Aprendizaje No Supervisado
parte II

César Gómez

5 de noviembre de 2020

Métodos de agrupamiento (clustering)

- *Clustering* o **agrupamiento** se refiere a un amplio conjunto de técnicas ideadas para encontrar *subgrupos* o *clusters* en los datos.

Métodos de agrupamiento (clustering)

- *Clustering* o **agrupamiento** se refiere a un amplio conjunto de técnicas ideadas para encontrar *subgrupos* o *clusters* en los datos.
- Cuando agrupamos en clusters las observaciones de un conjunto de datos, se busca particionar el conjunto de datos en *clusters*, de tal forma que las observaciones dentro de un mismo grupo sean lo "**más similares entre si**" en lo posible, mientras que 2 observaciones que se encuentren en dos *subgrupos* o *clusters* distintos sean más diferentes entre si.

- Para hacer esto más concreto, debe definirse lo que se entiende por que “2 observaciones o más ” sean *similares* o *diferentes*. En realidad, esta cuestión resulta de consideraciones *específicas de dominio* que deben basarse en conocimiento de los datos que están siendo estudiados.

- Un ejemplo de aplicación podría ser en el que se consideran n observaciones correspondientes a muestras de tejido de pacientes con cáncer de mama y los p atributos pueden corresponder a medidas clínicas sobre estos tejidos, tales como estado o grado del tumor, o pueden corresponder a medidas de expresión génica.
  Hay razones para creeer que hay heterogeneidad en los datos y que por ejemplo algunas de las muestras pueden corresponder a diferentes *subgrupos* de cáncer de mama desconocidos.

- Otra aplicación se origina en marketing. Donde se tiene acceso a una gran cantidad de medidas (por ejemplo, ingresos medios, ocupación, distancia desde el área urbana más cercana, etc) sobre una gran cantidad de personas.
  Acá el objetivo consiste en crear una *segmentación del mercado* identificando subgrupos entre las personas que pueden ser más receptivos a una forma particular de publicidad o pueden estar más propensos a comprar ciertos productos.

- Existe una gran cantidad de métodos de *clustering*, pero la exposición se centrará en,
  1. Clustering **K-medias**. Acá se comienza con un número pre-especificado de clusters.

- Existe una gran cantidad de métodos de *clustering*, pero la exposición se centrará en,
  1. Clustering **K-medias**. Acá se comienza con un número pre-especificado de clusters.
  2. Clustering **Jerárquico**. Acá se puede obtener un número cualquiera de clusters entre 1 y n (número de datos).
     Ambos algoritmos poseen sus ventajas y desventajas.

Clustering K-medias

- *Clustering* o agrupamiento por K-medias es una forma simple y elegante para particionar un conjunto en un número pre-especificado K de clusters distintos y disjuntos.

Clustering K-medias

- *Clustering* o agrupamiento por K-medias es una forma simple y elegante para particionar un conjunto en un número pre-especificado K de clusters distintos y disjuntos.
- **Notación**: $C_1, \ldots, C_K$ denotarán los conjuntos que contienen los índices de las observaciones en cada cluster. Estos conjuntos poseen 2 propiedades:
  1. $$C_1 \cup C_2 \cup \cdots \cup C_K = \{1, 2, \ldots, n\}$$

     Es decir, cada observación está en alguno de los K clusters.

Clustering K-medias

- *Clustering* o agrupamiento por K-medias es una forma simple y elegante para particionar un conjunto en un número pre-especificado K de clusters distintos y disjuntos.
- **Notación**: $C_1, \ldots, C_K$ denotarán los conjuntos que contienen los índices de las observaciones en cada cluster. Estos conjuntos poseen 2 propiedades:

1. $$C_1 \cup C_2 \cup \cdots \cup C_K = \{1, 2, \ldots, n\}$$

   Es decir, cada observación está en alguno de los K clusters.

2. $$C_k \cap C_{k'} = \emptyset \quad \text{para} \quad k \neq k^{'}.$$

   En otras palabras, los clusters son disjuntos, por lo tanto, una observación no puede estar en 2 clusters distintos a la vez.

