ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Introducción a la Ciencia de Datos con R

Miguel Jorquera

Educación Profesional
Escuela de Ingeniería

El uso de apuntes de clases estará reservado para finalidades académicas. La reproducción total o parcial de los mismos por cualquier medio, así como su difusión y distribución a terceras personas no está permitida, salvo con autorización del autor.

RESUMEN

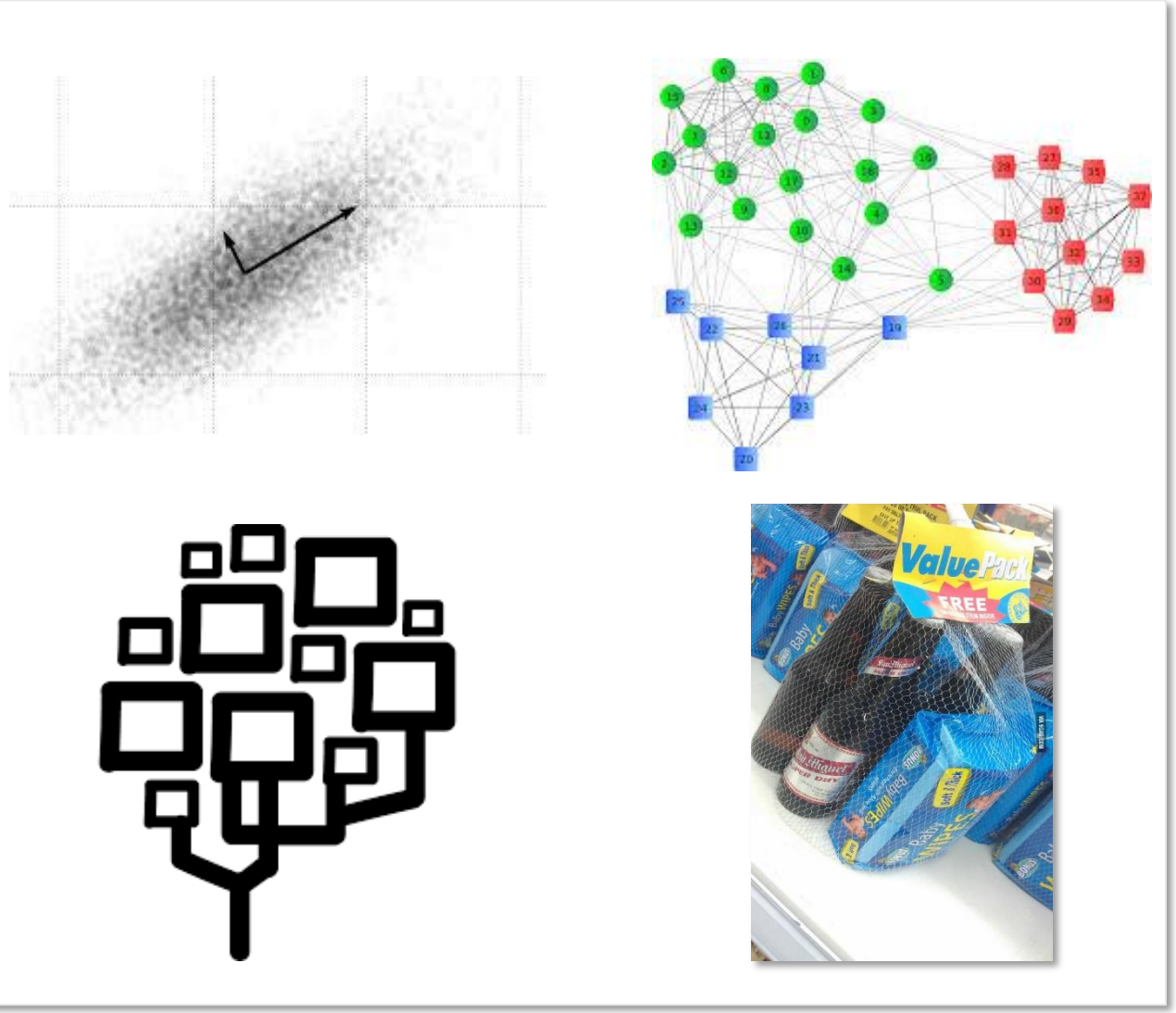
RESUMEN

EDUCACIÓN PROFESIONAL

Aprendizaje no supervisado

Algunas tareas usuales:

- Reducción de la dimensionalidad
- Generación de clusters
  - Comenzamos con K-means
- Reglas de asociación
  - Estudiamos el algoritmo apriori



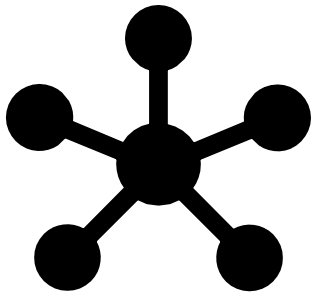
TEMAS PARA HOY

AGENDA

EDUCACIÓN
PROFESIONAL

Temas para hoy

- Clustering parte 2
- Introducción a modelos supervisados
  - Regresión Lineal
  - Validación y flexibilidad de métodos

