Análisis de Agrupamientos II: Reagrupamiento

Mario Hernández

Introducción

- Métodos reagrupamiento $\rightarrow$ Comenzar con una partición inicial e ir refinando la misma para optimizar una función objetivo $\rightarrow$ **K-medias (K-means)**

K-medias

Método K-medias

- Aproximación de clustering por reagrupamiento
- Cada cluster se asocia con un centroide (punto central)
- Las muestras se asignan al cluster cuyo centroide se encuentra más próximo
- Se debe especificar el número de clusters *K*.
- El algoritmo básico es muy simple
- Utilización de una matriz de pertenencia $\mathbf{W}_{kxm}$

$$w_{ij} = \begin{cases} 0 & \text{muestra } \mathbf{X}_j \notin \text{cluster } i \\ 1 & \text{muestra } \mathbf{X}_j \in \text{cluster } i \end{cases}$$

K-medias – Algoritmo Básico

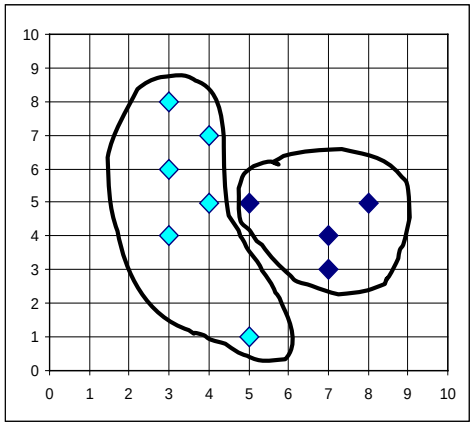
1. Distribuir las muestras aleatoriamente en K clusters
2. Calcular los centroides $\mathbf{Z}_i$

$$\mathbf{Z}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} w_{ij} \mathbf{X}_j \quad n_i \equiv \text{Núm. muestras cluster } i$$

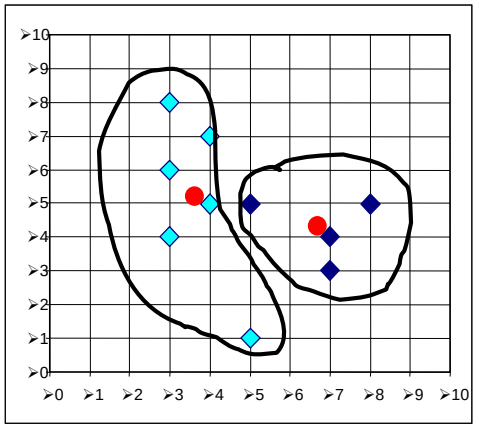
3. para todas las muestras $\mathbf{X}$ hacer
   1. Asignar la muestra $\mathbf{X}$ al centroide más cercano
   2. Actualizar matriz $\mathbf{W}$
   3. Recalcular los centroides $\mathbf{Z}_i$
4. hasta que no se mueva ninguna muestra de cluster

El Método *K-Medias*

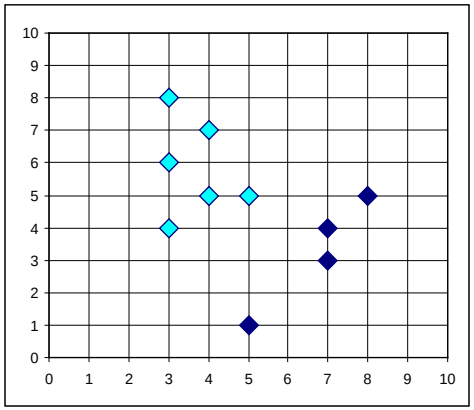
Fijar K y asignar las muestras aleatoria-mente a los K clusters



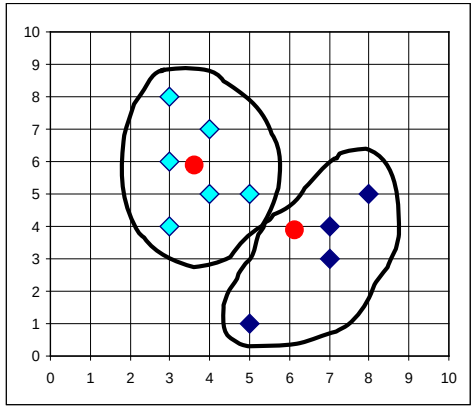
Actualizar los centroides de los cluster



