KIBERNUM
IT Academy

Módulo 6 – Aprendizaje de Máquina No Supervisado

Clustering Jerárquico

Especialización en Ciencia de Datos

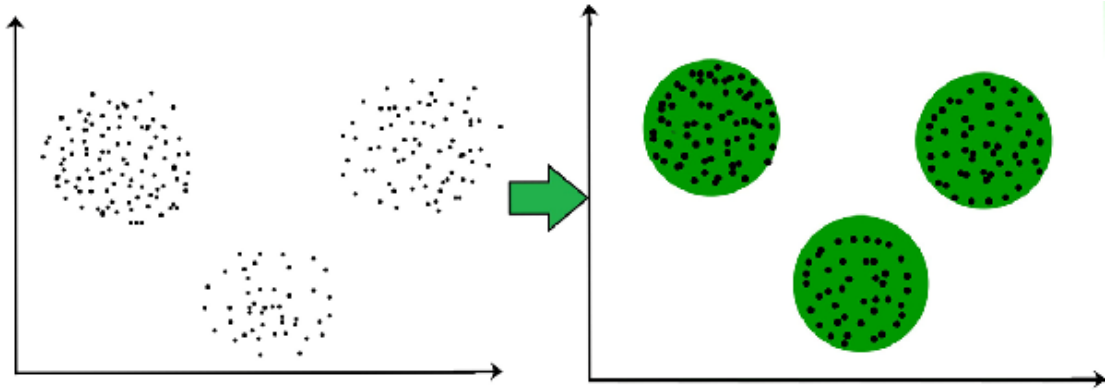
Contenido

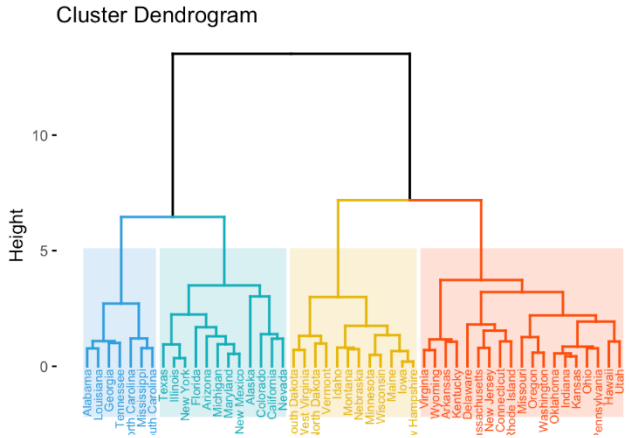
1. Utilizar los conceptos básicos de aprendizaje de máquinas no supervisado.
2. Conocer los distintos tipos de algoritmos para agrupamiento jerárquico.
3. Conocer sobre dendogramas.
4. Implementación en Python.
5. Ventajas y desventajas de la clusterización jerárquica.

KIBERNUM

Clustering Jerárquico

El clustering jerárquico es un método de clustering o agrupamiento en el que se crean jerarquías de clústeres. En este enfoque, los objetos o datos se agrupan en un árbol jerárquico, que se puede visualizar mediante un dendrograma. Es un método para agrupar datos en donde el proceso de agrupación se basa en la distancia entre los elementos.





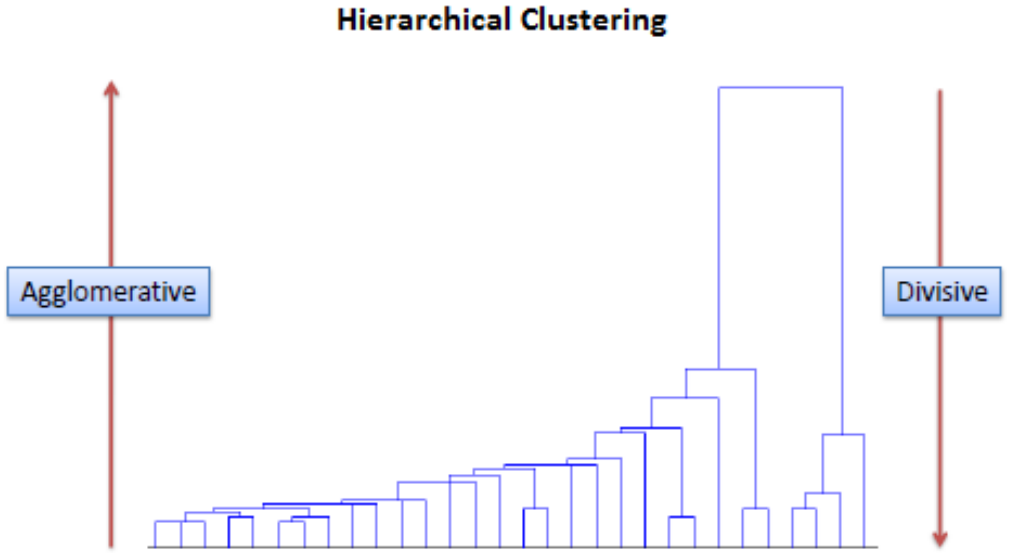
KIBERNUM

Clustering Jerárquico

Estrategias para agrupamiento jerárquico generalmente caen en dos tipos:

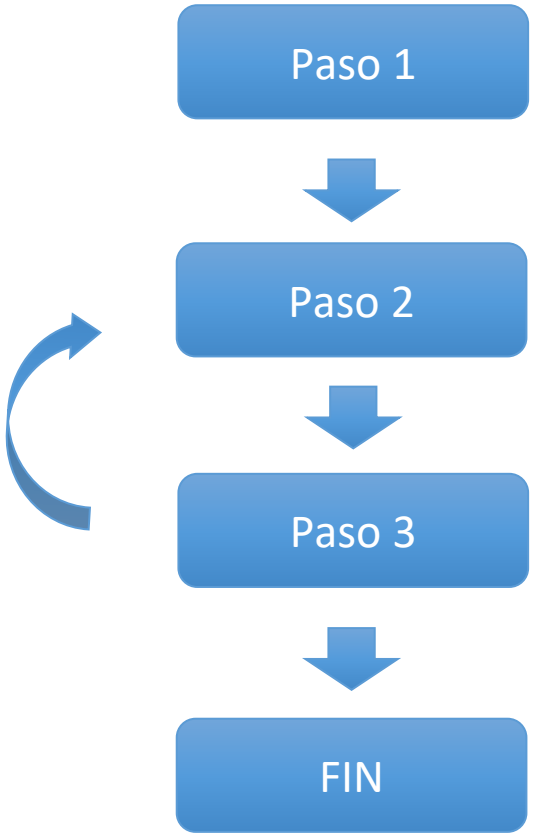
Aglomerativo: empezamos a agrupar desde cada elemento individual. Al inicio cada punto o dato está en un clúster separado. A cada paso, los dos clústers más cercanos se fusionan. Estas fusiones de clústers se siguen produciendo de forma sucesiva produciendo una jerarquía de resultados de clustering. Al final del proceso solo queda un único clúster que aglutina todos los elementos.

Divisivo: comenzamos a la inversa, partimos de un único clúster que aglomera todos los datos y vamos dividiendo en clústers más pequeños.



KIBERNUM

Algoritmo de Clustering Jerárquico Aglomerativo



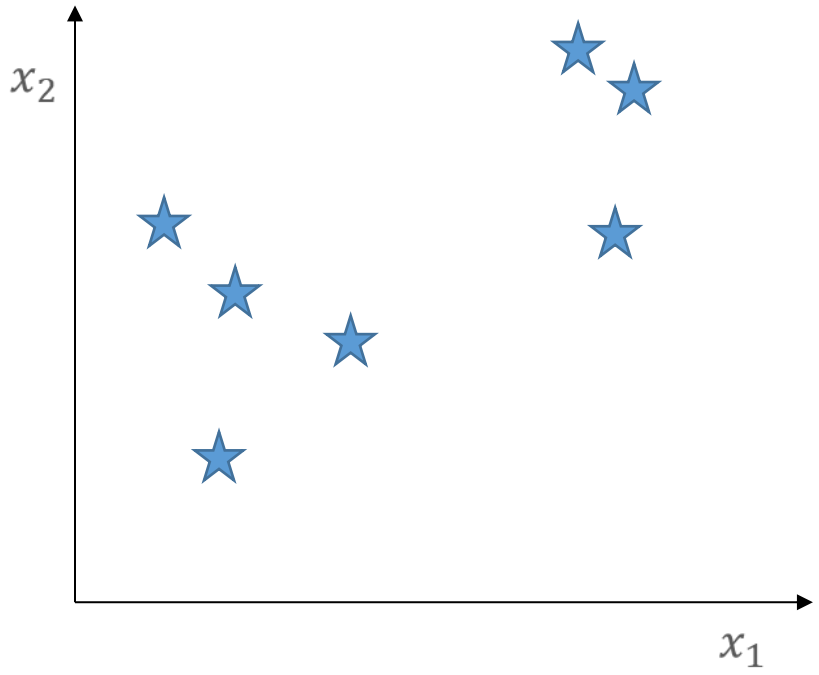
Hacer que cada punto sea un cluster (se forman N clusters).

