Introducción a las técnicas estadísticas de clasificación y regresión.

Aprendizaje no supervisado - *Clustering*

Laura Aspirot, Sebastián Castro

Universidad de la República (UdelaR)

Jueves 21 y viernes 22 de febrero de 2013

Índice general

Aprendizaje no supervisado

- Hasta aquí nos hemos concentrado en predecir el valor de una variable de respuesta Y, continua (regresión) o categórica (clasificación), para un conjunto dado de variables predictoras $X = (X_1, \ldots, X_p)$.
- Se dice que este problema es de *aprendizaje supervisado* ya que a partir de una muestra de entrenamiento donde se conocen los valores de (X, Y) se construye una función que permite predecir el valor desconocido de Y para una nueva observación X.
- En el caso de *aprendizaje no supervisado* se dispone únicamente de valores de X. Esto es, por ejemplo, que no hay etiquetas de clase que identifiquen las observaciones.

Agrupamiento o clustering

- En estos casos, uno de los objetivos del análisis consiste en investigar si las observaciones pueden ser organizadas en grupos o *clusters* "homogéneos".
- Es decir, donde las observaciones pertenecientes a un mismo grupo sean, en algún sentido, más similares o próximas que observaciones de distintos grupos.
- A diferencia de los problemas de clasificación entonces, la (posible) estructura de grupos es desconocida a priori, incluyendo el número de clases o clusters.

Agrupamiento o clustering (cont.)

- Posibles aplicaciones son:
  - en marketing, para segmentar el mercado en pequeños grupos homogéneos donde realizar campañas publicitarias específicas;
  - en biología, para dividir organismos en estructuras jerárquicas con el propósito de describir la diversidad biológica;
  - en medicina, para diseñar tratamientos específicos para distintos grupos de riesgo;
  - en psicología, para clasificar individuos en distintos tipos de personalidad, etc.
- Las técnicas de clustering esencialmente intentan extender y formalizar lo que los seres humanos observan muy bien en dos o tres dimensiones.

Agrupamiento o clustering (cont.)

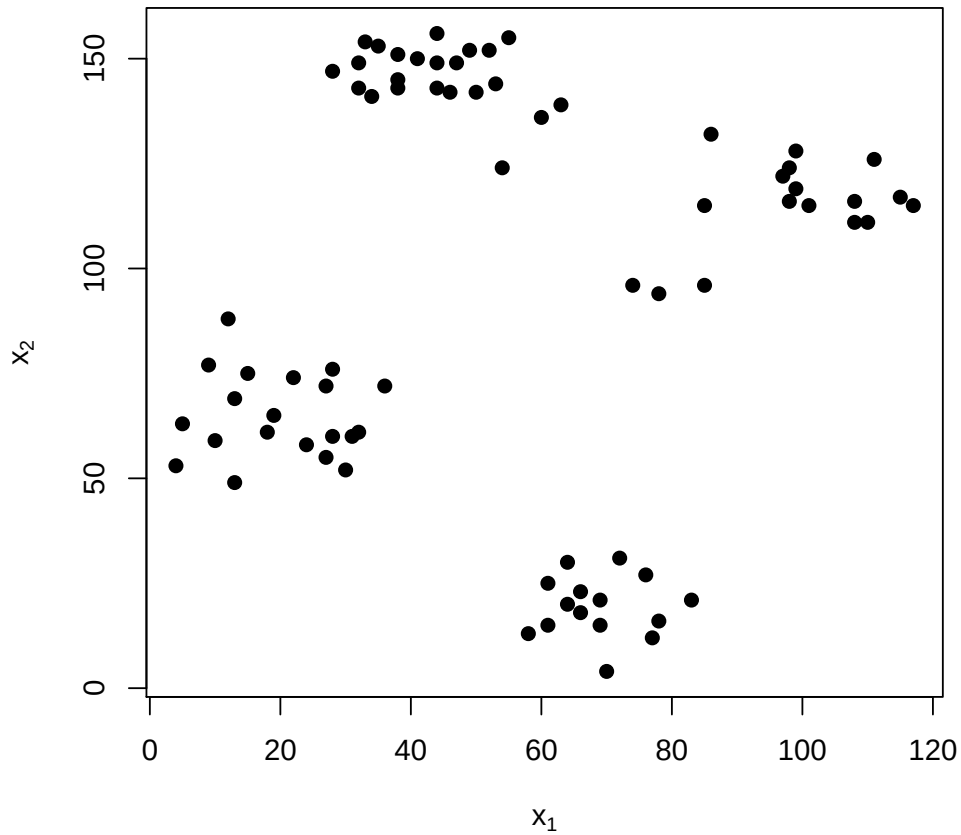

Figura : Ejemplo en dos dimensiones donde los grupos parecen "naturales".

Agrupamiento o clustering (cont.)

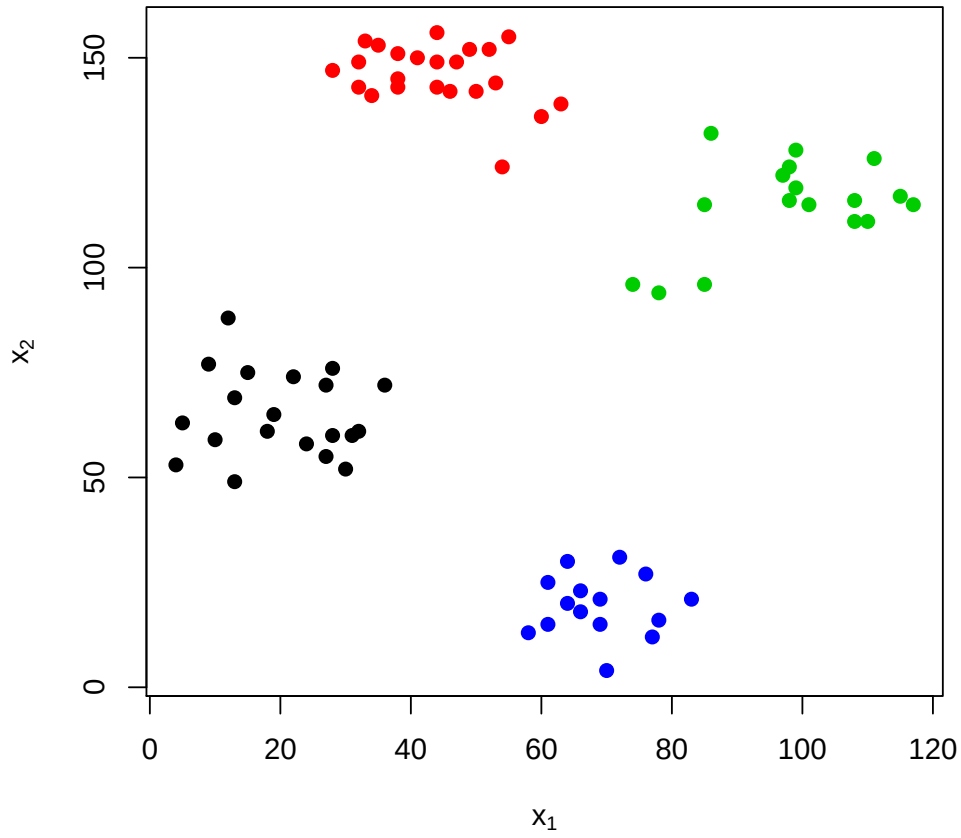

Figura : Grupos obtenidos mediante un algoritmo de clustering (PAM).

