MACHINE LEARNING: ALGORITMOS NO SUPERVISADOS

Clustering

Instructor: José Nelson Zepeda

San Salvador, Noviembre 2018

Clustering

Algoritmos No supervisados

Clustering

K-Means

Cuantos Clusters

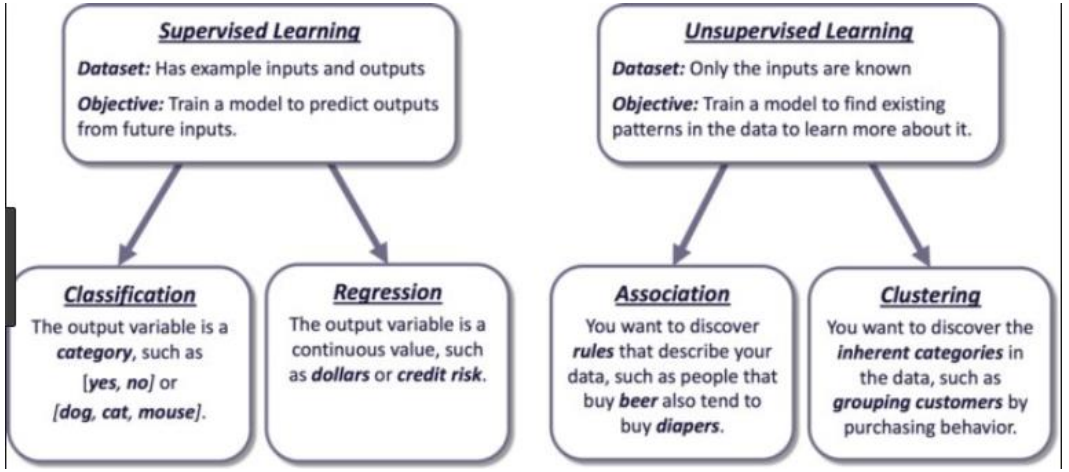
Algoritmos Jerárquicos

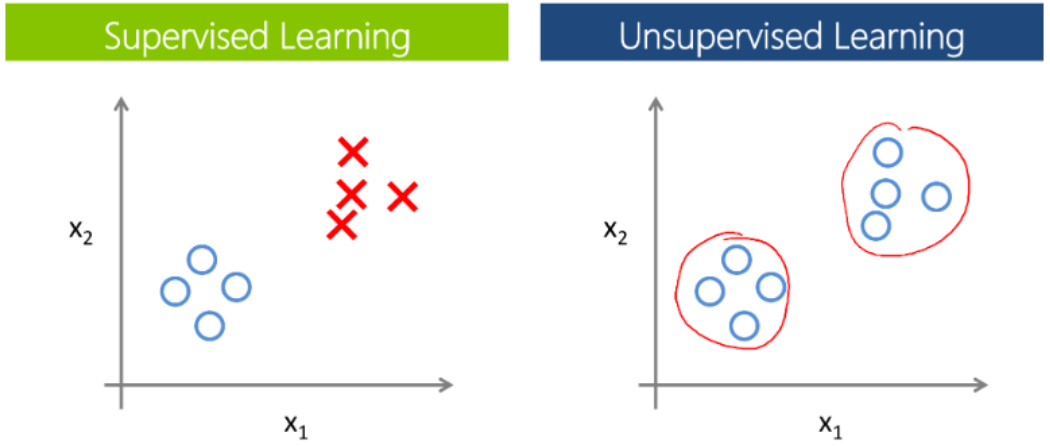
Conceptos Básicos

¿Qué es Aprendizaje No Supervisado?

Aprendizaje no supervisado es un método de Aprendizaje Automático donde un modelo es ajustado a las observaciones. Se distingue del Aprendizaje supervisado por el hecho de que no hay un conocimiento a priori.

La principal ventaja que presenta la clasificación no supervisada es que se puede obtener un conjunto de entrenamiento empleando muestras no etiquetadas valiéndose de algoritmos de agrupamiento, sin embargo, es necesario mencionar que el aprendizaje no supervisado es más subjetivo que el supervisando ya que su objetivo no se enfoca en una predicción o una respuesta.