Tomar los dos clusters más cercanos y formar un solo cluster (se forman N -1 cluster).

Repetir hasta que haya un solo cluster.

Algoritmo en Acción

Inicialmente, cada punto es un cluster.



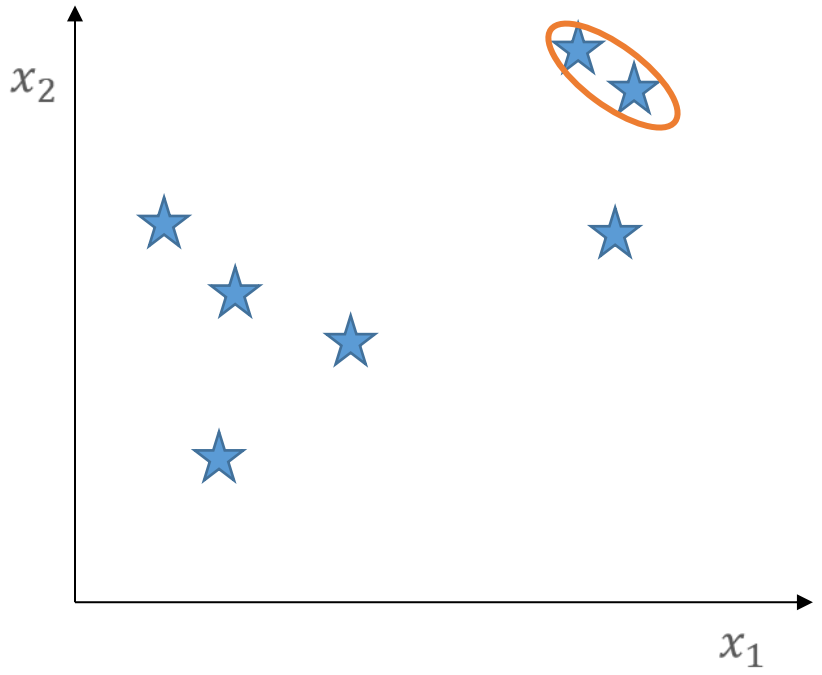
Cantidad de clusters

7

KIBERNUM

Algoritmo en Acción

Elegimos los puntos más cercanos y formamos un cluster entre ellos.



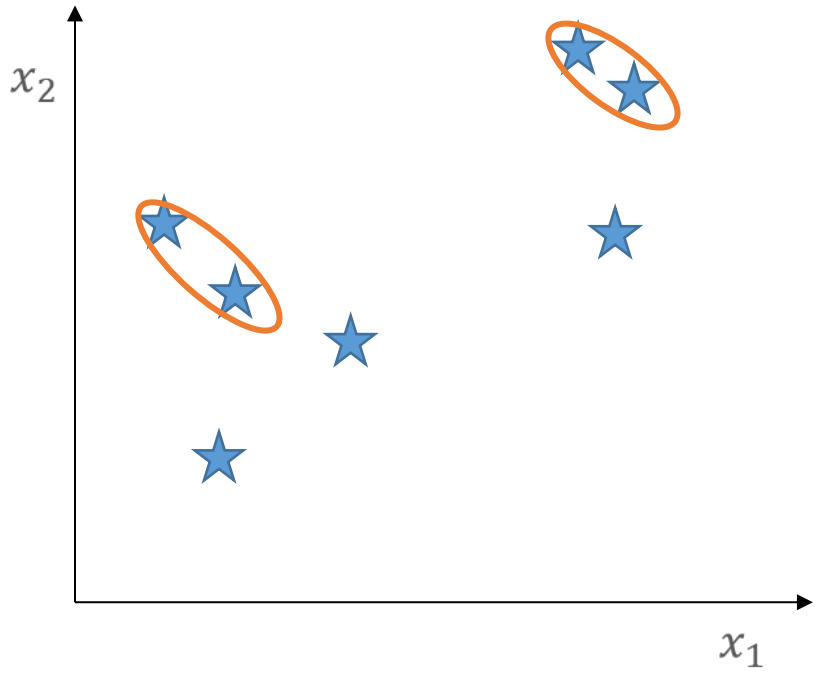
Cantidad de clusters

6

KIBERNUM

Algoritmo en Acción

Elegimos los puntos más cercanos y formamos un cluster entre ellos.



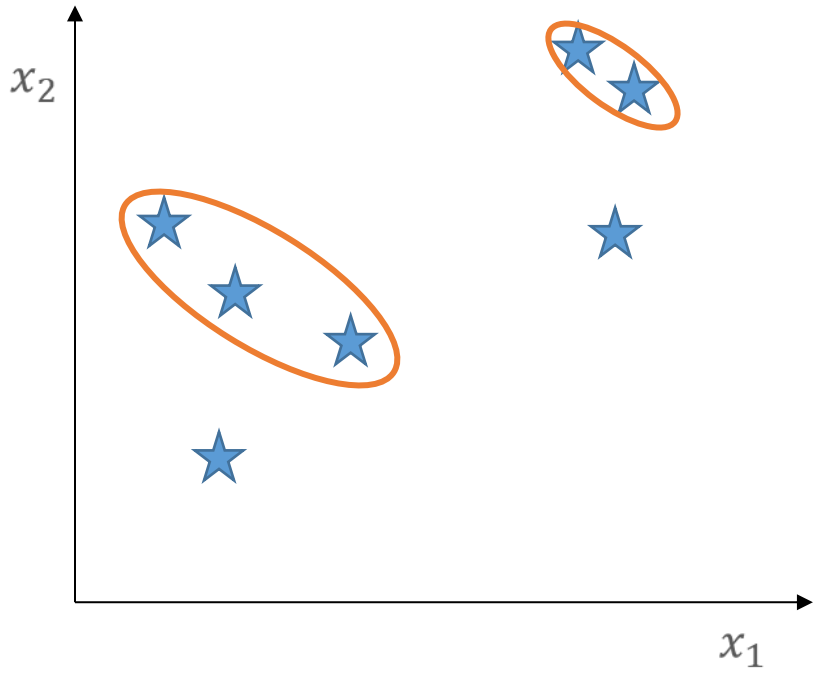
Cantidad de clusters

5

KIBERNUM

Algoritmo en Acción

Elegimos los puntos más cercanos y formamos un cluster entre ellos.



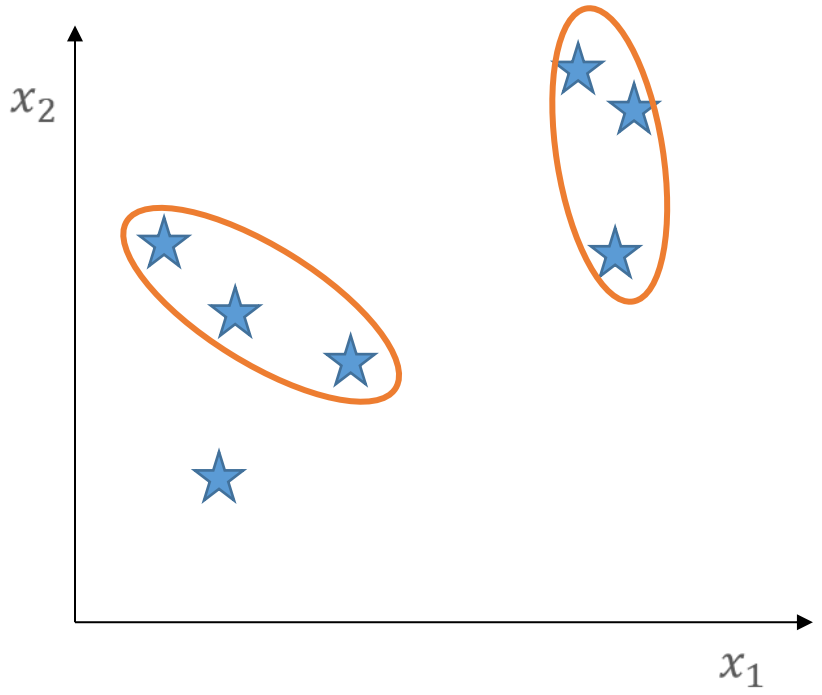
Cantidad de clusters

4

KIBERNUM

Algoritmo en Acción

Elegimos los puntos más cercanos y formamos un cluster entre ellos.