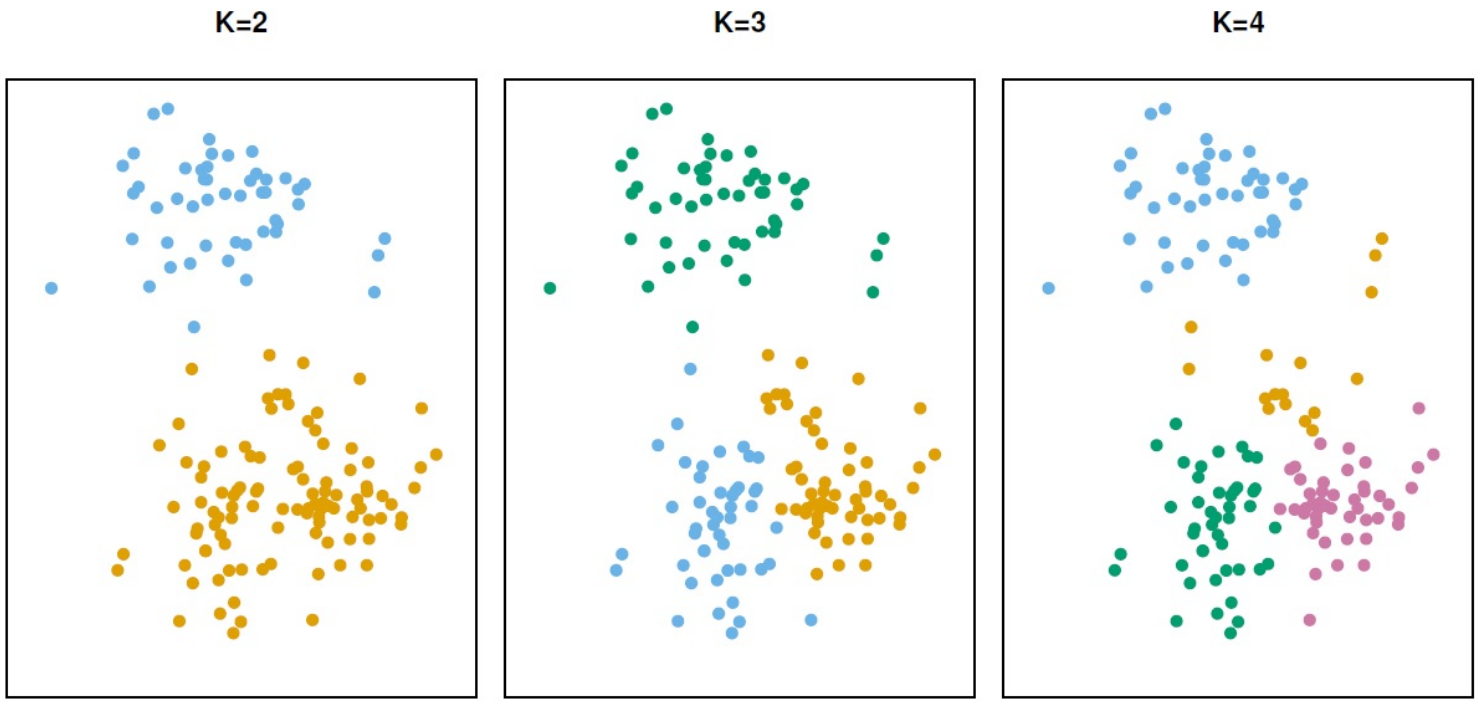

Figura 1: Un conjunto simulado de 150 observaciones en un espacio 2 dimensional, se ilustra el resultado del algoritmo de K-medias con $K = 2$, $K = 3$ y $K = 4$ clusters. La asignación de color es arbitraria y es un resultado del algoritmo.

- La idea tras el clustering de K-medias es que un *buen agrupamiento (clustering)* es uno para el cual la variación intra-cluster sea tan pequeña como sea posible.

- La idea tras el clustering de K-medias es que un *buen agrupamiento (clustering)* es uno para el cual la variación intra-cluster sea tan pequeña como sea posible.
- La *variación intra-cluster* para un cluster particular C_k es una medida $W(C_k)$ de la cantidad por la cual todas las observaciones en un mismo cluster difieren entre si.

- La idea tras el clustering de K-medias es que un *buen agrupamiento (clustering)* es uno para el cual la variación intra-cluster sea tan pequeña como sea posible.
- La *variación intra-cluster* para un cluster particular C_k es una medida $W(C_k)$ de la cantidad por la cual todas las observaciones en un mismo cluster difieren entre si.
- Encontrar un clustering optimo, consiste en resolver el siguiente problema de optimización

$$\underset{C_1,\ldots,C_K}{\text{Minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}. \tag{1}$$

- Resolver (1) parece una idea razonable, pero para llevarla a la práctica es necesario definir la *variación intra-cluster*, la elección más común envuelve la *distancia cuadrática Euclidea*

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \tag{2}$$

donde $|C_k|$ es el número de observaciones en el k-ésimo cluster.

- Resolver (1) parece una idea razonable, pero para llevarla a la práctica es necesario definir la *variación intra-cluster*, la elección más común envuelve la *distancia cuadrática Euclidea*

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \qquad (2)$$

  donde $|C_k|$ es el número de observaciones en el k-ésimo cluster.

- Combinando (1) y (2) se obtiene

$$\underset{C_1, \ldots, C_K}{\text{Minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}. \qquad (3)$$

Algoritmo de clustering K medias

1. Aleatoriamente asigne un número, de 1 a K, a cada una de las observaciones. Estos sirven como asignación inicial de los clusters para las observaciones.

Algoritmo de clustering K medias

1. Aleatoriamente asigne un número, de 1 a K, a cada una de las observaciones. Estos sirven como asignación inicial de los clusters para las observaciones.
2. Iterar los siguientes 2 pasos hasta que la asignación de clusters pare de cambiar:

Algoritmo de clustering K medias

1. Aleatoriamente asigne un número, de 1 a K, a cada una de las observaciones. Estos sirven como asignación inicial de los clusters para las observaciones.
2. Iterar los siguientes 2 pasos hasta que la asignación de clusters pare de cambiar:
   (a) Para cada uno de los K clusters, calcule el **centroide** del cluster. El *centroide* del k-ésimo cluster corresponde al vector de medias de los p atributos para las observaciones en el k-ésimo cluster.

Algoritmo de clustering K medias

1. Aleatoriamente asigne un número, de 1 a K, a cada una de las observaciones. Estos sirven como asignación inicial de los clusters para las observaciones.
2. Iterar los siguientes 2 pasos hasta que la asignación de clusters pare de cambiar:
   (a) Para cada uno de los K clusters, calcule el **centroide** del cluster. El *centroide* del k-ésimo cluster corresponde al vector de medias de los p atributos para las observaciones en el k-ésimo cluster.
   (b) Asignar cada observación al cluster cuyo centroide es el más próximo (en el sentido de la distancia Euclidea).

- Para el algoritmo de K medias se garantiza que el objetivo de optimización (3) siempre decrece en cada iteración.
  Para entender por qué, la siguiente expresión es ilustrativa

$$\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2, \quad (4)$$

donde $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$ es la media del atributo j en el cluster C_k.

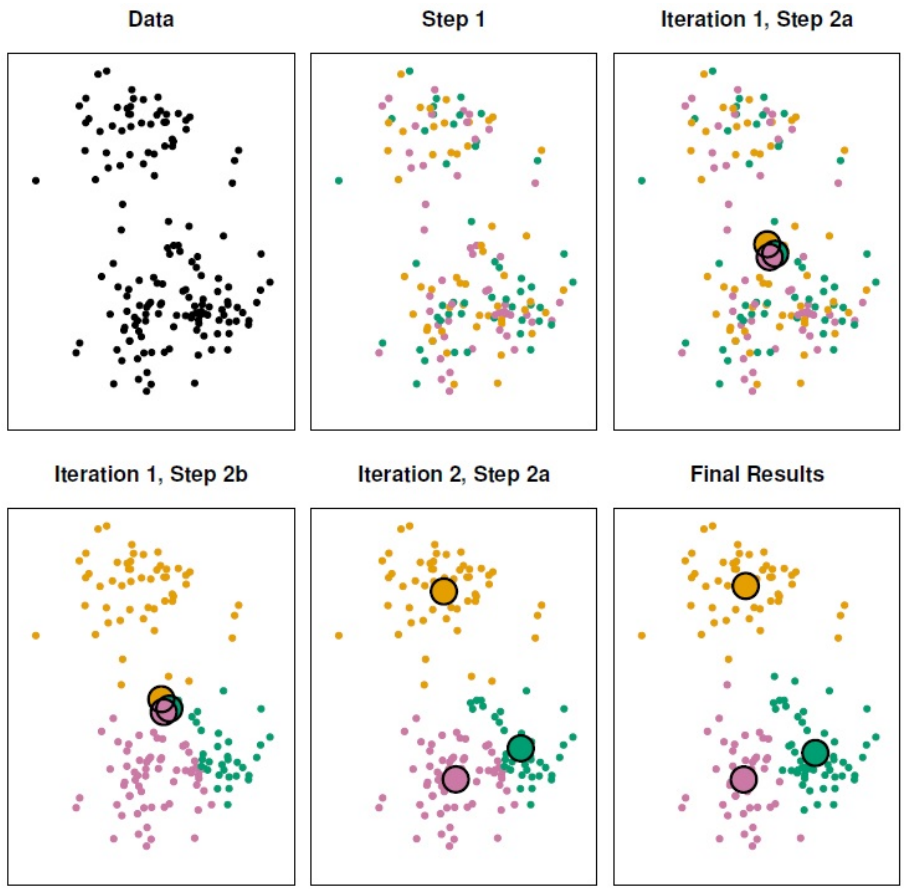
Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Figura 2

- Debido a que el algoritmo de K medias, encuentra un mínimo local en vez de uno global, los resultados obtenidos dependen de la asignación inicial de clusters.