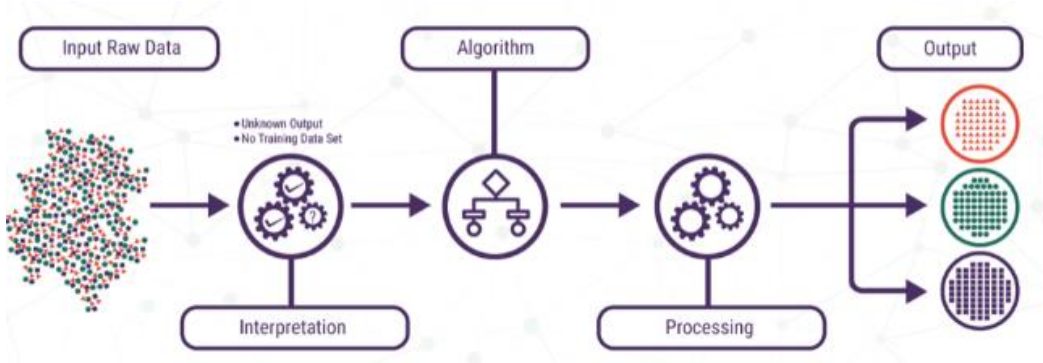




Usos del Aprendizaje No supervisado

Por sus características, los algoritmos de aprendizaje no supervisado orientados al agrupamiento son herramientas muy utilizadas en distintos contextos como

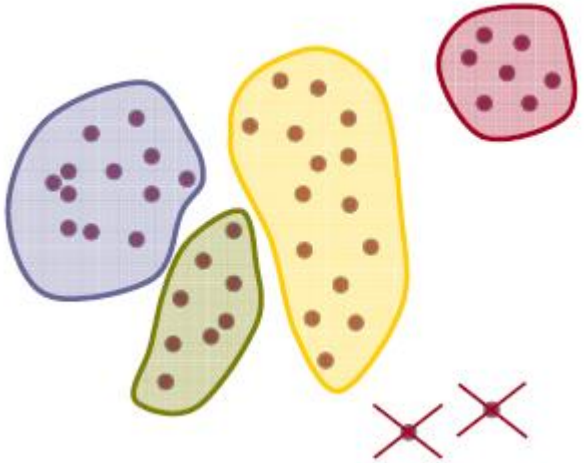
- la Recuperación de Información y la Minería de Textos
- El procesamiento de secuencias descriptoras de genes y proteínas
- El seguimiento y detección de sucesos en un flujo continuo de noticias
- La segmentación de imágenes
- La segmentación o perfilamiento
- La compresión de datos
- El procesamiento de bases de datos espaciales
- La clasificación de zonas geográficas
- La comprensión de imágenes de satélites
- La visualización de datos
- La prospección geológica
- La organización de documentos en bibliotecas
- y en muchas otras aplicaciones como la estructuración de grandes volúmenes de datos.



Usos más Frecuentes

A continuación se enumeran algunas aplicaciones de aprendizaje no supervisado:

1. Clustering: Permite dividir los datos en diferentes grupos en función de su similaridad.
2. Deteccion de Anomalias: Permite descubrir observaciones o datos inusuales, es muy útil en la búsqueda de actividad fraudulenta.
3. Asociación: Identifica objetos u eventos que frecuentemente ocurren juntos, el análisis de canasta es el ejemplo principal.
4. Variables Latentes: Técnicas que se aplican durante la fase de pre-procesamiento tales como la reducción de variables en un dataset



