Análisis de Agrupamientos I: Técnicas Heurísticas

Mario Hernández

Introducción

- Dpto de Ventas de cierta empresa con datos sobre las compras de sus clientes.
- Empresa Aseguradora con datos personales e índices de siniestralidad de sus asegurados

Clasificar en grupos según:

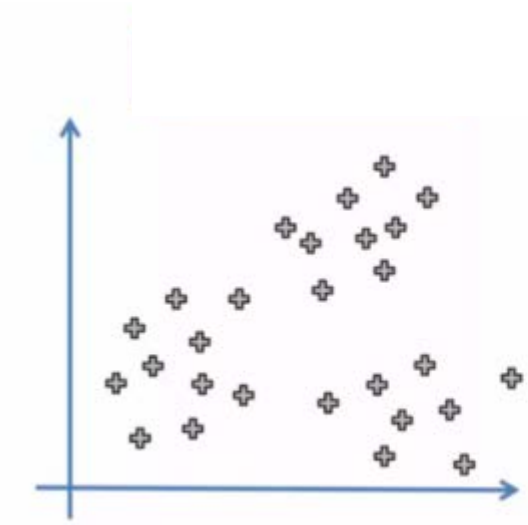
- Hábitos de compra $\rightarrow$ Diseñar campañas de marketing mejor adaptadas a los clientes
- Grupos de riesgo $\rightarrow$ Adaptar las tarifas a los riesgos potenciales de los asegurados

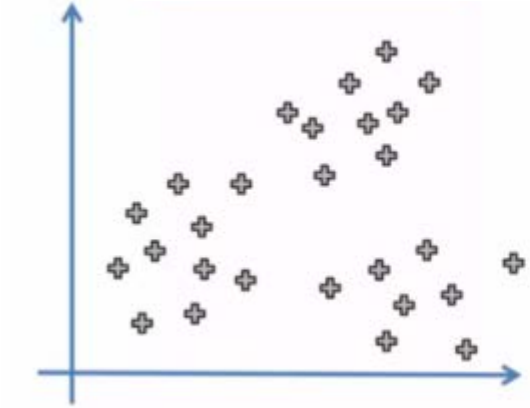
Introducción

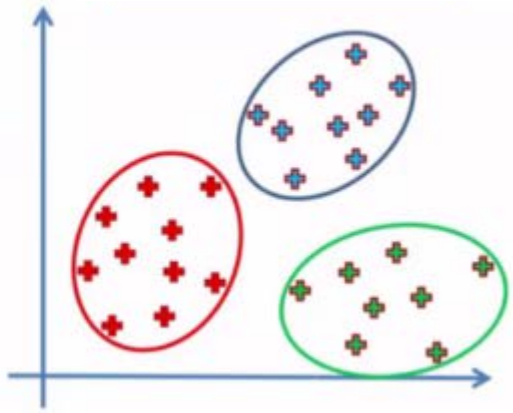
- Aprendizaje No Supervisado ≡Análisis de Agrupamientos
- Métodos de aprendizaje cuyo objetivo es obtener una descripción de los objetos en términos de grupos o clusters
- Los objetos en un grupo serán más similares o estarán más relacionados entre si que a los objetos pertenecientes a otros grupos.

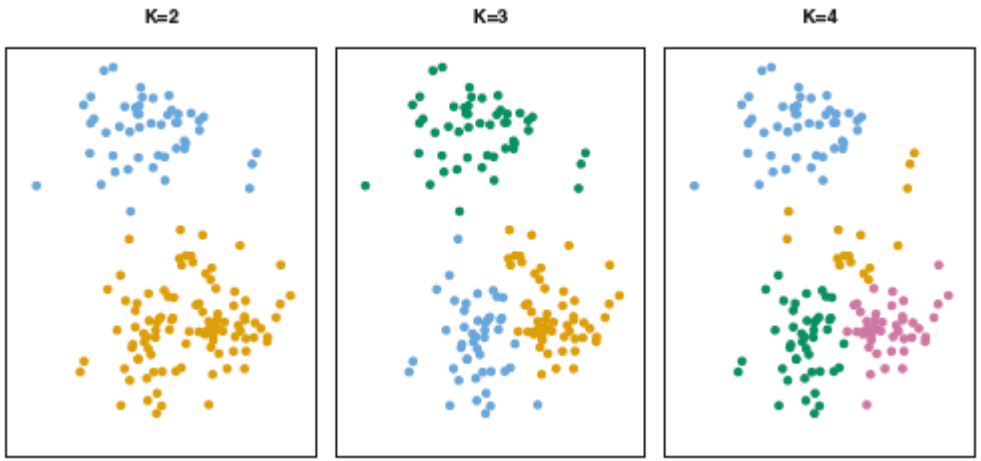
  a) Medición de la similitud o relación

