Trabajo Práctico 1 - Parte 2

Marcelo A. Soria – Mariana Landoni

La segunda parte del trabajo práctico tiene tres objetivos:

1. Realizar análisis de agrupamientos sobre el dataset que prepararon para la primera parte.
2. Extender este dataset con los datos originales y completos del estudio astronómico.
3. Realizar análisis de clusters con este dataset extendido.

A continuación les repetimos las indicaciones para realizar el TP y la fecha de entrega.

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 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.
- 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.
- El informe se realizará siguiendo los lineamientos de la **metodología CRISP**.
- Las fechas de entrega para la segunda parte son 17 y 20 de octubre hasta las 24 hs, respectivamente para la comisión de los martes y los viernes.

Tabla de puntos para el TP1:

- Cantidad máxima de puntos que se pueden obtener por la tarea obligatoria de la parte 1: 2.5
- Cantidad máxima de puntos que se pueden obtener por las tareas optativas de la parte 1: 3.0
- Cantidad máxima de puntos a obtener por la parte 2: 2.0
- Cantidad máxima de puntos que se pueden obtener por la tarea obligatoria de la parte 3: 3.0
- Cantidad máxima de puntos que se pueden obtener por la tarea obligatoria de la parte 3: 2.0
- Puntaje máximo posible: diez

Tarea 1. Análisis de agrupamientos

A partir del dataset procesado y limpio que prepararon para la primera parte realizar un análisis de agrupamientos por k-medias. Determinar el k óptimo y analizar la calidad del cluster con algunas de las técnicas que vimos en clase.

A partir de las correlaciones entre variables fotométricas observadas en la primera parte de este TP, no es razonable usar todas para el análisis de clusters. En cambio nos restringiremos a las variables del sistema de Johnson normalizadas por la banda S280MAG y la banda B (BjMag) sin normalizar. Agregar, además, Rmag, ApDRmag y Mcz.

Tareas optativas:

Realizar análisis de agrupamientos usando una o más de las técnicas que vimos en el curso (PAM, clustering por densidad, clustering difuso). Discutir las ventajas y desventajas para el caso particular.

Tarea 2. Ampliación del dataset

La descripción del proyecto y los archivos relacionados con los datos originales están en:

http://www.mpia.de/COMBO/combo_CDFSpublic.html

Desde el link http://www.mpia.de/COMBO/table3.dat se descargan los datos y en el archivo de texto http://www.mpia.de/COMBO/ReadMe está la descripción del archivo de datos.

La preparación de estos datos requiere algunos pasos especiales y por eso constituye una tarea separada.

El archivo table3.dat es un archivo de texto en el que cada línea es un registro separado. No hay separadores de campos, pero cada dato ocupa una posición fija. En el archivo ReadMe está la explicación de cómo se organizan los campos byte por byte. Para el TP nos interesan estos campos:

```
  1-  5 I5    ---     Seq      Sequential number (unique object ID)
 33- 39 F7.2  pix     x        x-coordinate on image cdfs_r.fit
 41- 47 F7.2  pix     y        y-coordinate on image cdfs_r.fit
 49- 54 F6.3  mag     Rmag     total magnitude in R
 69- 75 F7.3  mag     ApD_Rmag ? aperture difference of Rmag
104-107 I4    ---         phot_flag  flags on photometry (4)
117-131 A15   ---     MC_class multi-colour class
133-137 F5.3  ---     MC_z     ? mean redshift in distribution of p(z)
179-184 F6.2  mag     UjMag    ? Absolute Magnitude in Johnson U
192-197 F6.2  mag     BjMag    ? Absolute Magnitude in Johnson B
207-212 F6.2  mag     VjMag    ? Absolute Magnitude in Johnson V
308-313 F6.2  mag     S280Mag  ? Absolute Magnitue in 280/40
```

Las dos primeras columnas indican las posiciones de inicio y fin del campo, luego el tipo de dato, las unidades en que se miden, el nombre del campo y la descripción. Por ejemplo, el primer registro indica que desde el byte 1 al 5 se ubica un entero de cinco posiciones con el nombre de campo “Seq” y que se corresponde al ID del objeto.

Como es un archivo de texto, los bytes coinciden con posiciones en la línea de texto. Entonces una forma práctica de leer estos archivos es leer cada línea como un string y luego ir extrayendo los substrings que se correspondan con cada campo.