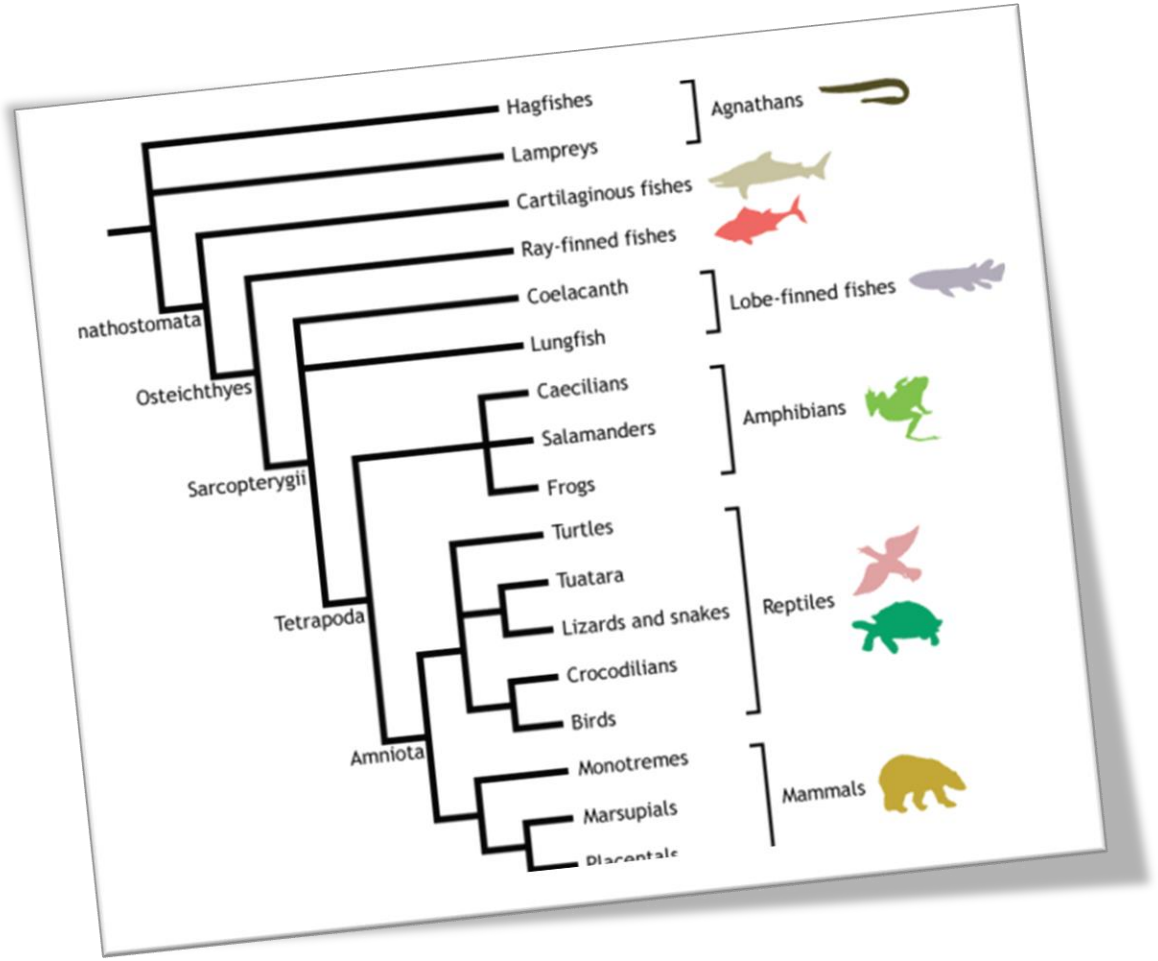


CLUSTERING (CONTINUACIÓN)

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

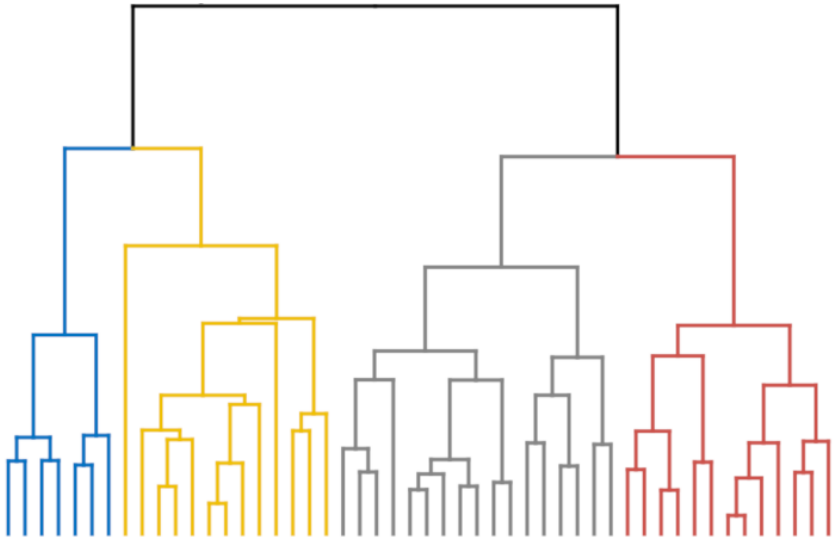
Clustering jerárquico

Tipos de clustering jerárquico:

- Aglomerativo
- Divisivo

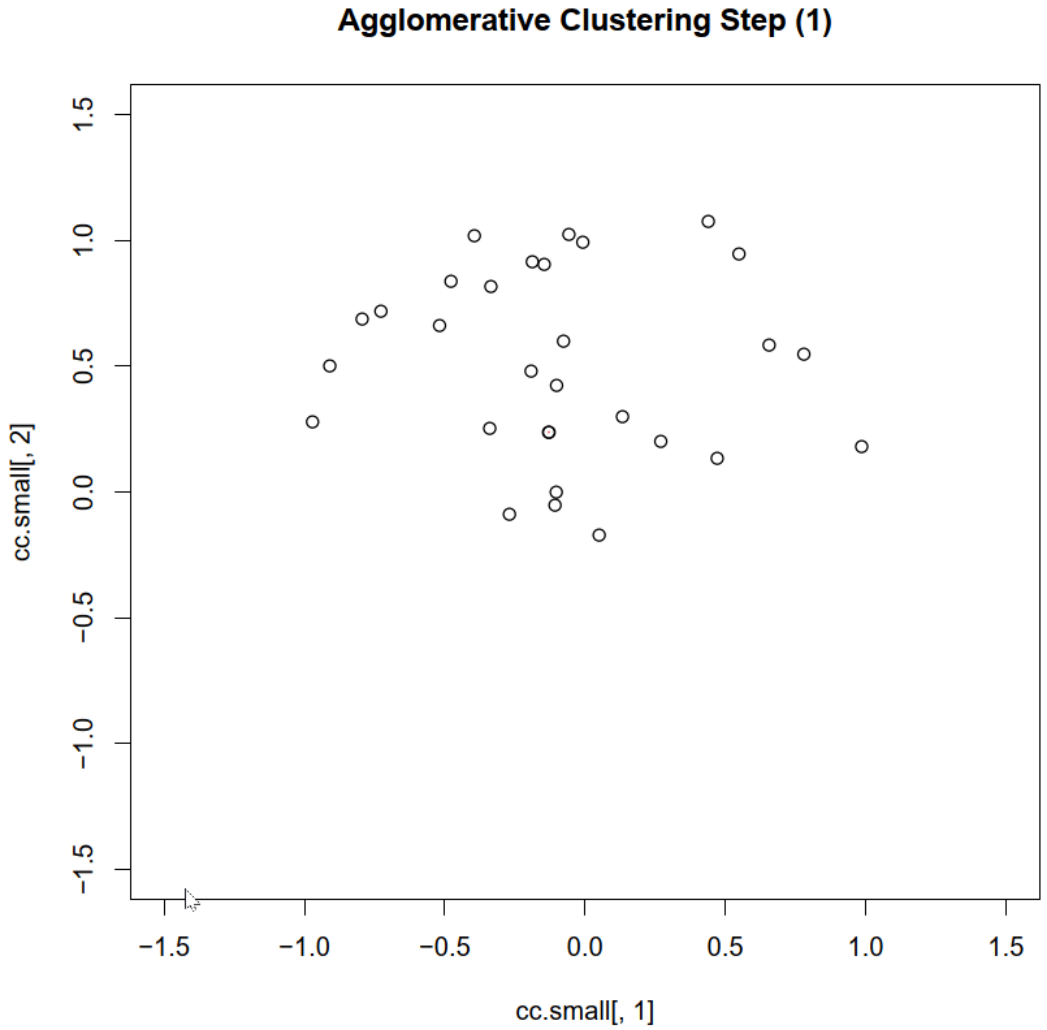
Idea del algoritmo:

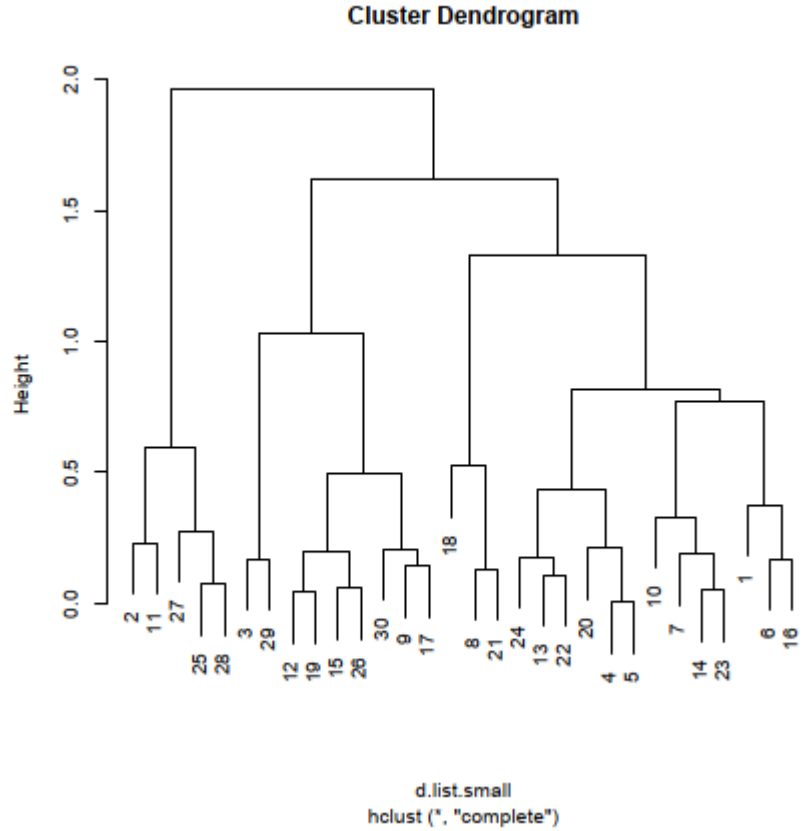
- Comenzar con todos los puntos como grupos individuales
- Repetir: Fusionar los dos grupos más “cercanos”
- Detener: cuando todos los grupos han sido fusionado en un solo gran grupo.



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

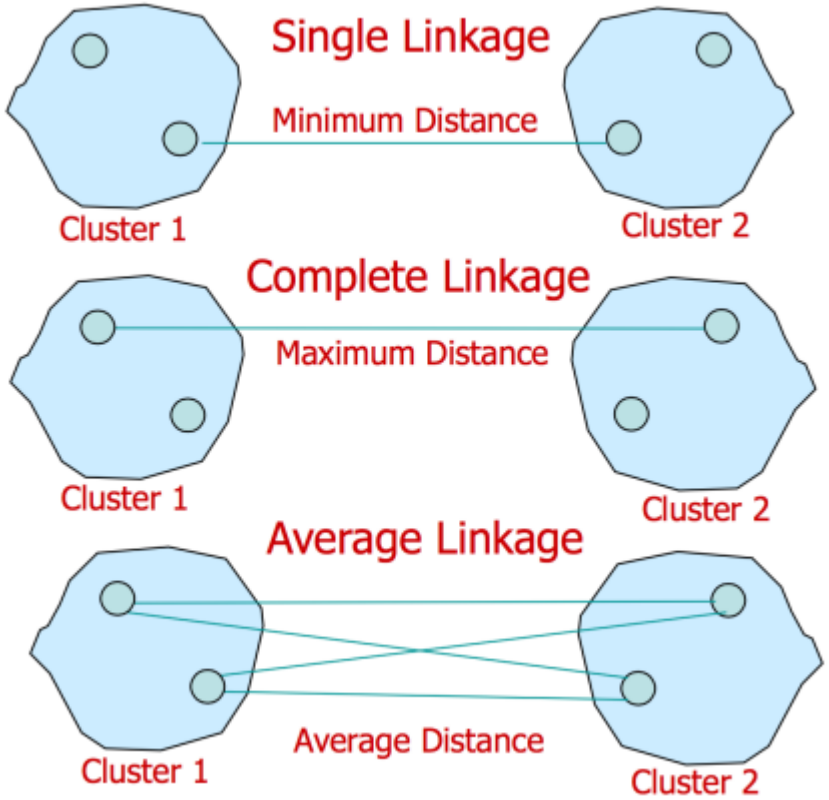
$$d_{CL}(G, H) = \max_{i \in G, j \in H} d(x_i, x_j)$$

