DSLAB
DATA SCIENCE LABORATORY

Aprendizaje Automático I

Ejercicio: Regresión Logística

DSLAB

noviembre, 2023

Estos datos abordan el rendimiento de los alumnos de enseñanza secundaria de dos centros portugueses. Podéis descargar los datos aquí: **Student Data**. Las características de los datos incluyen las calificaciones de los alumnos, características demográficas, sociales y relacionadas con el centro escolar) y se recogieron mediante informes escolares y cuestionarios. Se proporcionan dos conjuntos de datos relativos al rendimiento en dos asignaturas distintas: Matemáticas (mat) y Lengua portuguesa (por).

1. Leer los datos y unirlos en un único `data frame`.

```
df1=read.table("./datos/student-mat.csv",sep=";",header=TRUE)
df2=read.table("./datos/student-por.csv",sep=";",header=TRUE)

df3=merge(df1,df2,by=c("school","sex","age","address","famsize",
                       "Pstatus","Medu","Fedu","Mjob","Fjob","reason",
                       "nursery","internet"))
print(nrow(df3)) # 382 students
```

2. Análisis exploratorio de los datos. Visualiza la variable respuesta `G3.y`. Un valor de 0 indica que el alumno ha abandonado la asigantura. Estas observaciones han de ser elminadas.

3. Construye una nueva variable.

$$final = \begin{cases} "pass", & \text{si G3.y} \geq 10 \\ "fail", & \text{otro caso} \end{cases}$$

4. Divide la base de datos en train (60%) y test (40%).

5. Estudia la relación de la variable `final` con la variable `Fedu`. ¿Tiene sentido unir agrupar alguna de las características de la variable? Discute la misma cuestión cuando se plantea sobre la variable `Medu`.
6. Estudia mediante un modelo de regresión logística cómo se modifican la relación entre `Fedu` y `Medu` con la variable respuesta cuando se considera una relación multivariante.
7. Aplica un proceso de selección de variables mediante la función `step`. ¿Qué variables tiene el modelo final? Entrena el modelo empleando k-fold.
8. Evalua el rendimiento del modelo.