Práctica Análisis de Datos con R – Técnicas de modelado

Entrega

Fichero comprimido con tres entregables:

- **código R** necesario para hacer los modelos y pasos solicitados, comentado y entendible
- **respuestas** recopiladas en un documento de texto de forma breve
- **objeto R** final tras la ejecución del código R completo (salvar el *workspace* eliminando previamente la matriz de distancia euclídea para reducir el tamaño)

Datos: fichero "adult_sel.data"

Es una muestra aleatoria de 3,000 casos del conjunto de datos en https://archive.ics.uci.edu/ml/datasets/Census+Income (cabecera, separador ",")

El objetivo del análisis de estos datos es determinar, a partir de un conjunto de variables de entrada, si una persona genera ingresos mayores de 50,000 dólares al año.

Preprocesamiento de los datos

Paso 1: Selección de variables

- Variables de entrada seleccionadas: "age", "fnlwgt", "educationnum", "sex", "race", "capitalgain", "capitalloss", "hoursperweek"
- Variable de decisión: "ingresos"

Paso 2: Tratamiento de variables categóricas

- Tanto la variable "sex" (2 valores) como "race" (5 valores) deben ser transformadas a variables *dummy*.
- Asimismo, la variable de decisión "ingresos" (con valores "<=50K" y ">50K") será transformada a una variable numérica indicativa de la presencia (1) o no (0) de ingresos altos.

Paso 3: Separación conjunto de entrenamiento y de test

- Entrenamiento: 2/3 de los casos; Test: resto de casos

Paso 4: Normalización de las variables de entrada

- Primero el conjunto de entrenamiento: (x-media)/desviación
- Después, con la misma media y desviación, el conjunto de test

Técnicas de *Clustering*

Considerando sólo las variables de entrada y no la de decisión. Para el conjunto de entrenamiento.

Clustering jerárquico

- Realizar un *clustering* jerárquico con *linkage* simple basado en distancia euclídea
- Cortar el dendrograma a una altura igual a dos tercios de la altura máxima
- *Pregunta*
    - Tras cortar, ¿cuántos *clusters* quedan?

Clustering de repartición

- Realizar un *kmeans* de 4 *clusters* basado en distancia euclídea
- *Pregunta*
    - ¿A qué *clusters* pertenecerían los siguientes dos casos?

```
age,fnlwgt,educationnum,sex,race,capitalgain,capitalloss,hoursperweek,ingresosaltos
23,62278,9,Male,White,0,0,40,1
46,78022,1,Female,Black,0,0,40,0
```

Técnicas de Predicción

Consideraremos las variables de entrada para realizar una predicción de la variable de decisión.

Árboles de decisión

- Entrenar un árbol de decisión considerando el problema como de clasificación
- *Preguntas*
    - ¿Cuál es la variable más importante del árbol?
    - ¿Qué porcentaje de casos de test quedarían mal clasificados por el árbol construido?

Redes neuronales

- Entrenar una red neuronal (A) con 5 neuronas en la capa oculta, error "sse"
- *Preguntas*
    - ¿Qué error total (sse) se comete en el entrenamiento?
    - ¿Cuántos pasos han sido necesarios para entrenar la red?
- Generar la *cumulative gains chart* para el conjunto de test
- *Pregunta*
    - Si sólo se contacta al 20% de clientes del conjunto de test, aproximadamente, ¿cuántas personas tendrían ingresos >50K?
- Entrenar una red neuronal (B) con 5 neuronas en la capa oculta, error "sse", pero eliminando los atributos de condición relativos a la raza (5 variables *dummy*)
- Generar la *cumulative gains chart* para el conjunto de test
- *Preguntas*
    - En base a la curva y comparándola con la de la red neuronal anterior, ¿qué modelo dirías que es mejor?, ¿merece la pena utilizar la variable relativa a la raza?