
(Intercept)         mxPH         mnO2           Cl          NO3
  54.266167    -3.945670     0.909984    -0.047155    -1.699866
        NH4         oPO4          PO4         Chla
   0.001846    -0.005471    -0.045740    -0.091701
```

2. SE PIDE:
    1. Programar la función `miArbolModelos` que tome como entrada un dataframe con entradas discretas y contínuas y una salida contínua y produzca un árbol de modelos con el mejor atributo discreto en la raíz y modelos lineales con los atributos contínuos en las hojas. La salida de la función consiste en imprimir el árbol en modo texto, tal y como aparece en la sección anterior de objetivos. Para programar esta función es importante haber seguido el tutorial que acompaña a esta práctica.
    2. Comparar que en el caso de las algas, sale el árbol que se ve en la sección de objetivos
    3. Comprobar vuestra función con algún otro dominio de regresión que combine atributos contínuos y discretos. Podéis encontrar dominios de regresión aquí:
        a. http://archive.ics.uci.edu/ml/datasets.html
    4. Plantead un dominio de regresión sencillo (datos artificiales generados por vosotros) donde se vea claramente que vuestra función funciona
    5. Es necesario entregar el código R y una memoria donde describáis los resultados.