ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Introducción a la Ciencia de Datos con R

Miguel Jorquera

Educación Profesional
Escuela de Ingeniería

RESUMEN

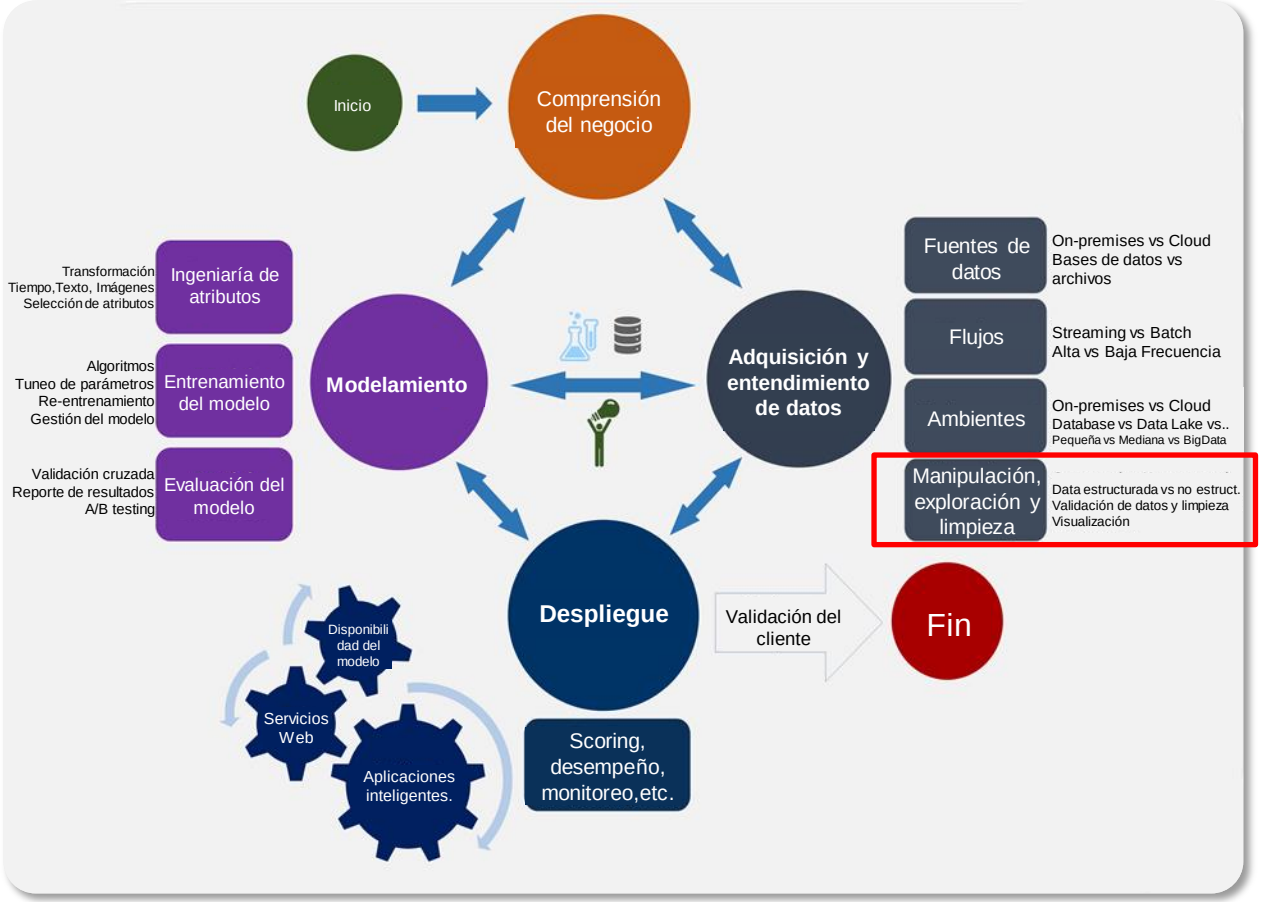
RESUMEN

EDUCACIÓN PROFESIONAL

TDSP

La metodología TDSP (Team Data Science Process) propuesta por Microsoft es una metodología ágil, iterativa y eficiente, que promueve la colaboración entre los distintos miembros del equipo de desarrollo así como la interacción permanente con el cliente.

Flujo de trabajo en Data Science



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

RESUMEN

Introducción a gráficos con Ggplot.

- Gramática de gráficos:

`ggplot()` + `aes()` + `geom_...()`

Recibe como argumento el origen de datos (data.frame).

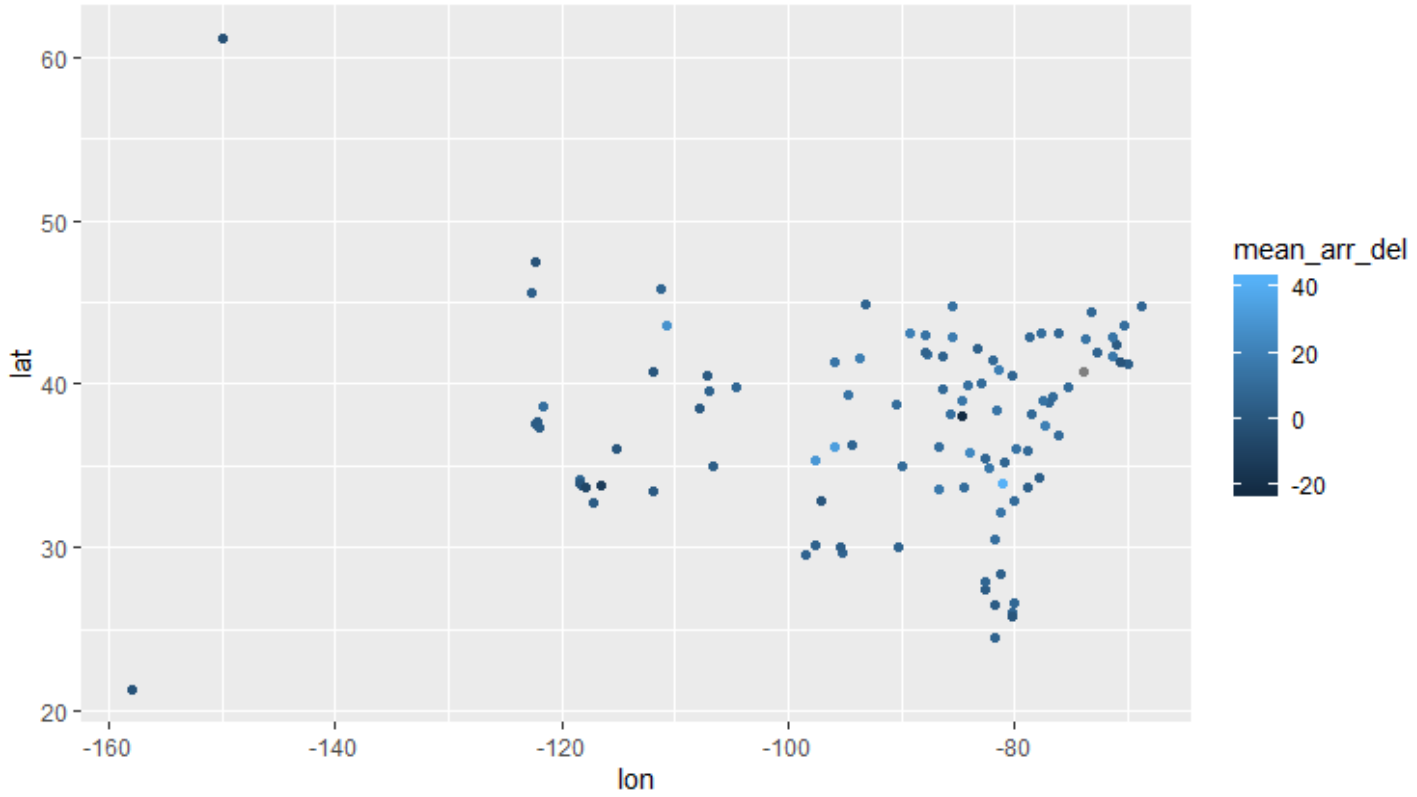
Se indican las variables a graficar (mapeo).

Se define el tipo de gráfico y otros parámetros visuales.