Agrupamiento o clustering (cont.)

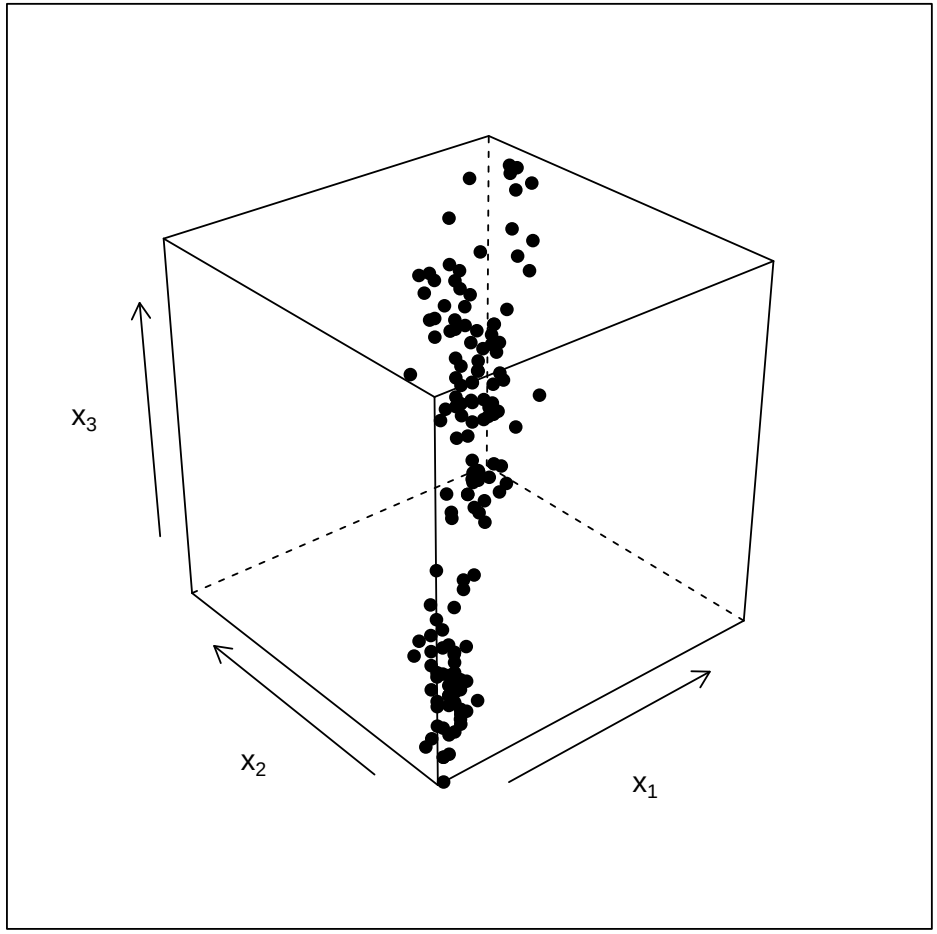

Figura : Ejemplo en tres dimensiones sobre los datos `iris`.

Agrupamiento o clustering (cont.)

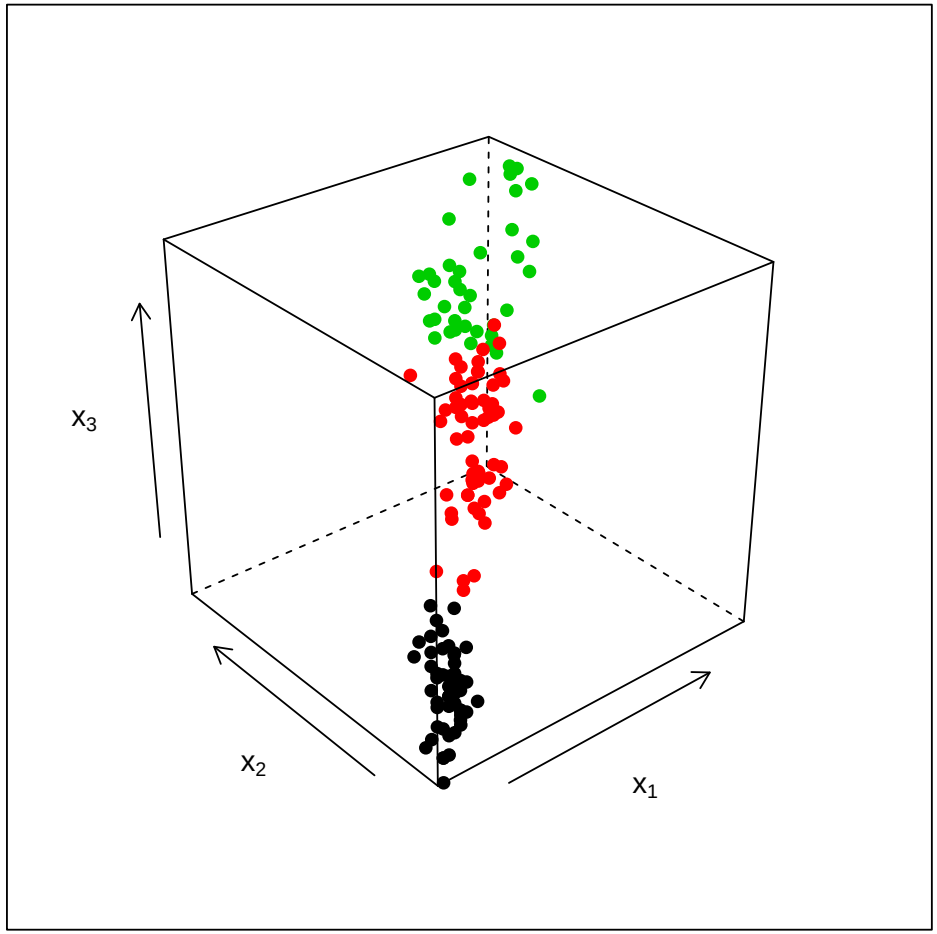

Figura : Grupos obtenidos mediante un algoritmo de clustering (PAM).