Reasignar



Actualizar los centroides de los cluster



K-medias - Consideraciones

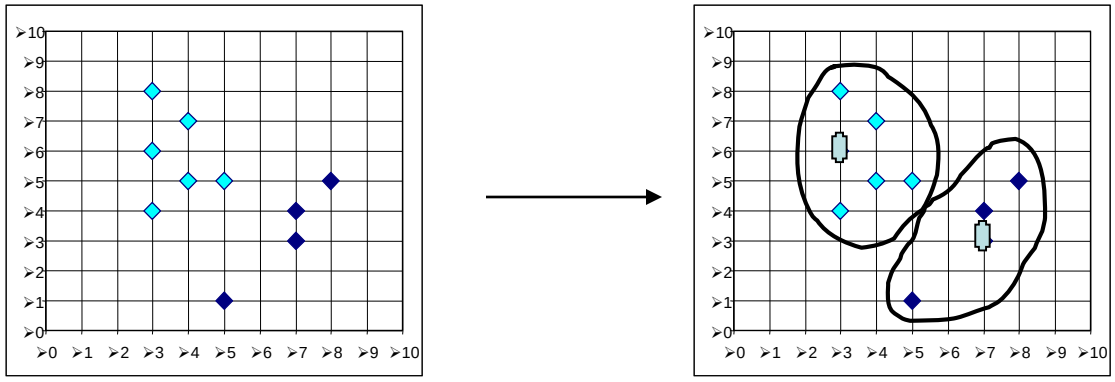
- Partición inicial es aleatoria $\Rightarrow$ Los resultados pueden variar de una ejecución del procedimiento a otra.
- 'más cercano' $\rightarrow$ utilizar cualquier medida de distancia, por ejemplo: euclídea o Manhattan$\Rightarrow$ el procedimiento converge a una solución
- Converge normalmente en pocas iteraciones
- A veces se relaja la condición de parada del algoritmo básico por:

  …

  4. <u>hasta</u> que pocas muestras se muevan de cluster

¿Problemas del K-Medias?

- El procedimiento es sensible a outliers !
  - Un objeto con un valor extremadamente separado del resto del cluster puede distorsionar la distribución de los datos..
- K-Medoides: En vez de usar el Centroide (valor medio) de las muestras del cluster como punto de referencia, usar el Medoide, que es la muestra localizada más centralmente en el cluster.



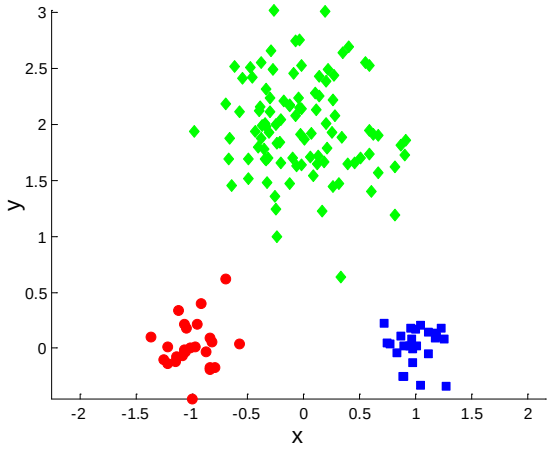
K-medias – Evaluación resultados

- Medida utilizada para la evaluación de resultados → Suma de los Errores al Cuadrado (Sum of the Squared Error - SSE)
  - Para cada muestra, el error se define como la distancia al centroide asociado a su cluster.
  - Para calcular el SSE, se obtiene la suma del cuadrado de los errores para todas las muestras en cada cluster:.

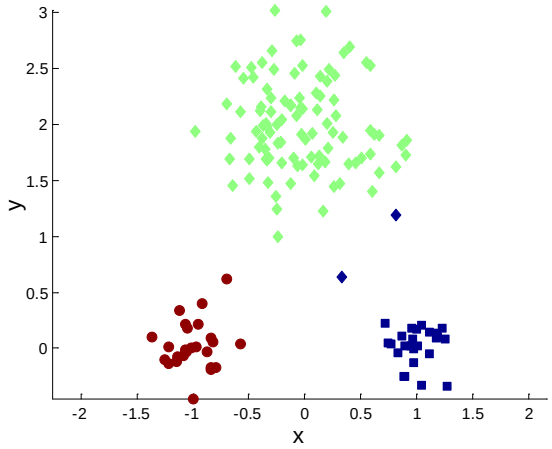
$$J_{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{m} w_{ij} \left\| \mathbf{X}_j - \mathbf{Z}_i \right\|^2$$

  - Dados dos resultados del algoritmo K-means, se puede elegir aquel con menor valor de SSE.

