# Métodos de clustering

Fernanda Mora
26 octubre, 2016

Diplomado de Minería de Datos
para la toma de decisiones

# Contenido

- Introducción
- Clustering jerárquico
- Clustering no-jerárquico
    - Ejercicio práctico 1

y = f(x1, ...,xn)

# Introducción

# Preliminares

- Dos enfoques diferentes al problema de clasificación
- Enfoque 1:
  - Grupos están bien definidos (i.e. se conocen a priori) y se trata de determinar un criterio para etiquetar cada individuo como perteneciente a alguno de los grupos
  - Vimos varios métodos: regresión logística, análisis discriminante, máquinas de soporte vectorial y redes neuronales
- Enfoque 2:
  - No se conocen los grupos y se quiere encontrarlos (si hay)
  - Métodos: análisis de clusters

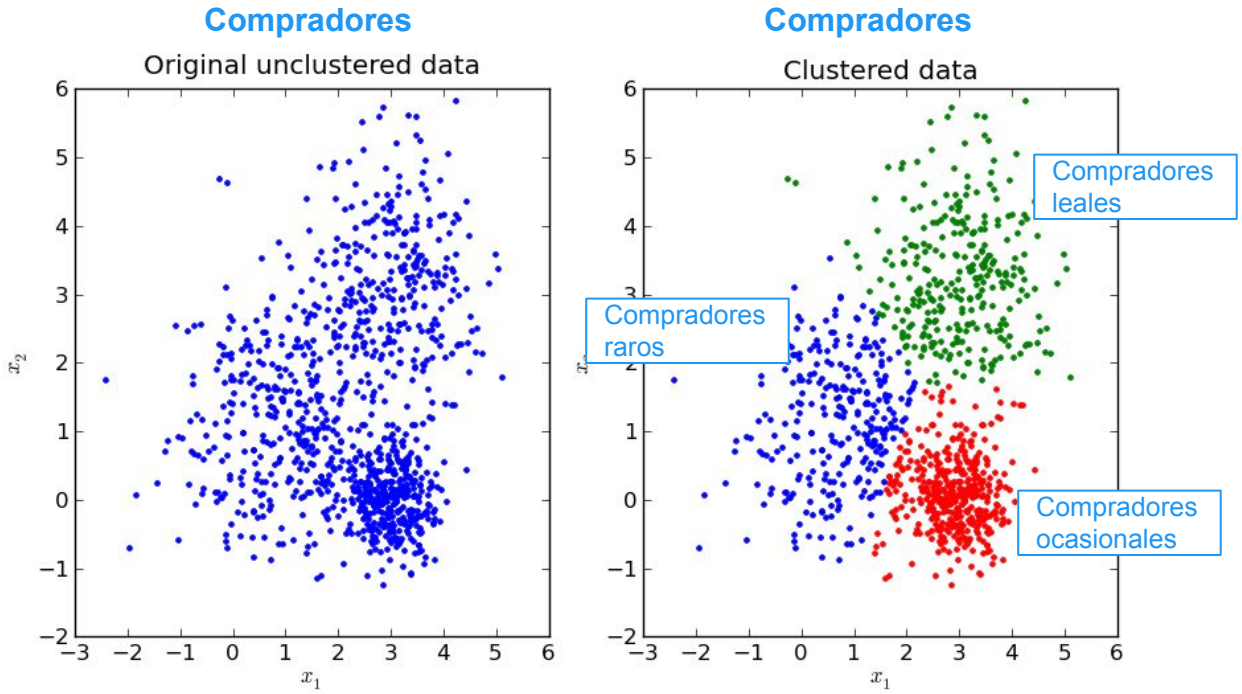
# Preliminares

- También se conoce como análisis de conglomerados o análisis de agrupaciones
- Estos modelos pertenecen al análisis no-supervisado, en donde no hay individuos etiquetados (clasificados) pero aún así se desea encontrar patrones en los datos
- La gran mayoría de los datos existentes NO están etiquetados (es costoso, difícil y tardado etiquetarlos)
- Aseguradora: muchos datos de los asegurados, al principio no se conoce a qué grupo de riesgo pertenecen: alto, medio, bajo