- *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

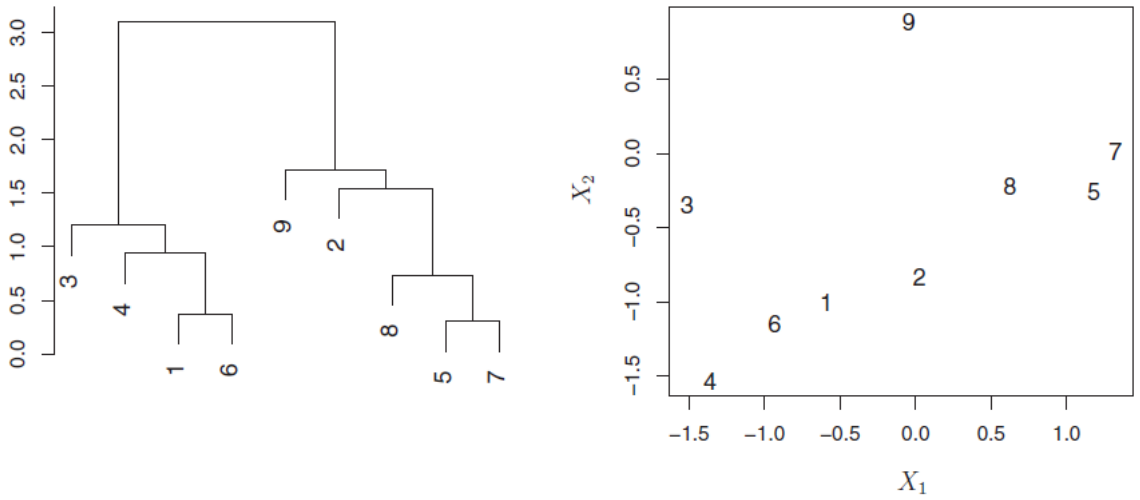


FIGURE 10.10. *An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space.* Left: *a dendrogram generated using Euclidean distance and complete linkage. Observations* 5 *and* 7 *are quite similar to each other, as are observations* 1 *and* 6. *However, observation* 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.

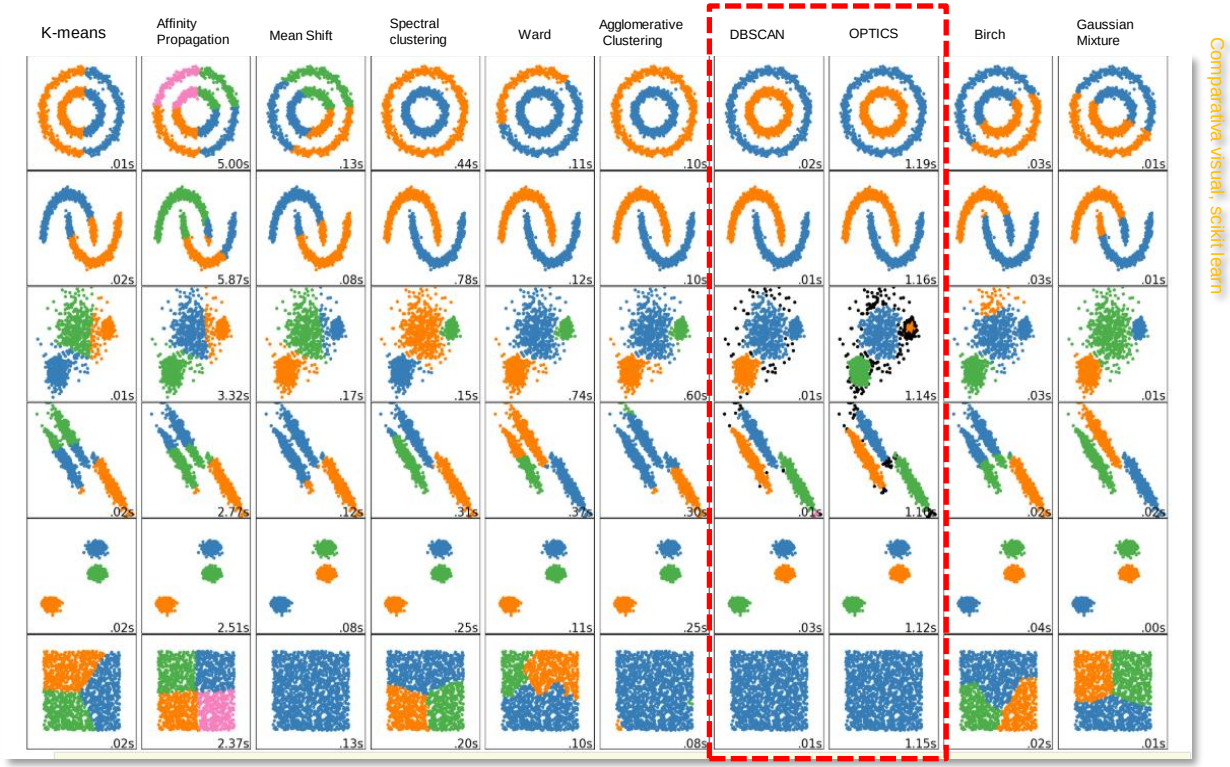
CLUSTERING

EDUCACIÓN PROFESIONAL

Otros algoritmos

- 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 métodos.

- 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.



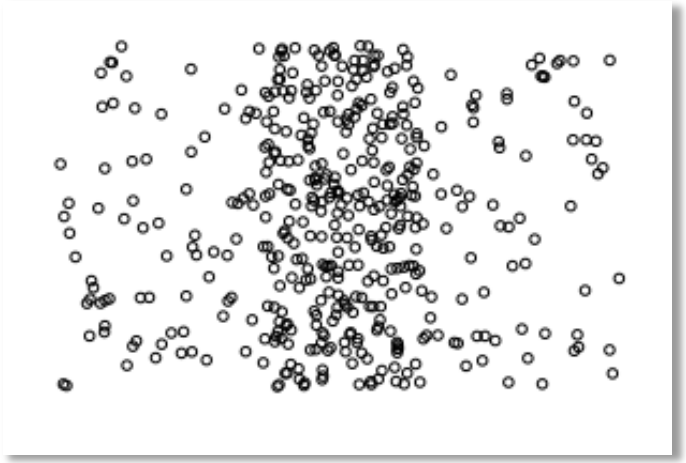
En esta imagen de referencia se destacan en color naranjo, azul y verde, los clústers generados por cada algoritmo, mientras que los puntos destacados en negro corresponden a puntos con baja densidad local,y por ende considerados como atípicos.