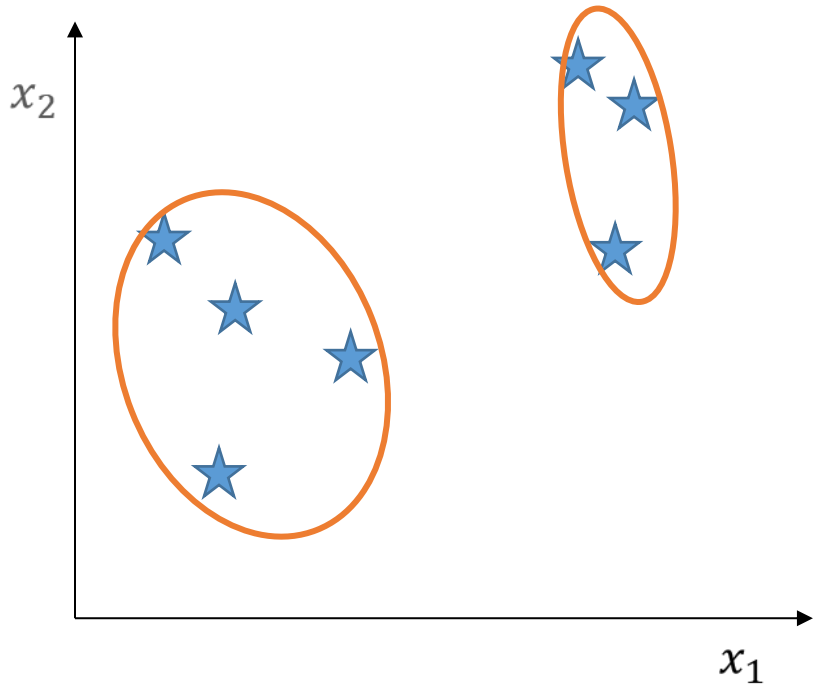
Cantidad de clusters

3

KIBERNUM

Algoritmo en Acción

Elegimos los puntos más cercanos y formamos un cluster entre ellos.



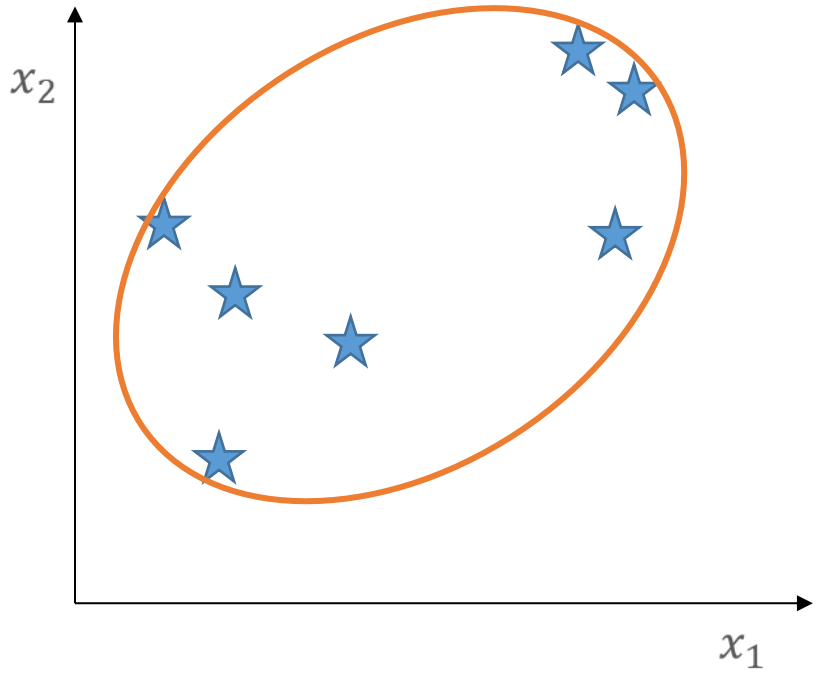
Cantidad de clusters

2

KIBERNUM

Algoritmo en Acción

Finaliza el proceso.



Cantidad de clusters

1

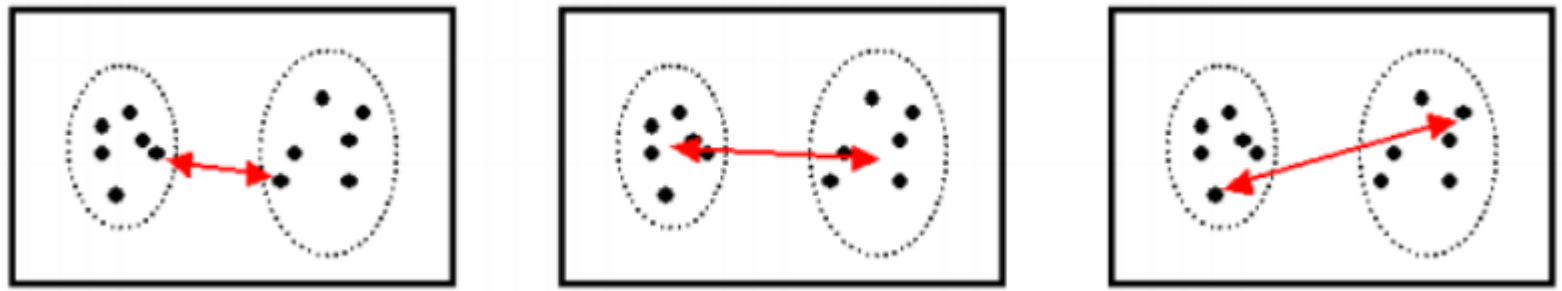
KIBERNUM

Distancia entre Clusters

- ¿Cómo medir la distancia entre dos clusters? Hay varios criterios:

- Considerar los puntos más cercanos
- Considerar los puntos más lejanos
- Considerar la distancia promedio
- Considerar la distancia entre sus centroides

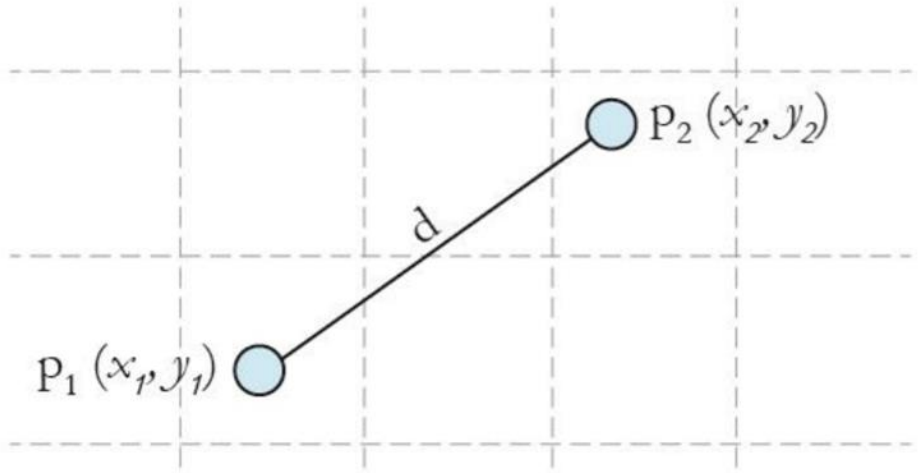
- Utilizar una u otra alternativa puede influir en el resultado de la clusterización.



KIBERNUM

Medida de la Distancia

El algoritmo utiliza la distancia Euclidiana para encontrar los clusters más cercanos, pero se podría elegir otra métrica de distancia dependiendo de las características del problema. Otra métrica podría ser, por ejemplo, la Distancia Manhattan, Minkowski, entre otras.



$$\text{Euclidean distance (d)} = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

KIBERNUM

Escalamiento de los Datos

El escalamiento de datos puede ser necesario en el clustering jerárquico dependiendo del algoritmo utilizado y la naturaleza de los datos. En general, es una buena práctica escalar los datos antes de aplicar cualquier método de clustering, incluyendo el clustering jerárquico.

La razón principal para escalar los datos es que algunos algoritmos de clustering, como el clustering basado en distancia, son sensibles a las diferencias de escala en las variables. Si no se realiza el escalamiento de datos, las variables con una mayor varianza pueden dominar el proceso de clustering y producir clústeres sesgados. Además, el uso de variables con diferentes unidades de medida puede generar distancias no comparables y, por lo tanto, afectar la calidad de los resultados de clustering.