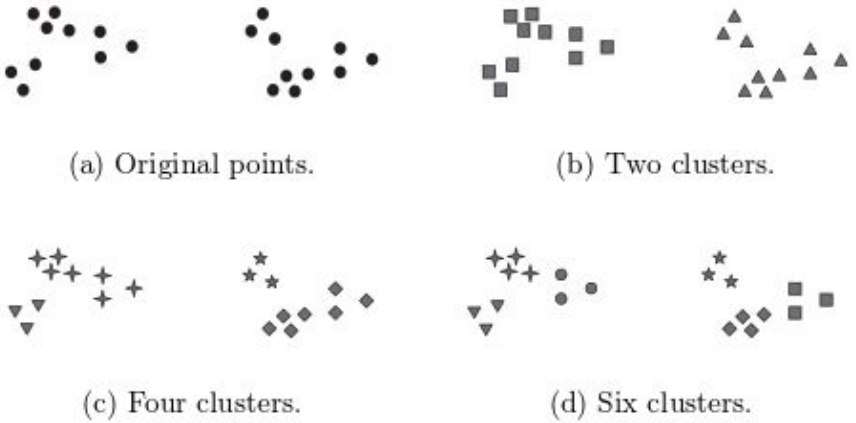
# Preliminares

- Teléfonos móviles: muchos datos, apps bajadas, apps usadas, compras, perfiles de facebook, clicks
- Datos de compradores de Amazon: ¿qué patrones hay?, ¿se pueden definir grupos?
- Muy útil para diseñar estrategias de marketing
- Encuentro los patrones de compra, los perfiles de los compradores y ahora puedo diseñar marketing inteligente
- No es lo mismo venderle al comprador leal a la marca a uno que no lo es
- Muy usado en business intelligence

# Preliminares



# Preliminares



(a) Original points. (b) Two clusters.

(c) Four clusters. (d) Six clusters.

**Diferentes maneras de hacer clusters de los mismos datos**

# Preliminares

- Las variables pueden ser de cualquier tipo: de razón o categóricas
- O sea que podemos trabajar en espacios no euclídeos
- Pero se complica porque hay que poder interpretar las distancias entre los elementos
- ¡Es bueno trabajar con variables de razón!
- Existen métodos que funcionan bien con variables categóricas: mapas autoorganizados