- Debido a que el algoritmo de K medias, encuentra un mínimo local en vez de uno global, los resultados obtenidos dependen de la asignación inicial de clusters.
- Por esta razón es importante correr el algoritmo, múltiples veces comenzando con diferentes configuraciones iniciales.

- Debido a que el algoritmo de K medias, encuentra un mínimo local en vez de uno global, los resultados obtenidos dependen de la asignación inicial de clusters.
- Por esta razón es importante correr el algoritmo, múltiples veces comenzando con diferentes configuraciones iniciales.
- Entonces se selecciona la mejor solución, aquella para la cual el objetivo (3) alcance el menor valor.

Figura 3: K medias llevado a cabo 6 veces con $K = 3$, los números corresponden al valor del objetivo en la ecuación (3).

Clustering jerárquico

- Una desventaja potencial del algoritmo de K medias consiste se comienza siempre un un número pres-especificado K de clusters.

Clustering jerárquico

- Una desventaja potencial del algoritmo de K medias consiste se comienza siempre un un número pres-especificado K de clusters.
- Clustering jerárquico es una alternativa que no requiere el compromiso de especificar un número de clústers desde el inicio.

Clustering jerárquico

- Una desventaja potencial del algoritmo de K medias consiste se comienza siempre un un número pres-especificado K de clusters.
- Clustering jerárquico es una alternativa que no requiere el compromiso de especificar un número de clústers desde el inicio.
- Una ventaja adicional que posee el clustering jerarquico sobre el algoritmo de K-medias consiste en que el resultado puede representarse en un diagrama de árbol denominado **dendrograma**.

Clustering jerárquico

- Una desventaja potencial del algoritmo de K medias consiste se comienza siempre un un número pres-especificado K de clusters.
- Clustering jerárquico es una alternativa que no requiere el compromiso de especificar un número de clústers desde el inicio.
- Una ventaja adicional que posee el clustering jerarquico sobre el algoritmo de K-medias consiste en que el resultado puede representarse en un diagrama de árbol denominado **dendrograma**.
- Se expondrá el **clustering aglomerativo**, que es el tipo de clustering jerárquico más común.

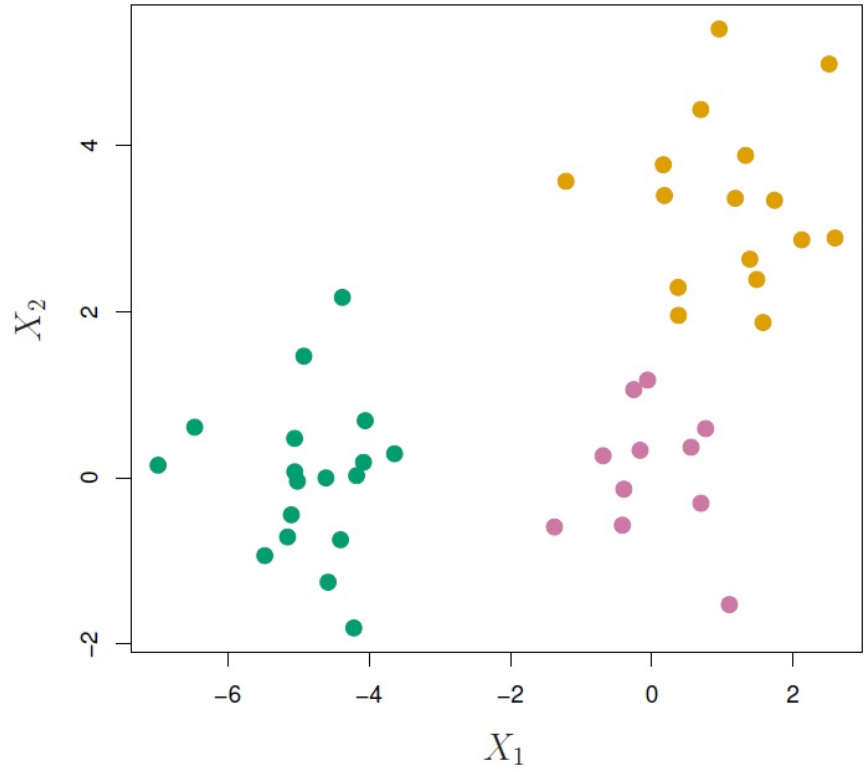

Figura 4: 45 observaciones generadas en un espacio 2 dimensional, en realidad hay 3 clases distintas, identificadas por los colores. Ahora bien se supondrán desconocidas estas clases y se buscará encontrar cluster para descubrir las clases.