MinMaxScaler

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

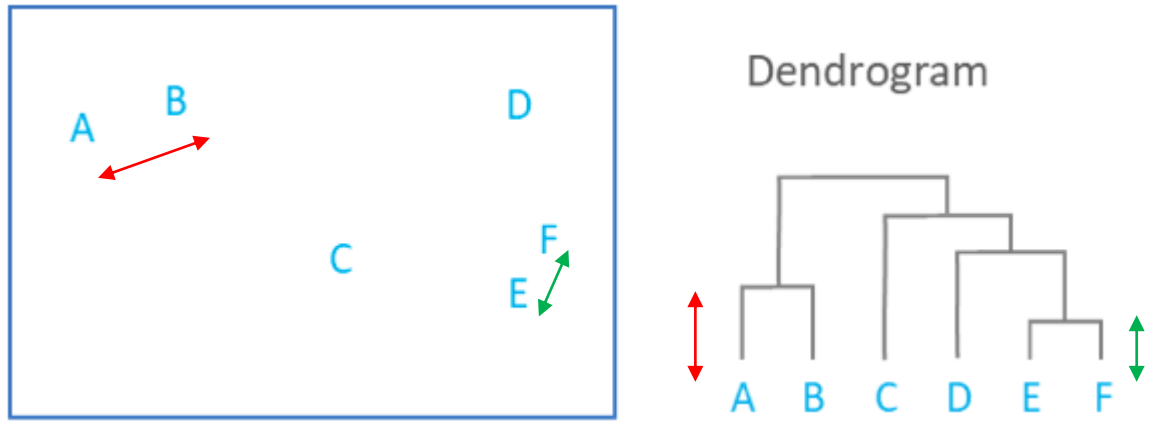
StandardScaler

$$x_{new} = \frac{x - \mu}{\sigma}$$

KIBERNUM

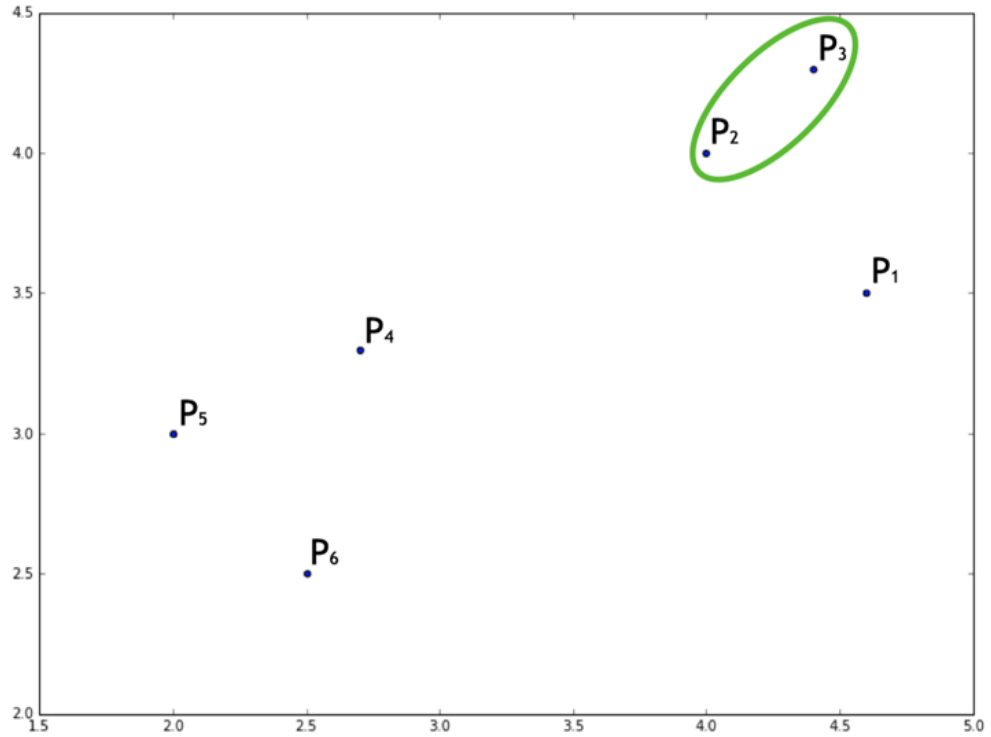
Representación: El Dendrograma

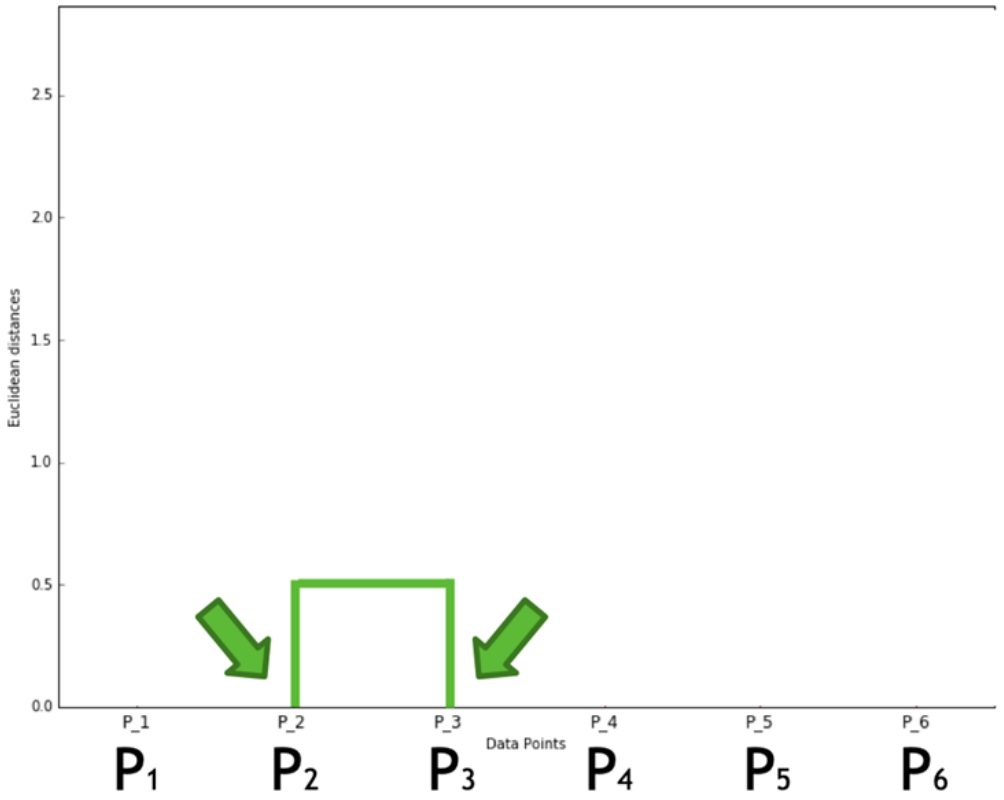
Un dendrograma, es un tipo de representación gráfica o diagrama de datos en forma de árbol que organiza los datos en subcategorías que se van dividiendo en otros hasta llegar al nivel de detalle deseado. El dendrograma representa la distancia entre los elementos clusterizados.



