Aprendizaje Automático I

Ejercicio: Particiones sobre los datos

DSLab

noviembre, 2023

En este ejercicio vamos a trabajar sobre las particiones de un conjunto de datos.

1. En primer lugar, debes leer el conjunto de datos `adult` de la página web de UC Irvine Machine Learning Repository.
2. ¿Qué tamaño tiene la base de datos que has leído?
3. Implementa tu propio código en R para dividir la base de datos en 3 muestras de:

- Entrenamiento 60%
- Prueba 20%
- Validación 20%

4. ¿Qué número de observaciones tienen cada una de las particiones?
5. Calcula la media de la variable `Area` en la muestra de entrenamiento. ¿Coincide con el valor en la muestra de validación?
6. Convierte la variable `Area` en la muestra de entrenamiento en otra variable de media 0 y varianza 1. Aplica la misma transformación en las otras particiones, pero cuidado, empleando la media y la varianza de la partición de entrenamiento. Recuerda que, cuando tienes un nuevo dato, no dispones de estadísticos en "todos" los nuevos datos. ¿Qué valores obtienes para la media y varianza de las nuevas variables en las particiones de validación y prueba?