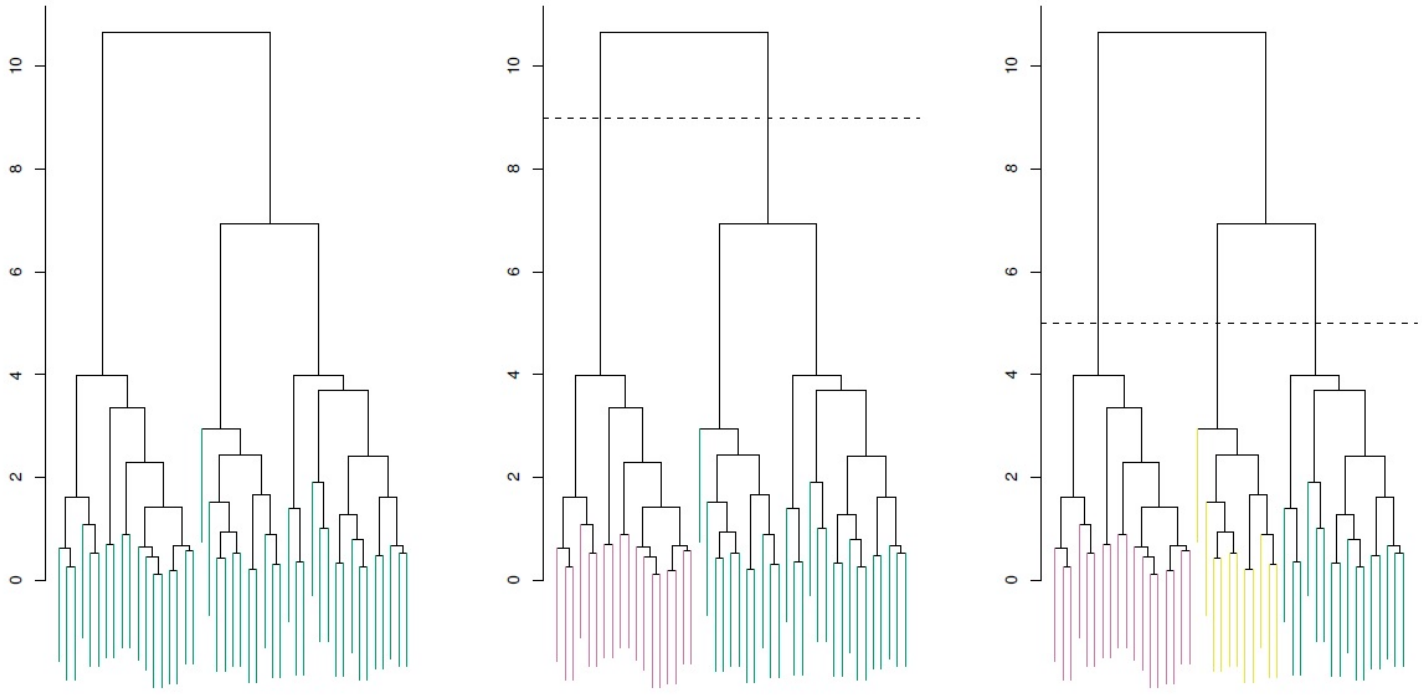

Figura 5: Dendrograma obtenido llevando a cabo clustering jerárquico en los datos de la figura (4), se obtiene utilizando un *enlace* (linkage) completo y la distancia Euclidea.

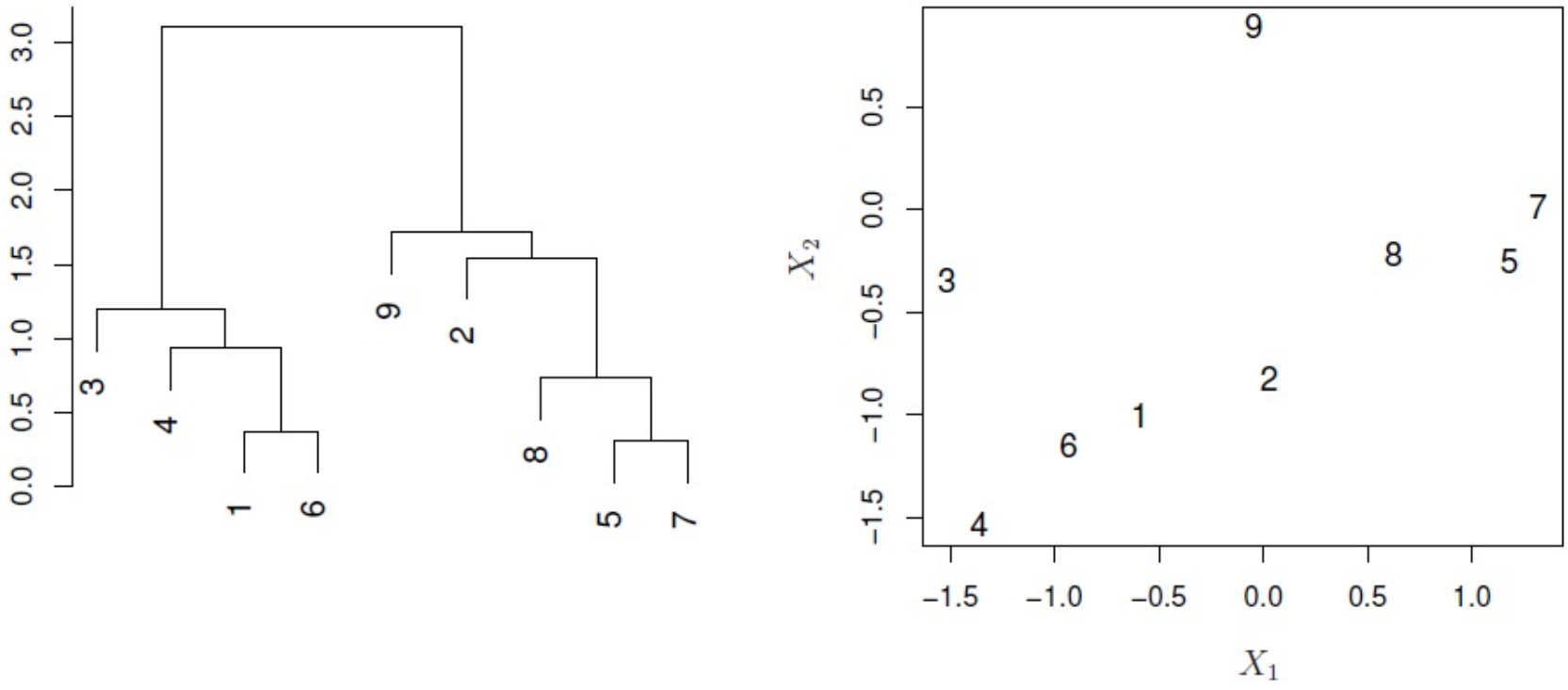

Figura 6: Una ilustración de cómo interpretar apropiadamente un dendrograma.**Izquierda**:Observaciones 5 y 7 son bastante similares entre si, como lo son las observaciones 1 y 6. Ahora bien Observación 9 no es más similar a la observación 2 de lo que lo es con las observaciones 8, 5 y 7. A pesar de que las observaciones 9 y 2 están más cerca una de la otra en distancia horizontal sobre en dendrograma. **Derecha**: Son los datos correspondientes al dendrograma de la izquierda para confirmar la similitud de las observaciones.

- El termino *jerárquico* hace referencia al hecho de que los clusters obtenidos cortando el dendrograma a determinada altura están necesariamente anidados dentro de los clusters que se obtienen cortando el dendrograma a una altura mayor.