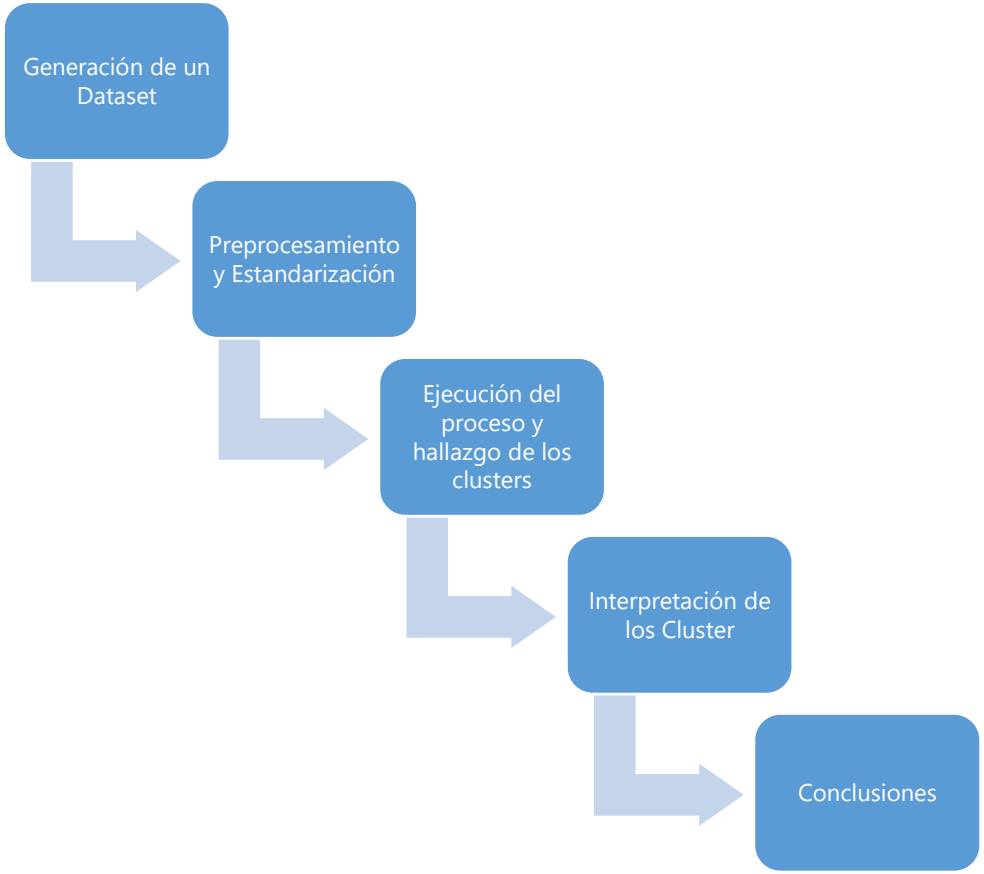
CLUSTERING

Análisis de Clusters

El análisis de clusters, se considera una técnica de aprendizaje no supervisado ya que su objetivo es encontrar las relaciones entre las diferentes variables de estudio teniendo en cuenta que las relaciones descubiertas no están en función de ninguna variable target.

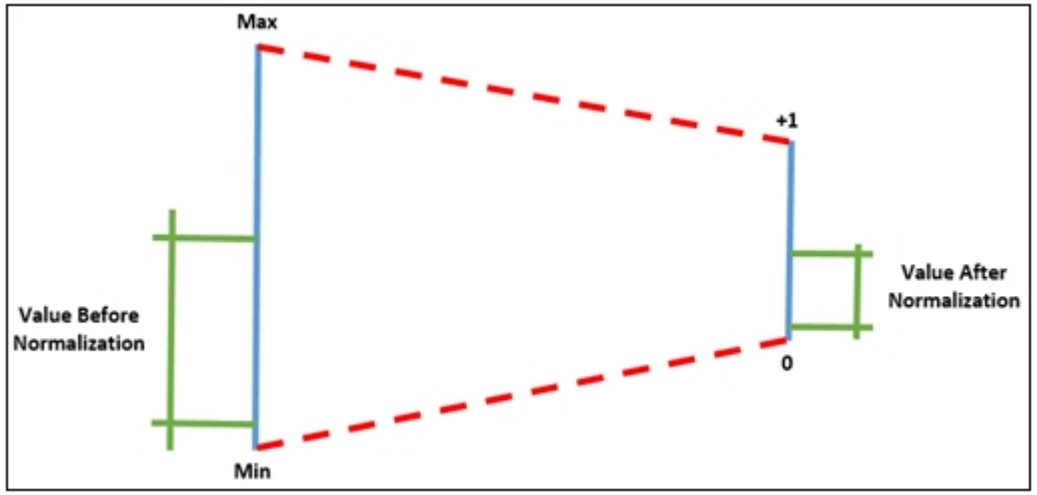
Los algoritmos de clusterización buscan cumplir con 3 requerimientos primordiales:

1. **Flexibilidad**: Se debe poder incluir atributos numéricos y categóricos.
2. **Robustez**: Estabilidad en los clusters ante cualquier ruido.
3. **Eficiencia**: Tiempos adecuados de procesamiento.



Transformación de Datos

Es normal que los analistas dediquen buena parte de su tiempo a comprender y jugar con los datos. Cuando la data es buena, la construcción de modelos será una tarea relativamente sencilla y también debemos tener en mente que cualquier mejora en los datos, traerá un impacto positivo sobre los modelos.



Transformación Lineal

El método más utilizado se conoce como el proceso de transformación lineal o recentralización (Recenter en Ingles), dicho proceso debe extraer primero la media y luego la desviación de cada variable.

Recenter – Z score

$$Z = \frac{X - \mu}{\sigma}$$

Transformación con mínimos y máximos

Este es otro método de transformación y su resultado nos devuelve un dataset cuyas variables numéricas están en el rango de 0 a 1.