Agrupamiento o clustering (cont.)

- Por ello, una posible solución al problema de encontrar grupos consiste en proyectar los datos en dos o tres dimensiones de forma tal que la estructura global se mantenga lo mejor posible y los grupos puedan ser identificados visualmente.
- Esto es lo que se hace con el *análisis de componentes principales* (ACP).
- Otra posibilidad consiste en representar gráficamente los datos multivariados de forma tal que puedan ser agrupados manualmente.
- Un ejemplo está dado por las *caras de Chernoff*, teniendo en cuenta que los humanos están por lo general bien entrenados para reconocer parecidos en los rostros (Dillon y Goldstein, 1984; Varmuza y Filmoser, 2009).

Ejemplo: 16 observaciones en 8 variables

$$
\begin{aligned}
x_1 &= (156, 592, 992, 263, 178, 553, 639, 75) \\
x_2 &= (968, 258, 357, 957, 539, 8, 618, 132) \\
x_3 &= (468, 632, 545, 834, 1, 340, 146, 329) \\
x_4 &= (775, 517, 239, 113, 307, 665, 799, 763) \\
x_5 &= (407, 864, 708, 519, 73, 384, 562, 740) \\
x_6 &= (537, 595, 678, 221, 563, 218, 171, 654) \\
x_7 &= (206, 724, 709, 570, 304, 812, 892, 870) \\
x_8 &= (186, 785, 771, 674, 735, 951, 761, 29) \\
x_9 &= (774, 881, 722, 613, 204, 886, 636, 451) \\
x_{10} &= (193, 91, 608, 177, 751, 147, 921, 497) \\
x_{11} &= (430, 560, 238, 956, 188, 924, 437, 891) \\
x_{12} &= (3, 974, 528, 67, 469, 673, 311, 565) \\
x_{13} &= (825, 918, 225, 949, 482, 872, 182, 550) \\
x_{14} &= (820, 213, 247, 581, 78, 362, 190, 292) \\
x_{15} &= (944, 464, 910, 159, 248, 371, 796, 166) \\
x_{16} &= (935, 44, 817, 137, 927, 556, 358, 191)
\end{aligned}
$$

Agrupamiento o clustering

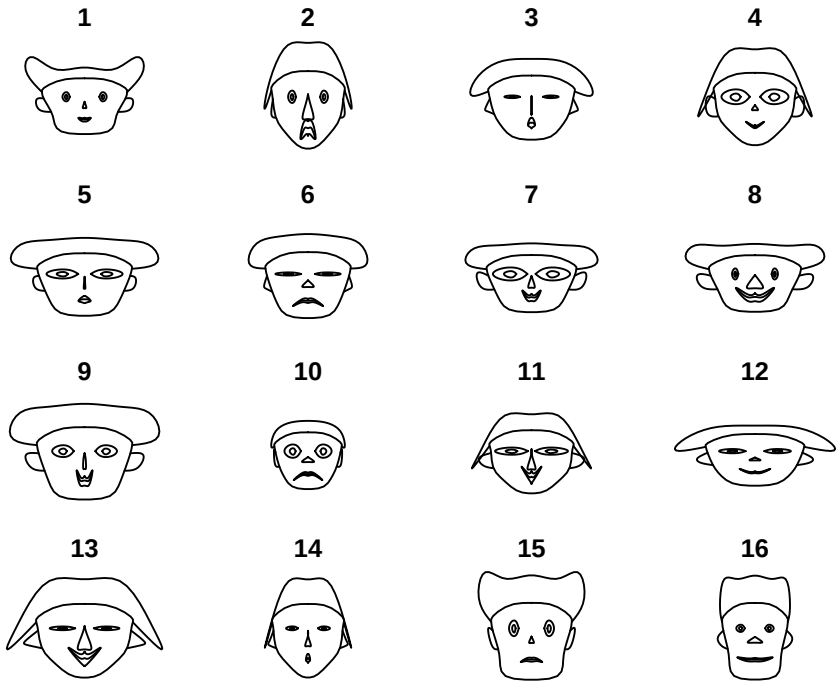

Figura : Ejemplo de representación de datos multivariados ($p = 8$) mediante *caras de Chernoff*.

Agrupamiento o clustering (cont.)

- Sin embargo, en esta exposición nos centraremos en describir técnicas "más automáticas" de clustering dado que son las más apropiadas para trabajar problemas en *grandes dimensiones* (número elevado de observaciones y/o variables).
- Nos concentraremos especialmente en tres tipos de técnicas:
  - de *particionamiento* (K-medias, PAM),
  - *jerárquicas* (algomerativas y divisivas),
  - *basadas en modelos* (distribuciones mezcladas).
- Un aspecto central en todas ellas es la noción de *similaridad* o *disimilaridad* entre los objetos a ser agrupados. Para ello es necesario en muchos casos utilizar medidas de distancia entre los datos.

Medidas de disimilaridad. Distancias.

- Para variables cuantitativas, las distancias entre $\mathrm{x}_i = (x_{i1}, \ldots, x_{ip})$ y $\mathrm{x}_j = (x_{j1}, \ldots, x_{jp}) \in \mathcal{R}^p$ más utilizadas son:
  - *euclídea* o L_2: $\mathrm{d}(\mathrm{x}_i, \mathrm{x}_j) = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right)^{1/2}$,
  - *Manhattan* o L_1: $\mathrm{d}(\mathrm{x}_i, \mathrm{x}_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$,
  - *Minkowski* o L_q $(q \geq 1)$: $\mathrm{d}(\mathrm{x}_i, \mathrm{x}_j) = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^q\right)^{1/q}$
- La distancia euclídea es la más utilizada en la práctica. La distancia L_1 es más robusta frente a la presencia de datos atípicos (*outliers*). La distancia de Minkowski es una generalización de las otras dos.
- Estas medidas de distancia no son invariantes frente a cambios de escala de las variables. Por ello, en algunos casos se estandarizan los datos previamente (media cero y variancia 1).

Ejemplo numérico en R

```
> (x1 <- c(-1, 2, 3.5, sqrt(2), -5))
[1] -1.000000  2.000000  3.500000  1.414214 -5.000000
> (x2 <- c(2, 8, 6, pi, -5))
[1]  2.000000  8.000000  6.000000  3.141593 -5.000000
>
> # distancia 'euclídea' o L2
> sqrt(sum( (x1 - x2)^2 ))
[1] 7.364363

> # distancia 'Manhattan' o L1
> sum( abs(x1 - x2) )
[1] 13.22738

> # distancia 'Minkowski' o Lq
> q <- 2.5 # por ejemplo
> (sum( (abs(x1 - x2))^q))^(1/q)
[1] 6.731693
```

Medidas de disimilaridad. Distancias.