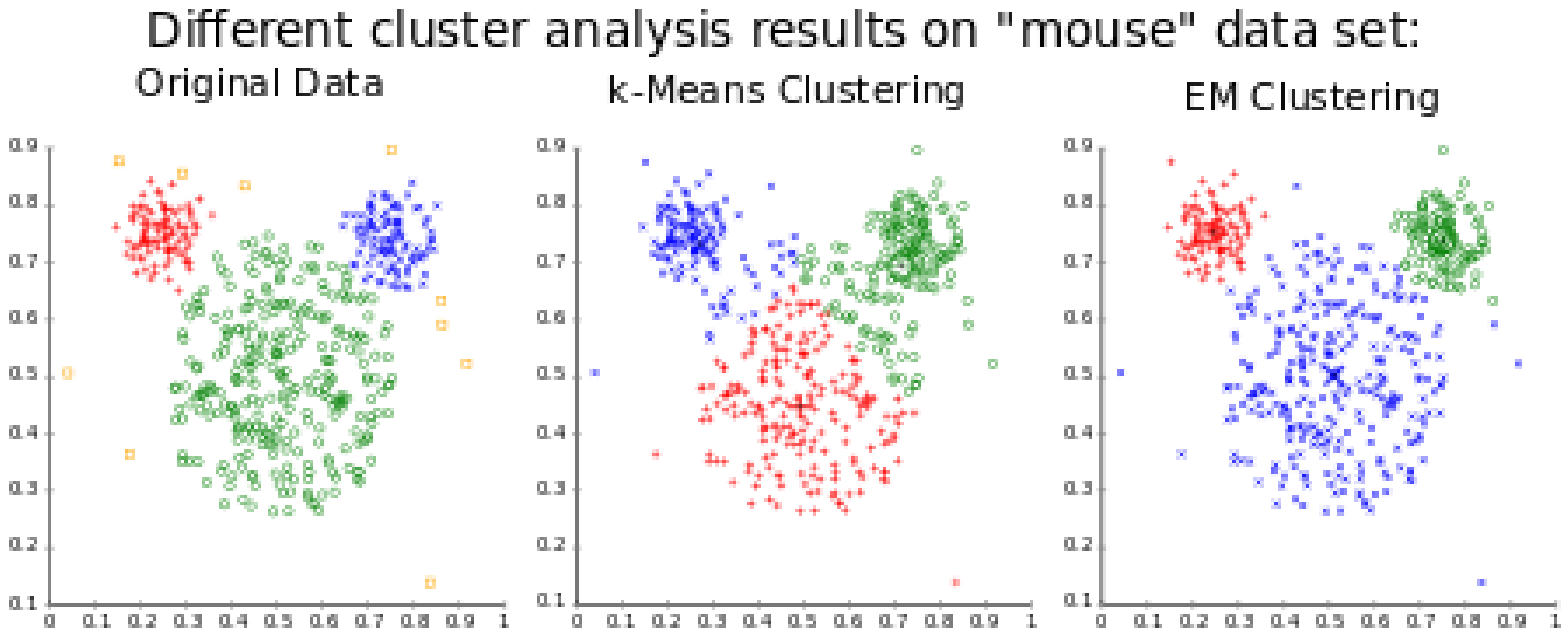
  b) Partición de los datos en grupos

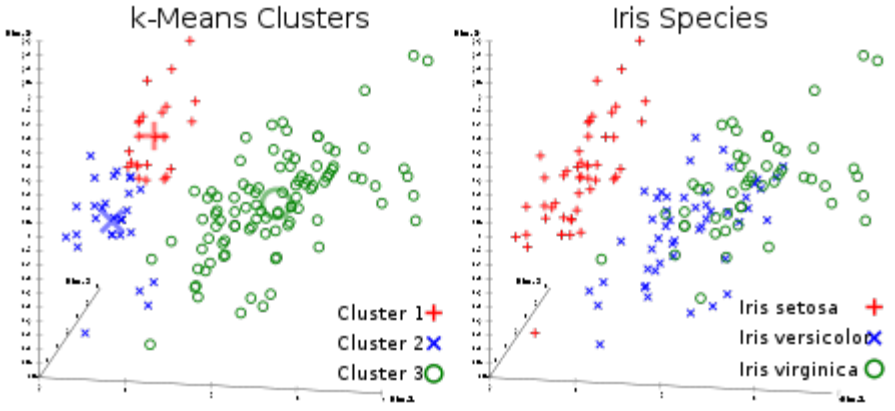







K=2
K=3
K=4


Different cluster analysis results on "mouse" data set:
Original Data
k-Means Clustering
EM Clustering


k-Means Clusters
Cluster 1
Cluster 2
Cluster 3
Iris Species
Iris setosa
Iris versicolor
Iris virginica