Scale [0-1]

$$Z_i = \frac{X_i - min\,(x)}{max(x) - min\,(x)}$$

Transformación de Datos

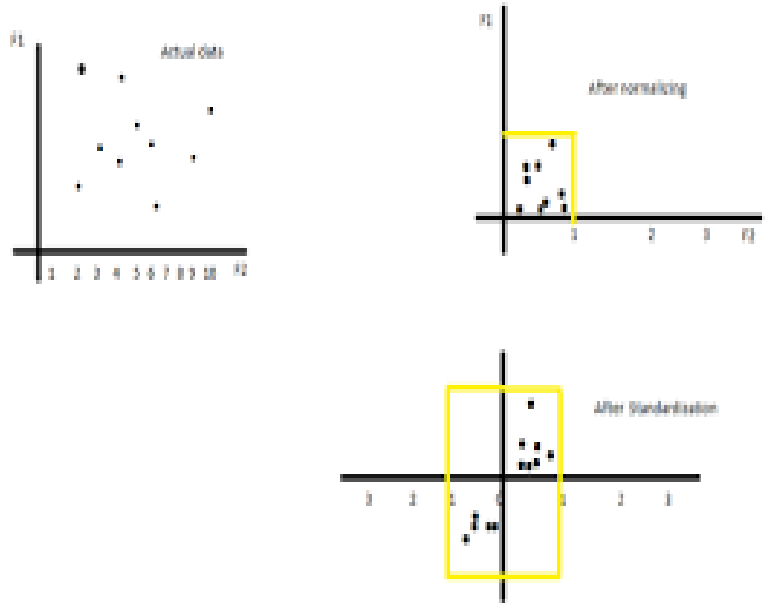
Transformación con la Mediana y MAD

Este método es más robusto que la transformación lineal (Z), se debe extraer la mediana de cada valor y luego hay que dividirlo entre la desviación absoluta media

Median / MAD

$$Z = \frac{X - Me}{MAD}$$

$$MAD = median(|x_i - \tilde{x}|)$$



Transformación Logarítmica

Este método es muy utilizado cuando estamos ante escenarios cuya distribución de datos presentan un sesgo elevado ya sea a la izquierda o a la derecha.

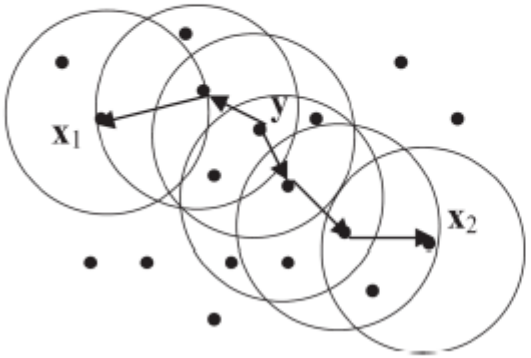
Esta transformación puede requerir más trabajo ya que los valores que produce pueden tender al infinito.

Clustering

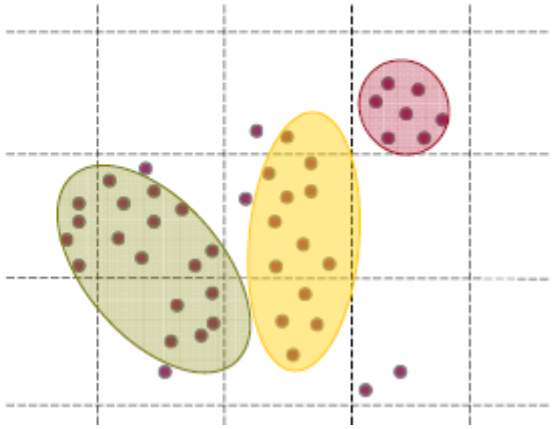
El clustering se basa en los conceptos de similaridad y distancia, en donde la proximidad entre los puntos es determinada por la función de la distancia.

Un aspecto importante a tener en cuenta es el hecho de que la persona que está haciendo el análisis debe indicar cuantos clusters se producirán y también deberá brindar soporte en la interpretación de los resultados.

Density-based methods
(DBSCAN)



Grid-based methods
(STING,CLIQUE)



A continuación se enumeran los diferentes métodos de clustering basados en su lógica de segmentación:

- Método por Partición: Los datos son divididos en un numero pre-calculado de clusters
- Método Jerárquico: Se construyen particiones basadas en una estructura de árbol.
- Métodos de Densidad: Construye los clusters tomando en cuenta la proximidad entre los diferentes puntos, es decir unifica entre vecinos.
- Método de Grilla (Grid): Construye las particiones basado en una estructura de grillas.

