9. Análisis de datos en ecología (IV): técnicas de clasificación

Ecología Metodológica y Cuantitativa (5C1)
Departamento de Ecología e Hidrología

Curso 2008–09

Índice

Antes de empezar

Iniciar R. Cargar el archivo de funciones de la asignatura, `funciones.R`:

```
source("http://www.um.es/docencia/emc/datos/funciones.R")
```

1. Introducción

Ejercicios. Bloque 1:

1. Utilizando el fichero `iris.dat` construir la matriz de datos `iris` y representar gráficamente cada uno de los individuos:

   ```
   read.table("http://www.um.es/docencia/emc/datos/iris.dat")->iris
   pairs(iris[,1:4],col=iris[,5])
   ```

   ¿Cómo se interpreta este gráfico? ¿Qué representa cada punto? ¿Puede describirse el comportamiento de cada uno de los ejemplares? ¿Pueden describirse las relaciones entre las variables? ¿Pueden describirse grupos de ejemplares en relación con cada una de las especies estudiadas? (negro=*I. setosa*, rojo=*I. virginica*, verde=*I. versicolor*)

2. Utilizando la función `stars()` representar los casos:

   ```
   stars(iris[,1:4])
   ```

   Para identificar el sentido del gráfico puede obtenerse una clave mediante:

   ```
   stars(iris[,1:4],key.loc = c(20,2))
   ```

   ¿Cómo se interpreta este gráfico? ¿Qué representa cada "rombo"? ¿Puede considerarse de ayuda este tipo de representación para encontrar grupos de individuos de características similares? ¿Qué efecto tendría un gran número de variables? ¿Y un gran número de objetos?

3. Los gráficos anteriores pueden modificarse las opciones para hacer los valores de las variables proporcionales a los radios de segmentos circulares, además, puede utilizarse la mitad del circulo para representar cada observación.

   ```
   stars(iris[,1:4],key.loc = c(20,2),draw.segments=T, full=F)
   ```

   ¿Puede mejorar la interpretación esta modificación gráfica?

---

2. Medidas de asociación y relación entre muestras y variables

2.1. Datos cualitativos y datos binarios

Ejercicios. Bloque 2:

1. Utilizando el fichero de datos `drapaces.dat` construir la matriz `rap`:

   ```
   read.table("http://www.um.es/docencia/emc/datos/drapaces.dat")->rap
   ```

   ¿Qué son las filas de la matriz de datos?

2. Calcular las matrices correspondientes al índice de Jaccard y de coincidencia simple. Para ello han de utilizarse las función `idb()` que por defecto calcula el índice de Jaccard; la opción `method="simmat"` permite calcular el índice de coincidencias simples:

   ```
   idb(rap)
   idb(rap,method="simmat")
   ```

   ¿Cómo es el resultado de aplicar la función a la matriz de datos? ¿Qué significan cada una de las filas y columnas de la matriz?

   ¿Coinciden ambas matrices? ¿cuales son las dos especies más similares? ¿y las más disimilares?

3. Transformar los resultados anteriores en distancias:

   ```
   sqrt(1 - idb(rap))
   sqrt(1 - idb(rap,method="simmat"))
   ```

   ¿Supone esta transformación alguna ventaja desde el punto de vista de la interpretación?

2.2. Datos cuantitativos

Ejercicios. Bloque 3:

1. Leeremos el fichero de datos `bioveg.dat` —recoge la biomasa (Kg m^{-2}) en el estrato arbóreo y herbáceo de 6 muestras— y construiremos la matriz `bioveg` y la representaremos gráficamente:

   ```
   read.table("http://www.um.es/docencia/emc/datos/bioveg.dat")->bioveg
   plot(bioveg,type="n"); text(bioveg,rownames(bioveg))
   ```

   ¿Cuantas variables y casos tiene la matriz de datos `bioveg`? ¿qué representa? ¿hay alguna característica destacable en los datos?

2. Calcular la matriz de varianzas-covarianzas (`cov`) y la matriz de correlaciones (`cor`).

   ```
   cov(bioveg)
   cor(bioveg)
   ```

   ¿Qué dimensiones tienen la matriz de datos, la matriz de varianzas-covarianzas y la matriz de correlaciones? ¿Qué significado tienen los elementos de estas matrices?

3. Calcular la matriz de distancias entre las muestras:

   `dist(bioveg)`

   ¿Qué dimensión tiene la matriz? ¿Qué significado tienen los elementos de esta matriz?