- Existen otras distancias (como la de *Mahalanobis*) para variables cuantitativas. Distintas distancias a menudo dan resultados diferentes (incluso para una misma técnica).
- Para otro tipo de variables (categóricas ordinales o nominales), se utilizan por lo general otro tipo de disimilaridades (basadas en el número de coincidencias, por ejemplo) (Varmuza y Filmoser, 2009).
- La elección de la medida de distancia a utilizar puede ser tan o más importante que la técnica de cluster (Hastie y otros, 2009).

Medidas de disimilaridad. Distancias. (cont.)

- Sin embargo, esta elección no es sencilla y requiere por lo general conocimiento específico del problema a resolver.
- Dadas n observaciones $x_1, \ldots, x_n \in \mathcal{R}^p$, algunas técnicas de clustering requieren el cálculo de las distancias entre todo par de observaciones x_i, x_j.
- Esta información puede representarse mediante una matriz de distancias, de dimensión $n \times n$, $D = ((d_{ij}))$, simétrica, donde $d_{ij} = d(x_i, x_j)$.

Ejemplo numérico en R

```
> X <- matrix(sample(-5:5, 200, rep = T), ncol = 20)

> dim(X)
[1] 10 20

> head(X, 5) # primeras 5 filas

     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,]    4    2    4    5   -4   -1   -1   -2    3     2    -1    -4    -4    -1
[2,]   -1    2    2    5   -3    5   -2   -4    3    -5    -1     4     0    -3
[3,]    4   -5    1   -3   -5    2    2    5    3     4    -3     2     5    -5
[4,]   -2   -1    5    4   -5   -4    0    2   -2    -5     3    -4    -3     0
[5,]   -3    5   -3   -2   -4   -2    2   -1    5     0     2     0    -4     3

     [,15] [,16] [,17] [,18] [,19] [,20]
[1,]    -2    -4    -1    -1     2    -4
[2,]    -5     0    -3     4    -1     3
[3,]     2    -3     3     4     2    -3
[4,]     0     4    -4    -2    -2     3
[5,]    -2    -3     3    -4    -3    -5
```

Ejemplo numérico en R (cont.)

```
> round(dist(X, 'euclídean', diag = T), 3)
        1      2      3      4      5      6      7      8      9     10
1   0.000
2  17.776  0.000
3  19.748 22.494  0.000
4  17.578 18.628 24.960  0.000
5  16.371 21.307 22.847 20.664  0.000
6  19.000 20.952 19.570 22.450 21.331  0.000
7  19.442 19.799 20.396 20.712 20.248 21.424  0.000
8  14.491 19.545 25.219 21.190 16.613 20.372 23.409  0.000
9  19.287 21.260 21.166 20.421 19.647 27.368 20.543 21.260  0.000
10 18.138 22.694 17.349 20.100 15.330 18.708 21.703 19.975 18.735  0.000
```

Técnicas de clustering

- Como se mencionó anteriormente, el objetivo del análisis de clustering es separar las observaciones en grupos (*clusters*) de manera que las disimilaridades entre observaciones del mismo cluster tiendan a ser más pequeñas que aquéllas de distintos clusters.
- Sin embargo, no existe una definición universalmente aceptada de qué es exactamente un cluster.
- Por eso, usualmente no es de esperar una única solución. Los resultados dependen de la medida de disimilaridad utilizada, el algoritmo de clustering y los parámetros elegidos.

Técnicas de particionamiento

- Estas técnicas buscan particionar el conjunto de datos en un número especificado de grupos K, $1 \leq K \leq n$, minimizando algún criterio o función objetivo que indica la "bondad" (en términos del objetivo del clustering) de cada partición.

- El enfoque más utilizado consiste en buscar la partición $C = \{C_1, \ldots, C_K\}$ de n individuos en K grupos de forma tal de minimizar la suma de las distancias respecto a un representante o *centroide* del grupo, $\mathrm{c}(k)$,:

$$SCD(C) = \sum_{k=1}^{K} \sum_{\mathrm{x}_i \in C_k} \mathrm{d}(\mathrm{x}_i, \mathrm{c}(k))$$

- Cuando la distancia utilizada es la euclídea y el centroide es el promedio de las observaciones del grupo, este es el criterio de *mínima suma de cuadrados* dentro de los grupos.

Técnicas de particionamiento (cont.)

- El problema parece entonces relativamente simple: considerar todas las particiones posibles de n individuos en K grupos y seleccionar aquélla con el valor más bajo de $SCD(C)$.
- Desafortunadamente, en la práctica esta solución no es viable. ¡El número total de particiones a considerar, $S(n, K)$, es por lo general tan elevado que la enumeración completa es imposible incluso para la computadora más rápida!
- Se puede demostrar que:

$$S(n, K) = \frac{1}{K!} \sum_{k=1}^{K} (-1)^{K-k} \binom{K}{k} k^n$$

(Hastie y otros, 2009).

Técnicas de particionamiento (cont.)

n	K	$S(n, K)$
15	3	2375101
20	4	$\approx 4.52 \times 10^{10}$
25	8	$\approx 6.9 \times 10^{17}$
100	5	$\approx 6.57 \times 10^{67}$

Cuadro : Número de posibles particiones según la cantidad de observaciones n y grupos K.

- Debido a este problema, los distintos algoritmos que se han desarrollado solo son capaces de examinar una fracción (a menudo muy) menor de las posibles particiones.

Técnicas de particionamiento (cont.)

- Estas estrategias están basadas en algoritmos iterativos del tipo “greedy”.
- Se especifica (o sortea) una partición inicial y en cada iteración se modifica la asignación de grupos de forma tal que el valor de la función objetivo disminuya.
- El objetivo es identificar y explorar un pequeño subconjunto de las particiones posibles que contenga la mejor o, al menos, una buena partición subóptima.

Técnicas de particionamiento (cont.)

- Estos algoritmos si bien pueden ser muy eficientes computacionalmente, no garantizan encontrar el óptimo global de $SCD(C)$ y pueden ser dependientes de la partición inicial elegida.
- La dificultad del problema y de encontrar algoritmos generales que sean eficientes, así como la gran disponibilidad de datos que existe actualmente, explican la cantidad de algoritmos que existen y de líneas de investigación abiertas.

Algoritmo de K-medias

- El algoritmo de K-medias es un algoritmo particional y fue propuesto en los '50 (Jain, 2009)
- A pesar de que su primera aparición es desde hace más de 50 años sigue siendo de los algoritmos más utilizados para clustering por su facilidad de implementación, simpleza y buenos resultados empíricos.

Algoritmo de K-medias (cont.)