RESUMEN

Ejemplo:

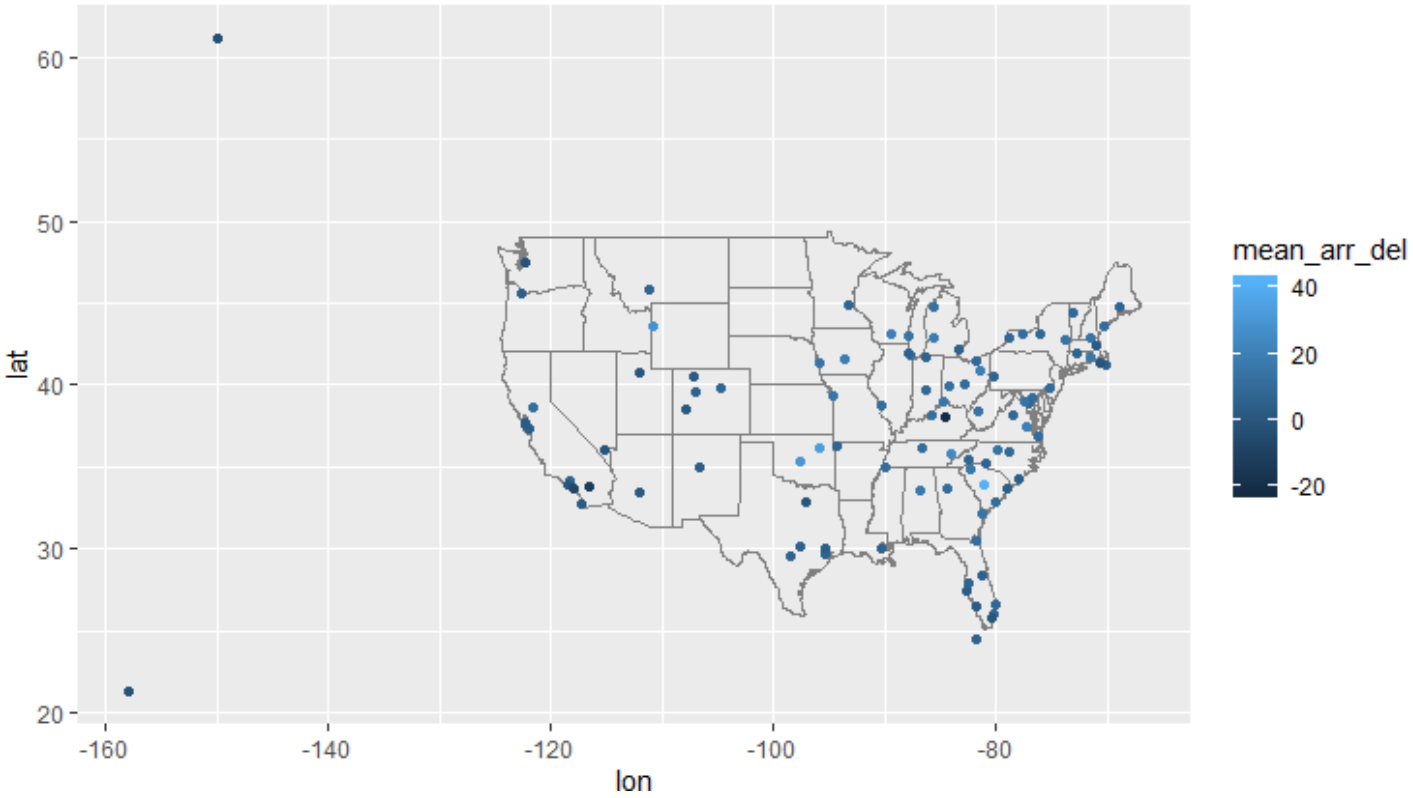
```
flights %>%
  group_by(dest) %>%
  summarise(mean_arr_del =
              mean(arr_delay,
                   na.rm = TRUE)) %>%
  left_join(airports,
            by = c("dest" = "faa")) %>%
  ggplot() +
  aes(x = lon,
      y = lat,
      color = mean_arr_del) +
  geom_point()
```



RESUMEN

Ejemplo:

```
flights %>%
  group_by(dest) %>%
  summarise(mean_arr_del =
              mean(arr_delay,
                   na.rm = TRUE)) %>%
  left_join(airports,
            by = c("dest" = "faa")) %>%
  ggplot() +
  aes(x = lon,
      y = lat,
      color = mean_arr_del) +
  geom_point() +
  borders("state") +
  geom_point()
```



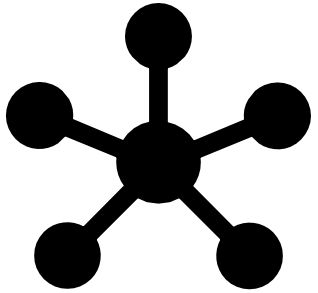
TEMAS PARA HOY

AGENDA

Temas para hoy

Introducción modelos no supervisados

- Análisis exploratorio mediante de reglas de asociación
- Clustering



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

MODELOS NO SUPERVISADOS

INTRODUCCIÓN

Aprendizae no suervisado

- Data no supervisada no contiene etiquetas y no existe una variable respuesta.
- Existe interés en encontrar una estructura escondida (que nunca es completamente observada).
- Validación de resultados compleja.

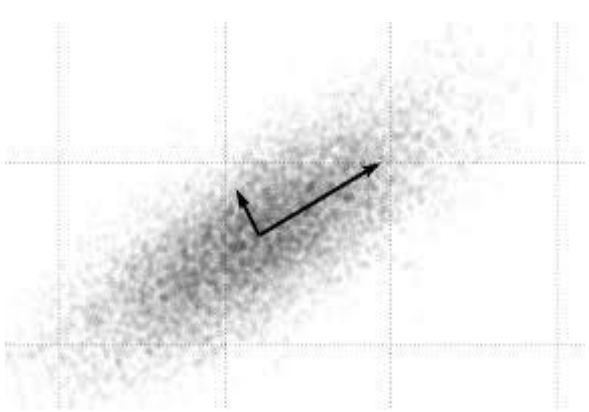
INTRODUCCIÓN

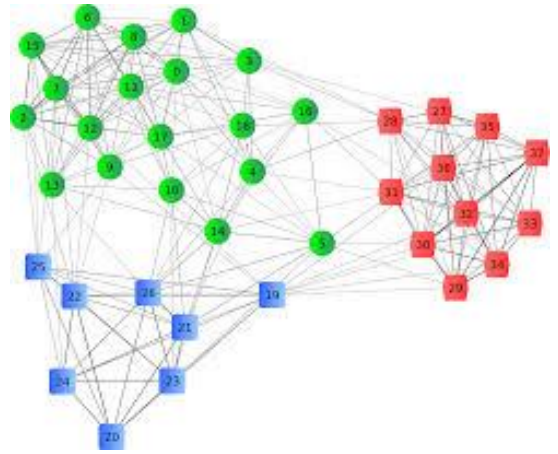
EDUCACIÓN PROFESIONAL

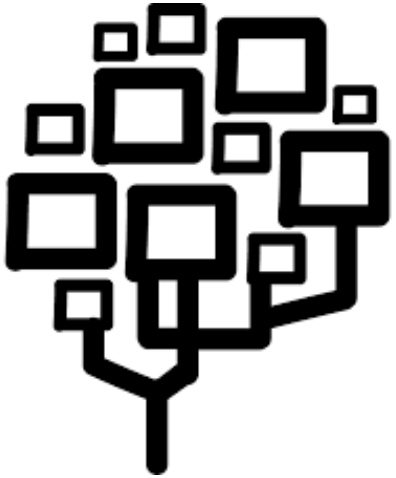
Aprendizaje no supervisado

Algunas tareas usuales:

- Reducción de la dimensionalidad
- Generación de clusters
- Reglas de asociación







REGLAS DE ASOCIACIÓN

REGLAS DE ASOCIACIÓN

Objetivo

- Generar “reglas” que asocien productos.
- Estas reglas deben ser:
  - Frecuentes
  - Razonables.

REGLAS DE ASOCIACIÓN

Definiciones

{Zapatos, cartera} ⟶ {Traje de Baño}

Conceptos claves:

- Item
- Itemset
- Antecedente
- Consecuente
- Regla de asociación

Métricas claves:

- Support
- Confidence
- Lift

REGLAS DE ASOCIACIÓN

Definiciones

Reglas basadas en probabilidades.

- $Supp(\{a,b\}) = \frac{\#\ Transacciones\ que\ contienen\ a\ \ y\ b}{\#\ Transacciones}$

- $Conf(\{a,b\} \rightarrow \{c\}) = \frac{Supp(\{a,b,c\})}{Supp(\{a,b\})} = \hat{P}(\{c\} \mid \{a,b\})$

REGLAS DE ASOCIACIÓN

Definiciones

¿Qué hace "buena" a una regla?

