Asociación Española para la Inteligencia Artificial (**AEPIA**)

UIMP Universidad Internacional Menéndez Pelayo

Practical: Preparación de Datos

Máster Universitario en Investigación en Inteligencia Artificial

Ciencia de Datos y Aprendizaje Automático

José Hernández-Orallo
Traducido de material original de M.José Ramírez-Quintana
ETSINF
Universitat Politècnica de València

September 20, 2016

- Ejercicio 1: Inspección de datos.
  El fichero "titanic.csv", disponible en la plataforma, contiene datos sobre el hundimiento del Titanic. Copia el fichero en tu directorio de trabajo. Ve al R y usa el comando

```
titanic <- read.csv(file.choose(),header=TRUE, sep=',')
```

y elige el fichero `csv`. Puedes también escribir directamente el nombre del fichero en vez de `file.choose()`, es decir, 'titanic.csv'. Puedes cambiar el carácter separador que se use en el fichero csv, de tal manera que se interprete bien. Muestra los nombres de las columnas. Observa que la primera columna (cuyo nombre es `"X"`) es redundante (denota el identificador de cada instancia), por lo que puede borrarse. para hacer eso, puedes usar la función `subset` como sigue:
`titanic<-subset(titanic,select=-X)`

Ahora prueba los siguientes comandos:

```
> titanic
> head(titanic)
> summary(titanic)
> plot(titanic)
```

¿Qué variables son cuantitativas y qué variables son categóricas? ¿Cómo podemos saberlo?

- Ejercicio 2: Trabajando con gráficos básicos.
  Descárgate el fichero "cars.csv" de la plataforma. Este fichero contiene información sobre la verlocidad y las distancias de frenada de varios coches.

  2.1 Realizar una gráfica del campo distancia frente al campo velocidad (usa la sintaxis $).

  2.2 Haz un histograma de la variable distancia.

  2.3 Haz un histograma de la variable velocidad.

  2.4 Modifica los gráficos anteriors para mostrar el nombre de las variables ("speed" o "distance") en los ejes. Cambia también el título de los tres gráficos, y también usa colores para los histogramas y los títulos. Guarda los nuevos gráficos en ficheros pdf.

- Ejercicio 3: Transformaciones de las variables y los datasets.
  Elimina la primera columna del data frame 'cars'. Ahora, asume que los datos de otros dos coches están disponibles:

speed	dist
21	47
34	87

  3.1 Construye el nuevo data frame con los datos anteriores.

  3.2 Añade el data frame construido al data frame de `cars`.

3.3 Ordena los datos en el dataset resultante por la columna `speed` (en orden ascendente). Hay dos maneras de hacer esto: usando el comando `order()` o combinando el comando `with` con el comando `order()`. (Sugerencia: busca en Internet "how to sort a data frame by columns").

- Ejercicio 4: Manipulación de datos. Baja el fichero "airquality.csv" de la plataforma. Este dataset contiene medidas de calidad del aire de Nueva York. Resuelve las siguientes cuestiones:

  1. Extrae las dos primeras filas del data frame y muéstralas por consola. ¿ Cómo es la salida?
  2. ¿Cuántas observaciones (filas) hay en este data frame?
  3. ¿Cuál es el valor de `Ozone` en la fila 40?
  4. ¿Cuántos valores faltantes hay en la columna `Ozone` del data frame?
  5. ¿Cuál es la media de la columna `Ozone` en este dataset? Excluye los valores faltantes (codificados como NA) en el cálculo de la media.
  6. Extrae el subconjunto de filas del data frame donde los valores `Ozone` estén por encima de 31 y los valores `Temp` estén por encima de 90. ?Cúal es la media de `Solar.R` en este conjunto?

- Ejercicio 5: Transformación de datos (2).

  Con el data frame"airquality.csv" resuelve los siguientes ejercicios:

  1. Discretiza la columna Ozone en cinco bins ('bin1', 'bin2', ...) del mismo ancho y un sexto bin ('binNA') para NA.
  2. Discretiza la columna Solar en cuatro bins del mismo tamaño y un quinto bin para NA.
  3. Crea una nueva columna AbsDay a partir de las columnas Month y Day de tal manera que cuente el número de días que han pasado desde Month=5 y Day=1.

- Ejercicio 6: Transformación de datos (3).

  Con el data frame "titanic" resuelve los siguiente ejercicios:

  1. Numeriza la columna de clase, donde Crew=4, 1st=3, 2nd=2 and 3rd=1.
  2. Transforma el data frame del titanic en un nuevo data frame (titanic2) con tantos ejemplos como pasajeros usando la columna Freq. En otras palabras, no debe haber filas resultantes para aquellas filas originales cuya Freq=0 y debe haber 35 filas replicadas para aquellas filas originales cuya Freq=35.

3. Compara las gràficas del data frame titanic original con el nuevo.

- Ejercicio 7: Selección de datos.

1. Calcula la matriz de correlaciones para el dataset air. ?Ves un par de atributes redundantes?
2. Calcula la matriz de correlación para el dataset cars. ¿Ves algún par de atributes que sea redundante?
3. Usando el conjunto de datos 'air', realiza un muestreo simple aleatorio de 50 ejemplos.
4. Usando el conjunto de datos 'air', realiza un muestreo estratificado aleatorio de 5 ejemplos de cada mes.