- El termino *jerárquico* hace referencia al hecho de que los clusters obtenidos cortando el dendrograma a determinada altura están necesariamente anidados dentro de los clusters que se obtienen cortando el dendrograma a una altura mayor.
- Ahora bien en un conjunto de datos arbitrario esta estructura jerárquica no siempre es consistente o realista.

- Supóngase, por ejemplo, que se tiene un conjunto de observaciones que corresponden a un grupo de personas 50-50 hombres y mujeres. Supóngase también que el grupo se encuentra dividido uniformemente entre Americanos, Japoneses y Franceses.

- Supóngase, por ejemplo, que se tiene un conjunto de observaciones que corresponden a un grupo de personas 50-50 hombres y mujeres. Supóngase también que el grupo se encuentra dividido uniformemente entre Americanos, Japoneses y Franceses.
  Se puede imaginar un escenario en el cual la mejor división en 2 subgrupos de este conjunto de datos es por género y la mejor división del mismo conjunto de datos en 3 subgrupos es por nacionalidad. Pero en este caso los verdaderos clusters no están anidados, en el sentido de que la mejor división en trés subgrupos resulta de primero encontrar la mejor división en 2 subgrupos.

- Consecuentemente esta situación no puede ser bien representada por medio de un clustering jerárquico.

- Consecuentemente esta situación no puede ser bien representada por medio de un clustering jerárquico.
- En situaciones como estas, el clustering jerárquico puede arrojar peores (menos precisos) resultados que el clustering basado en K medias.

Algoritmo de clustering jerárquico

El dendrograma producido por el clustering jerárquico es conseguido por medio de un simple algoritmo.

1. Se comienza definiendo alguna clase de medida de *disimilitud* entre cada par de observaciones. Usualmente se escoge la distancia Euclidea (más sobre este punto adelante).

Algoritmo de clustering jerárquico

El dendrograma producido por el clustering jerárquico es conseguido por medio de un simple algoritmo.

1. Se comienza definiendo alguna clase de medida de *disimilitud* entre cada par de observaciones. Usualmente se escoge la distancia Euclidea (más sobre este punto adelante).
2. El algoritmo procede iterativamente. Comenzando en el extremo inferior del dendrograma, donde cada una de las n observaciones es su propio cluster.
   Los 2 clusters más similares entre si, se fusionan para formar un nuevo cluster, por lo que ahora hay $n - 1$ clusters.

Algoritmo de clustering jerárquico

El dendrograma producido por el clustering jerárquico es conseguido por medio de un simple algoritmo.

1. Se comienza definiendo alguna clase de medida de *disimilitud* entre cada par de observaciones. Usualmente se escoge la distancia Euclidea (más sobre este punto adelante).
2. El algoritmo procede iterativamente. Comenzando en el extremo inferior del dendrograma, donde cada una de las n observaciones es su propio cluster.
   Los 2 clusters más similares entre si, se fusionan para formar un nuevo cluster, por lo que ahora hay $n - 1$ clusters.
3. Ahora entre estos $n - 1$ clusters, se seleccionan los 2 más similares y se fusionan para formar un nuevo cluster, ahora se tienen $n - 2$ clusters.

Algoritmo de clustering jerárquico

El dendrograma producido por el clustering jerárquico es conseguido por medio de un simple algoritmo.

1. Se comienza definiendo alguna clase de medida de *disimilitud* entre cada par de observaciones. Usualmente se escoge la distancia Euclidea (más sobre este punto adelante).
2. El algoritmo procede iterativamente. Comenzando en el extremo inferior del dendrograma, donde cada una de las n observaciones es su propio cluster.
   Los 2 clusters más similares entre si, se fusionan para formar un nuevo cluster, por lo que ahora hay $n - 1$ clusters.
3. Ahora entre estos $n - 1$ clusters, se seleccionan los 2 más similares y se fusionan para formar un nuevo cluster, ahora se tienen $n - 2$ clusters.
4. El algoritmo continua de esta forma, hasta que solo quede un solo cluster.

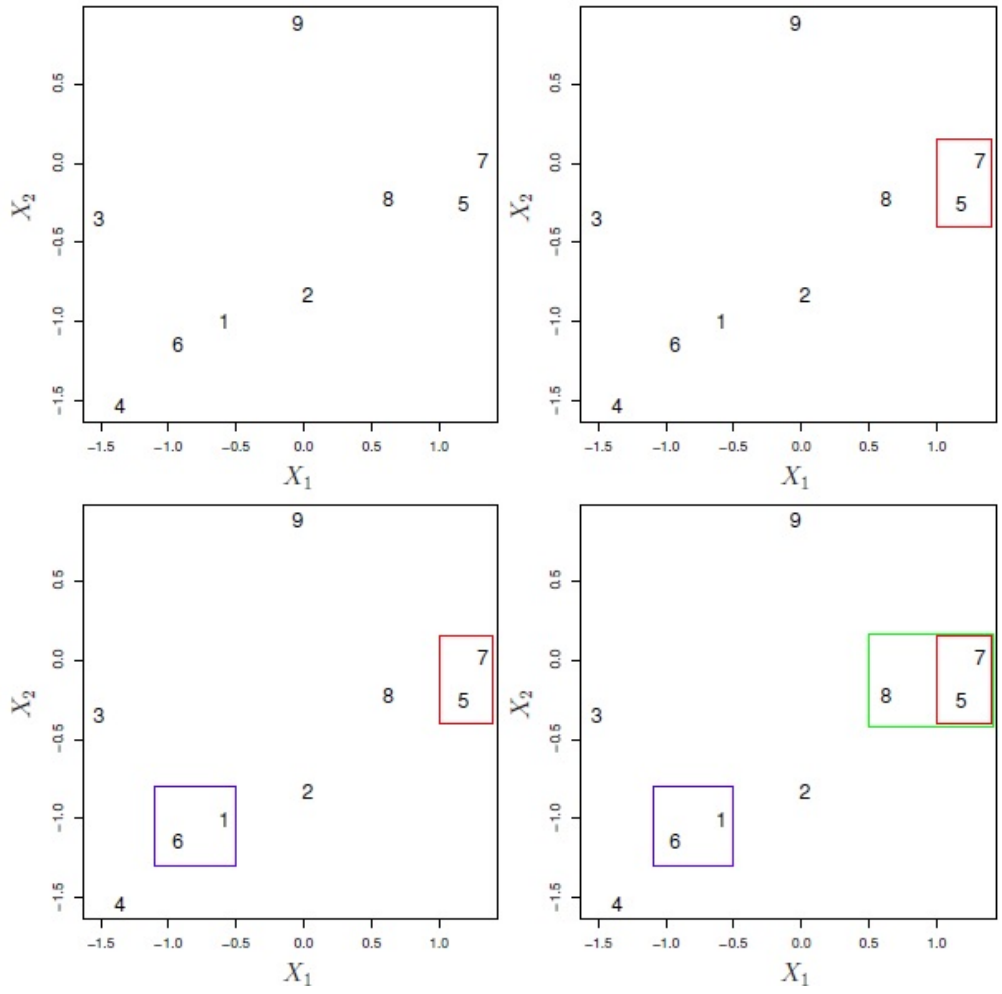

Figura 7: Ilustración del algoritmo con los datos de la figura 6 , enlace completo y distancia Euclidea.