Debe ser común:

$$Supp(\{a, b\}) \geq \theta$$

¿Cómo generar las reglas?

Debe ser razonable:

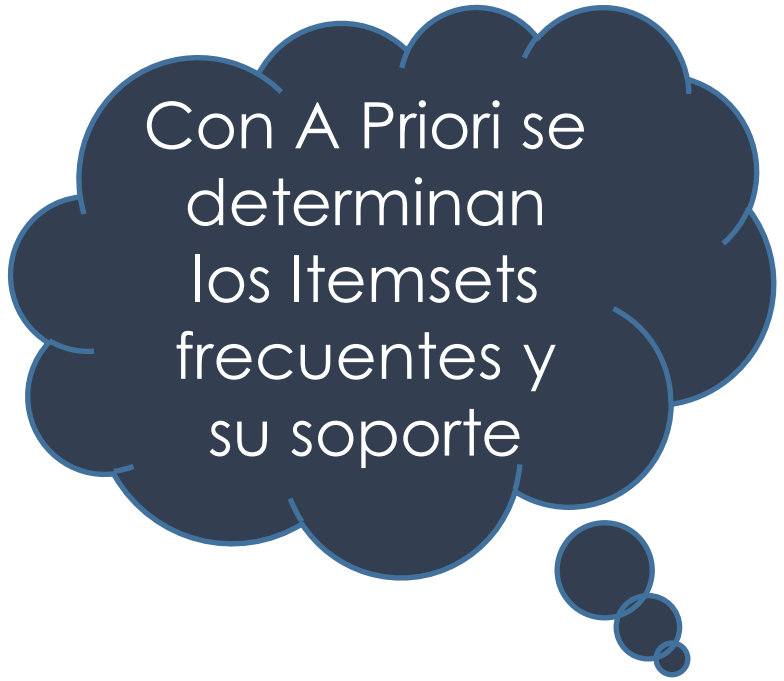
$$Conf(\{a, b\} \rightarrow \{c\}) \geq \text{minconf}$$

REGLAS DE ASOCIACIÓN

Algoritmo apriori

Algoritmo:

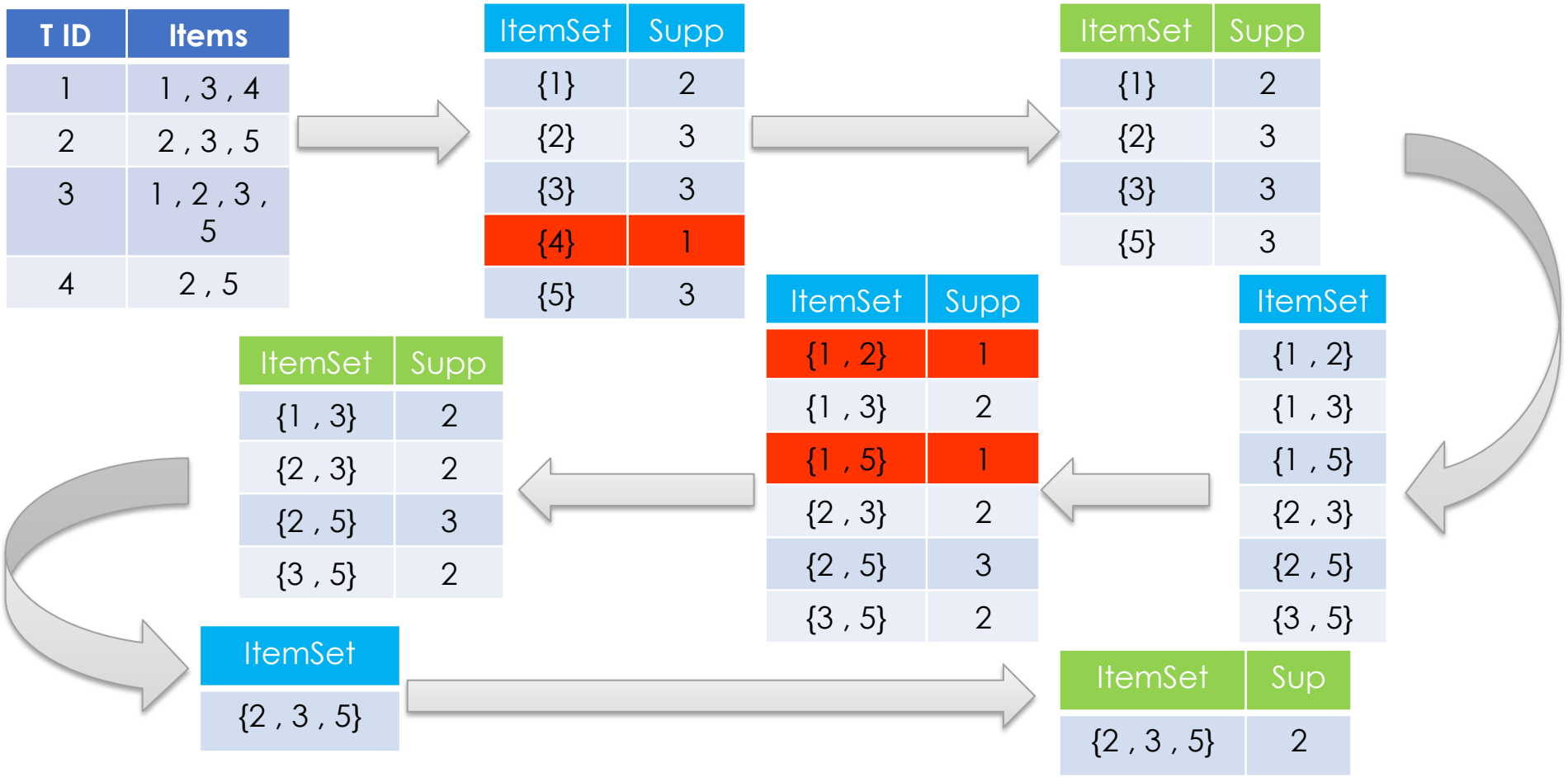
- Se buscan los itemset de un item y se filtran aquellos con soporte mayor o igual que θ
- Repetir hasta que no se puedan formar nuevos Itemsets:
  - Crea itemsets candidatos: Para cada par de itemsets ya listados con k elementos, combinarlos si comparten k-1 elementos.
  - Poda: Retener candidato si tiene un soporte de al menos θ para definir la lista con itemset con k+1 elementos.
  - Fin: si la lista de itemsets con k+1 elementos es vacía.



REGLAS DE ASOCIACIÓN

Algoritmo apriori

$$\theta = \frac{1}{4} = 0.25$$



REGLAS DE ASOCIACIÓN

Algoritmo apriori

Itemsets

Itemset	Supp
{1}	2
{2}	3
{3}	3
{5}	3
{1 , 3}	2
{2 , 3}	3
{2 , 5}	3
{3 , 5}	2
{2 , 3 , 5}	2

¿Qué reglas escogemos?

Reglas de asociación

Regla	Confidence	Regla	Confidence
$1 \rightarrow 3$	2/2 = 1	$5 \rightarrow 3$	2/3 = 0.66
$2 \rightarrow 3$	3/3 = 1	$\{2,3\} \rightarrow 5$	2/3 = 0.66
$2 \rightarrow 5$	3/3 = 1	$\{3,5\} \rightarrow 2$	2/2 = 1
$3 \rightarrow 5$	2/3 = 0.66	$\{2,5\} \rightarrow 3$	2/3 = 0.66
$3 \rightarrow 1$	2/3 = 0.66	$5 \rightarrow \{2,3\}$	2/3 = 0.66
$3 \rightarrow 2$	3/3 = 1	$2 \rightarrow \{3,5\}$	2/3 = 0.66
$5 \rightarrow 2$	3/3 = 1	$3 \rightarrow \{2,5\}$	2/3 = 0.66

REGLAS DE ASOCIACIÓN

Algoritmo apriori

¿Qué reglas son preferibles?

- Ordenar por confidence:

$$Conf(a \rightarrow b) = \hat{P}(b|a) = \frac{Supp(a \cup b)}{Supp(a)}$$

- Ordenar por lift:

$$Lift(a \rightarrow b) = \frac{Conf(a \rightarrow b)}{Supp(b)} = \frac{\hat{P}(a \cup b)}{\hat{P}(a)\hat{P}(b)}$$

REGLAS DE ASOCIACIÓN

Algoritmo apriori

¿Qué reglas son preferibles?