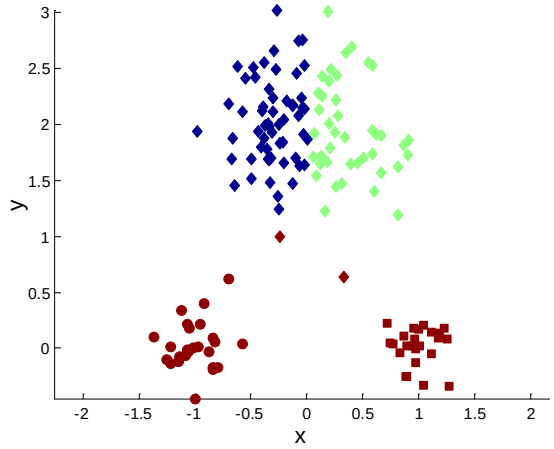
K-medias



Muestras

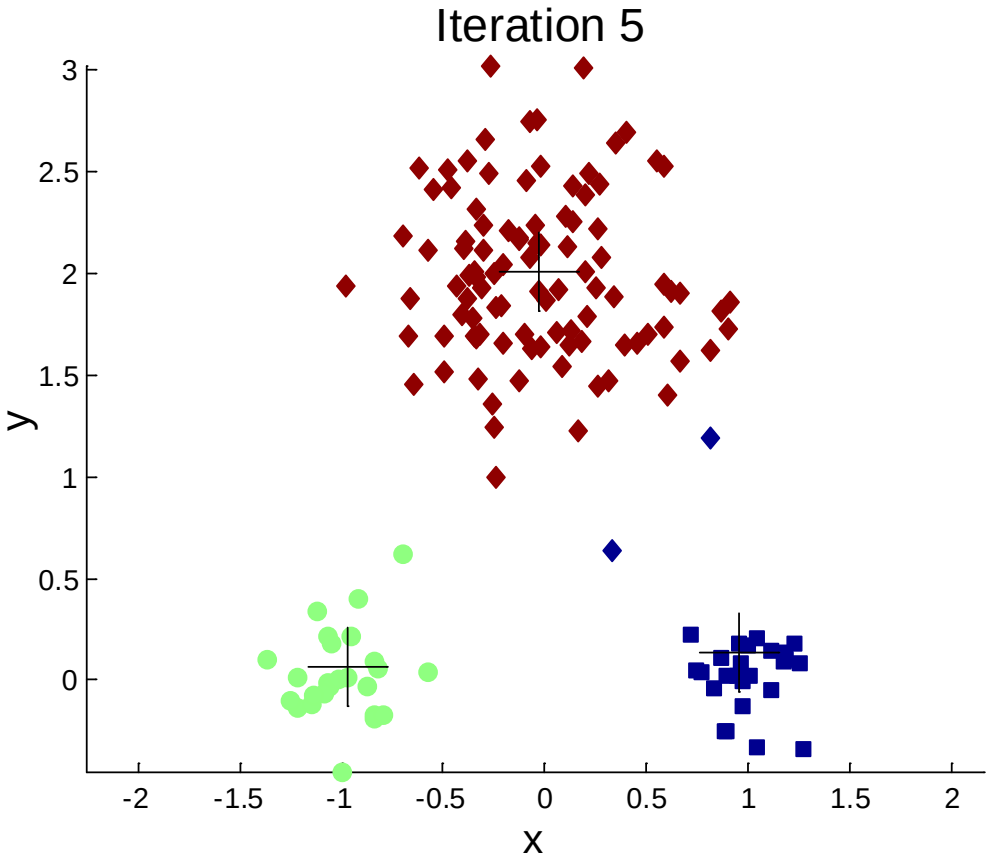


Clustering óptimo

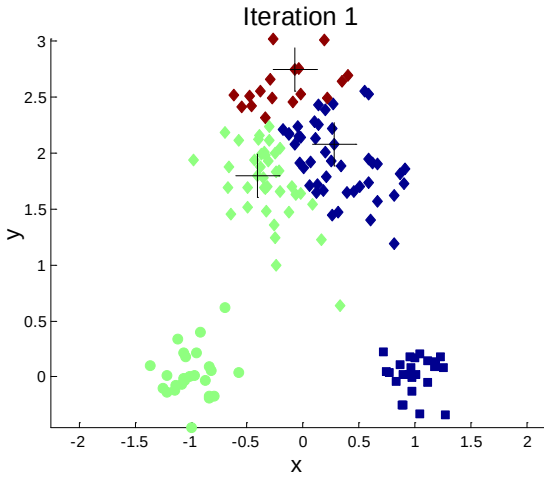


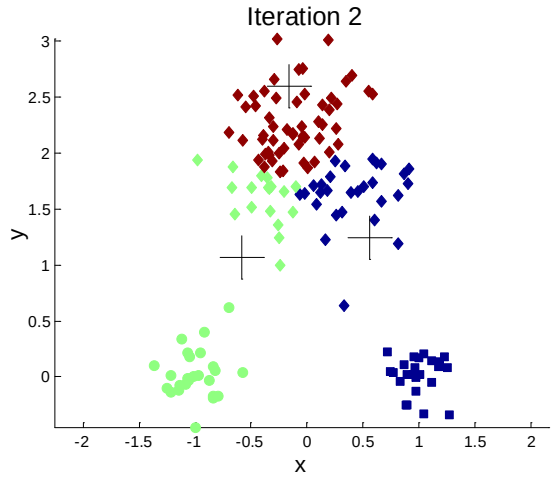
Clustering sub-óptimo