- En la discusión que expone el algoritmo de clustering jerárquico hay que agregar la forma en que se determina la *disimilitud* entre 2 clusters en el caso en que ambos o uno solo pueden ser compuestos (es decir compuestos por más de una observación).

- En la discusión que expone el algoritmo de clustering jerárquico hay que agregar la forma en que se determina la *disimilitud* entre 2 clusters en el caso en que ambos o uno solo pueden ser compuestos (es decir compuestos por más de una observación).
- Esto conlleva a considerar la noción de **enlace** (*linkage*) que define la *disimilitud* entre 2 grupos de observaciones.

Enlace	Descripción
Completo	Disimilitud inter-cluster máxima. Calcule todas las disimilitudes a pares, entre las observaciones en el cluster A y las observaciones en el cluster B y registre el máximo de estas disimilitudes.
Single	Disimilitud inter-cluster mínima. Calcule todas las disimilitudes a pares, entre las observaciones en el cluster A y las observaciones en el cluster B y registre el mínimo de estas disimilitudes.
Average	Disimilitud inter-cluster promedio. Calcule todas las disimilitudes a pares, entre las observaciones en el cluster A y las observaciones en el cluster B y registre el **promedio** de estas disimilitudes.
Centroide	Disimilitud entre centroide del cluster A (vector de medias de longitud p) y el centroide del cluster B.

Cuadro 1

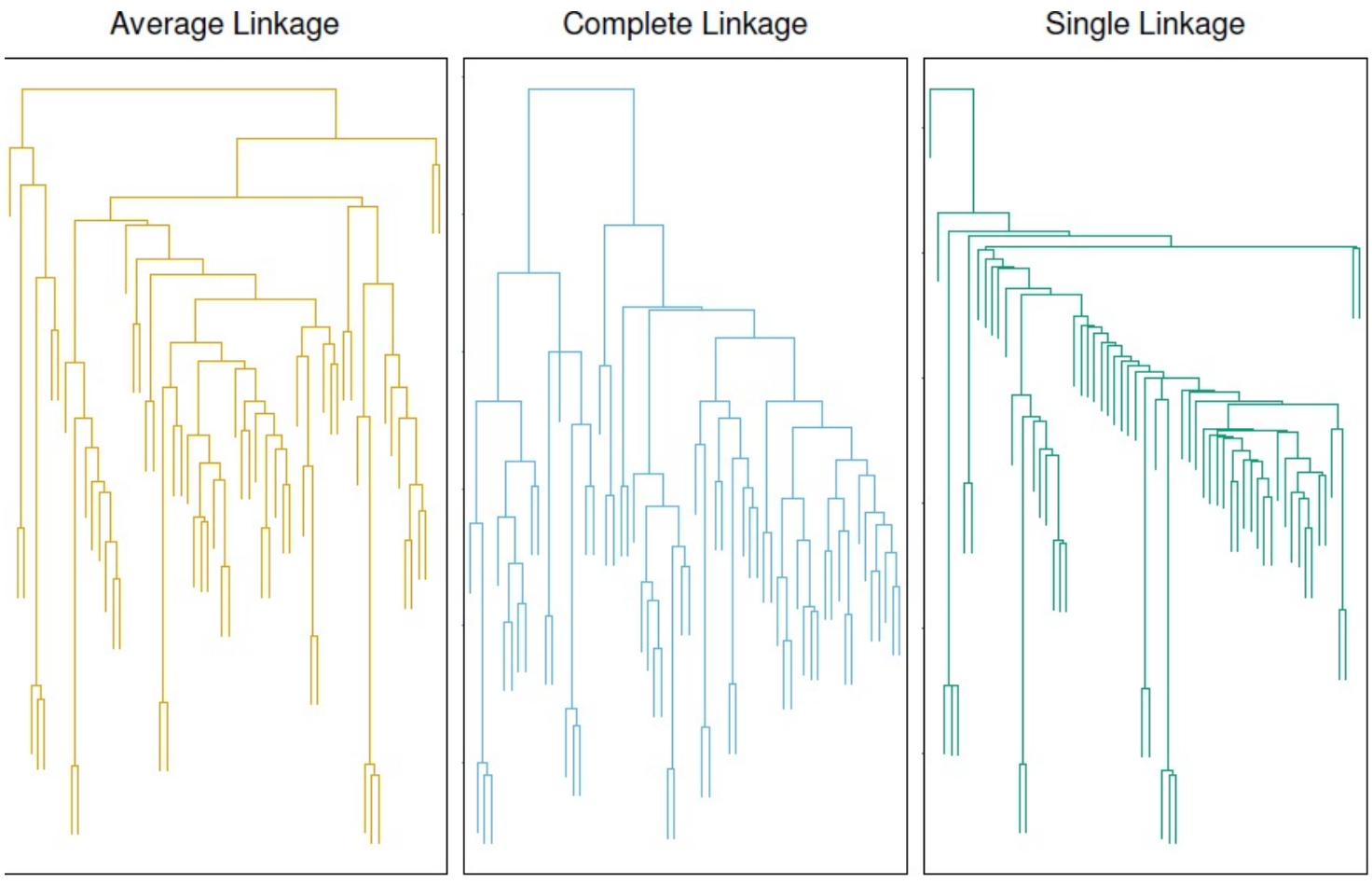

Figura 8: Enlace average, complete y single. El enlace average en general produce dendrogramas más balanceados.

Elección de la medida de disimilitud

- Hasta el momento hemos utilizado la distancia Euclidea como medida de *disimilitud*, pero en algunos contextos otras medidas pueden ser más preferibles.

Elección de la medida de disimilitud

- Hasta el momento hemos utilizado la distancia Euclidea como medida de *disimilitud*, pero en algunos contextos otras medidas pueden ser más preferibles.
- Por ejemplo una *distancia basada en correlación* coloca el énfasis en las formas de las observaciones en vez de sus magnitudes.