- Ordenar por confidence:

$$Conf(a \rightarrow b) = \hat{P}(b|a) = \frac{Supp(a \cup b)}{Supp(a)}$$

- Ordenar por lift:

$$Lift(a \rightarrow b) = \frac{Conf(a \rightarrow b)}{Supp(b)} = \frac{\hat{P}(a \cup b)}{\hat{P}(a)\hat{P}(b)}$$

REGLAS DE ASOCIACIÓN

Algoritmo apriori

Wikipedia:

"Lift is a measure of the performance of a targeting model (association rule) at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model. A targeting model is doing a good job if the response within the target is much better than the average for the population as a whole. Lift is simply the ratio of these values:"

$$Lift = \frac{target\ response}{average\ response}$$

REGLAS DE ASOCIACIÓN

Algoritmo apriori

Wikipedia:

Por ejemplo,

En una población la tasa de respuesta es de un 5%, pero cierto modelo (o regla) logra identificar un segment con una tasa de resúesta de un 20%. Entonces dicho segment tiene un lift de 4.0 (20%/5%).

REGLAS DE ASOCIACIÓN

EDUCACIÓN
PROFESIONAL

Vamos!

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

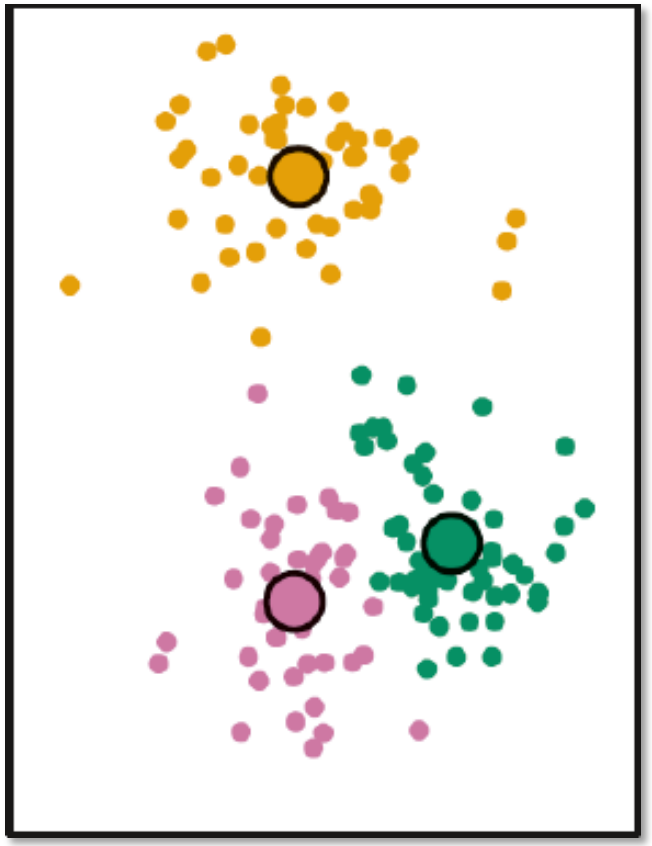
CLUSTERING

CLUSTERING

Objetivo

Buscamos subgrupos homogéneos de observaciones

- Enfoque de particiones
  - K-means, k-medioids, CLARANS
- Enfoque jerárquico
  - Diana, Agnes, BIRCH, CAMELEON
- Enfoque basado en densidad
  - DBSCAN, OPTICS, DenClue

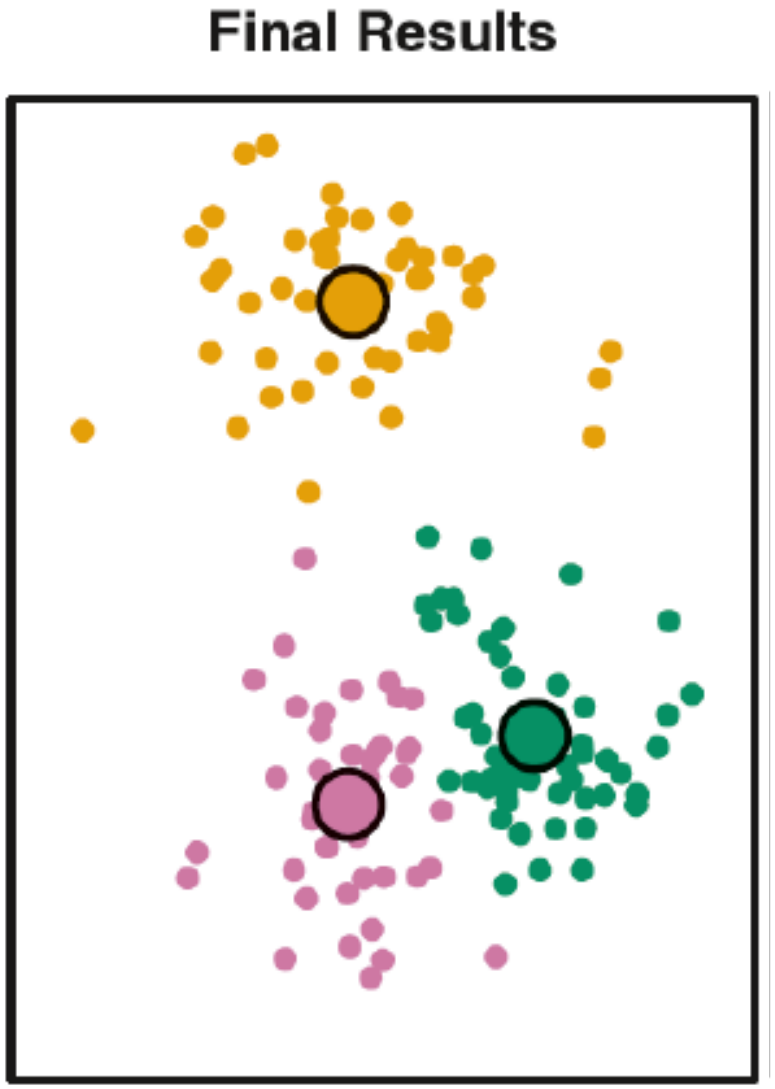


CLUSTERING

K-means

Idea del algoritmo :

1. Suponer (por ejemplo), K=3 grupos y asignar aleatoriamente las obs. a ellos.
2. Definir los centroides.
3. Reasignar la data a al centroide más cercano.
4. Redefinir centroides
5. Repetir proceso hasta que no haya cambio en la asignación



K-means

¿Cómo determinar el número óptimo de clústers?

Minimizando la suma de cuadrados totales de todos los grupos (total within sum squares)