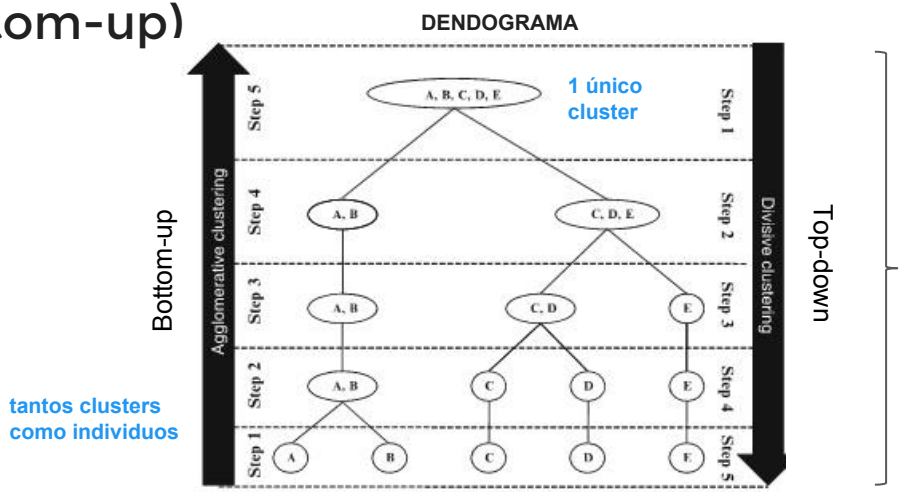
# Preliminares

- ¿Cómo construimos estos grupos?
- Parecido a Análisis Discriminante: queremos que los individuos dentro los grupos se parezcan y fuera de los grupos no
- Necesitamos una noción general de similitud/disimilitud: ¡distancia!
- Individuos distantes serán diferentes (dispersión), individuos cercanos serán parecidos (cohesión)
- No hay un método único para hacer análisis de clústers
- Clustering jerárquico (clústers, luego # de clústers), k-means (# de clústers, luego clústers), mapas auto-organizados

# Clustering jerárquico

# Clustering jerárquico

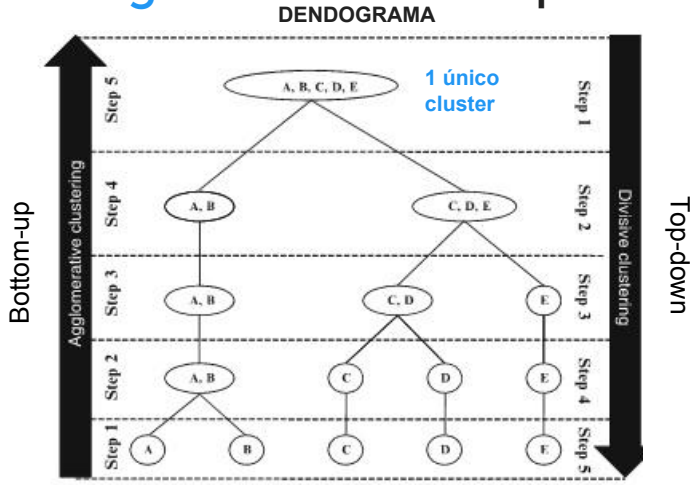
- La similaridad sería la distancia entre los individuos
- Con clustering jerárquico se puede hacer clustering de variables
- Hay dos tipos de clustering jerárquico: divisivo (top-down) y aglomerativo (botom-up)



# Clustering jerárquico

- En clustering aglomerativo un cluster no puede dividirse, sólo puede combinarse con otros
- La complejidad del aglomerativo es $O(n^2 log(n))$, del divisivo $O(2^n)$
- En la práctica se usa más el aglomerativo porque es más rápido

**Necesitamos una condición de paro en ambos casos: un único clúster no nos sirve, tantos clústers como individuos tampoco**



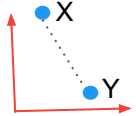
# Clustering jerárquico

- 3 pasos:
  - Criterio para determinar la similaridad
  - Criterio para decidir qué clusters se combinan en cada paso
  - Criterio para decidir el número de clústers
- NO hay un número de clústers óptimo (no es trivial decidirlo)
- En cada iteración debemos ver los clústers formados y al final decidir cuál es el número de clústers más adecuado a nuestras necesidades
- También necesitamos tener variables relevantes al problema que queremos resolver

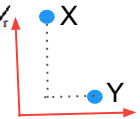
# Clustering jerárquico

## Similaridad

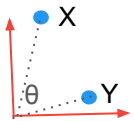
- La noción de distancia es MUY general
- Euclideana: $d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$




- Minkowski: $d(x,y) = \left(\sum_{i=1}^{m}|x_i - y_i|^r\right)^{1/r}$




- Manhattan: $d(x,y) = \sum_{i=1}^{m}|x_i - y_i|$
- Coseno:



# Clustering jerárquico

## Similaridad

- Mahalanobis(vector a vector): $\sqrt{(a-b)^\top S^{-1}(a-b)}$ Distancia de un vector a al vector b con S la matriz de Covarianza
  - Sin unidades y toma en cuenta correlaciones
  - Si S es la matriz identidad entonces corresponde a la distancia euclidea usual
- Mahalanobis (vector a media): cuántas desviaciones estándar está un punto del vector de medias: $D_M(\vec{x}) = \sqrt{(\vec{x}-\vec{\mu})^T S^{-1}(\vec{x}-\vec{\mu})}.$
- Euclideana normalizada: quitamos el efecto de la desviación estándar: $d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i^2}},$

# Clustering jerárquico

## Conformación de clústers

- Lo anterior nos permite calcular la distancia entre individuos: ¿qué hay de los clústers?
- El criterio de “enlace” (linkage) determina la distancia entre clústers usando la distancia predefinida entre individuos.