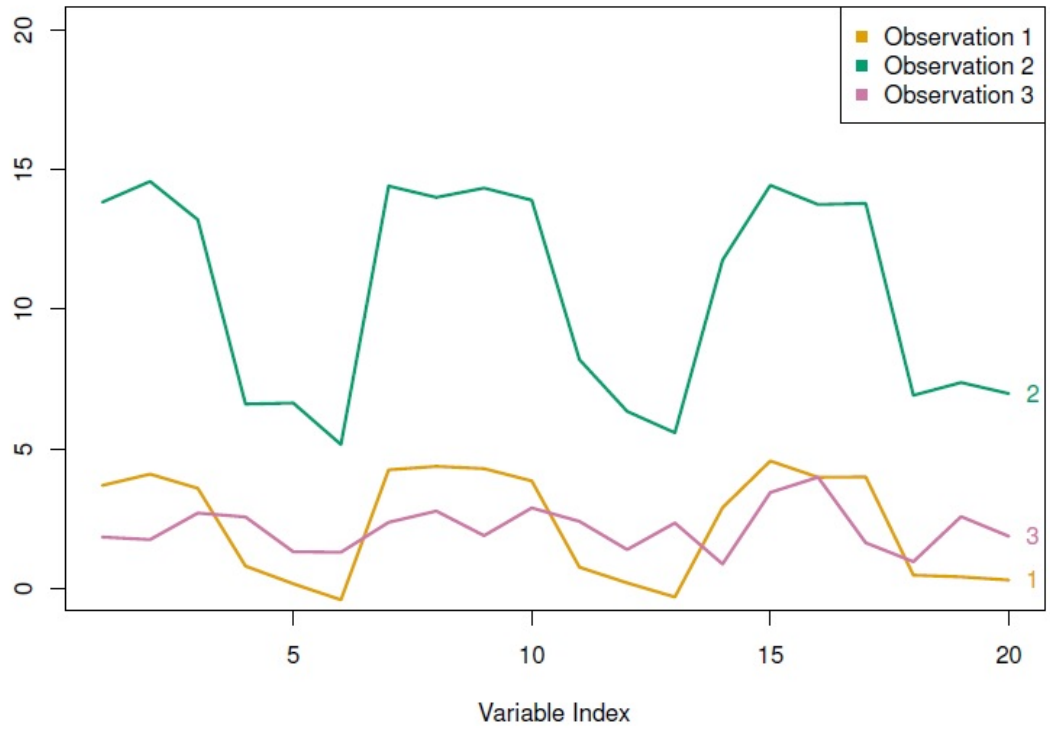

Figura 9: Tres observaciones con medidas en 20 variables son ilustradas. Observaciones 1 y 3 poseen valores similares de las variables, por lo tanto, hay una distancia Euclidea pequeña entre ellas, sin embargo, ellas están débilmente correlacionadas. De otro lado, observaciones 1 y 2 presentan valores bastante distintos atravez de las variables pero hay una alta correlación entre ellas por lo que la distancia (disimilitud basada en correlación) entre ellas es pequeña.

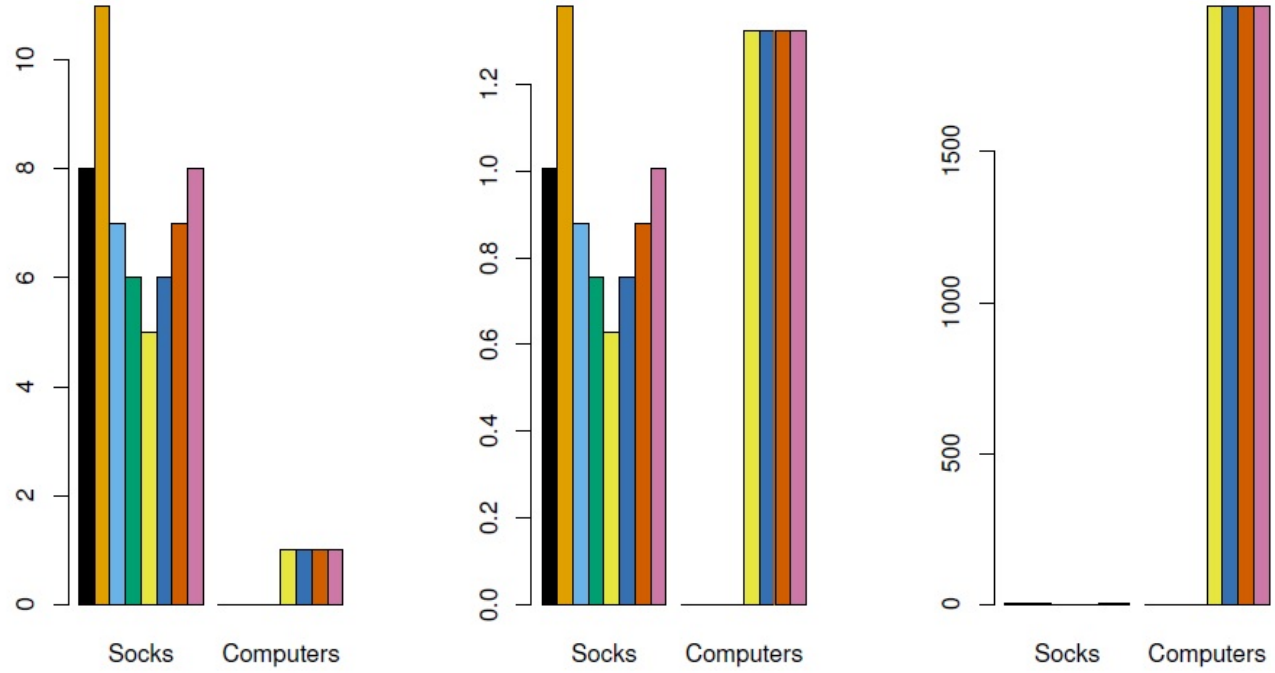

Figura 10: Un minorista en línea vende dos artículos: calcetines y computadoras. **Izquierda**: Se muestra la cantidad de pares de calcetines y computadoras, comprados por ocho compradores en línea . Cada comprador se muestra en un color diferente. Si las diferencias entre observaciones se calculan utilizando la distancia euclidiana en las variables tal como están especificadas, entonces la cantidad de calcetines comprados por un individuo impulsará las diferencias obtenidas, y la cantidad de computadoras compradas tendrá poco efecto (continua en el siguiente slide).