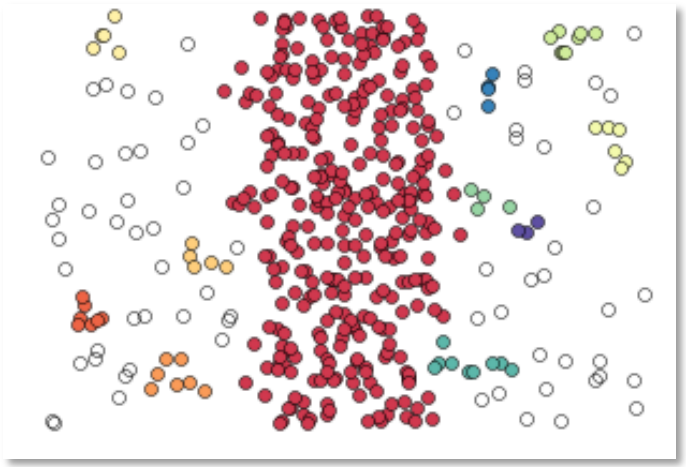
CLUSTERING

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

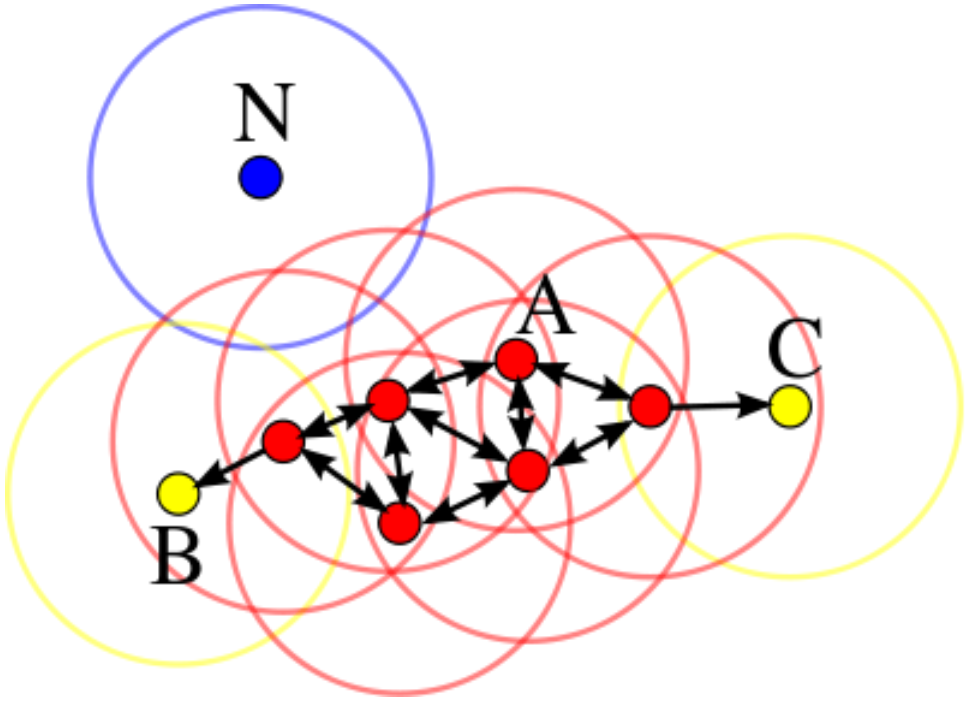
Conceptos claves

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.



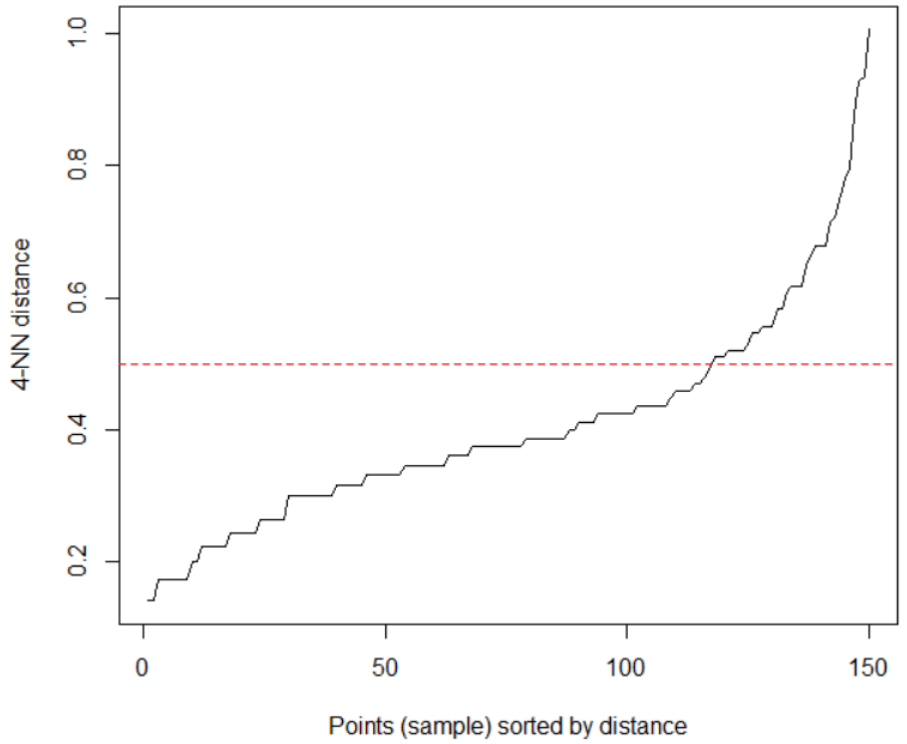
minPts = 4
A: Puntos núcleos
B, C: Puntos densamente alcanzables
N: Ruido

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.