Medición de la similitud

- Cómo medir la similitud entre muestras:
  - Utilización de medidas de similitud o distancia.
  - Adecuación de las medidas a la estructura de los datos.

Partición de los datos en grupos

- Cómo realizar la partición del conjunto de muestras en distintos grupos:
  - Exhaustiva: Conjunto de m datos $\Rightarrow$ Comprobar todos los posibles grupos de tamaño 1, 2, …, m

$$s = \sum_{k=1}^{m} \binom{m}{k}$$

para $m=50$ se obtiene $s=1.1259e+015$

Partición de los datos en grupos

- Evitar el fenómeno de explosión combinatoria con el particionado exhaustivo
  - Utilización de heurísticas $\rightarrow$ Incorporación de reglas obtenidas por experiencia del diseñador.
  - Optimización iterativa de una función objetivo $\rightarrow$ Comenzar en un particionado de los datos inicial y luego mover/unir/dividir las muestras entre los grupos hasta alcanzar un mínimo de una función objetivo.

Cuando se obtiene un buen resultado?

- Un buen resultado en análisis de agrupamientos se producirá cuando los clusters generados posean:
  - alta similitud intra-cluster
  - baja similitud inter-cluster
- La calidad de los clusters resultantes dependerán de la medida utilizada y de la estrategia implementada
- La calidad de un método de clustering también se puede medir por la capacidad de descubrir relaciones ocultas en los datos.

Clasificación de los métodos de análisis de agrupamientos

- Metodos de reagrupamiento vs Métodos jerárquicos.
- Métodos jerárquicos aglomerativos vs divisivos.
- Métodos jerárquicos multicaracterísticas vs tipológicos.
- Métodos de agrupamiento exclusivo vs solapado.
- Métodos directos vs iterativos
- Métodos secuenciales vs simultaneos
- Métodos adaptativos vs no adaptativos

Ejemplos de Análisis de Agrupamientos

- Marketing: Ayuda a las empresas a descubrir tipos de clientes en sus bases de datos y utilizar este conocimiento para desarrollar programas de marketing específicos para cada tipología de cliente.
- Utilización del terreno: Identificar áreas de uso parecido del terreno en bases de datos de observación de la tierra.
- Seguros: Identificar grupos de usuarios de pólizas de seguro de automóviles con una tasa de siniestralidad determinada.
- Planificación en ciudades: Agrupar las casas de acuerdo a su tipo, valor y localización geográfica.
- Estudio terremotos: Obtener zonas de epicentros de terremotos en las zonas de fallas continentales

LEADER

Procedimiento LEADER

- Método de agrupamiento heurístico iterativo.
- Algoritmo:

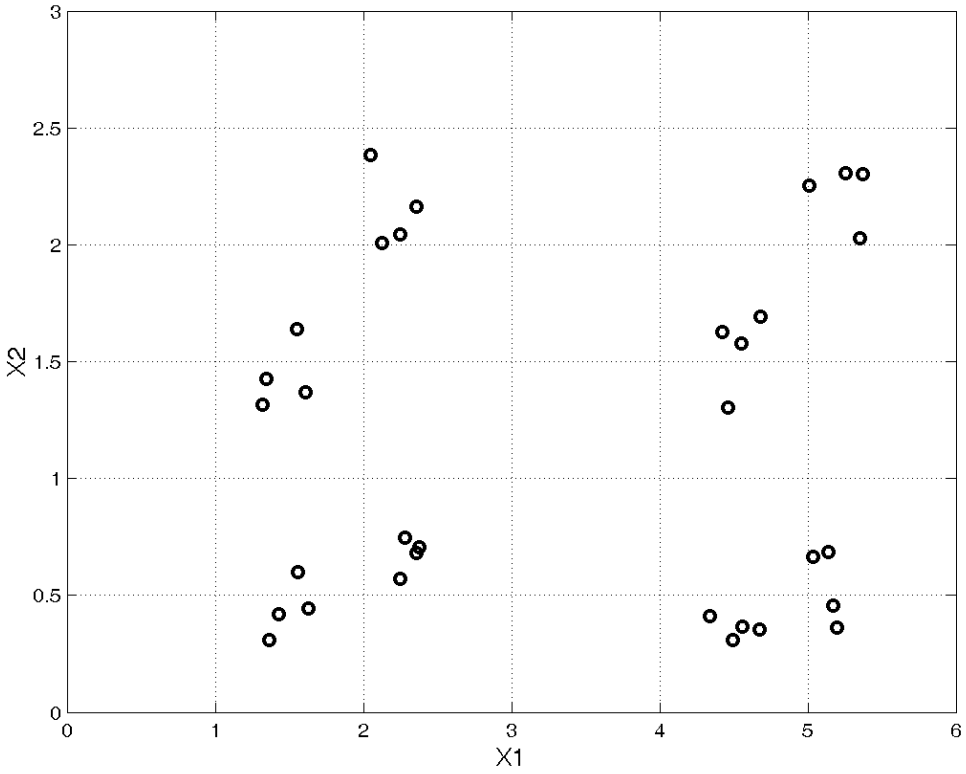
1. Considerar la primera muestra como el primer cluster
2. Asignar como centroide del primer cluster la primera muestra
3. Siguiente muestra:
   - Calcular la distancia a los centroides de todos los cluster
   - d_{min}=distancia al cluster más cercano
   - si d_{min} < umbral entonces
     - asignar la muestra al cluster más cercano
     - actualizar el centroide del cluster más cercano
   - si no
     - Generar un nuevo cluster con la muestra
4. Repetir el paso 3 hasta que todas las muestras estén asignadas a un cluster

Procedimiento LEADER

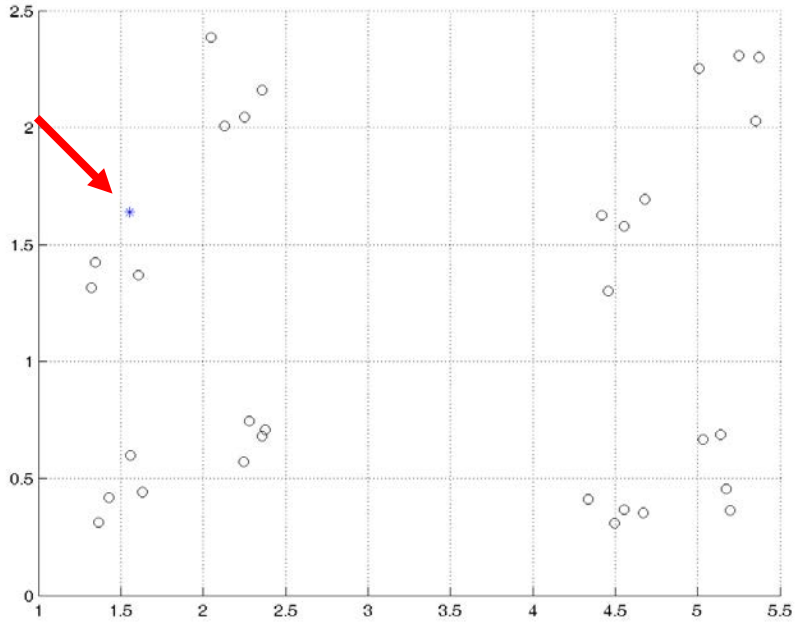
- Los resultados del procedimiento LEADER dependen de:
  - La primera muestra escogida
  - El orden en que se visitan las muestras
  - El umbral de distancia para generar un nuevo cluster
  - Las propiedades geométricas de las clases