- Esto podría ser indeseable, ya que (1) las computadoras son más caras que los calcetines y por eso el minorista puede estar más interesado en alentar a los compradores a comprar computadoras que calcetines, y (2) una gran diferencia en la cantidad de calcetines comprados por dos compradores puede ser menos informativa sobre las preferencias generales de compra de los compradores que un pequeña diferencia en la cantidad de computadoras compradas.

- Esto podría ser indeseable, ya que (1) las computadoras son más caras que los calcetines y por eso el minorista puede estar más interesado en alentar a los compradores a comprar computadoras que calcetines, y (2) una gran diferencia en la cantidad de calcetines comprados por dos compradores puede ser menos informativa sobre las preferencias generales de compra de los compradores que un pequeña diferencia en la cantidad de computadoras compradas.
- **Centro**: los mismos datos se muestra, después de escalar cada variable por su desviación estándar. Ahora el numero de las computadoras compradas tendrá un efecto mucho mayor en la disimilitudes entre observaciones obtenidas.

- Esto podría ser indeseable, ya que (1) las computadoras son más caras que los calcetines y por eso el minorista puede estar más interesado en alentar a los compradores a comprar computadoras que calcetines, y (2) una gran diferencia en la cantidad de calcetines comprados por dos compradores puede ser menos informativa sobre las preferencias generales de compra de los compradores que un pequeña diferencia en la cantidad de computadoras compradas.
- **Centro**: los mismos datos se muestra, después de escalar cada variable por su desviación estándar. Ahora el numero de las computadoras compradas tendrá un efecto mucho mayor en la disimilitudes entre observaciones obtenidas.
- **Derecha**: se muestran los mismos datos, pero ahora el eje Y representa la cantidad de dólares gastados por cada comprador en línea en calcetines y en ordenadores. Como las computadoras son mucho más caras que las medias, ahora el histórico de compras de computadoras dominará las disimilitudes obtenidas entre las observaciones.

Algunas consideraciones prácticas del clustering

1. **Pequeñas decisiones con grandes consecuencias**
   - En el caso de clustering jerárquico
     - Deben las observaciones o los atributos estandarizarse en alguna forma? Por ejemplo deben centrarse las variables y escalarse para que tengan desviación estándar de uno?.
     - Que medida de *disimilitud* se debe utilizar?
     - Que tipo de enlace, (*complete, average o single*), se debe utilizar?

Algunas consideraciones prácticas del clustering

1. **Pequeñas decisiones con grandes consecuencias**
   - En el caso de clustering jerárquico
     - Deben las observaciones o los atributos estandarizarse en alguna forma? Por ejemplo deben centrarse las variables y escalarse para que tengan desviación estándar de uno?.
     - Que medida de *disimilitud* se debe utilizar?
     - Que tipo de enlace, (*complete, average o single*), se debe utilizar?
   - En el caso de K medias, que número de clusters se debe buscar en los datos?

Algunas consideraciones prácticas del clustering

1. **Pequeñas decisiones con grandes consecuencias**
   - En el caso de clustering jerárquico
     - Deben las observaciones o los atributos estandarizarse en alguna forma? Por ejemplo deben centrarse las variables y escalarse para que tengan desviación estándar de uno?.
     - Que medida de *disimilitud* se debe utilizar?
     - Que tipo de enlace, (*complete, average o single*), se debe utilizar?
   - En el caso de K medias, que número de clusters se debe buscar en los datos?
   - Cada una de estas decisiones puede tener un fuerte impacto en los resultados. En la práctica es conveniente intentar diferentes alternativas y buscar la solución más útil, la más interpretable.

Algunas consideraciones prácticas del clustering

1. **Pequeñas decisiones con grandes consecuencias**
   - En el caso de clustering jerárquico
     - Deben las observaciones o los atributos estandarizarse en alguna forma? Por ejemplo deben centrarse las variables y escalarse para que tengan desviación estándar de uno?.
     - Que medida de *disimilitud* se debe utilizar?
     - Que tipo de enlace, (*complete, average o single*), se debe utilizar?
   - En el caso de K medias, que número de clusters se debe buscar en los datos?
   - Cada una de estas decisiones puede tener un fuerte impacto en los resultados. En la práctica es conveniente intentar diferentes alternativas y buscar la solución más útil, la más interpretable.
   - Con estos métodos no hay una única respuesta 100 % cierta. Cada solución puede resaltar algún aspecto interesante de los datos analizados.

2 Validando los clusters obtenidos

- Cada vez que se lleva a cabo un proceso de *clustering* sobre un conjunto de datos, se obtendrán algunos clusters.

3 Validando los clusters obtenidos

- Cada vez que se lleva a cabo un proceso de *clustering* sobre un conjunto de datos, se obtendrán algunos clusters.
- Pero se hace mandatorio saber si los *clusters* obtenidos representan verdaderos subgrupos de los datos analizados o simplemente son el resultado que consigue el algoritmo al intentar *agrupar* el ruido.

4 Validando los clusters obtenidos

- Cada vez que se lleva a cabo un proceso de *clustering* sobre un conjunto de datos, se obtendrán algunos clusters.
- Pero se hace mandatorio saber si los *clusters* obtenidos representan verdaderos subgrupos de los datos analizados o simplemente son el resultado que consigue el algoritmo al intentar *agrupar* el ruido.
- Por ejemplo, si el algoritmo se aplica sobre un conjunto independiente de observaciones(de la misma naturaleza de los datos que se están analizando) se encontrará el mismo conjunto de clusters?.

5 Otras consideraciones en el empleo de clustering

- Los métodos de *clustering* están sujetos a problemas de robustez, es decir pueden ser fuertemente influenciados por la presencia de outlyers u observaciones atípicas.

5 Otras consideraciones en el empleo de clustering

- Los métodos de *clustering* están sujetos a problemas de robustez, es decir pueden ser fuertemente influenciados por la presencia de outlyers u observaciones atípicas.
- Se recomienda realizar clustering en subconjuntos aleatoriamente extraídos de los datos para hacerse una idea de la robustez del agrupamiento que se está llevando a cabo.

5 Otras consideraciones en el empleo de clustering

- Los métodos de *clustering* están sujetos a problemas de robustez, es decir pueden ser fuertemente influenciados por la presencia de outlyers u observaciones atípicas.
- Se recomienda realizar clustering en subconjuntos aleatoriamente extraídos de los datos para hacerse una idea de la robustez del agrupamiento que se está llevando a cabo.
- Los resultados que se obtengan, no pueden ser tomados como una verdad absoluta sobre los datos que están siendo analizados. Más bien, estos resultados deben constituir el punto de partida para el desarrollo de una hipótesis y su estudio, preferiblemente en un conjunto independiente de datos.