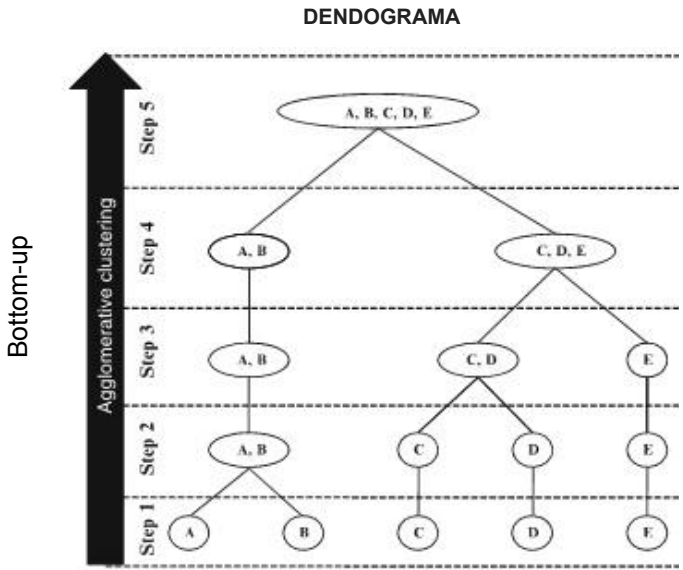
| Names | Formula |
|---|---|
| Maximum or complete-linkage clustering | $\max\{\, d(a,b) : a \in A,\, b \in B \,\}$. |
| Minimum or single-linkage clustering | $\min\{\, d(a,b) : a \in A,\, b \in B \,\}$. |
| Mean or average linkage clustering, or UPGMA | $\frac{1}{\lvert A\rvert \lvert B\rvert} \sum_{a \in A} \sum_{b \in B} d(a,b)$. |
| Centroid linkage clustering, or UPGMC | $\lVert c_s - c_t \rVert$ where $c_s$ and $c_t$ are the centroids of clusters $s$ and $t$, respectively. |

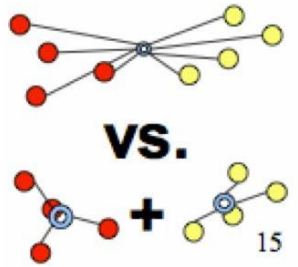
- Minimizar la varianza intra-cluster
- La disminución en la varianza por el cluster a unir (método de Ward)

# Clustering jerárquico

## Conformación de clústers

- En cada paso el par de clusters con la distancia más pequeña se fusiona en un único cluster



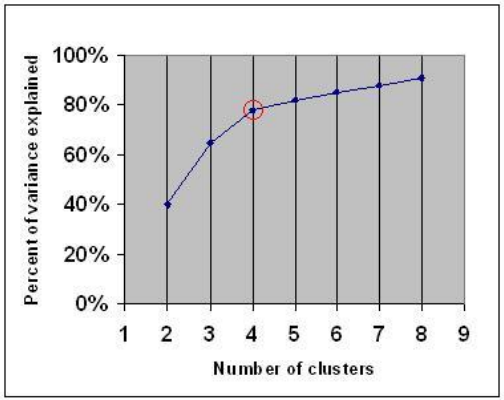

$$TD_{c_1 \cup c_2} = \sum_{x \in c_1 \cup c_2} D(x, \mu_{c_1 \cup c_2})^2$$

Si uniéramos a $c_1$ con $c_2$ ¿cómo cambiaría la distancia total a los centroides?

