

Titanic

Alberto Torres Barrán

21/3/2019

Vamos a intentar predecir la supervivencia de las víctimas del Titanic a partir de las siguientes variables:

- survival: Supervivencia (0 = No; 1 = Si)
- pclass: Clase de pasajero (1, 2, 3)
- name: Nombre
- sex: Sexo
- age: Edad
- sibsp: Número de hermanos/esposos/as a bordo.
- parch: Número de padres/hijos a bordo
- ticket: Número de ticket
- fare: Coste del billete
- cabin: Cabina
- embarked: Puerto de embarque

Con el conjunto de datos anterior:

1. Cargar el fichero `titanic.csv` en R.
2. Ver cuantos valores missing tiene cada variable con `summary`.
3. Eliminar la variable `cabin` (¿por qué?).
4. Eliminar también las variables `name` y `ticket` (¿por qué?).
5. Eliminar las filas que contengan algún NA (función `na.omit`).
6. Dividir datos en 80% entrenamiento y 20% test, aleatoriamente.
7. Ajustar un modelo KNN para distintos valores de K y calcular la tasa de acierto (función `knn()` de la librería `class`)
8. Ajustar un modelo regresión logística regularizada y comparar con el modelo KNN
9. Calcular los *odds ratio* e interpretarlos