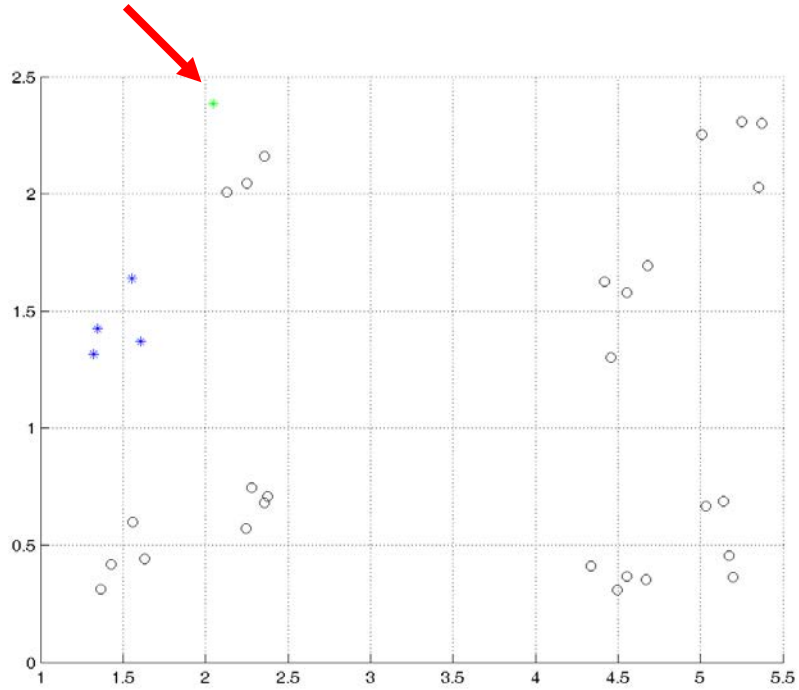
Procedimiento LEADER - Ejemplo

- Conjunto de muestras

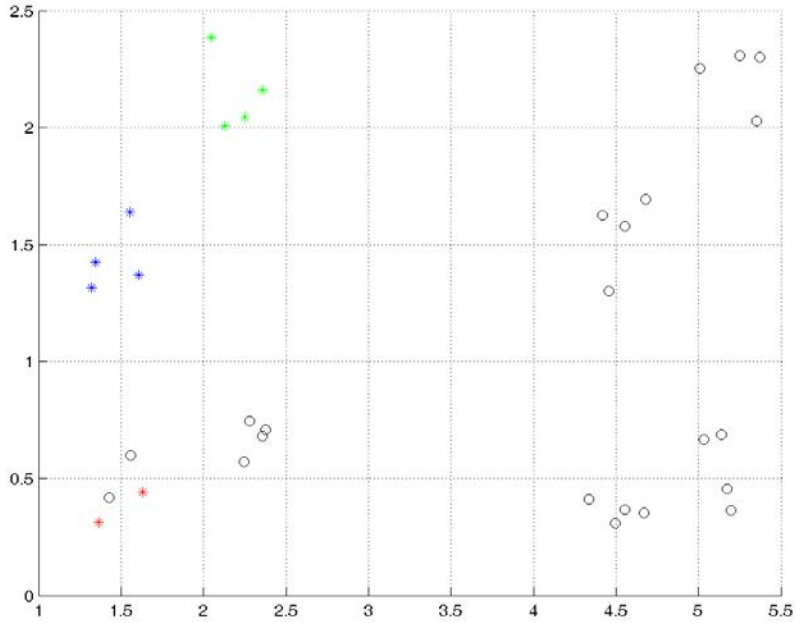


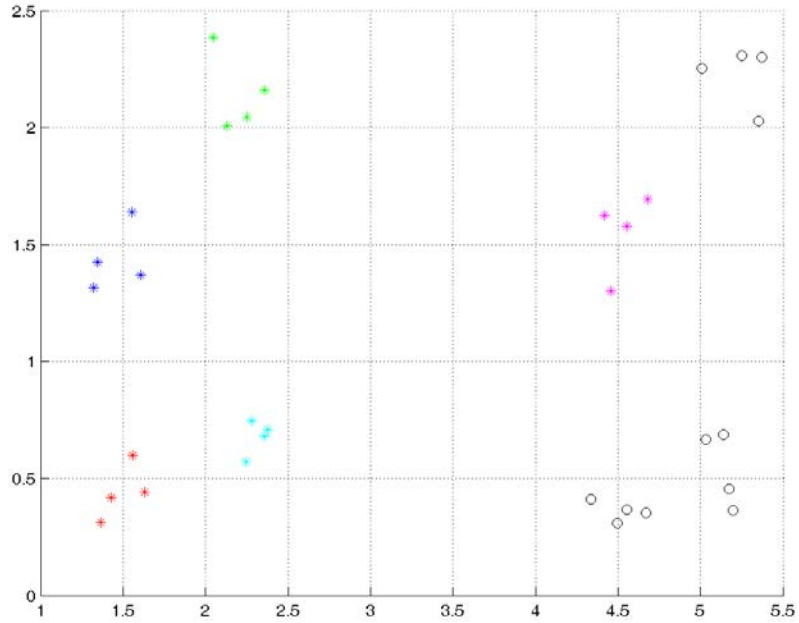
Procedimiento LEADER - Ejemplo



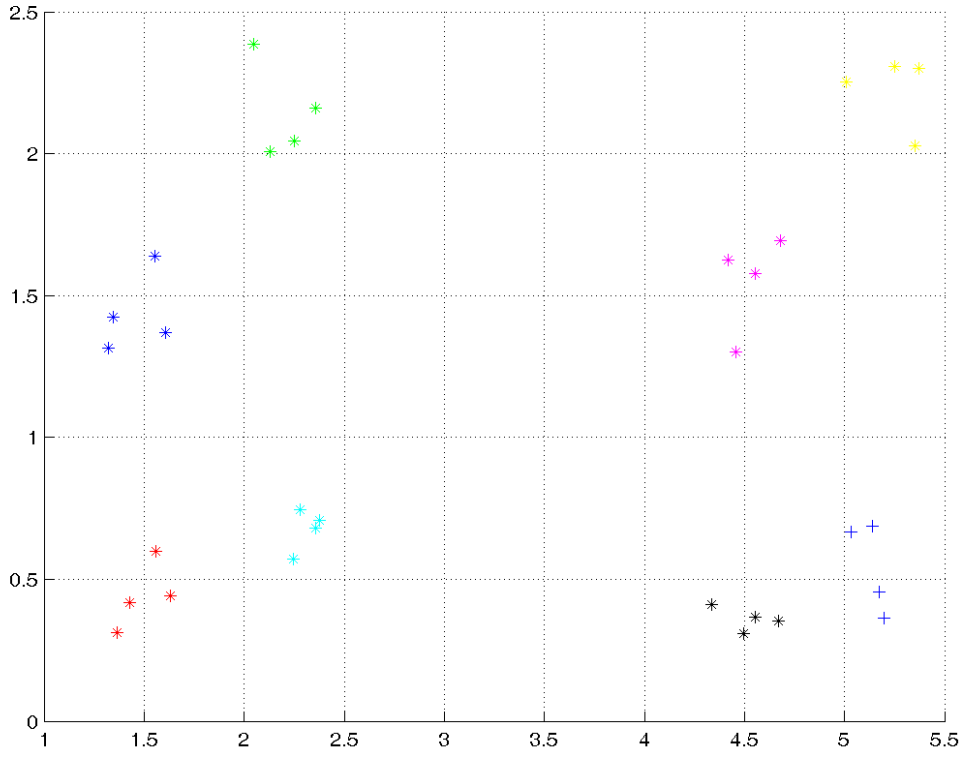


Procedimiento LEADER - Ejemplo