- Los principales pasos del algoritmo son los siguientes:
  1. seleccionar una partición inicial (determinada por los centros de los clusters) con K clusters, repetir los pasos 2 y 3 hasta que los clusters se estabilicen;
  2. generar una nueva partición asignando cada dato al cluster cuyo centro está más cercano;
  3. calcular los nuevos centros de los clusters $c(1), \ldots c(K)$ (promediando los datos asignados a ese cluster en el paso anterior si la distancia es la euclídea);

- El algoritmo K-medias requiere del usuario los siguientes parámetros:
  - número de clusters,
  - incialización de los clusters (centros),
  - distancia (en general la distancia euclídea).

Algoritmo de K-medias (cont.)

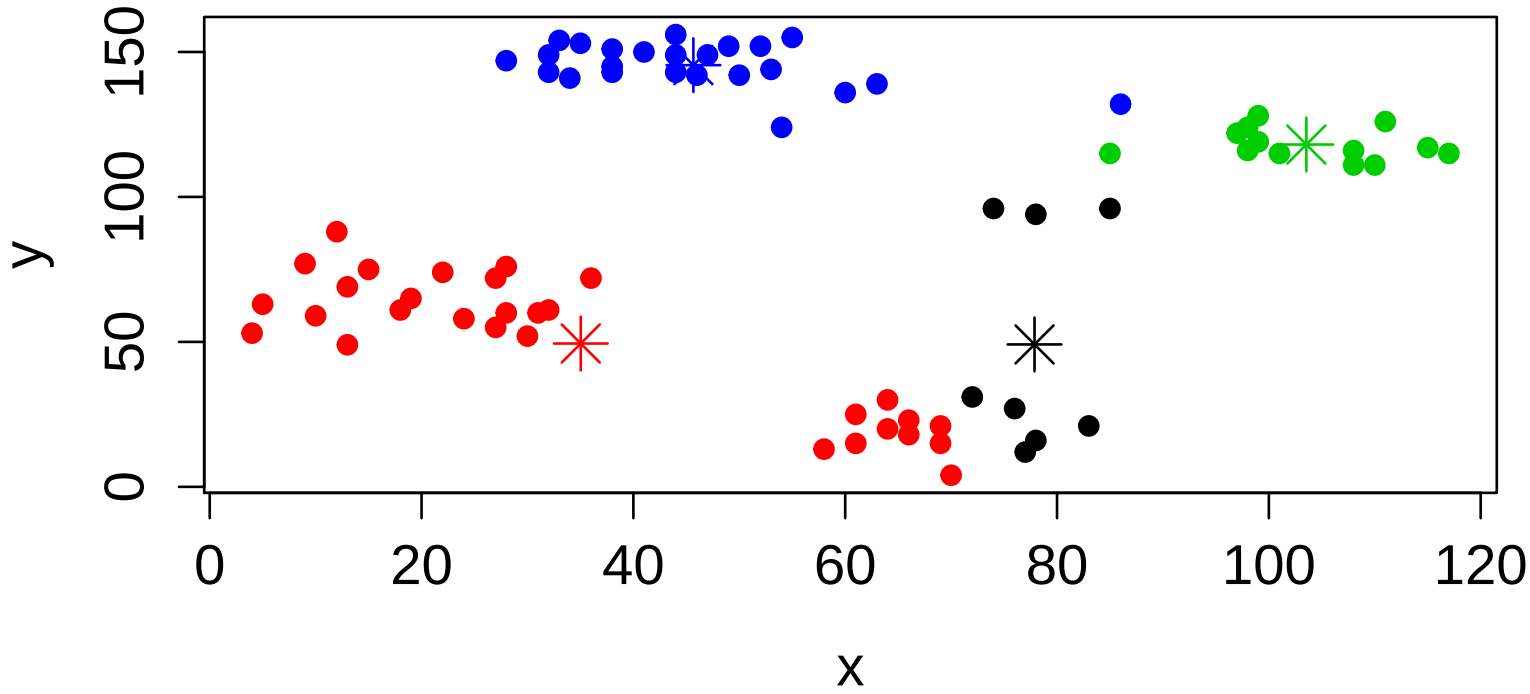

Figura : Grupos obtenidos en la iteración 1 de K-medias.

Algoritmo de K-medias (cont.)

