Trabajo Práctico 1 - Parte 1

Marcelo A. Soria – Mariana Landoni

Preparación de los datos

El objetivo de esta primera parte del trabajo práctico es conocer los datos con que van a trabajar, realizar un análisis exploratorio y detectar y solucionar varios problemas que presentan los datos.

El dataset que van a utilizar es un subconjunto de 3462 galaxias captadas en una región del cielo y unas 64 variables. El dataset completo, que van a utilizar más tarde, tiene unos 64000 objetos. Las variables incluyen estimadores del tamaño de la galaxia, de su corrimiento al rojo y datos y flujos y magnitudes absolutas obtenidas a diferentes longitudes de onda.

Indicaciones para la realización del TP:

- El TP se realizará en grupos de **tres o cuatro** personas.
- El TP consiste de una serie de tareas, que pueden consistir en mostrar un análisis o contestar una pregunta. Algunas de estas preguntas o tareas están indicadas como **optativas**. Realizar estas tareas suma puntos pero no son obligatorias. En esta primera etapa, todas las tareas son obligatorias.
- Se puede usar **cualquier** herramienta de análisis o combinación de herramientas, debiendo indicarlas en el informe. Los ejemplos de esta guía están en R, pero eso no es excluyente.
- La fecha de entrega de esta primera parte es el lunes 14 de setiembre.

Obtener los datos

El dataset se puede descargar directamente desde la web con este comando:

```
glx <- read.csv("http://astrostatistics.psu.edu/datasets/COMBO17.csv", he
ader = T, stringsAsFactors = F)
```

Análisis preliminar

Todas las variables del dataset son numéricas. consisten en mediciones y en la mayoría de los casos están acompañadas por su error de medición. Así, por ejemplo, *UjMAG* es la magnitud absoluta obtenida con el filtro U del sistema Johnson y e.UjMAG es su error de medición.

Con la función *str()* se pueden ver los nombres de las variables, sus tipos y algunos valores representativos:

```
str(glx)
```

Tarea 1

Hay una variable con una anomalía en el tipo. ¿Cuál es y qué es lo que está causando este problema?

Definiciones de las variables

Las variables que nos interesan para este TP son el identificador de la galaxia, el corrimiento al rojo de la galaxia, la estimación de tamaño y todas aquellas variables que registran los valores de magnitud absoluta a varias longitudes de onda en reposo, esto es, corregidas por el corrimiento al rojo.

Las magnitudes absolutas en reposo están registradas para tres sistemas de filtros: Johnson, SDSS-III y Bessell. Además hay una determinación de magnitud absoluta en reposo en el ultravioleta, a 280 nm.

Los nombres de las variables de interés y su significado son:

- **Nr**: ID de la galaxia
- **Rmag** y **e.Rmag**: Magnitud total en la banda R y su error de medición. En general, cuanto menor es el valor de magnitud, más brillante es la estrella.
- **ApDRmag**: Es la diferencia entre la magnitud total y la magnitud de apertura medidas en la banda R. Es un estimador del tamaño de una galaxia. Un valor de cero indica que es una fuente puntual de luz. Los valores negativos se generan por errores de medición.
- **Mcz**: Estimación del corrimiento al rojo
- Magnitudes absolutas y corregidas por corrimientos al rojo. Sistema de fotometría Johnson:
  - **UjMAG**: banda U, 364 nm
  - **BjMAG**: banda B, 442 nm
  - **VjMAG**: banda V, 540 nm
- Magnitudes absolutas y corregidas por corrimientos al rojo. Sistema de fotometría SDSS-III:
  - **usMAG**: banda u, 355 nm
  - **gsMAG**: banda g, 469 nm
  - **rsMAG**: banda v, 617 nm
- Magnitudes absolutas y corregidas por corrimientos al rojo. Sistema de fotometría Bessell:
  - **UbMAG**: banda U, 366 nm
  - **BbMAG**: banda B, 420 nm
  - **VnMAG**: banda V, 545 nm (el nombre de esta variable debería ser *VbMag*)
- **S280MAG**: Magnitud absoluta en el ultravioleta a 280 nm.

Ejemplos de gráficos de diagnóstico

Existen varias opciones de gráficos de diagnóstico para analizar los datos. R cuenta con al menos tres sistemas gráficos. El más básico es *graphics*, que viene incorporado a R base, a esto se suma *lattice* y el más moderno *ggplot2*.

A continuación, algunos ejemplos con *graphics* y *ggplot2* para la variable *ApDRmag*.

```
# Histograma y boxplot del estimador del tamaño de una galaxia
hist(glx$ApDRmag)
```

```
boxplot(glx$ApDRmag)
```

```
# Un gráfico de densidad kernel:
plot(density(glx$ApDRmag))
```

Los mismos gráficos con ggplot2.

Importante: si el paquete no está instalado, instalar con:

```
install.packages("ggplot2")

library(ggplot2)

qplot(ApDRmag, data=glx)
```

```
qplot(factor(0), ApDRmag, geom = "boxplot", xlab="", data=glx)
```

La función *qplot()* de *ggplot2* sirve para hacer gráficos sencillos y siguiendo una lógica similar a las funciones de *graphics*. Para gráficos más elaborados hay que utilizar la función *ggplot()*:

```
# Un histograma con ggplot()
ggplot(glx, aes(x = ApDRmag)) +  geom_histogram()
```

```
# El gráfico de densidad kernel
ggplot(glx, aes(x = ApDRmag)) +  geom_density(kernel = "epanechnikov")
```

```
# Ambos combinados
ggplot(glx, aes( x = ApDRmag) ) +
```

```
  geom_histogram( aes(y = ..density..), color = "black", fill = "beige")
+
  geom_density( kernel = "epanechnikov", color="red")
```

Otros gráficos y técnicas de diagnóstico.