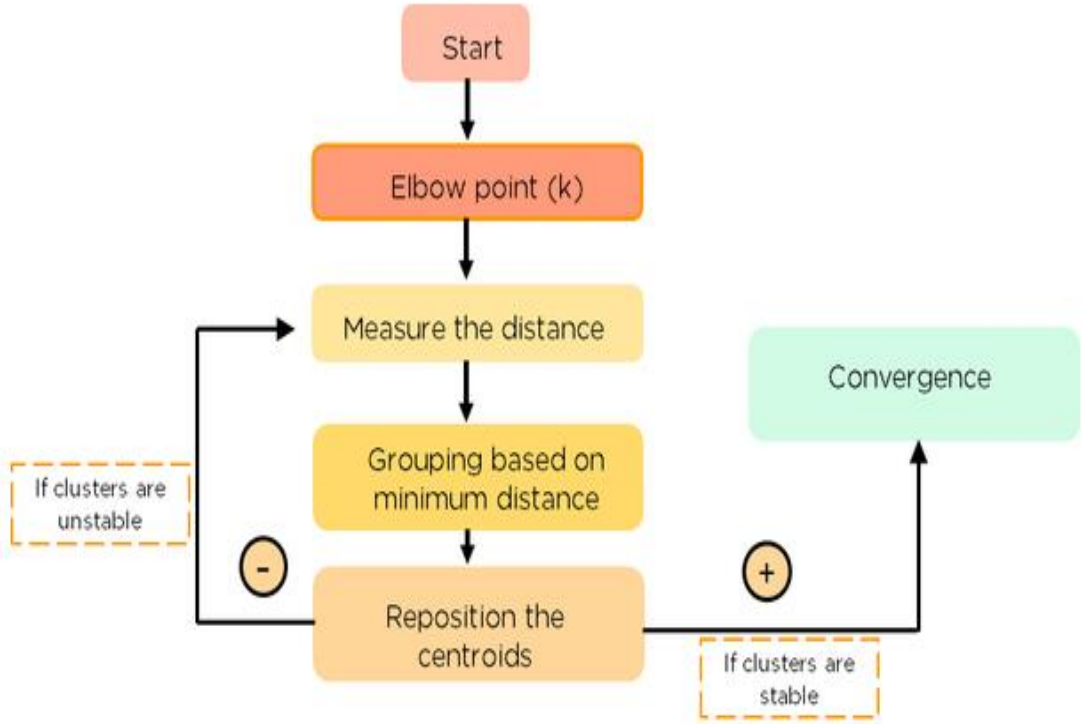
K-Means

El algoritmo de las K-means (presentado por MacQueen en 1967) es uno de los algoritmos de aprendizaje no supervisado más simples para resolver el problema de la clusterización.