Procedimiento LEADER - Ejemplo



MAXIMIN

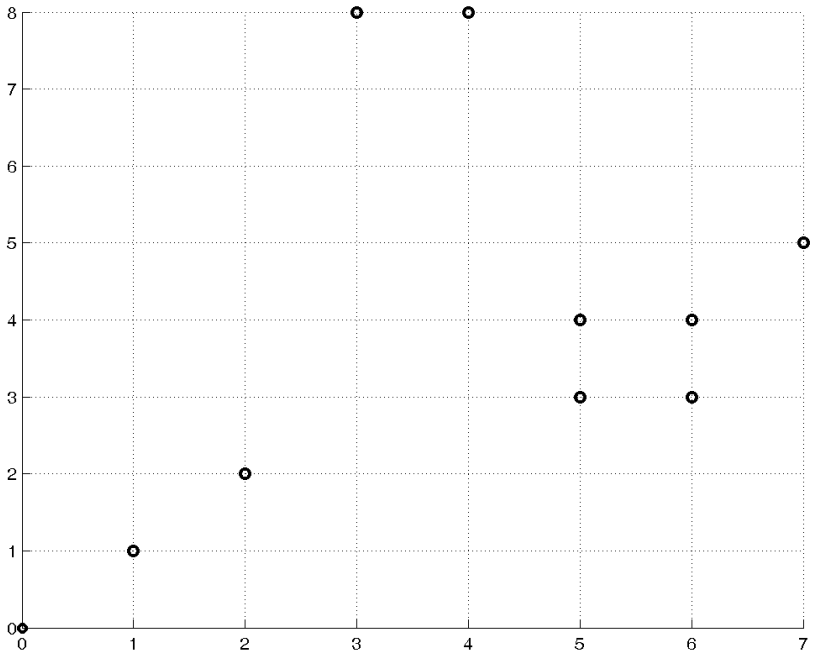
Procedimiento MAXIMIN

- Método iterativo que busca cluster apartados entre si.
- Termina cuando todas las muestras se encuentran a una distancia por debajo de un umbral fijado.