KIBERNUM

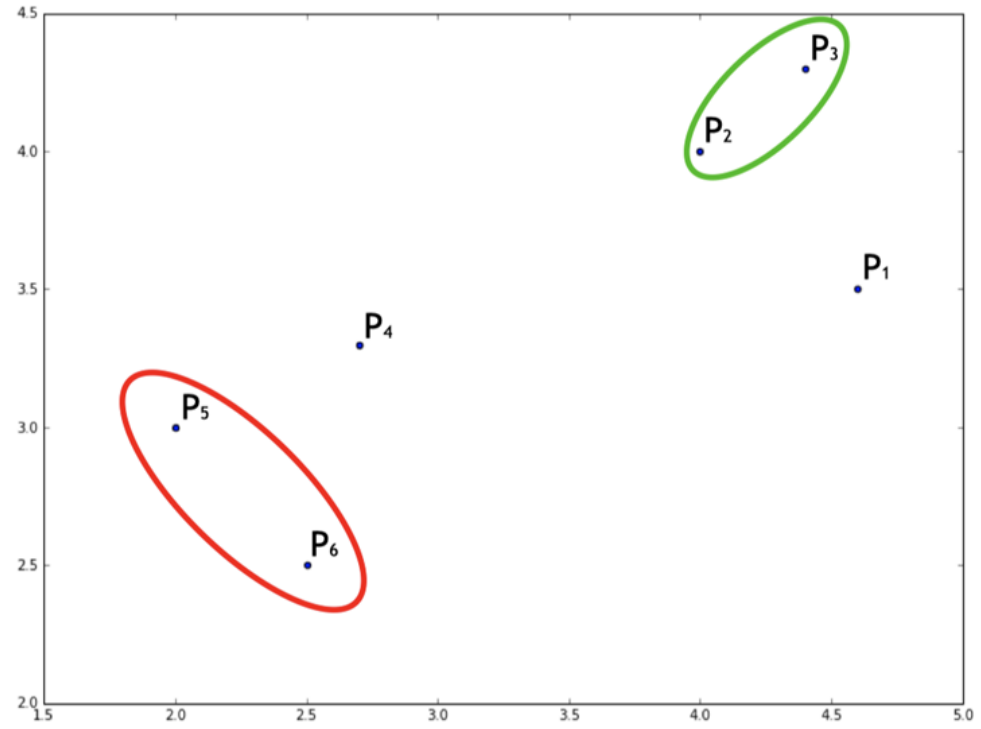
Dendogramas

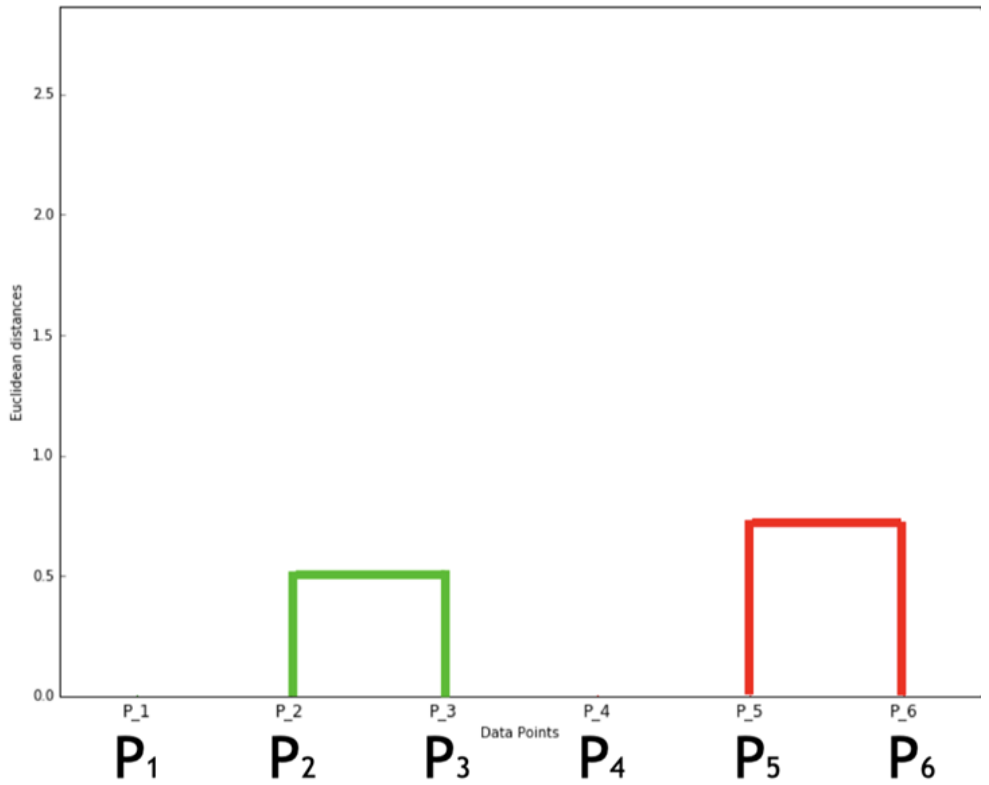




KIBERNUM

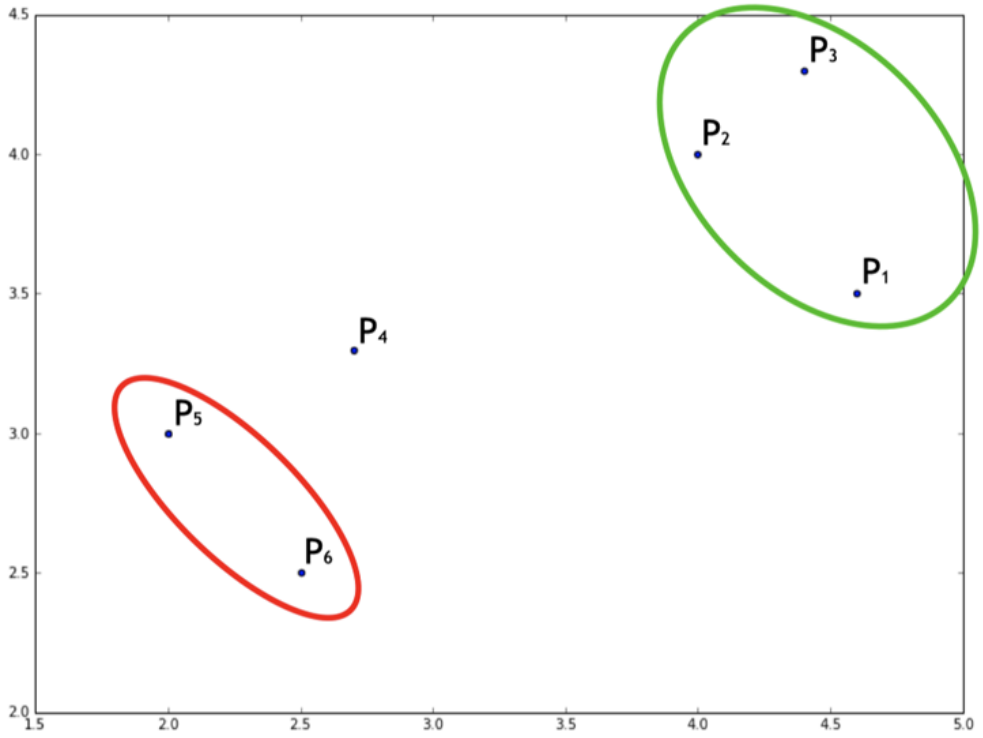
Dendogramas

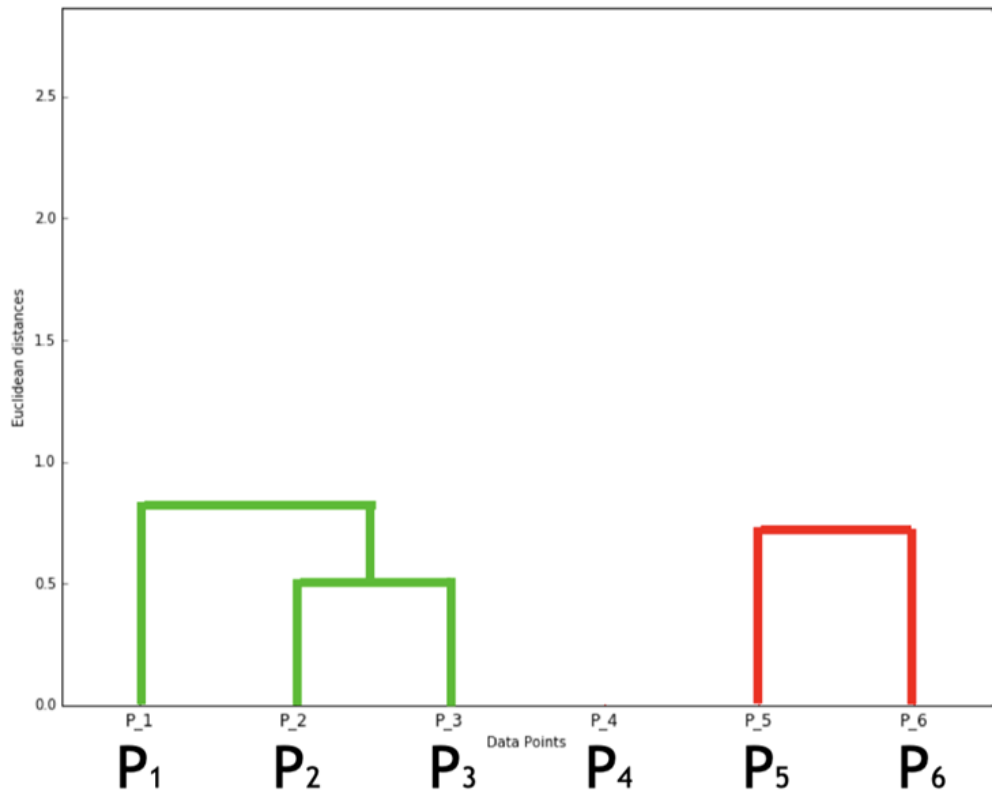




KIBERNUM

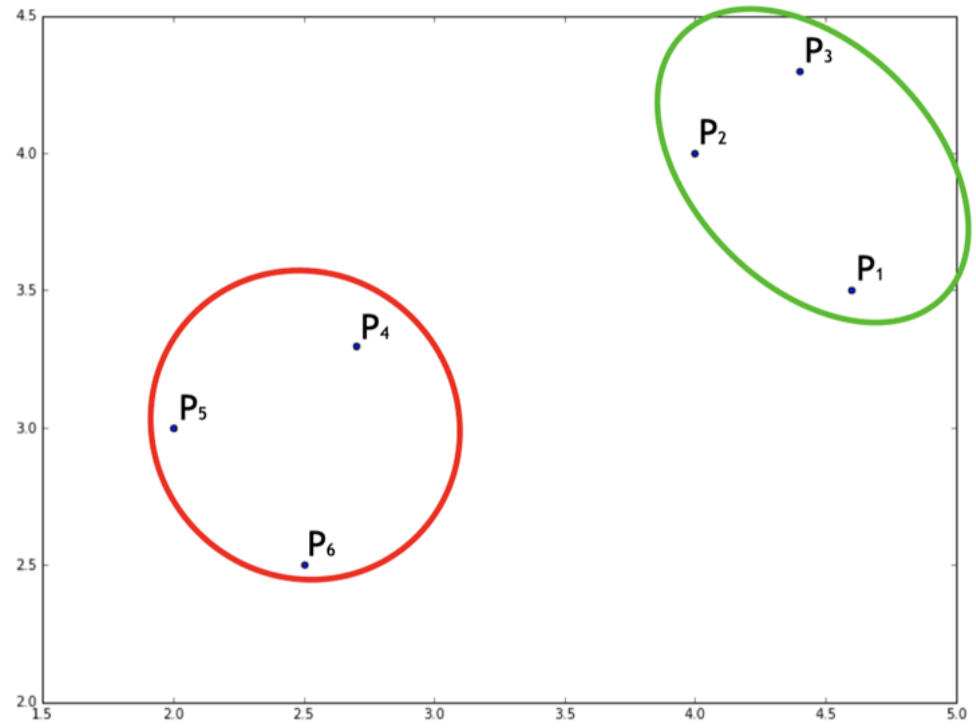
Dendogramas

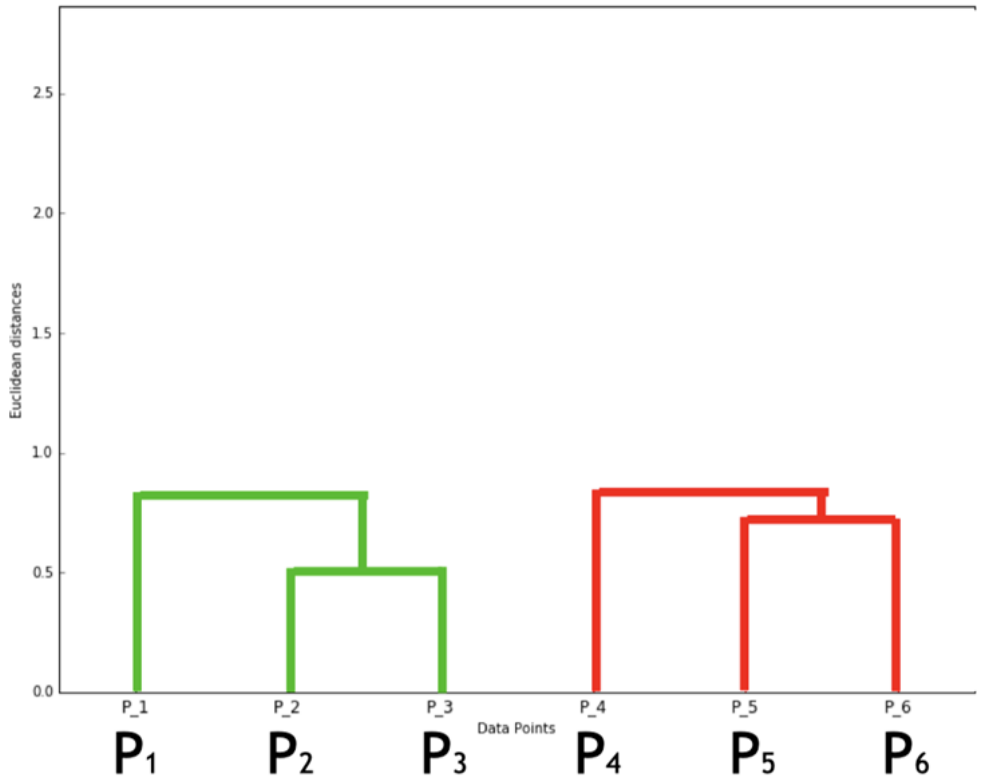