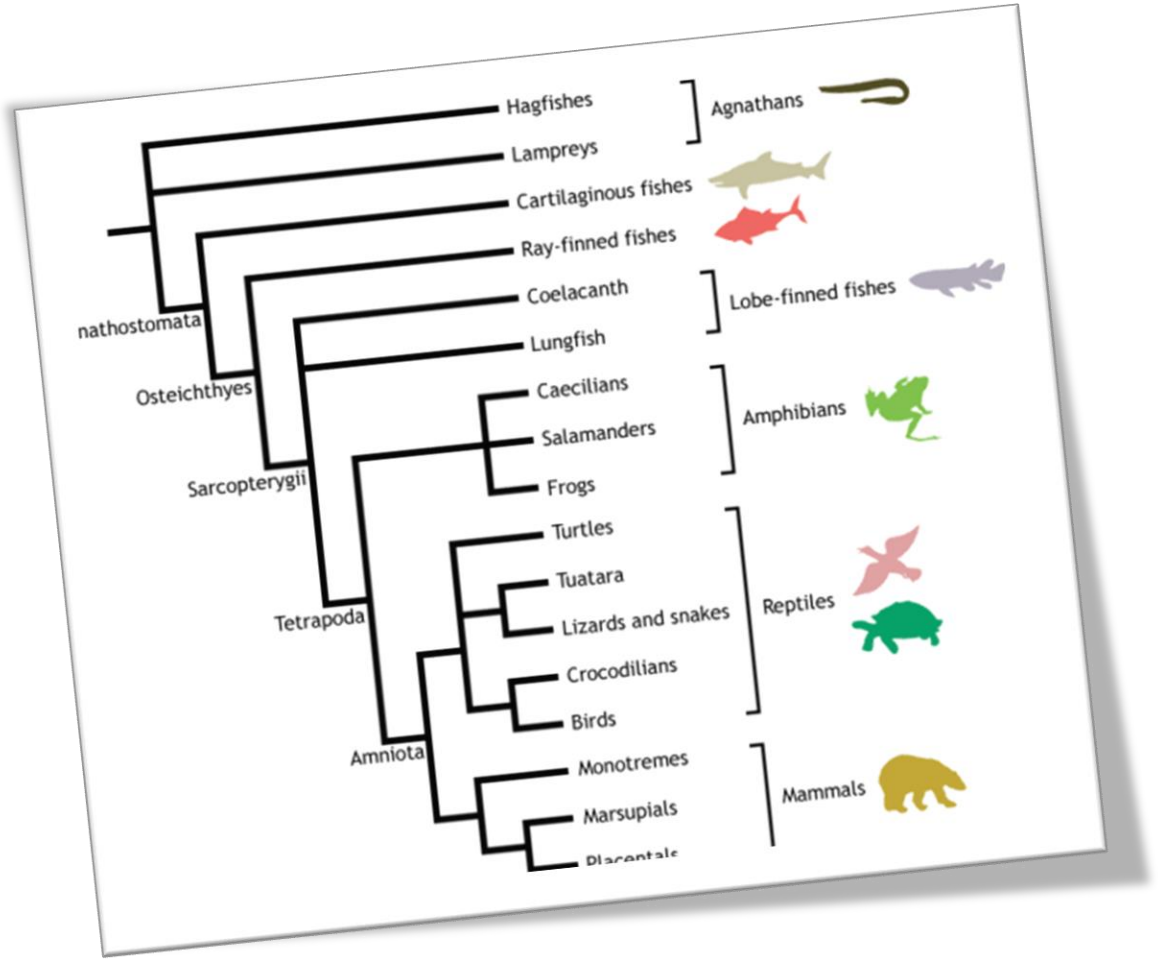
$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

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

CLUSTERING

Clustering jerárquico

- Idea:
  - Construir un árbol binario para representar los grupos.
  - Grupos separados se fusionan sucesivamente.
  - Fácil representación a través de un árbol.



Clustering jerárquico

¿Qué necesitamos?

- K-means requiere:
- Número de clusters K.
- Una agrupación inicial (al azar)
- Una métrica para definir la distancia entre dos puntos.

Clustering jerárquico:

- Una métrica para definir la distancia entre dos conjuntos de puntos.

CLUSTERING

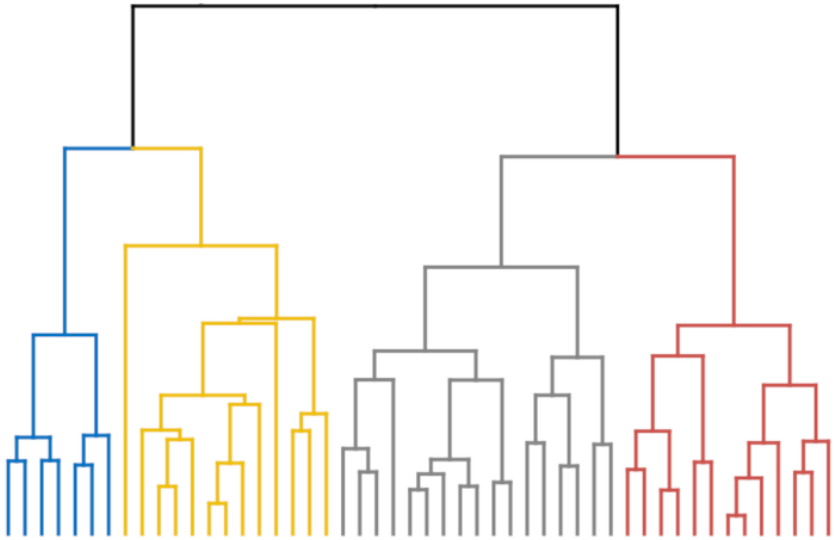
Clustering jerárquico

Tipos de clustering jerárquico:

- Aglomerativo
- Divisivo

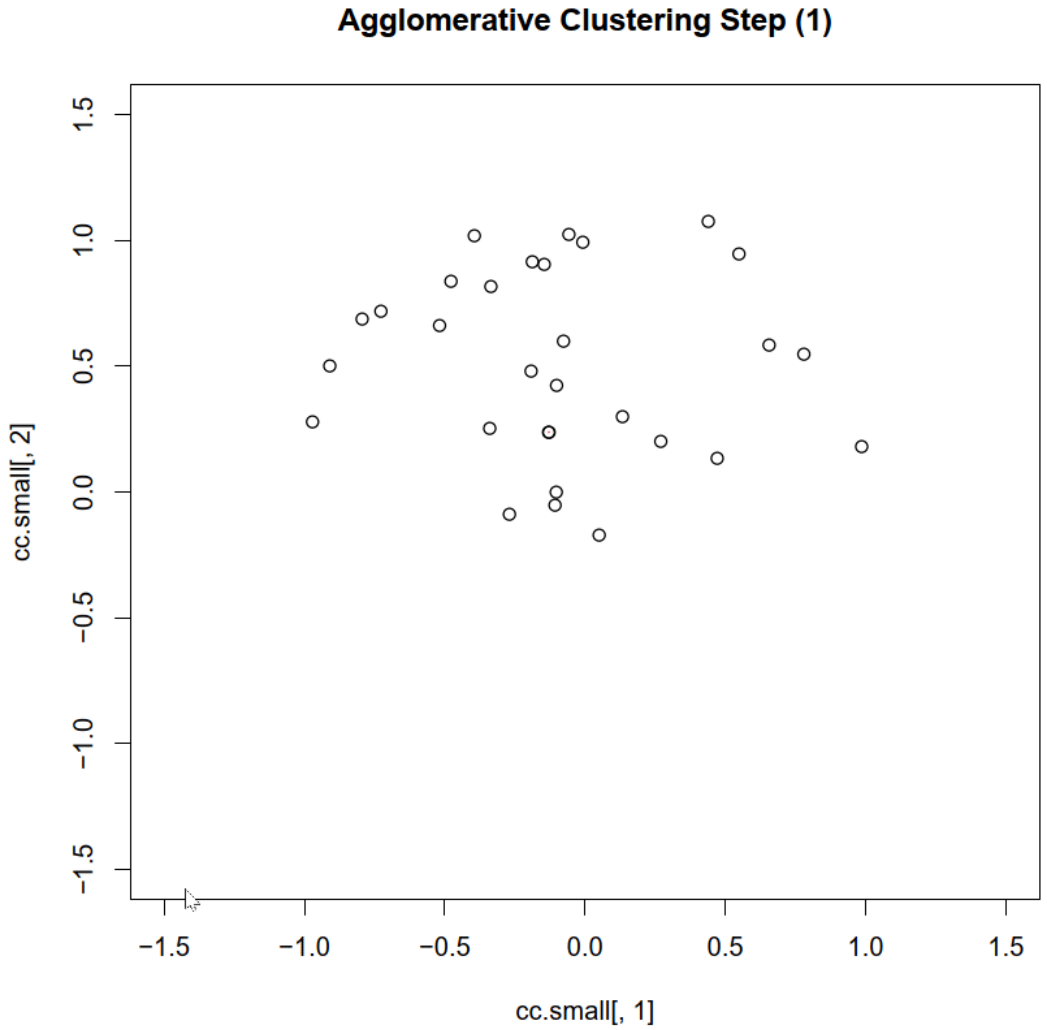
Idea del algoritmo:

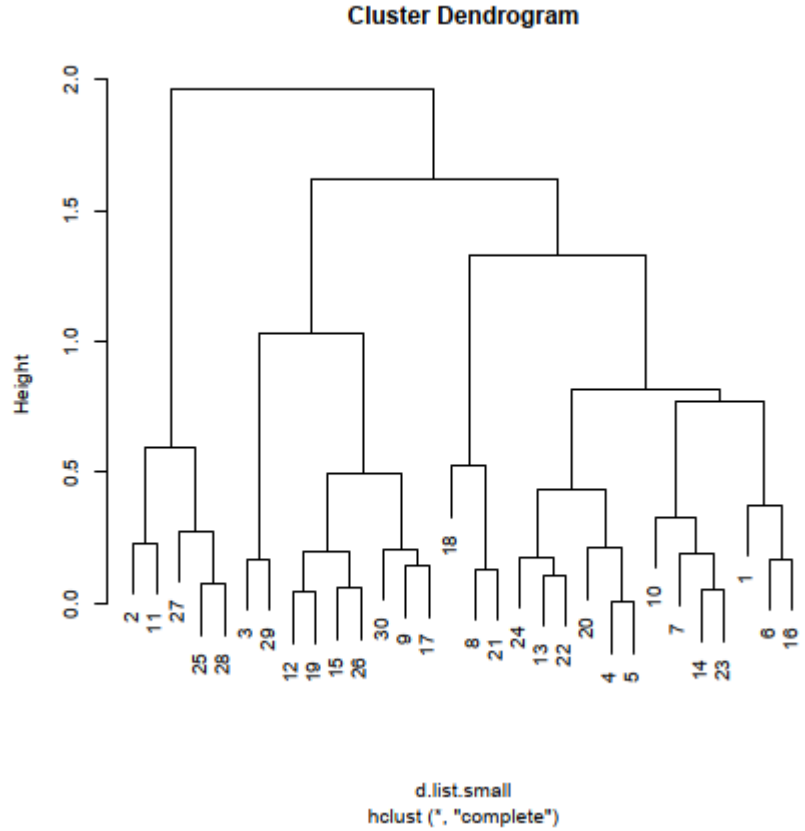
- Comenzar con todos los puntos como grupos individuales
- Repetir: Fusionar los dos grupos más “cercanos”
- Detener: cuando todos los grupos han sido fusionado en un solo gran grupo.



CLUSTERING

Clustering jerárquico





CLUSTERING

Clustering jerárquico

¿Cómo entendemos la distancia entre dos grupos?

The most popular choices:

- *Single-linkage*: the distance between the closest pair

$$d_{SL}(G, H) = \min_{i \in G, j \in H} d(x_i, x_j)$$

- *Complete-linkage*: the distance of the furthest pair

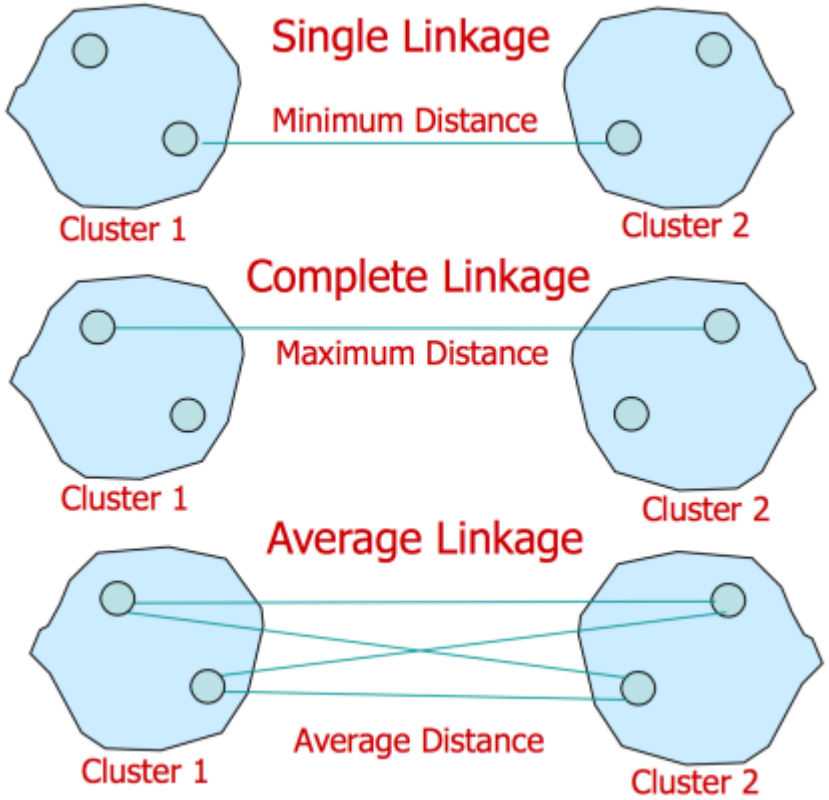
$$d_{CL}(G, H) = \max_{i \in G, j \in H} d(x_i, x_j)$$

- *Group-average*: the average distance between groups

$$d_{GA}(G, H) = \frac{1}{N_G N_H} \sum_{i \in G, j \in H} d(x_i, x_j)$$

- *Metroid:* the distance between the means or metroids of groups

$$d_{ME}(G, H) = d(\mu_G, \mu_H)$$



CLUSTERING

Clustering jerárquico

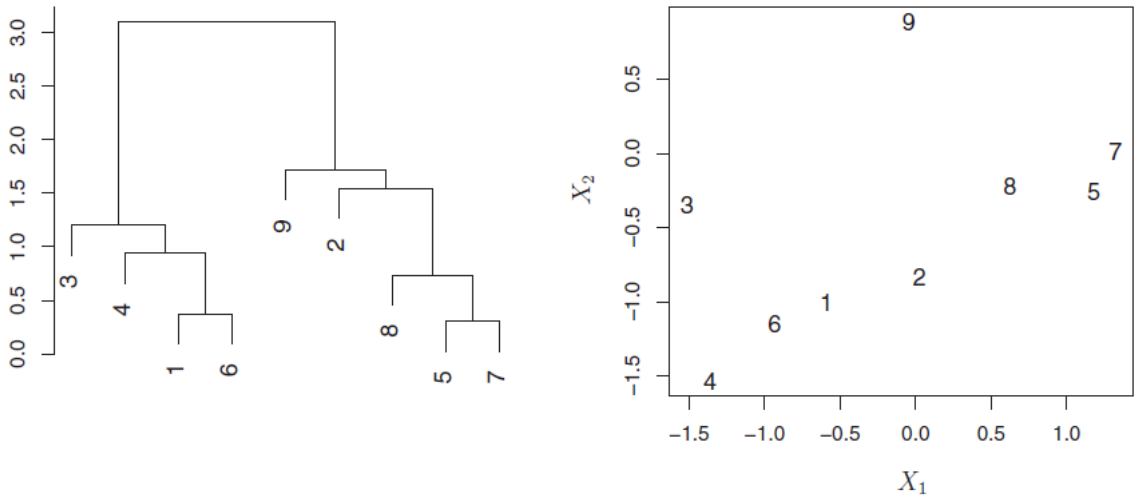


FIGURE 10.10. *An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space.* Left: *a dendrogram generated using Euclidean distance and complete linkage. Observations* 5 *and* 7 *are quite similar to each other, as are observations* 1 *and* 6. *However, observation* 9 *is* no more similar to *observation* 2 *than it is to observations* 8, 5, *and* 7, *even though observations* 9 *and* 2 *are close together in terms of horizontal distance. This is because observations* 2, 8, 5, *and* 7 *all fuse with observation* 9 *at the same height, approximately* 1.8. Right: *the raw data used to generate the dendrogram can be used to confirm that indeed, observation* 9 *is no more similar to observation* 2 *than it is to observations* 8, 5, *and* 7.