Procedimiento MAXIMIN

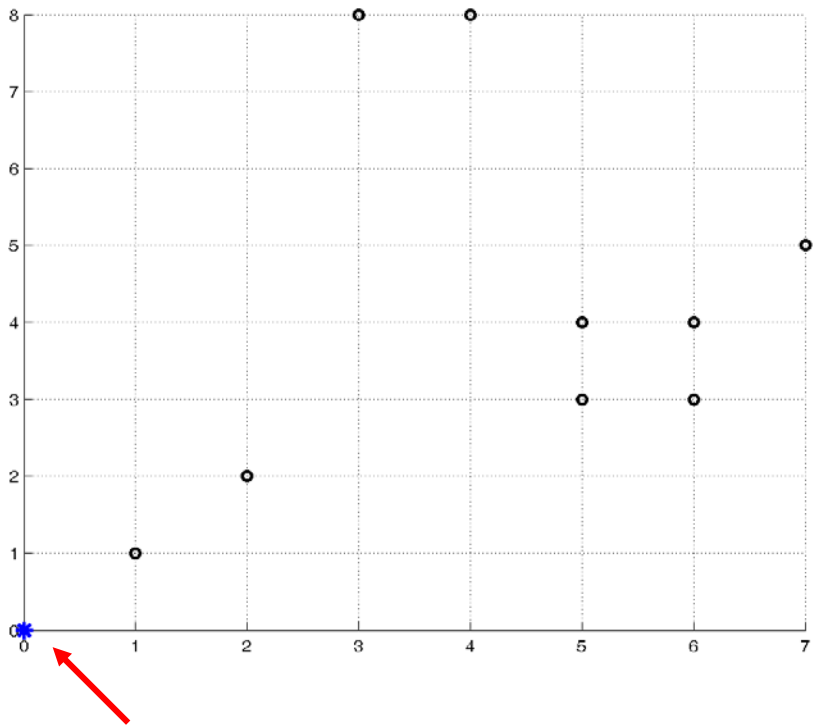
1. Elegir aleatoriamente una muestra y asignarla como centro $\boldsymbol{Z}_1$ del primer cluster .
2. Buscar la muestra más lejana a $\boldsymbol{Z}_1$ y asignarla como centro $\boldsymbol{Z}_2$ del segundo cluster.
3. Calcular la distancia de todas las muestras a los centroides existentes $\Rightarrow$ Tantas distancias como centroides por cada muestra.
4. Guardar la mínima distancia de cada muestra a los centroides.
5. Buscar la máxima de las distancias anteriores, *maxdist*
6. si *maxdist* > umbral entonces
   - hacer la muestra $\mathbf{X}$ un nuevo centro
7. sino
   - Asignar las muestras a los clusters más cercanos representados por los centro obtenidos
   - Finalizar
8. fin si
9. Ir al paso 3