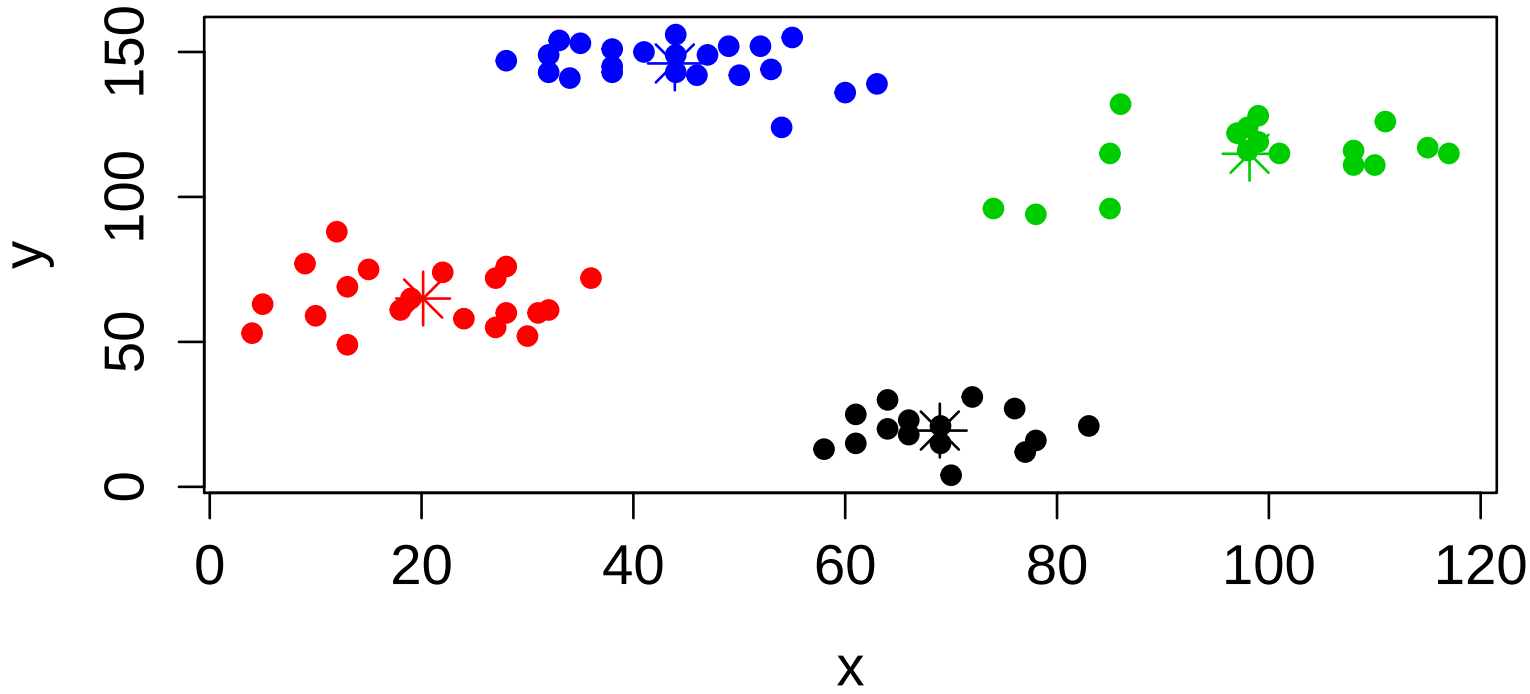

Figura : Grupos obtenidos en la iteración 2 de K-medias.

Algoritmo de K-medias (cont.)

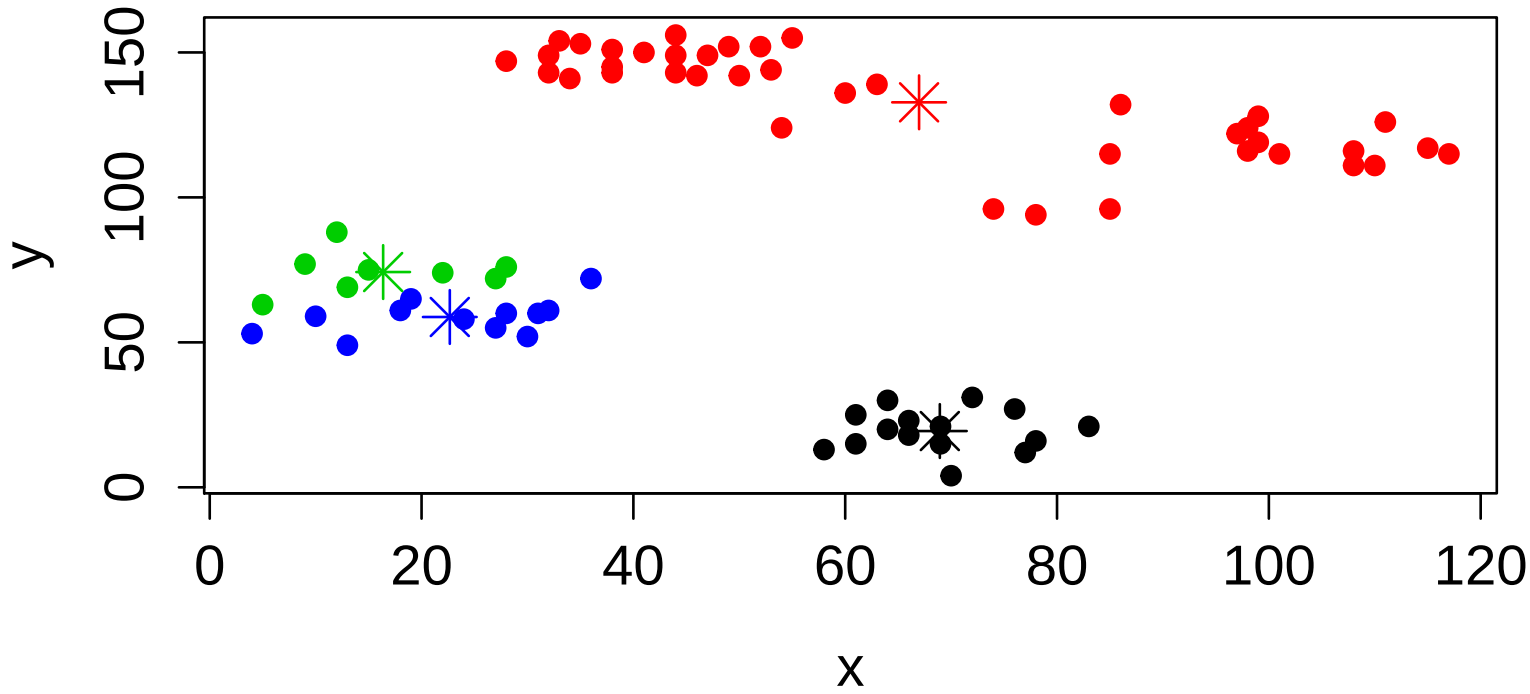

Figura : Grupos obtenidos en la iteración 10 de K-medias. Los grupos se estabilizan en un óptimo local.

Algoritmo de K-medias (cont.)

- Algunos incovenientes de K-medias y los métodos particionales:
  - hay que elegir el número de clusters
  - son sensibles a la inicialización
- Algunas soluciones:
  - para el número de clusters: visualización, consideraciones sobre el problema, criterios de penalización (penalizan la cantidad de clusters)
  - inicialización: como el algoritmo puede converger a mínimos locales, una solución es correr el algoritmo con diferentes particiones iniciales aleatorias y elegir el resultado con menor función objetivo.

Algoritmo de K-medias (cont.)

- Otro de los problemas que surge es el de validación de los clusters obtenidos. Por ejemplo el algoritmo de K-medias siempre encuentra clusters, no importa si los hay o no.
- Algunos criterios de validez de los clusters:
  - criterios internos: solo se basan en los datos y en el algoritmo usado. Una medida interna es la noción de estabilidad de los clusters, donde se mide la variablidad de los clusters al aplicar el algoritmo a diferentes submuestras de los datos;
  - criterios relativos: se comparan diferentes estructuras, por ejemplo se hacen clusters con diferentes algoritmos y se elige cuáles son mejores en algún sentido;
  - criterios externos: se valida en base a información a priori de etiquetas de los datos (aunque en el caso de existir etiquetas se podrían usar métodos supervisados).

Una variante robusta: algoritmo PAM

- Al igual que K-medias, el algoritmo de *partitioning around medoids* (PAM) necesita una configuración (partición) inicial y un número preespecificado de grupos.
- Por otro lado, PAM busca los K "individuos representativos" (o *medoids*) entre el conjunto de observaciones (mientras que K-medias utiliza los promedios del grupo), que minimizan la suma de las disimilaridades al resto de los integrantes.
- En general PAM es más robusto que K-medias y requiere como argumento de entrada solamente la matriz de disimilaridades entre observaciones y no los datos originales.
- Como contrapartida es más intensivo computacionalmente, debido principalmente a la búsqueda de *medoids* (Izenman, 2008).