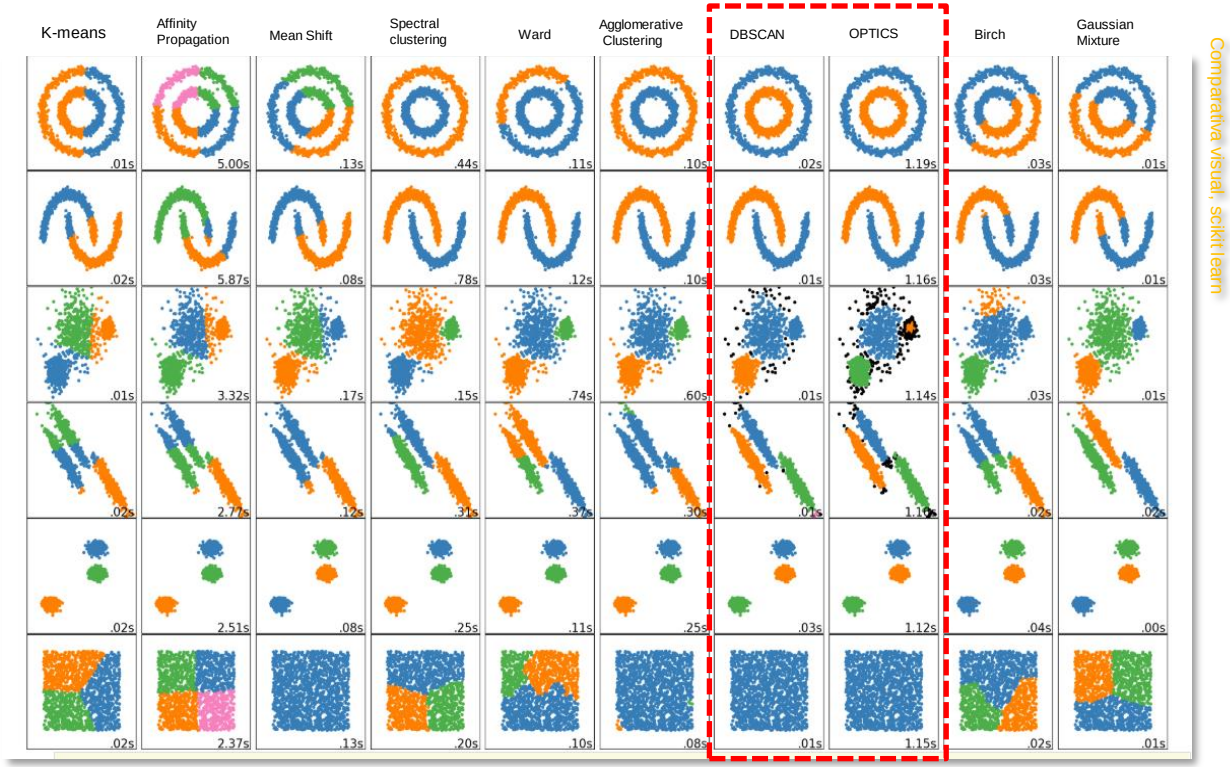
CLUSTERING

EDUCACIÓN PROFESIONAL

Clustering jerárquico

- Los métodos de clustering son sensibles a la forma de los conjuntos que buscamos agrupar.
- La siguiente comparativa muestra la forma en que logran caracterizar los grupos los distintos métoos.

- Otro enfoque que permite aislar los puntos que están fuera de zonas "densas" son los algoritmos DBSCAN y OPTICS, los que a su vez pueden ser utilizados como un método de detección de anomalías.



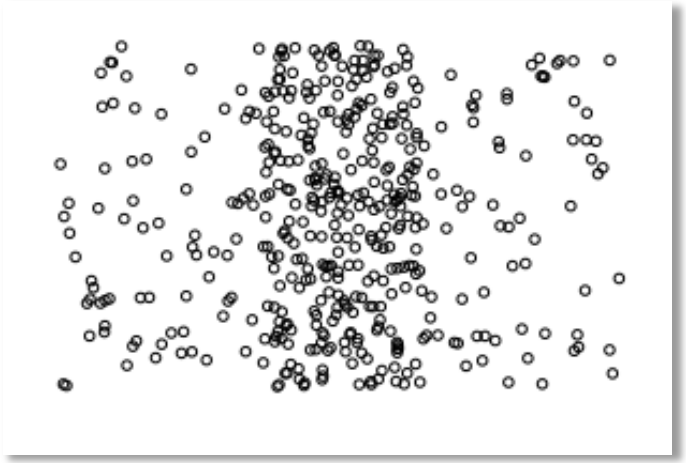
En esta imagen de referencia se destacan en color naranjo, azul y verde, los clústers generados por cada algoritmo, mientras que los puntos destacados en negro corresponden a puntos con baja densidad local,y por ende considerados como atípicos.

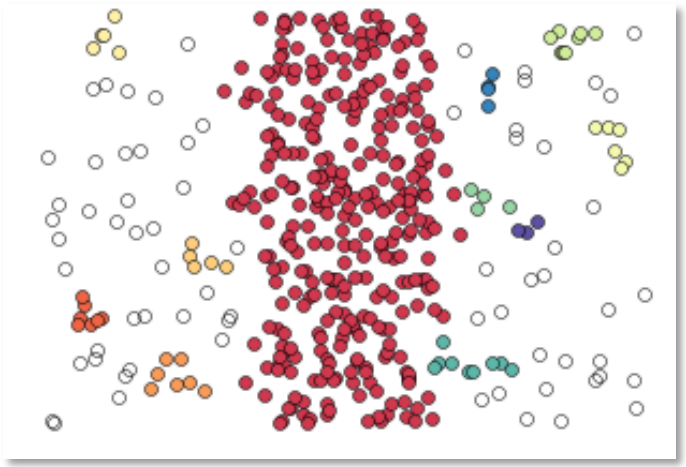
CLUSTERING

DBSCAN
Density-based spatial clustering of applications with noise

Idea del algoritmo

1. Se eligen 2 parámetros y un valor épsilon (eps) y un número de puntos (minPts).
2. Se elige un punto aleatorio del espacio
3. Se buscan todos aquellos puntos, que estén a una distancia igual o menor a eps del punto inicial.
1. Si hay más de minPts puntos (incluyendo el inicial), se eligen sólo minPts
4. Se expande el cluster , revisando todos los nuevos puntos, y se ve si ellos forman un cluster, incrementando el cluster recursivamente.
5. Cuando se acaban los puntos, se toma un nuevo punto y se comienza de nuevo.
6. Aquellos puntos que no entran (por estar más distantes de e de los otros) se les considera "ruido" y quedan aislados.





CLUSTERING

DBSCAN
Density-based spatial clustering of applications with noise

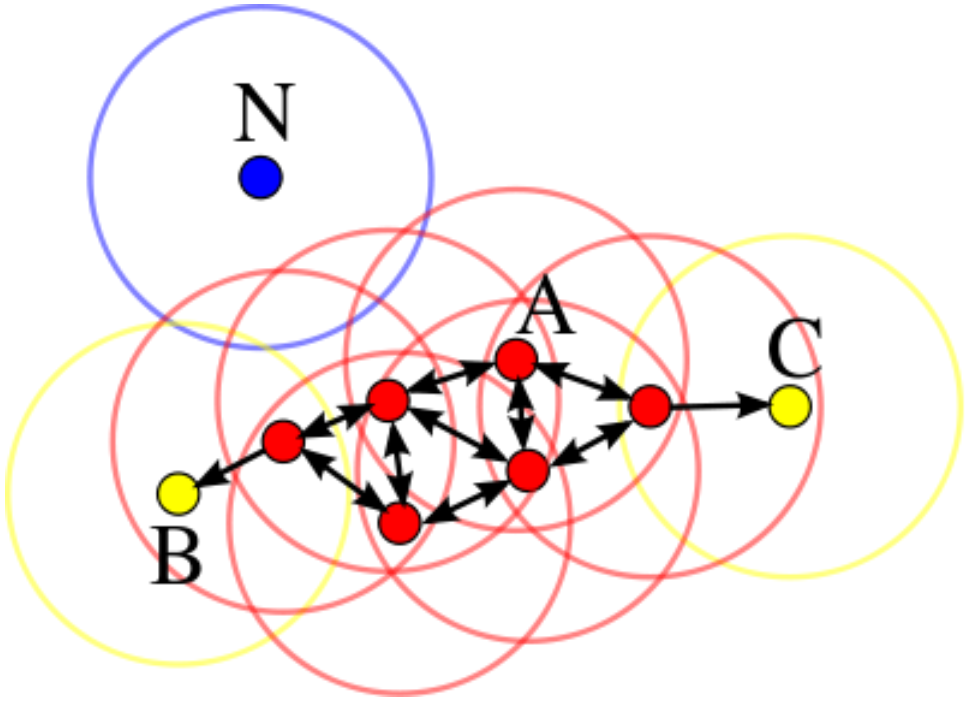
Conceptos claves

1. Core-point: Si al menos minPts están en la vecindad de radio eps de p

2. Directly Reachable-point: Un punto q es alcanzable desde p (core-point) si está a distancia menor que eps

3. Reachable-point: un punto q es densamente alcanzable desde p, si existe una ruta de puntos p_1,..., p_n , con p_1= p y p_n=q, donde p_(i+1) es directamente alcanzable desde p_i .Notar que p_i es punto núcleo con la posible salvedad de q.