KIBERNUM

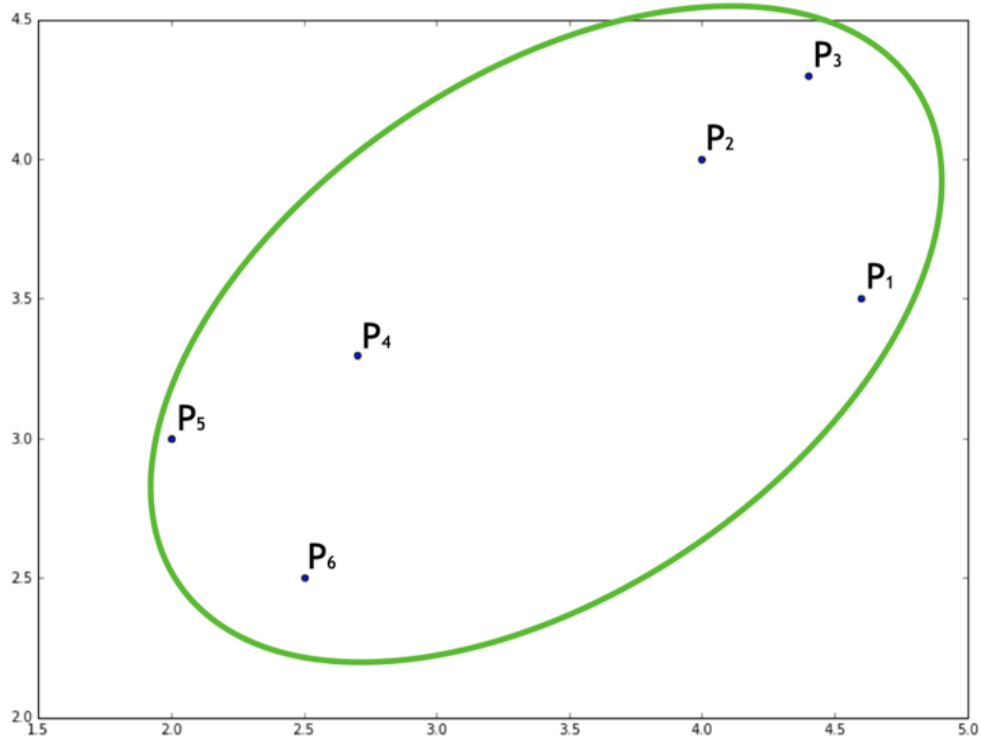
Dendogramas

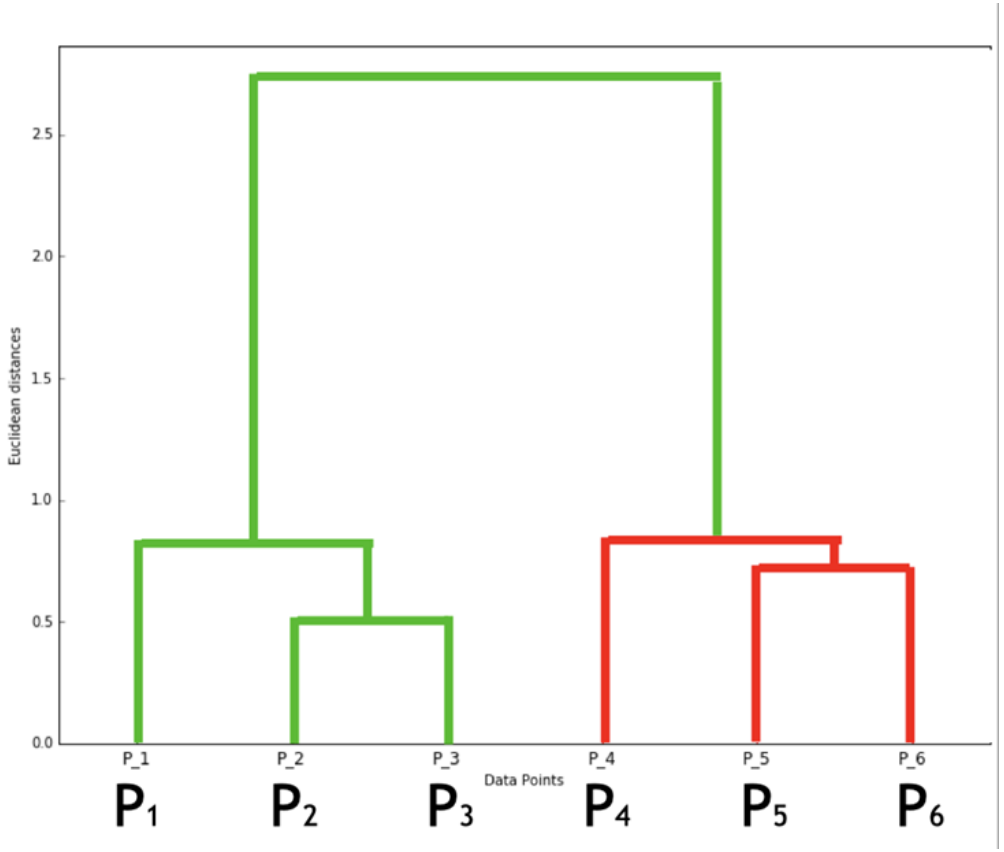




KIBERNUM

Dendogramas

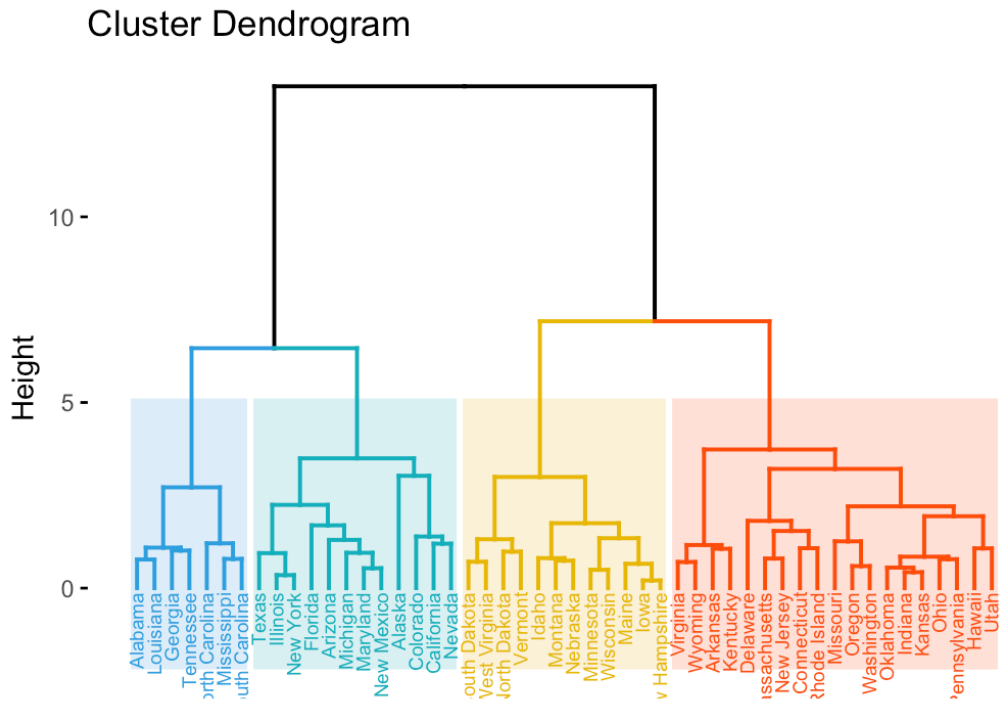




KIBERNUM

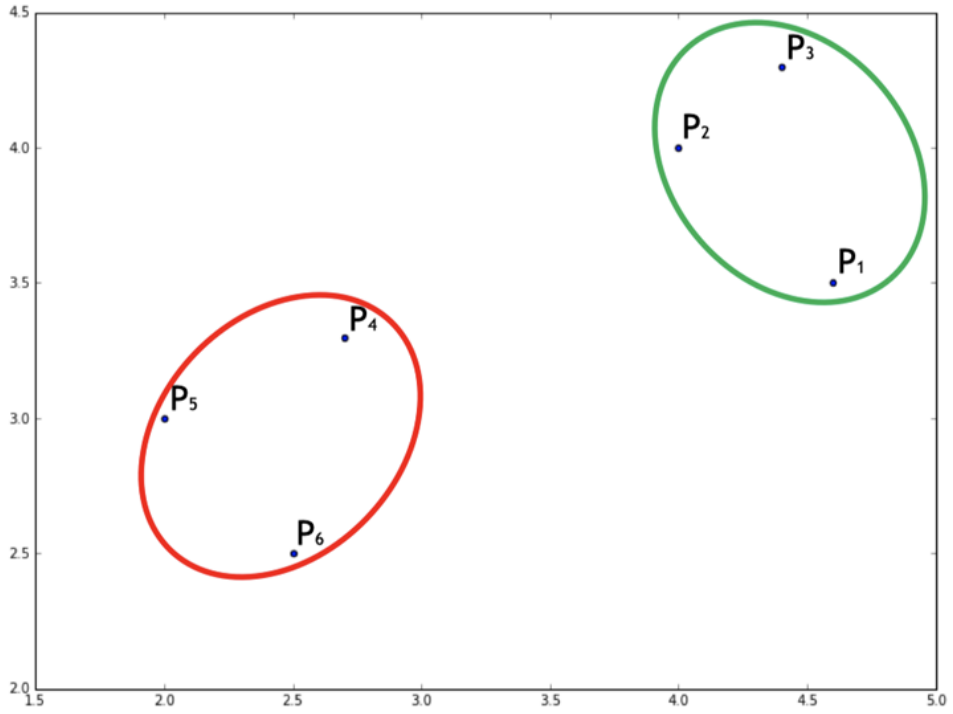
Elección de Clusters

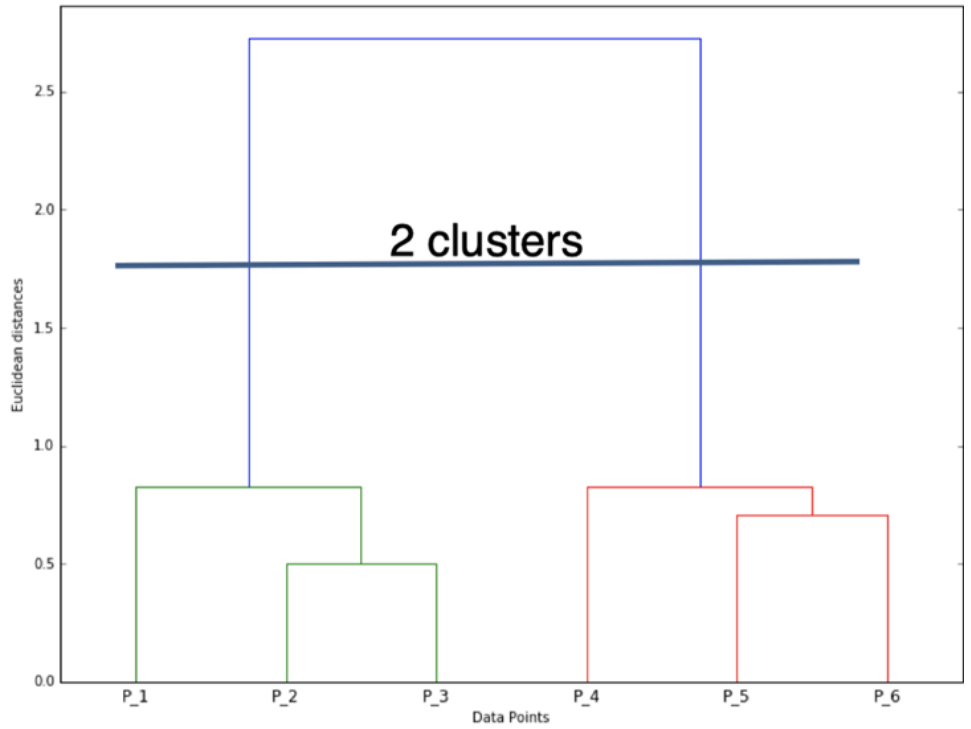
- A pesar que existen técnicas para sugerir una cantidad de clusters, la forma más efectiva es la observación y la interpretación.
- Definimos un **umbral** de distancia y computamos los clusters formados. En este ejemplo, se definió un umbral de valor 5 que identificó 4 clusters, y un umbral de 10 que identificó 2 clusters.