Técnicas jerárquicas

- Los resultados de aplicar las técnicas de particionamiento (K-medias o PAM por ejemplo), dependen de la elección del número de clusters y de una configuración inicial.
- En cambio, los métodos de *clustering jerárquico* no requieren estas especificaciones.
- En su lugar, estas técnicas necesitan que el usuario especifique una *medida de disimilaridad entre grupos* (disjuntos), basada en las disimilaridades entre las observaciones de los grupos.

Técnicas jerárquicas (cont.)

- Como su nombre sugiere, estas técnicas producen representaciones jerárquicas en las cuales los clusters en cada nivel de la jerarquía son creados uniendo clusters del siguiente nivel inferior.
- En el nivel más bajo cada cluster contiene una única observación y en el más alto hay un sólo cluster con todas las observaciones.
- Las estrategias para el clustering jerárquicos se dividen en dos paradigmas básicos: *algomerativos* y *divisivos*.

Métodos aglomerativos

- Estos métodos comienzan con cada observación representando un sólo cluster.
- En cada uno de los siguientes pasos los dos cluster más cercanos (menos disímiles) son unidos en un único grupo, produciendo un cluster menos que en el nivel inmediato anterior.
- Por lo tanto, es necesario definir una medida de disimilaridad entre clusters o grupos de observaciones.
- Distintas medidas dan lugar a variantes que pueden producir resultados diferentes.

Métodos aglomerativos (cont.)

Las tres medidas de disimilaridad más comunes entre grupos de observaciones A y B son:

- *single linkage* o vecino más cercano:

$$d_{\text{mín}}(A, B) = \min_{x_i \in A, x_j \in B} d(x_i, x_j)$$

- *complete linkage* o vecino más lejano:

$$d_{\text{máx}}(A, B) = \max_{x_i \in A, x_j \in B} d(x_i, x_j)$$

- *average linkage*:

$$d_{\text{media}}(A, B) = \frac{1}{|A||B|} \sum_{x_i \in A, x_j \in B} d(x_i, x_j)$$

Métodos aglomerativos (cont.)

- siendo $|A|, |B|$ el número de observaciones en los clusters respectivos.
- Ninguna de estas tres medidas es *uniformemente* mejor que las otras para todos los problemas de clustering (Izenman, 2008).
- El vecino más cercano a menudo produce largas "cadenas" de clusters, unidas por unos pocos puntos cercanos (lo cual podría no ser deseado), mientras que el vecino más lejano tiende a producir muchos clusters, pequeños y compactos.
- La distancia promedio es dependiente del tamaño de los clusters, a diferencia de las otras dos.
- Si los datos presentan una estructura de grupos bien diferenciados, entonces los tres métodos tenderán a dar los mismos resultados.

Métodos divisivos

- Los algoritmos de clustering divisivo comienzan con un único cluster con todas las observaciones y recursivamente dividen uno de los clusters existentes en dos clusters "hijos" hasta obtener tantos grupos como observaciones.
- Dividir un cluster es computacionalmente más demandante que unir dos, dado que no sólo se debe encontrar el cluster a ser dividido sino que también las observaciones que formarán los dos nuevos grupos deben ser identificadas.
- Por este motivo, los métodos divisivos son menos utilizados en la práctica (Varmuza y Filmoser, 2009).

Representación gráfica: *dendograma*

- Los resultados del clustering jerárquico suelen ser representados mediante un diagrama de árbol jerárquico, conocido como *dendograma*.
- Grupos u observaciones que son más similares son combinados a bajas alturas, mientras que los más disímiles lo hacen a alturas grandes.
- Una partición de los datos en un número dado de grupos puede obtenerse "cortando" el dendograma en un nivel apropiado de altura.

Ejemplo numérico en R

```
X <- rbind(c(1, 3), c(2, 4), c(1, 5), c(5, 5), c(5, 7),
 c(4, 9), c(2, 8), c(3, 10))
> X
     [,1] [,2]
[1,]    1    3
[2,]    2    4
[3,]    1    5
[4,]    5    5
[5,]    5    7
[6,]    4    9
[7,]    2    8
[8,]    3   10
> (D <- dist(X))
         1        2        3        4        5        6        7
2 1.414214
3 2.000000 1.414214
4 4.472136 3.162278 4.000000
5 5.656854 4.242641 4.472136 2.000000
6 6.708204 5.385165 5.000000 4.123106 2.236068
7 5.099020 4.000000 3.162278 4.242641 3.162278 2.236068
8 7.280110 6.082763 5.385165 5.385165 3.605551 1.414214 2.236068
```

Ejemplo numérico en R (cont.)

```
> par(mfrow = c(2, 2), mex =  .8)
>
> plot(X, ylim = c(.8 * min(X[, 1]), 1.2 * max(X[, 2])), pch = 19, cex = 1.5,
+ xlab = expression(x[1]), ylab = expression(x[2]))
> text(X[, 1], 1.25 + X[, 2], 1:8, cex = 1.5)
>
> plot(hclust(dist(X), 'single'), main = 'Single linkage', xlab = '',
+ cex = 1.5)
> plot(hclust(dist(X), 'complete'), main = 'Complete linkage', xlab = '',
+ cex = 1.5)
> plot(hclust(dist(X), 'average'), main = 'Average linkage', xlab = '',
+ cex = 1.5)
>
> par(mfrow = c(1, 1))
```

Ejemplo numérico en R (cont.)