4. Outlier (noise point): Un punto que no sea directa o densamente alcanzable desde cualquier otro punto.



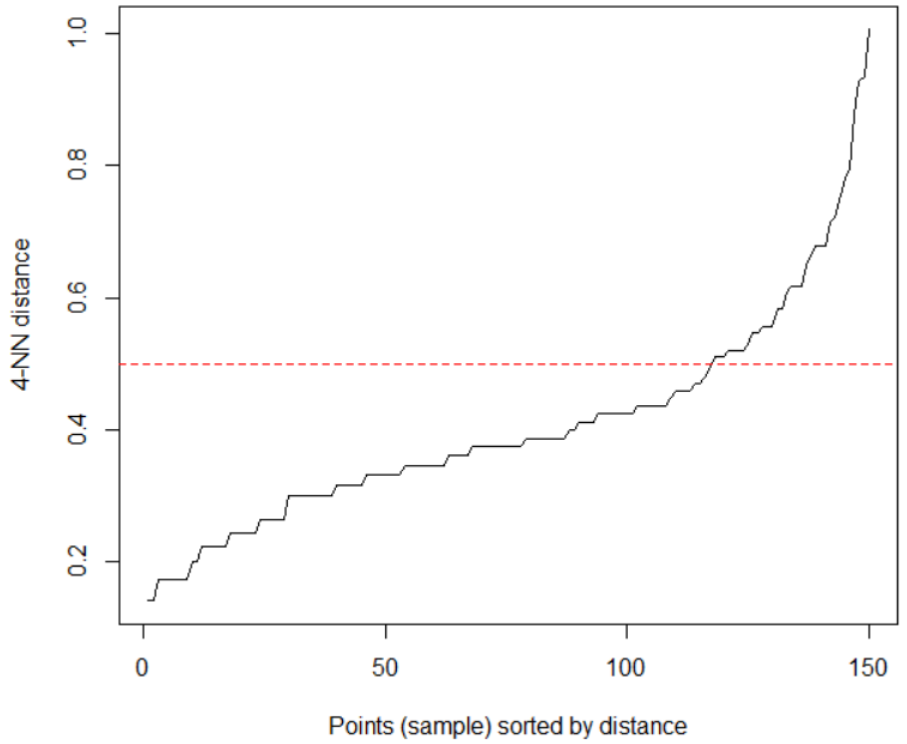
minPts = 4
A: Puntos núcleos
B, C: Puntos densamente alcanzables
N: Ruido

CLUSTERING

DBSCAN
Density-based spatial clustering of applications with noise

Notar que los dos parámetros eps y minPts, deben ser definidos inicialmente. Estos determinarán la calidad de la partición generada.

- Usualmente se considera minPts como p+1, donde p es la dimensión del espacio de atributos de nuestro dataset.
- Una manera de escoger el valor de eps es mediante la visualización de las KNN distancias de cada punto.
  - Para cada punto se calcula la distancia al k = minPts vecino más cercano.
  - Se ordenan las distancias de menor a mayor.
  - Se busca el punto de mayor crecimiento (regla del "codo")



CLUSTERING

OPTICS
Ordering Points To Identify Clustering Structure

Una extensión del algoritmo DBSCAN corresponde al algoritmo OPTICS

- Comparte los mimos conceptos que DBSCAN, pero a cada punto se le asignan nos nuevas distancias que lo carcterizan:
  - $Core{-}dist\_(\epsilon,minPts)\ (p)$= distancia al minPts-ésimo punto más cercano dentro de $N\_\epsilon\ (p)$
    Sólo definida si p es punto núcleo.
  - $Reachability{-}dist\_(\epsilon,minPts)\ (o,p)$=max($Core{-}dist\_(\epsilon,minPts)\ (p)$, $dist(p,o)$).
    Sólo definida si p es un punto núcleo.

- El algoritmo propone un ordenamiento de la base de datos, a modo de poder calcular bajo dicha indexación la distancia de alcance de cada punto.

- Las distancias de alcance funcionan como una especie de dendograma, y permitirá establecer la cantidad de clusters a generar.

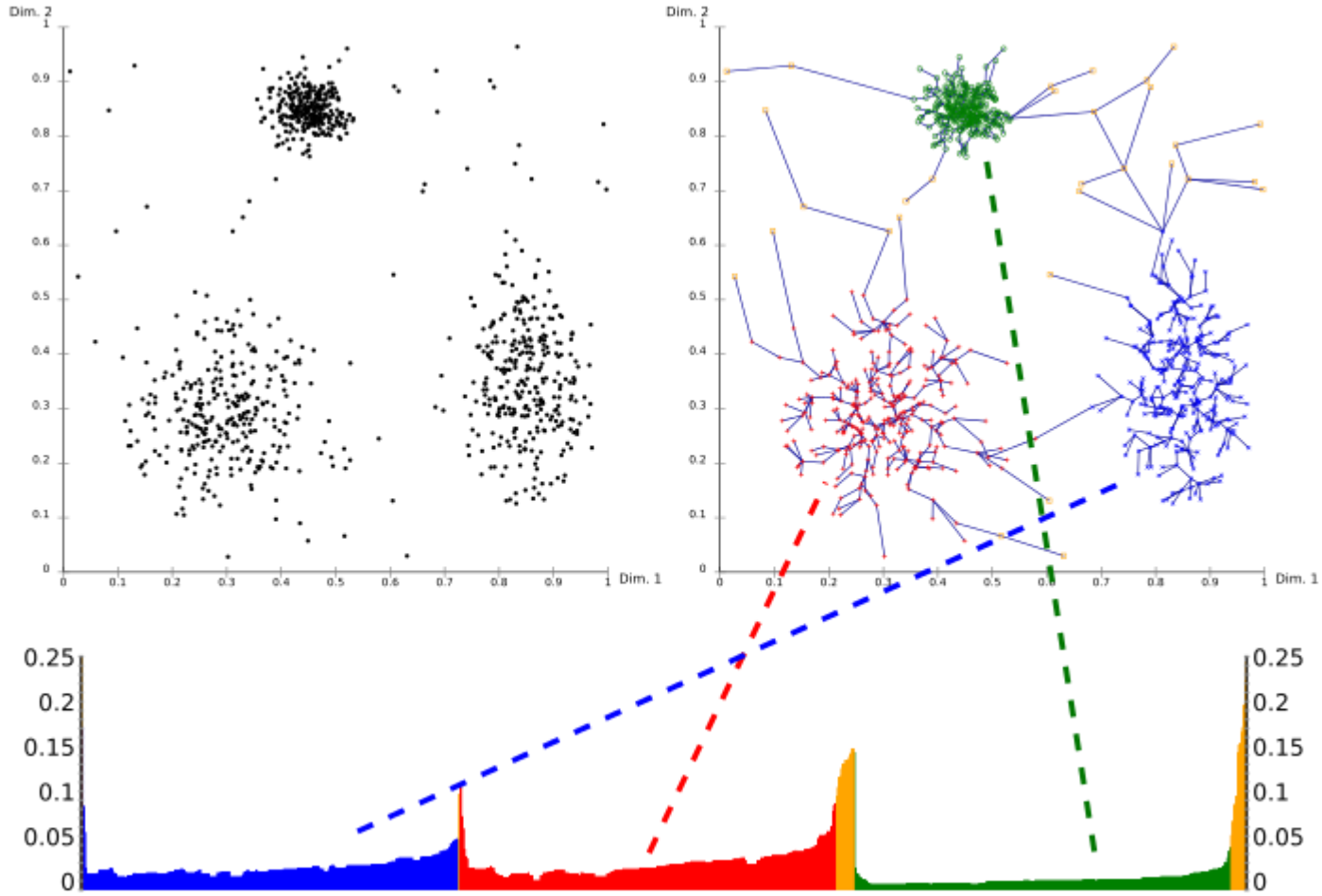
- A diferencia de DBSCAN, el parámetro eps no es requerido (gracias a la inclusión de la distancia de alcance), pero se recomienda su uso para efectos de costo computacional.

CLUSTERING

EDUCACIÓN PROFESIONAL

OPTICS
Ordering Points To Identify Clustering Structure

- El gráfico de las distancias de alcance (reachability plot), contiene los puntos bajos el odenamiento propuesto por el algoritmo en el eje-x , mientras que en el eje-y la distancia de alcance respectiva.
- Puntos que pertenezcan al mismo cluster, tendrán una menor distancia de alcance a sus vecinos cercanos.
- Los valles representan los clusters.
- Mientras más pronunciado es el valle, mayor densidad del cluster



CLUSTERING

EDUCACIÓN PROFESIONAL

OPTICS
Ordering Points To Identify Clustering Structure

Trabajo original

- https://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=C75424AECC04C36AC911CCCC7DB41238?doi=10.1.1.129.6542&rep=rep1&type=pdf

Extensión OPTICS-OF para la detección de outliers

- https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.46.6586&rep=rep1&type=pdf

Manual de referencia package dbscan

- https://cran.r-project.org/web/packages/dbscan/dbscan.pdf