   Utilizando la ayuda de `dist` ¿Qué distancia entre objetos se calcula por defecto?

2.3. Datos cuantitativos y estandarización

Ejercicios. Bloque 4:

1. Estandarizar los datos `bioveg` con ayuda de la función `scale()` y comprobar el efecto mediante la representación de los datos.

   ```
   biovege<-scale(bioveg)
   plot(bioveg,type="n"); text(bioveg,rownames(bioveg))
   x11(); plot(biovege,type="n"); text(biovege,rownames(bioveg))
   ```

   ¿Cuál es la principal diferencia entre los dos gráficos? ¿Cómo actúa R en la representación de los datos?

2. Comparar la matriz de varianzas-covarianzas de los datos estandarizados con la matriz de correlaciones de los datos originales y los datos estandarizados.

   ```
   cor(bioveg); cov(biovege)
   cor(biovege); cov(biovege)
   ```

   ¿Cuál ha sido el efecto de la estandarización?

3. En relación a la representación gráfica realizada anteriormente, comparar las matrices de distancias de los datos originales y de los estandarizados:

   ```
   dist(bioveg); dist(biovege)
   ```

   ¿Cuál es el par de muestras más próximo con los datos en bruto? ¿Y con los datos estandarizados? ¿Se mantienen en términos relativos las mismas distancias entre los objetos? ¿Qué opción es más razonable: trabajar sobre los datos originales o sobre los estandarizados?

---

3. Clasificación

3.1. Clasificación no jerárquica

Ejercicios. Bloque 5:

1. Aplicar la técnica de *k-means* a la matriz `bioveg` para construir dos grupos:

   `kmeans(bioveg,2)`

   ¿Qué significan los términos `$cluster`,`$centers` y `$size` obtenidos como resultado?

2. Representar los resultados de la clasificación utilizando como ejes de coordenadas las variables originales, y representando las muestras (coloreadas según el grupo al que pertenecen y la localización de los centroides de cada grupo:

   ```
   kmeans(bioveg,2)->bioveg.cl
   plot(bioveg,col=bioveg.cl$cl)
   text(bioveg,rownames(bioveg),pos=3)
   points(bioveg.cl$ce,pch=17,col=3)
   text(bioveg.cl$ce[,1], bioveg.cl$ce[,2],pos=1,cex=3,col=2)
   ```

   De forma opcional puede utilizarse la función `dibuja.kmeans()`:

`dibuja.kmeans(bioveg,3)`

¿Coinciden los grupos obtenidos con los que se aprecian a simple vista? ¿Ocupan los centroides de los grupos el lugar esperado?

3. Repetir la clasificación pero utilizando tres grupos y considerando la matriz de datos original y la matriz estandarizada?

   `kmeans(bioveg,3)`

   `kmeans(biovege,3,iter.max=100)`

   ¿Se obtienen los mismos resultados? ¿Qué ocurre al repetir los análisis?

   En el caso de que haya diferencias: ¿a qué se pueden deber? ¿Cuál sería el mejor procedimiento: utilizando datos originales o estandarizados?

3.2. Clasificación aglomerativa jerárquica

Ejercicios. Bloque 6:

1. Para realizar la clasificación jerárquica utilizaremos la función `hclust()`. No presenta en pantalla una gran cantidad información. Conviene representarla gráficamente para valorar los distintos elementos que intervienen en la construcción del dendrograma.

   `bioveg.cl<-hclust(dist(bioveg))`

   `plot(bioveg.cl,hang=-1)`

   `bioveg.cl$merge`

   ¿que significan `$merge`, `$height` y `$method`?

2. Comparar el resultado anterior con el que se obtiene considerando la matriz estandarizada utilizando una nueva ventana para la representación:

   `x11()`

   `plot(hclust(dist(biovege)),hang=-1)`

   ¿Por qué no se obtienen los mismos resultados?

3. Tras el análisis del dendrograma es posible decidir el número de grupos que se desea considerar. Podemos basar nuestro criterio en la distancia o en el número de grupos deseado. Para obtener un vector que describa la pertenencia de cada muestra a una de las clases finales a considerar utilizaremos:

   `cutree(bioveg.cl,h=5)`

   `cutree(bioveg.cl,k=2)`

   ¿Coinciden los resultados? ¿Por qué?