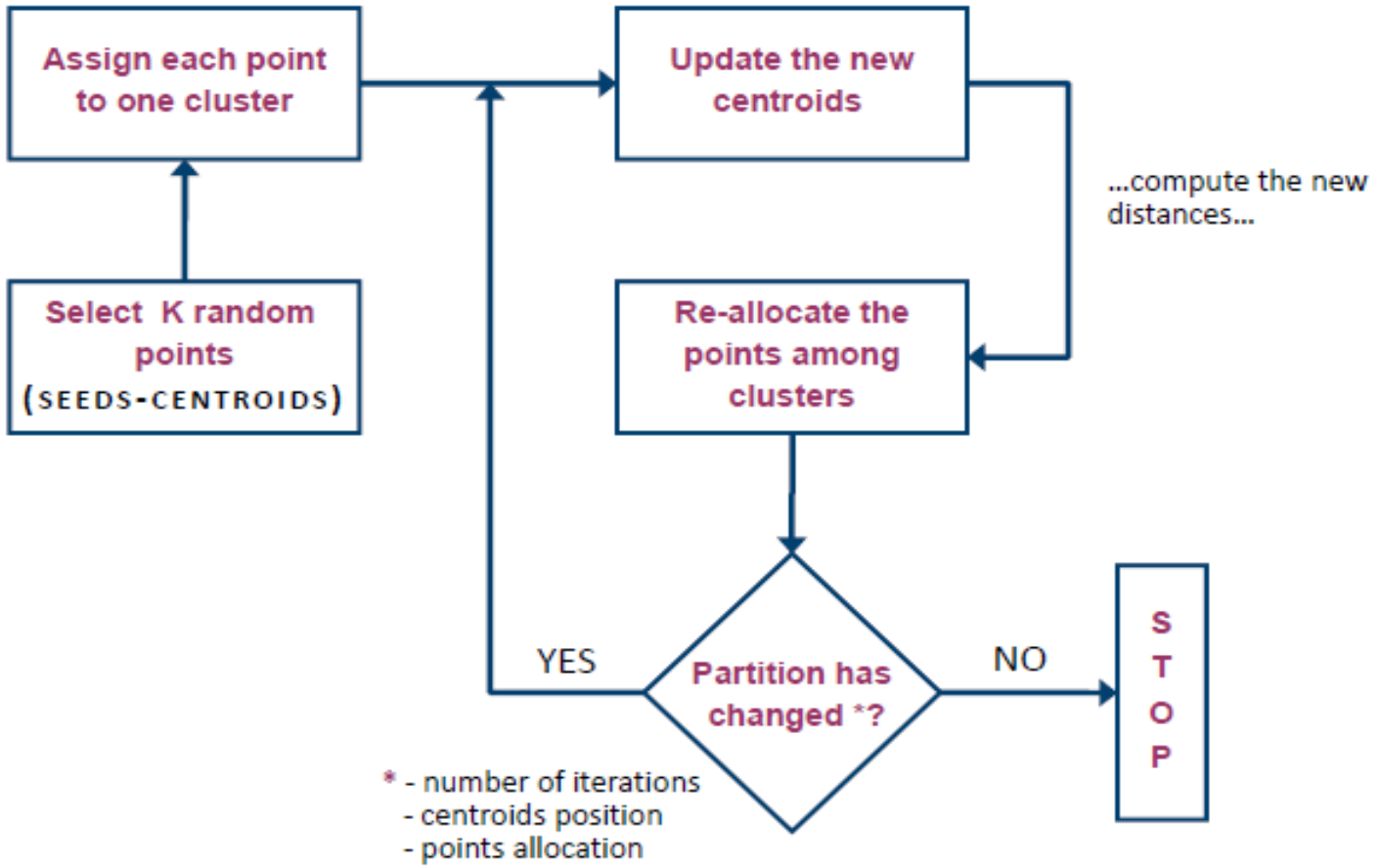
K-means es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.

El procedimiento aproxima por etapas sucesivas un cierto número (prefijado) de clusters haciendo uso de los centroides de los puntos que deben representar.

https://es.wikipedia.org/wiki/K-means

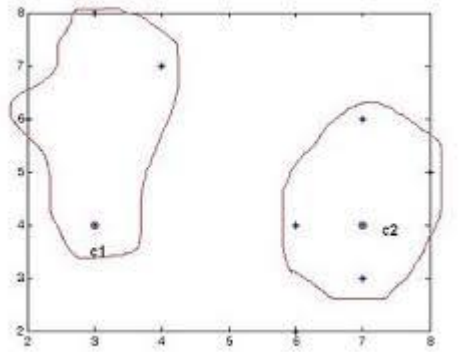


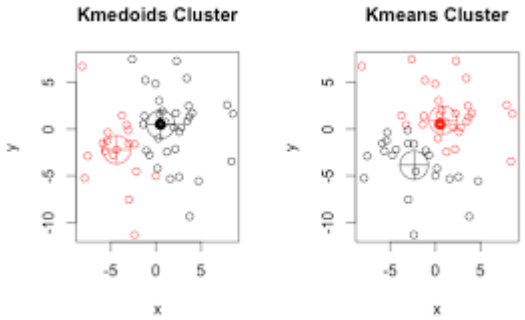
K-Means


Assign each point to one cluster
Update the new centroids
...compute the new distances...
Select K random points (SEEDS-CENTROIDS)
Re-allocate the points among clusters
YES
Partition has changed *?
NO
STOP
* - number of iterations
- centroids position
- points allocation

K-Means Variantes

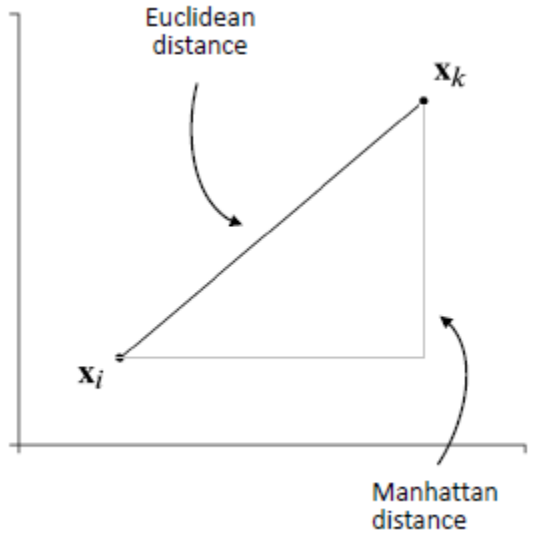
- K –medianas: Este algoritmo funciona de forma similar al k-means y es también sensible a la selección de centroides iniciales, continua sustituyendo el valor de promedios por el vector de medianas del grupo de datos y utiliza una distancia manhattan como una medida de disimilitud.
- K-medoids: Fue introducido por Kaufman y Rousseeuw en 1987. Este algoritmo está basado en un conjunto de datos localizados muy en el centro de cada clusters, los puntos restantes del grupo son agrupados con el medoids más cercano.
- Fuzzy c-means: Es un algoritmo que fue desarrollado para solucionar los datos que pueden pertenecer parcialmente a más de un clusters.





