Trabajo Práctico 1 - Parte 2

Marcelo A. Soria

La segunda parte del trabajo práctico tiene tres objetivos, el primero es realizar análisis de agrupamientos sobre el dataset que prepararon para la primera parte. El segundo objetivo es extender este dataset con los datos originales y completos del estudio astronómico. El tercero es intentar realizar análisis de clusters con este dataset extendido.

A continuación les repito 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 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 martes 13 de octubre.

Tabla de puntos para el TP1:

- Cantidad máxima de puntos a obtener con la parte 1: cuatro
- Cantidad máxima de puntos a obtener con las tareas obligatorias de la parte 2: cuatro
- Cantidad máxima de puntos a obtener con la tareas opcionales de la parte 2: cuatro
- 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.

Tareas optativas:

Realizar análisis de agrupamientos usando una o más de las técnicas que vimos en el curso. 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)
  7-  8 I2    h       RAh      Right ascension (J2000)
 10- 11 I2    min     RAm      Right ascension (J2000)
 13- 18 F6.3  s       RAs      Right ascension (J2000)
     20 A1    ---     DE-      Declination sign
 21- 22 I2    deg     DEd      Declination (J2000)
 24- 25 I2    arcmin  DEm      Declination (J2000)
 27- 31 F5.2  arcsec  DEs      Declination (J2000)
 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
 56- 60 F5.3  mag     e_Rmag   mean error (1-sigma) of Rmag
 62- 67 F6.3  mag     Ap_Rmag  ? aperture magnitude in R
 69- 75 F7.3  mag     ApD_Rmag ? aperture difference of Rmag
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
222-227 F6.2  mag     usMag    ? Absolute Magnitude in SDSS u
235-240 F6.2  mag     gsMag    ? Absolute Magnitude in SDSS g
250-255 F6.2  mag     rsMag    ? Absolute Magnitude in SDSS r
265-270 F6.2  mag     UbMag    ? Absolute Magnitude in Bessell U
278-283 F6.2  mag     BbMag    ? Absolute Magnitude in Bessell B
293-298 F6.2  mag     VbMag    ? Absolute Magnitude in Bessell 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 linea como un string y luego ir extrayendo los substrings que se correspondan con cada campo.

Por ejemplo, en R:

```
t3 <- readLines("table3.dat")
# t3 es un vector de strings.
# Con el comando que sigue se lee el campo MC_class
#donde se indica el tipo de objeto:
substr(t3[1],117,131)
[1] "Galaxy          "
ejemplo <- substr(t3[1],117,131)
# Con el comando que sigue se eliminan los espacios en blanco
# al inicio y al fin usando expresiones regulares
ejemplo <- gsub("(^\\s+|\\s+$)", "", ejemplo)
ejemplo
[1] "Galaxy"
```

Los campos de variables numéricas hay que convertirlos de tipo texto a númericos

```
ejemplo2 <- substr(t3[1], 49, 54)
ejemplo2 <- as.numeric(ejemplo2)
ejemplo2
[1] 25.898
```

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 RAh, RAm, Ras, DE, DEd, DEm, DEs, x, y son variables nuevas que indican la ubucación espacial de los objetos.

El resto de las variables tienen nombres que conciden con las que ya estuvieron trabajando. 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 (si usan R o Pandas en Python) o tabla (si trabajna con SQL), 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 que desarrollaron para el primer dataset y repetirlo en table3.

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 porqué. ¿Qué sucede con muestras 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 sobre la imagen original 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 debería verse si hay “manchones”de galaxias más cercanas o más lejanas.

Importante

Enviar el informe de este trabajo a soria@agro.uba.ar antes de las 24 hs. del martes 13 de octubre. Los trabajos que ingresen después no serán considerados.