# Clustering jerárquico

## Número de clústers

- ¿Cómo saber cuántos clusters son?
- Problema MUY difícil, No hay una solución óptima
- Idea: debemos escoger un # de clústers tal que al añadir otro clúster no obtenemos una mejora considerable
- Criterio del codo, pero hay otros
- A veces es mejor explorar visualmente



# Clustering no-jerárquico

# Clustering NO jerárquico

## K-means

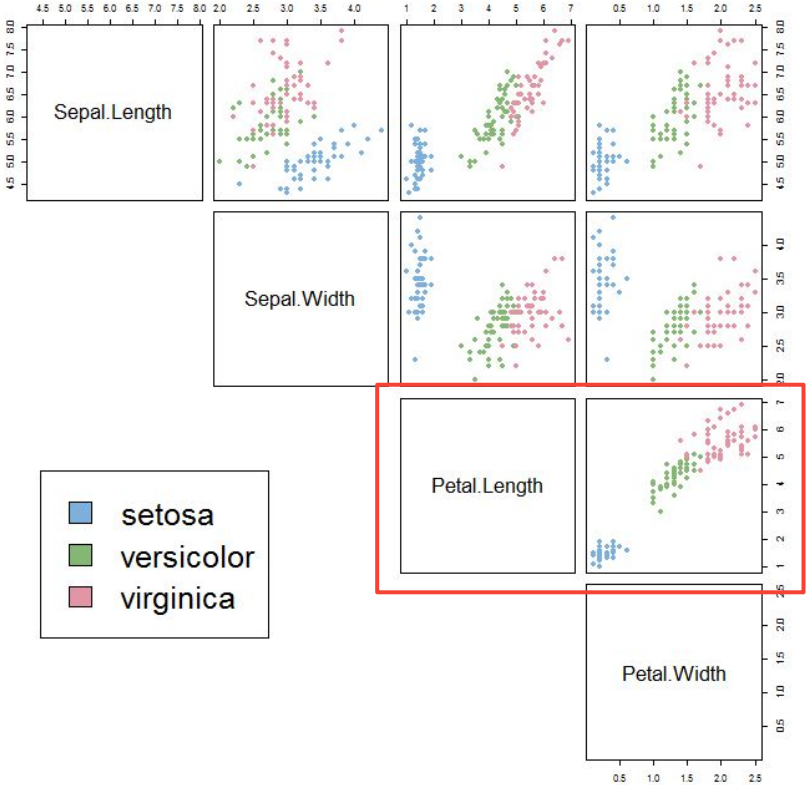
1. Especificar k, el número de clusters
2. Escoger k puntos al azar como centroides (hay que definir una semilla)
3. Asignar a cada individuo a su cluster más cercano
4. Calcular los centroides para cada clúster (medias)
5. Esos centroides serán los nuevos centros de los clústers
6. Iterar hasta que nos centros de los clústers "no cambien"

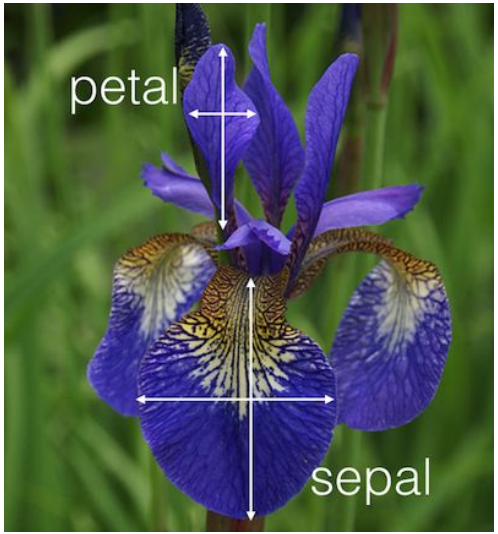
Desventaja: Los resultados dependen de la semilla, es un método local

# Ejercicio práctico

Iris Dataset

# Graficando las observaciones de iris



# Gracias