Procedimiento MAXIMIN - Ejemplo



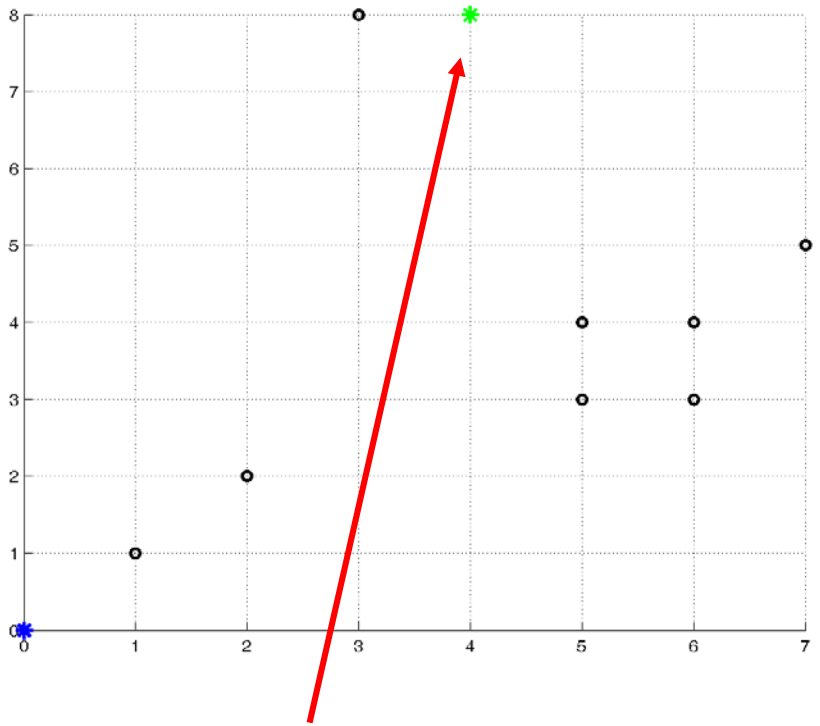
Conjunto de muestras

Procedimiento MAXIMIN - Ejemplo



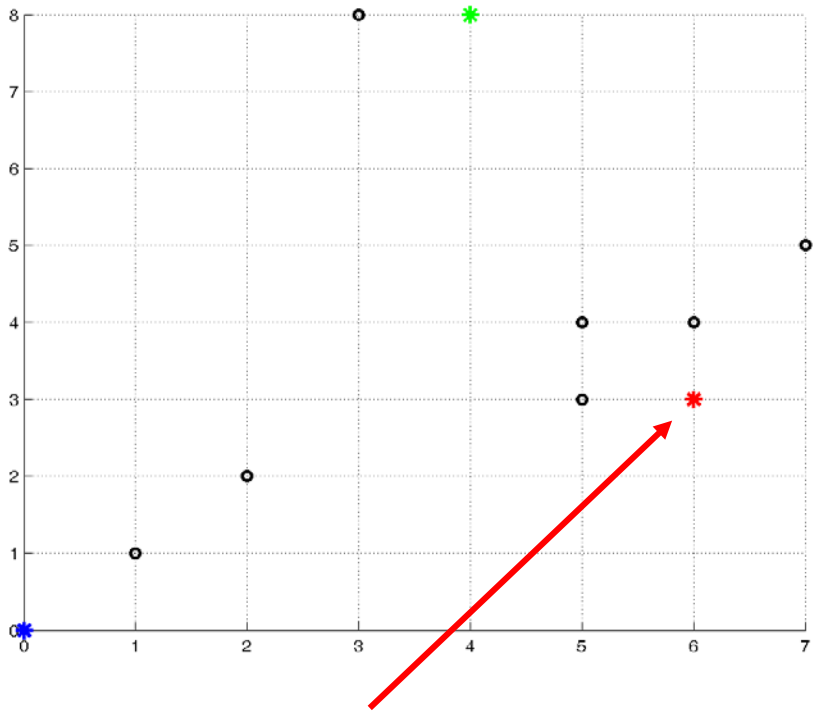
Primer centroide

Procedimiento MAXIMIN - Ejemplo



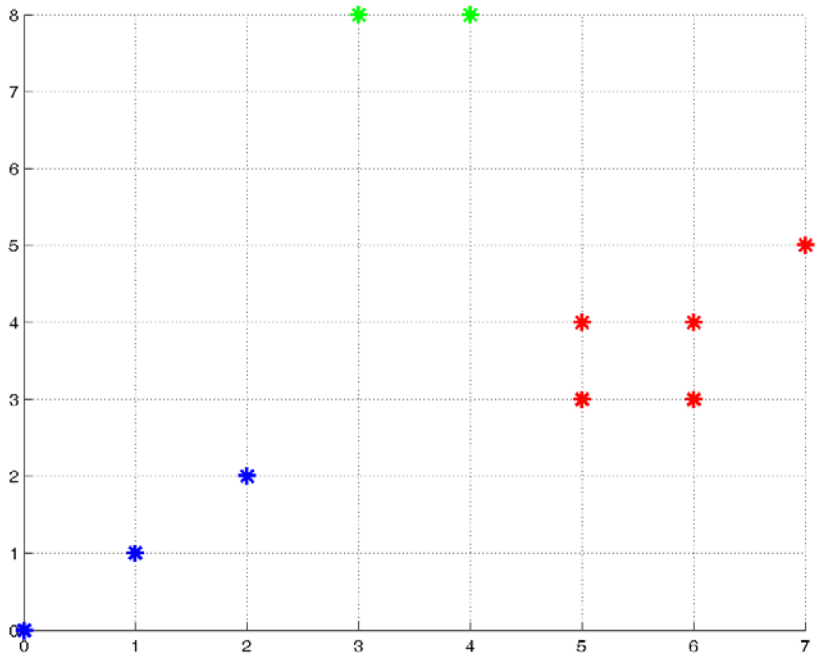
Segundo centroide

Procedimiento MAXIMIN - Ejemplo



Tercer centroide

Procedimiento MAXIMIN - Ejemplo



Asignación de muestras
al centroide más cercano

Fin Agrupamiento Heurístico