3.3. Criterios de agregación

Ejercicios. Bloque 7:

Utilizando los datos de la matriz de distancias `eurodist` que reflejan las distancias geográficas entre distintas capitales de Europa. Compararemos los distintos métodos de agregación `single`, `complete` (que se toma como opción por defecto), `centroid`, `mcquitty` y `ward`, analizando el efecto que producen en los resultados obtenidos.

1. Cargar los datos de `eurodist` que se encuentran entre los ejemplos proporcionados por el sistema:

   `data(eurodist)`

   ¿Qué son la filas y columnas de la matriz `eurodist`?

2. Realizar una clasificación mediante el criterio del vecino más próximo (*single linkage*):

   `plot(hclust(eurodist,method="single"),hang=-1)`

   ¿Qué aspecto general tiene el dendrograma obtenido?

   Realizar una clasificación mediante el criterio del vecino más alejado (*complete linkage*):

   `plot(hclust(eurodist,method=çomplete"),hang=-1)`

   ¿A qué distancia está Madrid de Lisboa? ¿A qué distancia está Madrid de París? Según este dendrograma ¿qué está mas cerca Marsella de Estocolmo o de Hamburgo? ¿que está más cerca Marsella de Milán o de Viena?

   Realizar una clasificación mediante el criterio del centroide o UPGMC (*unweighted pair group method using median centroid*):

   `plot(hclust(eurodist,method=çentroid"),hang=-1)`

   ¿Presenta el dendrograma alguna "anomalía"? ¿Cómo puede explicarse la jerarquía para Madrid-Lisboa-Gibraltar?

   Realizar una clasificación mediante el criterio WPGMA (*weighted pair group method using median average*):

   `plot(hclust(eurodist,method="mcquitty"),hang=-1)`

   ¿Se obtiene resultados "políticamente correctos"?

   Realizar una clasificación mediante el criterio de mínima varianza:

   `plot(hclust(eurodist,method="ward"),hang=-1)`

   ¿A qué distancia está Madrid de Lisboa? ¿A qué distancia está Madrid de París? ¿Qué variación se ha producido en el eje de ordenadas con respecto a los dendrogramas anteriores? ¿Refleja el dendrograma grupos de ciudades de regiones reales?

---

4. Ejercicios adicionales

1. En el fichero `biom2003.dat` disponemos de la información morfológica correspondiente a los alumnos de la asignatura EMC durante el curso 2003–2004- Consideramos para cada individuo las siguientes variables cuantitativas: *peso* (`Peso`), *altura* (`Altura`), *talla del calzado* (`Pie`), *anchura de hombros* (`Hombros`), *longitud de brazo* (`Brazo`) y *perímetro de caderas* (`Caderas`); y las siguientes variables cualitativas: *sexo* (`Sexo`, ♀=1, ♂=2), *color de ojos* (`Ojos`, claros=1, oscuros=2) y *morfotipo* (`Tipo` G=1, R=2, S=3).

   Además existe una variable de control que indica el *grupo de prácticas* (`Grupo`) del que proceden los individuos.

   Realizar una clasificación por el método de K–means, considerando sólo las variables cuantitativas (de la 2 a la 8).

   Contrastar la eficacia de la técnica para separar los individuos de sexo masculino y femenino, utilizando 2, 5 y 8 grupos. Para este objetivo se cruzará la variable `Sexo` con el resultado de la clasificación. Dada el comportamiento de las variables conviene repetir los ensayos para verificar la estabilidad de los resultados.

   `table(Sexo,grupos)`

   ¿Se obtiene grupos formados por alumnos o alumnas de forma exclusiva? ¿Cuales son las principales diferencias estadísticas entre los grupos obtenidos para las distintas variables? ¿En que casos se obtienen los mejores resultados? ¿Se da algún grupo de características "especiales"?

   Incluya sus datos en la matriz:

   `rbind(biom,c(0,` su peso, su estatura, ... `))`

   sus datos aparecen asignados al individuo 99. Puede repetirse la orden para incorporar otros individos. Obtener ahora la clasificación y comprobar que sus datos lo situan el el grupo adecuado. ¿Qué condiciones morfológicas lo habrían situado en un grupo "inadecuado"?

2. Utilizando la matriz de datos del fichero `iris.dat` realizar una clasificación por $K - means$ y posteriormente realizar una clasificación jerárquica de los centroides.

   ¿Muestran los resultados los grupos de ejemplares la adecuada clasificación en relación con la variable `especie`? ¿Cuál es la proporción de fallos?

   ¿Permite la clasificación jerárquica obtenida una adecuada interpretación de los grupos proporcionados por la partición previa?