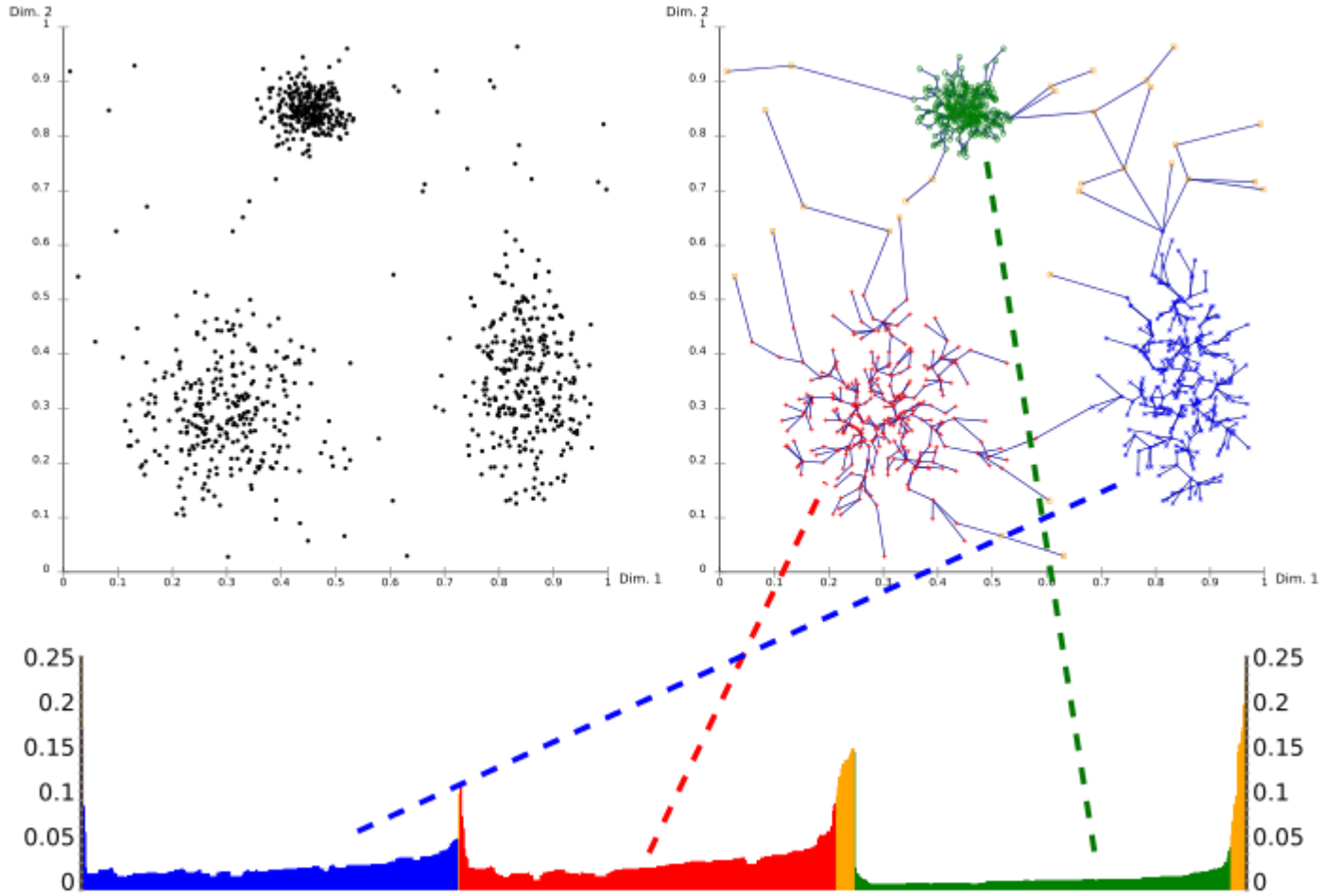
- A diferencia de DBSCAN, el parámetro eps no es requerido (gracias a la inclusión de la distancia de alcance), pero se recomienda su uso para efectos de costo computacional.