KIBERNUM

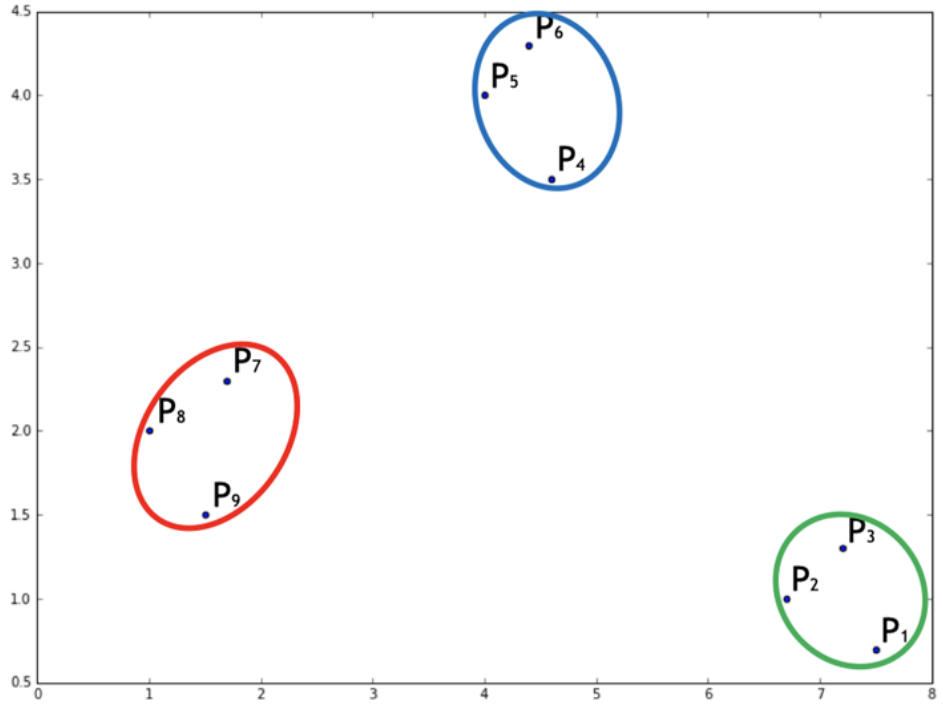
Elección de Clusters

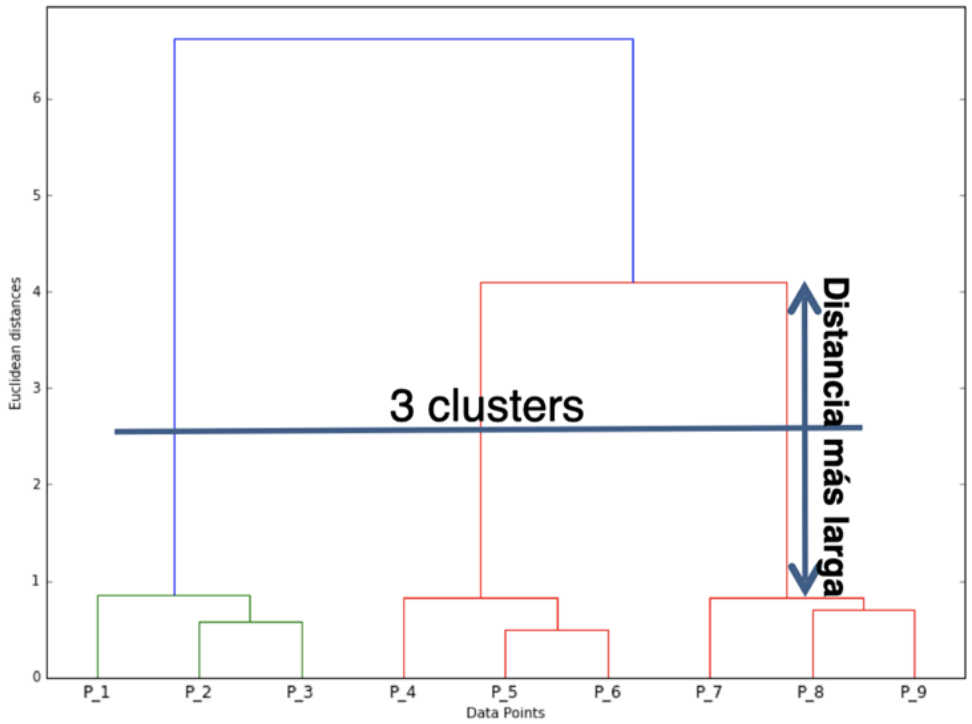




KIBERNUM

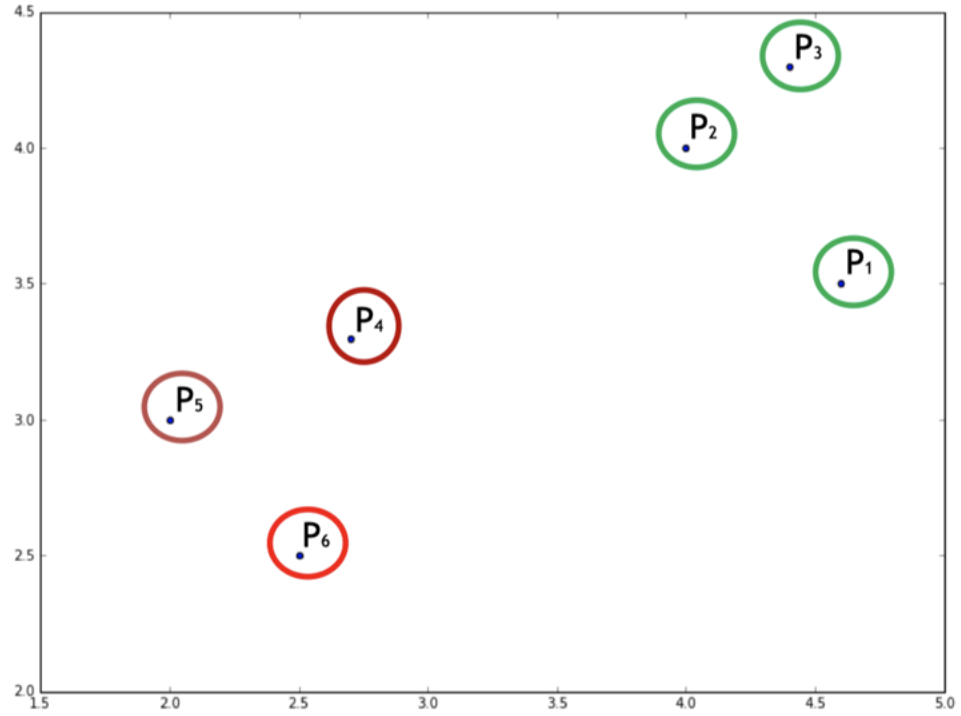
Elección de Clusters

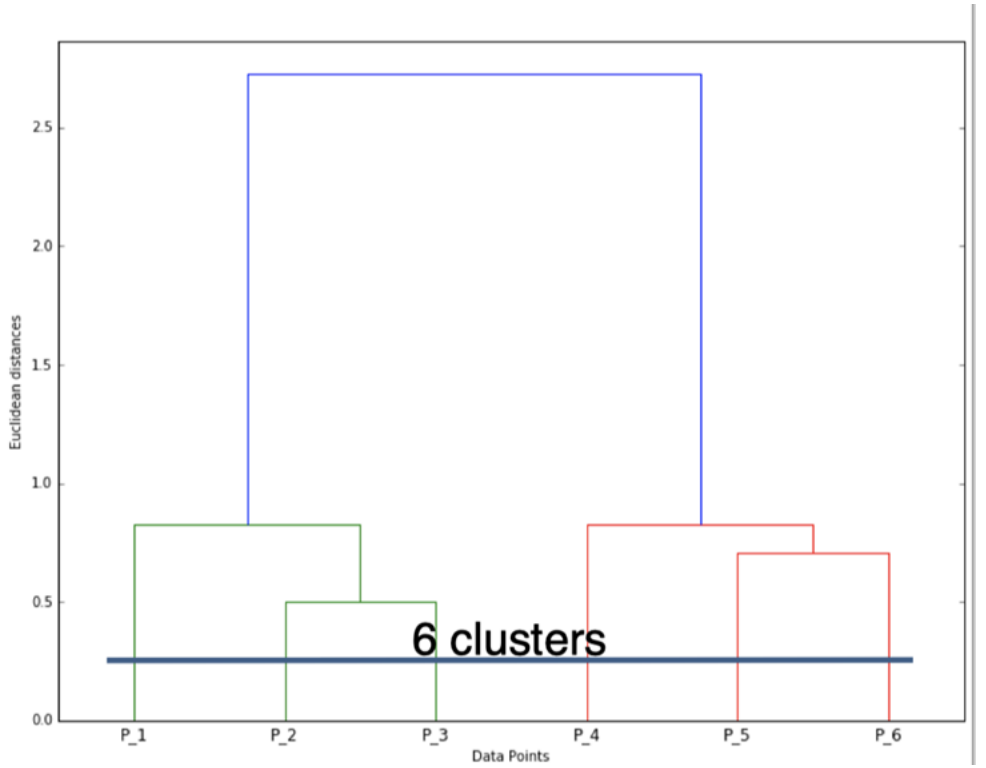




KIBERNUM

Elección de Clusters





KIBERNUM

Comentarios sobre Clustering Jerárquico

Es importante considerar lo siguiente:

- Este método no es apropiado para datasets muy grandes.
- Es importante recordar que debemos estandarizar las variables antes de aplicar el algoritmo.
- La elección de la medida de distancia podría generar distintos agrupamientos.

Clustering Jerárquico

La clusterización jerárquica es un método de análisis de datos que agrupa objetos en función de su similitud o proximidad. Esta técnica tiene ventajas y desventajas que se describen a continuación:

Ventajas

1. No se requiere especificar el número de clústeres de antemano.
2. Se puede visualizar la estructura jerárquica de los clústeres.
3. Puede ser útil para identificar patrones en los datos, como grupos de objetos que son similares entre sí.

Desventajas

1. La clusterización jerárquica puede ser computacionalmente costosa cuando se trata de grandes conjuntos de datos.
2. El método jerárquico puede no ser el más adecuado para todos los conjuntos de datos. En algunos casos, la clusterización basada en particiones puede ser más efectiva.
3. La clusterización jerárquica puede ser sensible a los valores atípicos y a los errores de medición en los datos.
4. No permite reajustar el número de clústeres una vez que se ha creado la jerarquía.

Dudas y consultas

Fin de la Presentación

KIBERNUM