Tipos de Distancias

Distancia	Definición
Distancia Euclideana	Distancia proveniente de la raíz cuadrada entre 2 vectores.
Distancia Máxima	Distancia máxima entre 2 componentes de X y Y
Distancia Manhattan	Distancia absoluta entre 2 vectores
Distancia Canberra	Distancia Manhattan ponderada
Distancia Binaria	Los vectores son tratados como bits, si un elemento tiene valor se representa con un 1, de lo contrario son 0
Distancia Pearson	Distancia de tipo Euclidea, conocida como Pearson No Centrada sum(x_i y_i) / sqrt [sum(x_i^2) sum(y_i^2)]
Distancia por Correlación	Tambien conocida como Pearson Centrada. 1 - corr(x,y)
Distancia Spearman	Calcula la distancia basada en un ranking



Criterios de Validación

- **Cohesión:** El miembro de cada clúster debe ser lo más cercano posible a los otros miembros del mismo clúster, la afinidad entre las observaciones se puede medir mediante el coeficiente de Jaccard o bien el coeficiente de afinidad.

- **Separación:** Los clústeres deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster:
  - Distancia entre el miembro más cercano,
  - Distancia entre los miembros más distantes
  - Distancia entre los centroides.

Métricas de Validación

- **Sum of Squared Within (SSW)**: Medida interna especialmente usada para evaluar la Cohesión de los clústeres que el algoritmo de agrupamiento generó.

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon C_i} dist^2(m_i, x)$$

- **Sum of Squared Between (SSB)**: Es una medida de separación utilizada para evaluar la distancia interclúster (Separación)

$$SSB = \sum_{j=1}^{k} n_j \, dist^2 \, (c_j - \bar{x})$$

Otros Indices de Validación

- **Indice de Davies-Bouldin (DB):** Valores pequeños para el índice DB indica clústeres compactos, y cuyos centros estas bien separados los unos de los otros. Consecuentemente el número de clústeres que minimiza el índice DB se toma como el óptimo.

$$DB = \frac{1}{k} \sum_{i=1, i \neq j}^{k} \max \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

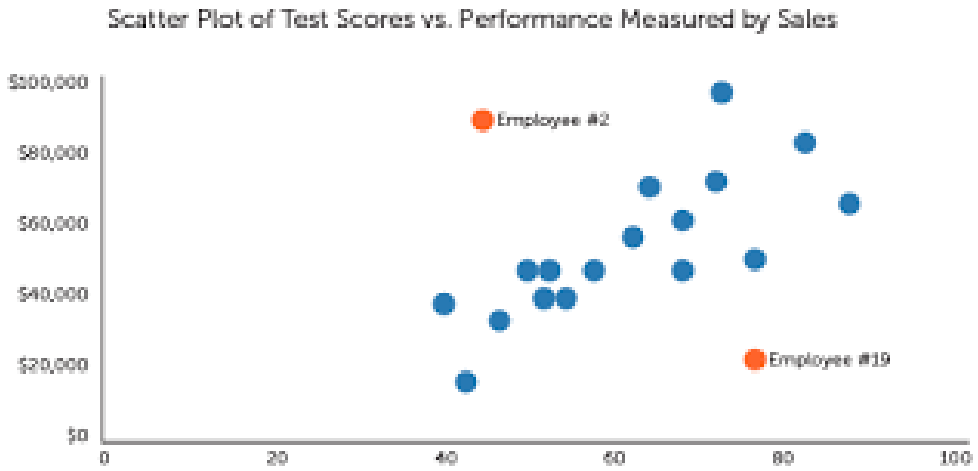
- **Coeficiente de Silhouette:** Contrasta la distancia promedio de elementos en el mismo cluster con la distancia promedio de elementos en otros clusters. Los elementos con alto valor se consideran bien agrupados, mientras que objetos con medidas bajas se consideran outliers.