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

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

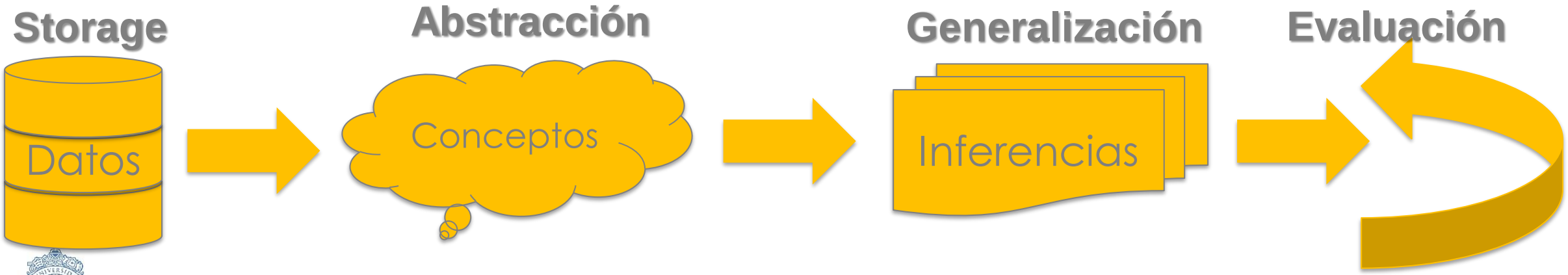
EDUCACIÓN
PROFESIONAL

INTRODUCCIÓN MACHINE LEARNING
Modelos supervisados y regresión

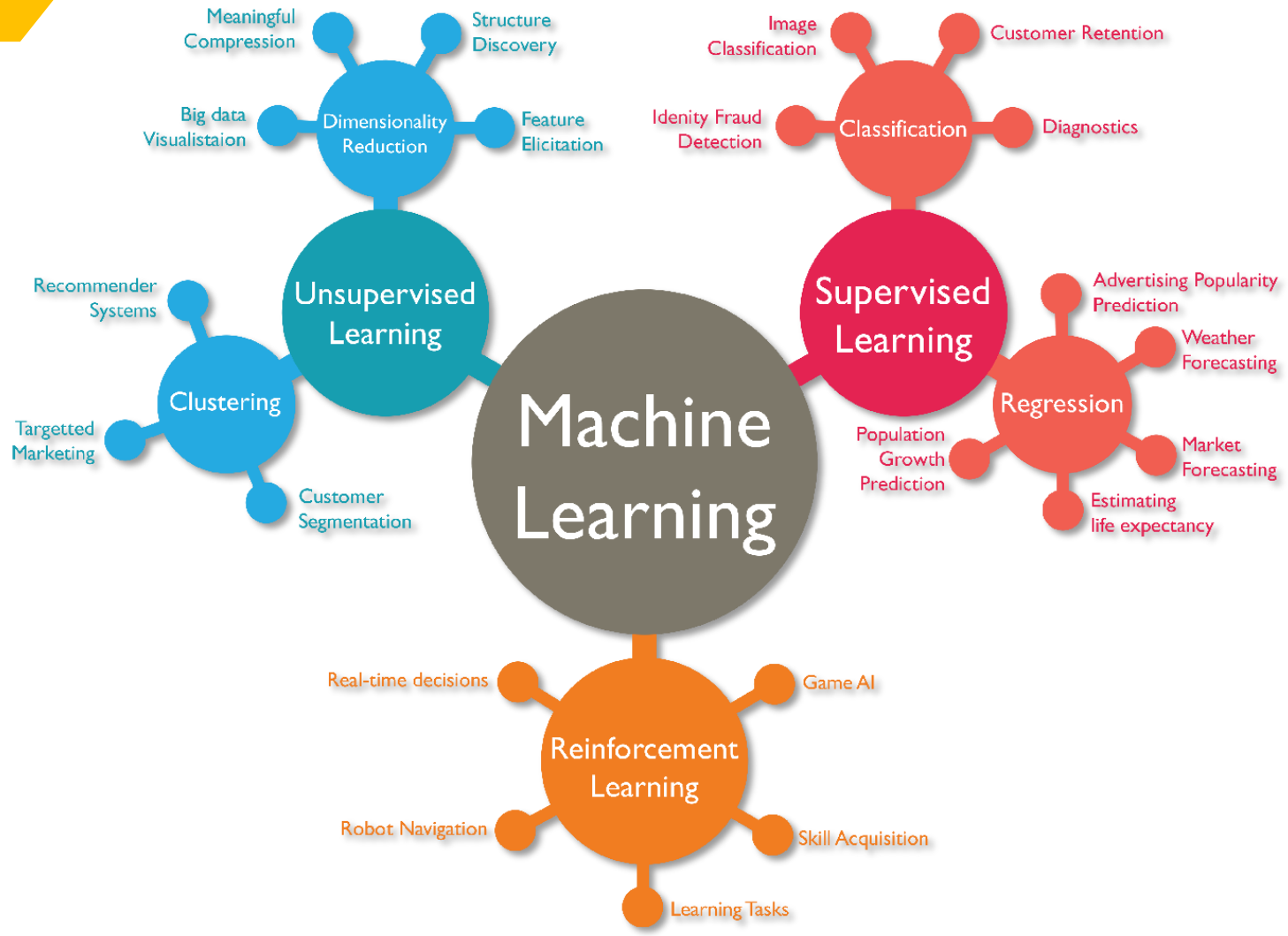
Machine Learning

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

Tom M. Mitchell (1997)



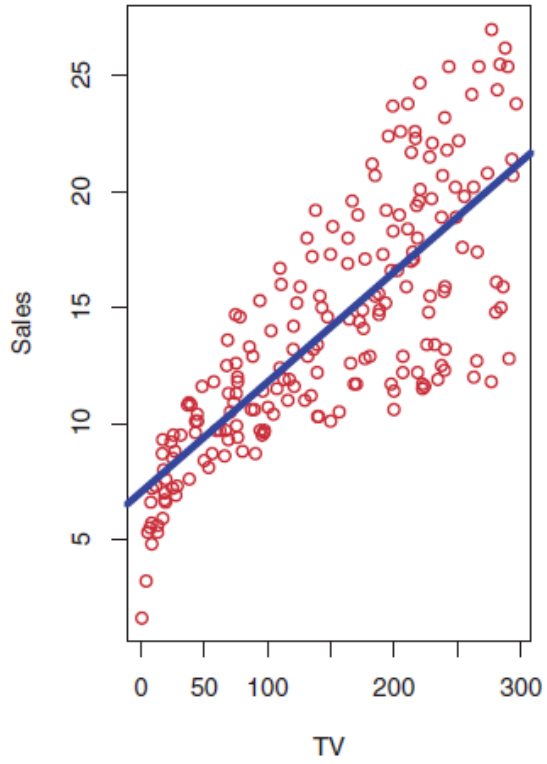
Machine Learning

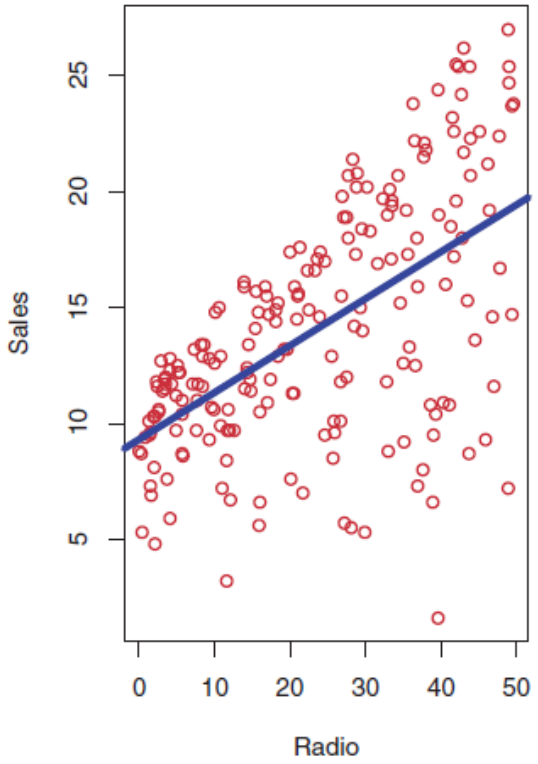


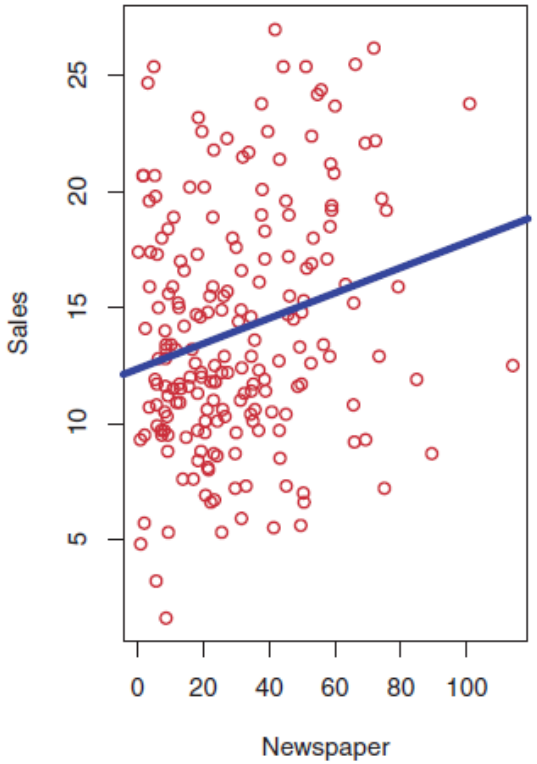
MODELOS DE REGRESIÓN

Regresión lineal simple

Supongamos que interesa predecir el nivel de ventas de un determinado producto en función de los montos invertidos en distintos medios de publicidad (tv, radio, periódico).