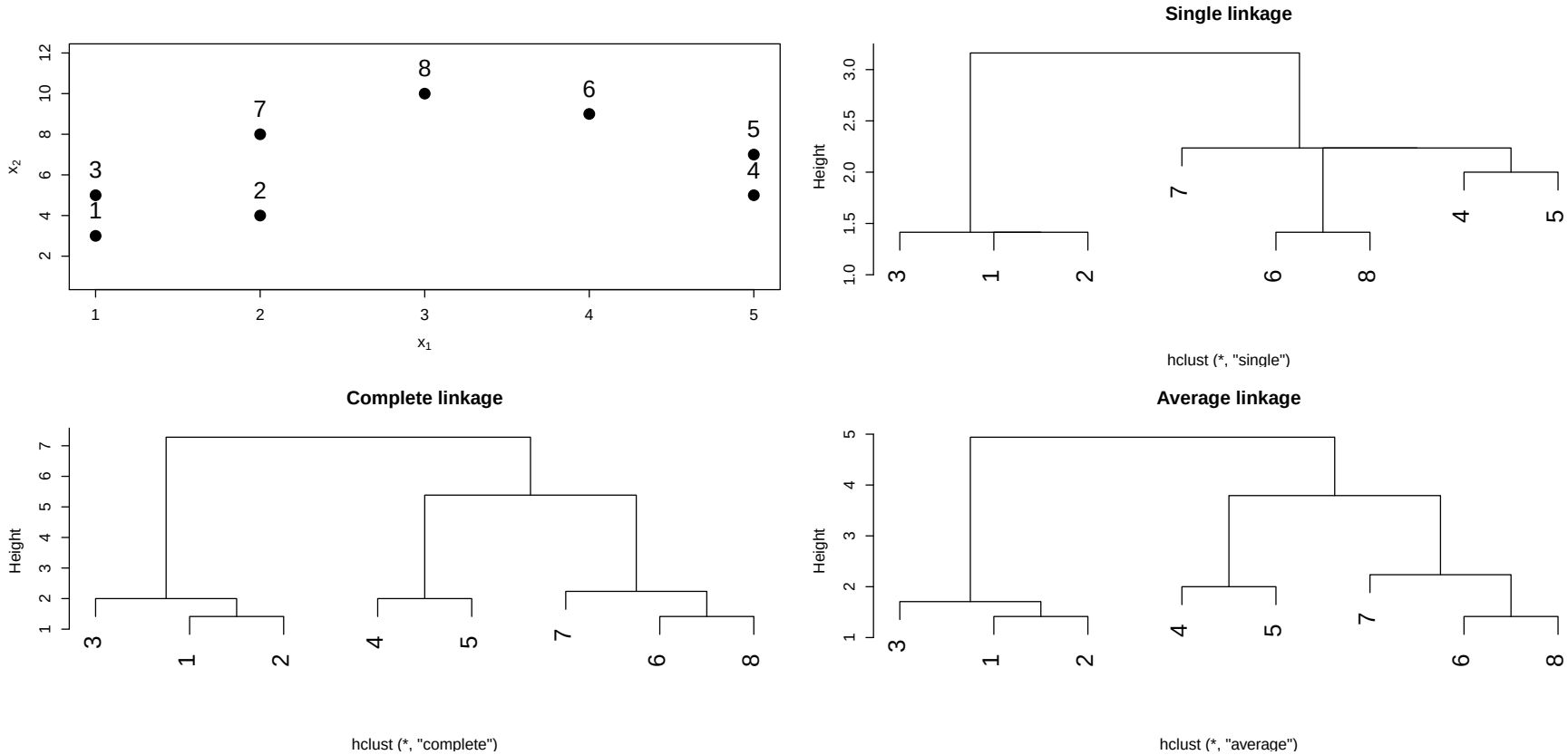

Figura : Pequeño ejemplo de clustering jerárquico aglomerativo.

Técnicas jerárquicas, algunos comentarios

- Las fusiones o divisiones, una vez que son hechas, son irreversibles. De manera que, por ejemplo, cuando un algoritmo aglomerativo puso dos observaciones en un mismo grupo ellas no podrán aparecer en diferentes grupos en etapas posteriores.
- Estos métodos imponen una estructura jerárquica independientemente de que la misma exista o no en los datos.
- Aunque se los ha utilizado en diversas áreas, es en las aplicaciones biológicas donde son más relevantes y justificados (Everitt y Hothorn, 2008).

Clustering basado en modelos

- En las técnicas de clustering presentadas anteriormente no se hace ningún supuesto acerca de la distribución de los datos (excepto que se asume una estructura de grupos).
- En cambio, en las técnicas de clustering *basado en modelos* se asume un modelo estadístico para los clusters, donde cada uno de ellos puede ser representado por una distribución de probabilidad multivariada.
- La distribución conjunta de los datos $x = (x_1, \ldots, x_p)$ se dice que es una *mezcla* de distribuciones con componentes dadas por cada uno de los clusters:

$$f(x|\theta) = \sum_{j=1}^{K} \pi_j f(x|\theta_j)$$

con $\pi_j \geq 0$ y $\sum_{j=1}^{K} \pi_j = 1$.

Clustering basado en modelos (cont.)

- Luego, el algoritmo debe buscar la estimación de los parámetros que "mejor" se adapten a los datos y al modelo propuesto (*máxima verosimilitud*), así como el grado de pertenencia de cada observación a los distintos grupos.
- La estimación de los parámetros del modelo asumido está basada en el algoritmo EM (*Expectation Maximization*).
- Este algoritmo busca maximizar la verosimilitud $f(x|\theta)$ alternando entre un paso de *esperanza* (donde las pertenencias a los grupos son estimadas) y un paso de *maximización* (donde los parámetros del modelo son estimados).

Clustering basado en modelos (cont.)

- El ajuste global de los distintos modelos propuestos puede ser evaluado utilizando el valor de la *función de verosimilitud*, el cual debería ser lo más grande posible.
- En la práctica ésto se hace a través del criterio BIC (*Bayesian Information Criterion*) que penaliza además por la complejidad del modelo.
- Con esta información es posible decidir cuál modelo y número de clusters K, es el más adecuado.

Clustering basado en modelos (cont.)

- En el modelo más simple se asume que cada uno de los K clusters están representados por distribuciones normales (gaussianas) multivariadas con diferentes medias pero iguales matrices de covarianzas de la forma $\sigma^2 \mathbf{I}$.
- Este supuesto produce clusters *esféricos* y de igual tamaño.
- En un modelo un poco más general se plantea que los clusters tienen matrices de covarianza $\sigma_j^2 \mathbf{I}$, para $j = 1, \ldots, K$. Los clusters resultantes son aún esféricos pero pueden tener distintos tamaños.
- En el modelo más general, las matrices de covarianza Σ_j, no tiene porqué ser diagonales (como en los casos anteriores) lo cual permite modelar clusters *elípticos*.

Otros temas no tratados

- *Clustering difuso (fuzzy)*.
  - A diferencia de K-medias, donde cada dato pertence a un único cluster, permite asignar más de un cluster al mismo dato pero con distintos *coeficientes de pertenencia*.
- *Clustering de variables*.
  - En principio se pueden utilizar las mismas técnicas que para agrupar observaciones, siendo la principal diferencia la medida de distancia utilizada (por ejemplo, basada en *correlaciones*).
  - Una de las principales aplicaciones ha sido en el *clustering de genes* en experimentos de microarreglos (Izenman, 2008).

Bibliografía


Dillon, W. y Goldstein, M. 1984 *Multivariate Analysis: Methods and Applications*. Wiley.


Everitt, B. y Hothorn, T. 2008 *A Handbook of Statistical Analyses Using R*, second edition. CRC Press.


Hastie, T., Friedman, J. and Tibshirani, R. 2009 *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, second edition. Springer.


Izenman, A. 2008. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer-Verlag.

Bibliografía (cont.)


Jain, Anil K. 2010. *Data clustering: 50 years beyond K-means*. Pattern Recognition Letters. 31, 8, 651-666.


Varmuza, K. y Filzmoser, P. 2009 *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press.