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

Sugerencias de Pre y Post-Procesamiento

Pre-procesamiento:

- Estandarizar los datos
- Descartar los outliers



Post-procesamiento:

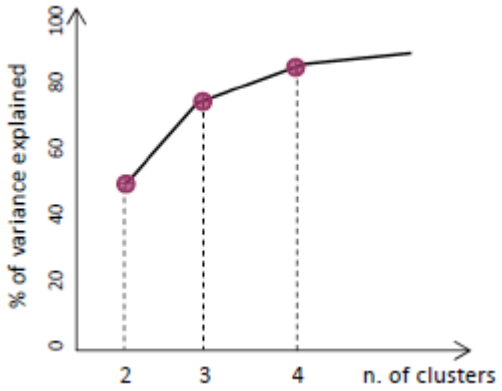
- Descartar los clusters pequeños que pueden representar outliers
- Re-clusterizar los clusters que son de gran tamaño
- Unificar los clusters que son cercanos

Determinando el número de Clusters

- Regla del Pulgar: Corresponde a la raíz cuadrada del total de observaciones dividido por 2.

$$K \approx \left(\frac{m}{2}\right)^{1/2}$$

- Método del Codo (Elbow Method): El porcentaje de varianza que se puede explicar esta en función del número de clusters.



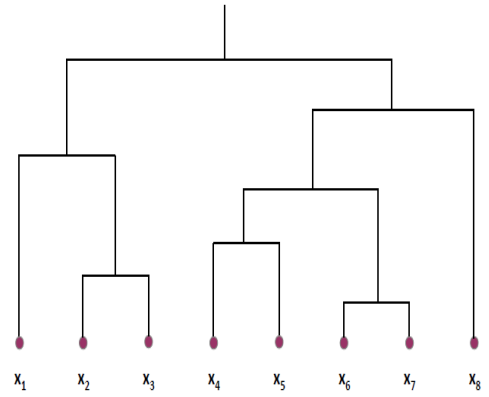
- Método de Silhouette: Silhouette indica que tan similar es una observación con respecto a las demás observaciones del mismo clusters con respecto a otros clusters.

$$\text{silh}\,(\mathbf{x}_i) = \frac{v_i - u_i}{\max\,(u_i, v_i)}$$

- within [-1,1]
- the closer to 1 the better
- average silhouette

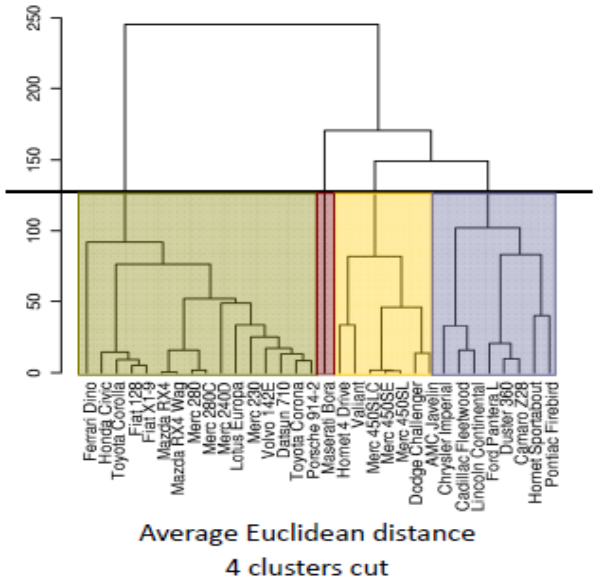
Métodos Jerárquicos

- Estan basados en una estructura de árbol (dendograma)
- Utiliza las distancias entre los diferentes puntos para unificar o particionar
- No se necesita el número de clusters como entrada del proceso



Los algoritmos a utilizar pueden ser de tipo aglomerativos o divisivos:

- **Aglomerativos**: Inicialmente cada observación representa un cluster, por cada iteración se van agrupando los 2 cluster más cercanos y el algoritmo se detiene cuando cada observación esta categorizada dentro de un cluster.
- **Divisivos**: Inicialmente todas las observaciones están dentro de un solo cluster, por cada iteración se generan 2 clusters en función de su distancia máxima, el algoritmo se detiene cuando cada observación representa un cluster único.



Bibliografía

Unsupervised Learning with R

By Erik Rodríguez Pacheco, 2015

Data Clustering

By Charu C. Aggarwal; Chandan K. Reddy, 2016

A Framework for Analysis of Data Quality Research

by Richard Y. Wang,1995

An Introduction to Data Cleaning with R

by Edwin de Jonge & Mark van der Loo,2013