MODELOS DE REGRESIÓN

Regresión lineal simple

- En general, el problema podría expresarse matemáticamente de la siguiente manera:

$$Y = f(X) + \epsilon \qquad (1)$$

- Donde X contiene a las variables explicativas (monto en tv, radio y periódico en el ejemplo), y ϵ es un error aleatorio no observable.
- En esta especificación, f es una función desconocida y que buscamos estimar.
- En términos simples, diremos que un modelo es de regresión, cuando en la expresión (1), la variable de interés a predecir, Y, es una variable numérica.

Regresión lineal simple

- Cuando el modelo f a estimar, se asume como una función lineal, diremos que (1) es un modelo de regresión lineal. En tal caso, el modelo matemático queda expresado de la siguiente manera:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \ + \epsilon \qquad (1)$$

- Donde $\beta_i, i = 0, \ldots, p$ son los parámetros a estimar (estos parámetros definen al modelo), y ϵ es un error aleatorio no observable, típicamente siguiendo una distribución aleatoria normal $N(0, \sigma^2)$.

REGRESIÓN LINEAL

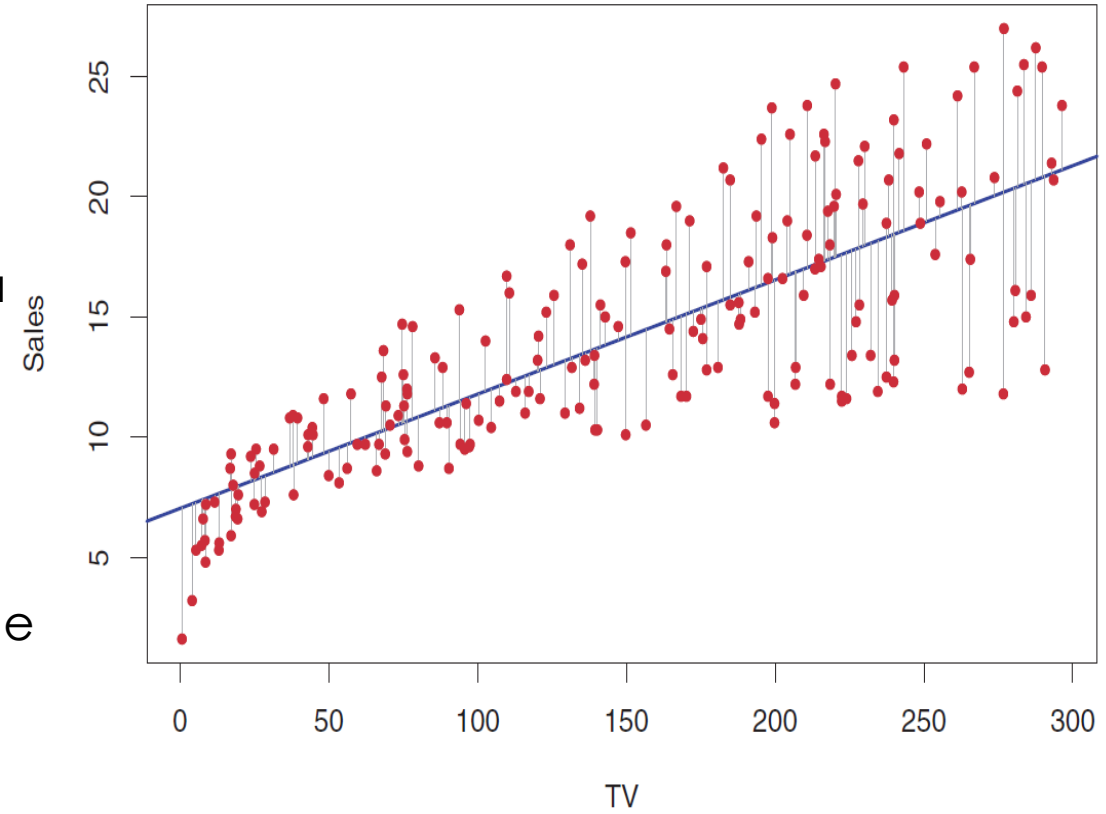
Regresión lineal simple

¿Cómo se estiman los parámetros en una regression lineal?

- Tanto en R como en la mayoría de los softwares, la manera estándar de estimar los coeficientes en un modelo de regresión, es mediante la estimación vía mínimos cuadrados, donde se busca minimizar la suma residual:

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

- No entraremos en detalle, respecto de las bondades de esta estimación y que coincide con otros estimadores en el caso de la regresión lineal con errores normales.



REGRESIÓN LINEAL

Regresión lineal simple

- Sin entrar en más detalles técnicos, veamos como podemos ajustar una regresión lineal en R.
- Para ello podemos utilizar la función `lm()`, del paquete base. Esta recibe como argumento principal una formula y un datasets, del siguiente modo.

```
lm(formula = y ~ x1+x2+…+xp, data = dataset)
```

- Generemos nuestra primera regresión lineal con el dataset "Advertising". El cual contiene las ventas totales de un producto y los montos invertidos en tres tipos de publicidad (tv, radio, periódico). En esta primera iteración consideremos solamente la variable newspapper

```
lm(formula = sales ~ newspapper, data = Advertising)
```

- Hablaremos sobres los coeficientes estimados y la salida que genera R en el notebook

REGRESIÓN LINEAL

Regresión lineal simple

¿Cómo interpretamos los parámetros de una regresión lineal (simple)