El dataset parcial que están utilizando en esta primera parte del TP fue desarrollado para un curso de análisis de datos en astronomía. En la página donde se describe el dataset, se muestra un gráfico que muestra la magnitud en la banda B contra la magnitud de esa banda normalizada por la magnitud absoluta a 280 nm. No se indica a qué sistema fotométrico pertenece la banda B, Johnson o Bessell. Podemos reproducir el gráfico con ambos:

```
# Si no está instalado, instalar el paquete gridExtra
# install.packages("gridExtra")
library(gridExtra)
p1 <- qplot(BjMAG, S280MAG-BjMAG, data = glx)
p2 <- qplot(BbMAG, S280MAG-BbMAG, data = glx)
grid.arrange(p1, p2, ncol=2)
```

```
## Warning: Removed 24 rows containing missing values (geom_point).
```

```
## Warning: Removed 24 rows containing missing values (geom_point).
```

Evidentemente en BjMAG hay un outlier, un valor inusualmente alto. Podemos determinar cuál es:

```
which.max(glx$BjMAG)
```

```
## [1] 2
```

```
# Es el registro 2. Para ver las 13 primeras variables de ese registro:
glx[2,1:13]
```

```
##   Nr   Rmag e.Rmag ApDRmag  mumax   Mcz e.Mcz MCzml chi2red  UjMAG e.U
jMAG
## 2  9 25.013  0.181  -0.135 25.303 0.927 0.122 0.864    0.41 -18.28
0.22
##   BjMAG e.BjMAG
## 2 17.86    0.55
```

La salida del gráfico también nos alerta que existen valores faltantes.

Tarea 2

Encontrar otros casos con valores extremos en las variables de interés y eliminarlos. Se podrían imputar esos valores, pero en este caso es preferible eliminarlos, porque son pocos.

Tarea 3

Determinar qué variables tienen datos faltantes y cuáles son los casos con datos faltantes. Igual que en la tarea anterior, eliminar los casos con datos faltantes. Proporcionalmente son pocos.

Ayuda: algunas funciones de R para datos faltantes:

- *is.na(x)*: determina si el vector o dataframe x tiene valores faltantes. Devuelve un vector o matriz de valores TRUE/FALSE de la misma dimension que el vector o dataframe original.
- *anyNA(x)*: si al menos un valor de x es faltante, devuelve TRUE
- *apply(glx, 2, function(x) anyNA(x))*: para aplicar anyNA a cada variable del datagrame por separado
- *complete.cases(x)*: indica qué registros tienen datos completos o al menos un dato faltante
- *which(is.na(x))*: Devuelve el número de registro de los casos faltantes.

Correlaciones entre variables

Las variables de magnitud absoluta en reposo están muy correlacionadas entre sí. En astronomía una forma común de analizar los datos espectrales es normalizándolos con respecto a alguna banda de referencia. Ya vimos un caso así al reproducir el gráfico que se presenta en la descripción del dataset. En este caso se graficó la banda B versus la banda B menos la banda de 280 nm. La normalización se hace restando porque estas variables son logarítmicas.

Ayuda: se puede crear un vector con los números que le corresponden a las variables que registran datos de magnitudes absolutas en reposo para facilitar la construcción de las matrices de correlación:

```
espectrales <- c(10,12,14,16,18,20,22,24,26,28)
head( glx[, espectrales] )
```

```
##    UjMAG  BjMAG  VjMAG  usMAG  gsMAG  rsMAG  UbMAG  BbMAG  VnMAG S280M
AG
## 1 -17.67 -17.54 -17.76 -17.83 -17.60 -17.97 -17.76 -17.53 -17.76  -18.
22
## 2 -18.28  17.86 -18.20 -18.42 -17.96 -18.43 -18.36 -17.85 -18.19  -17.
97
## 3 -19.75 -19.91 -20.41 -19.87 -20.05 -20.71 -19.82 -19.89 -20.40  -19.
77
## 4 -17.83 -17.39 -17.67 -17.98 -17.47 -17.89 -17.92 -17.38 -17.67  -18.
12
## 5 -17.69 -18.40 -19.37 -17.81 -18.69 -19.88 -17.76 -18.35 -19.37  -13.
93
## 6 -19.22 -18.11 -18.70 -19.34 -18.27 -19.05 -19.30 -18.08 -18.69  -19.
18
```

Varias bandas normalizadas tienen una explicación física. Para el trabajo práctico se puede calcular y evaluar cualquier combinación, y de esta manera crear nuevas variables, sin que necesariamente esas combinaciones tengan una explicación astronómica.

Tarea 4

Estimar las correlaciones de las variables espectrales sin normalizar entre sí, y por otro lado las correlaciones de las variables normalizadas por la magnitud absoluta en reposo a 280 nm.

Importante

Subir los informes al campus virtual (http://datamining.dc.uba.ar/campus/) antes de las 24 hs. del lunes 12 y jueves 16 de setiembre para las comisiones de los martes y viernes respectivamente.