Para leer un archivo en R que está constituido por cadenas de caracteres sin información a priori sobre cómo se distribuyen las variables:

```
t3 <- readLines("table3.dat")
```

Para extraer los variables podemos escribir una función que a partir de la tabla y de las posiciones de inicio y fin dentro del string, levante la información:

```
extraer.v <- function(tabla.dat, inicio, fin){
  v.cruda <- substr(tabla.dat,inicio,fin)
  # Lo que sigue saca los espacios en blanco al inicio y fin
  v.proc <- gsub("(^\\s+|\\s+$)", "", v.cruda)
  return(v.proc)
}
```

Dos ejemplos de uso de la función:

```
MC_class <- extraer.v(t3, 117, 131)
RMag <- as.numeric( extraer.v(t3, 49, 54) )
```

Los campos de variables numéricas hay que convertirlos de tipo texto a numéricos. Una vez extraídas las variables, se agrupan en un dataframe, y es conveniente borrar las variables individuales.

El campo “Seq” de table3 almacena la misma información de ID que “Nr” del dataset que prepararon para la primera parte del TP.

Los campos x, y son variables que no estaban en el dataset de la primera parte de este TP, y que indican la ubicación espacial de los objetos. Se prepararon a partir de las coordenadas expresadas en declinación y ascensión recta.

El resto de las variables tienen nombres que coinciden con las que ya estuvieron trabajando (aunque cambia el uso de mayúsculas y minúsculas). Excepto que en este dataset hay una variable correctamente nombrada VbMag, que en el anterior se llama VnMAG.

Una vez que se leen los datos en un dataframe, tienen que eliminar todos los registros que para la clase MC_class no sean estrictamente “Galaxy”. Esto es, también tienen que eliminar “Galaxy (Uncl!)”.

Luego tienen que aplicar los mismos criterios de eliminación de registros con outliers en las variables espectrométricas que desarrollaron en la primera parte del TP y repetirlo en table3. Además, los autores del trabajo original informan que las galaxias con Rmag < 24 son las que tienen datos más confiables, porque el resto tienen brillos muy tenues y los datos asociados tienen mucho error. Los autores también señalan que es conveniente trabajar con registros que para la variable *phot_flag* tomen valores menores de 8. Esta variable registra diversos errores en las mediciones fotométricas usando un sistema de *flags*. Los errores que toman un valor mayor o igual que 8 son serios y para un análisis de rutina se deberían descartar.

La variable ApD_Rmag no debería tomar valores menores que cero. En los casos que sea negativa, se deberían convertir los valores a cero, sin descartar el registro.

El dataset contiene los datos para los sistemas fotométricos Johnson, SDSS-III y Bessell, pero en la primera parte práctico habíamos visto que estas variables están bastante correlacionadas. Solo mantendremos ahora el sistema de Johnson.

Tarea 3. Análisis de agrupamientos sobre el dataset extendido

En este punto deben volver a utilizar el o los algoritmos que hayan utilizado para la tarea 1. Pero ahora en el dataset extendido. Es posible que algunos algoritmos no puedan correr bien sobre este dataset. El propósito de esta tarea es justamente determinar qué algoritmos fallan y en lo posible entender por qué. ¿Qué sucede con muestras progresivamente más pequeñas? ¿Es posible que el algoritmo de clustering funcione, pero después no puedan calcular Silhouette?

Tareas optativas:

1. ¿Los clusters detectados corresponden a grupos de galaxias próximas? Con los datos de ascensión recta y declinación, o las coordenadas x, y que extrajeron en la tarea 1 se puede construir una especie de imagen sintética, donde se muestran las galaxias según su ubicación y coloreadas de acuerdo al cluster que pertenecen. Se puede usar el dataset construido para la primera parte, o el derivado de table3, pero en este caso posiblemente sea necesario trabajar con muestras.

2. ¿Las galaxias se distribuyen uniformemente de acuerdo a su corrimiento al rojo? El corrimiento al rojo indica la distancia a la galaxia. Si se ubican las galaxias en un mapa 2D (como el del punto anterior) y se colorean de acuerdo a su corrimiento al rojo se debería determinar si hay o no “manchones” de galaxias más cercanas o más lejanas.

Importante

Enviar el informe de este trabajo antes de la hora y fecha de cierre. Los trabajos que ingresen después no serán considerados.