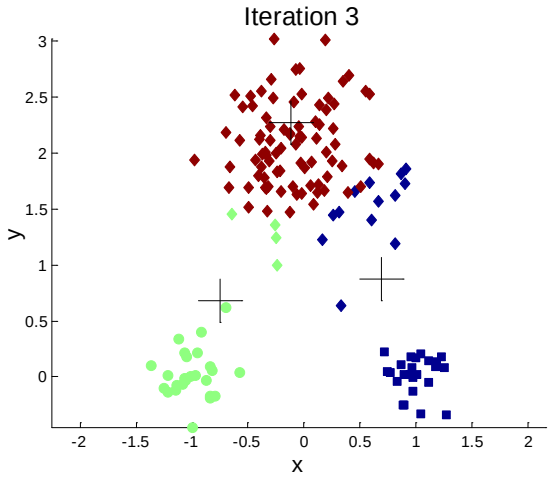
K-medias – Dependencia cluster inicial

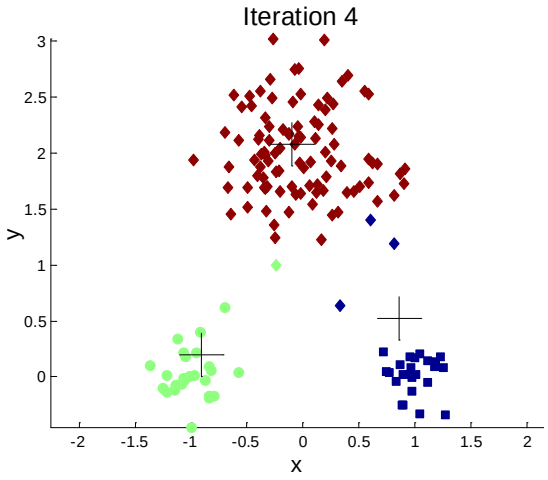


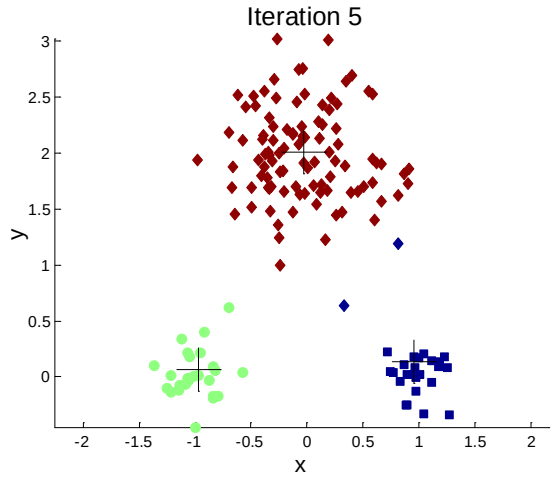
K-medias – Dependencia cluster inicial

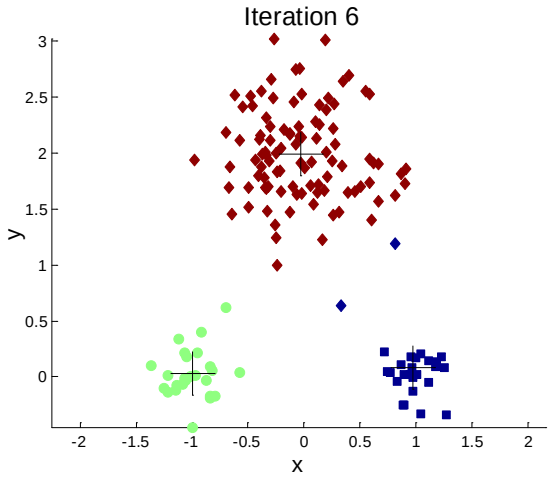




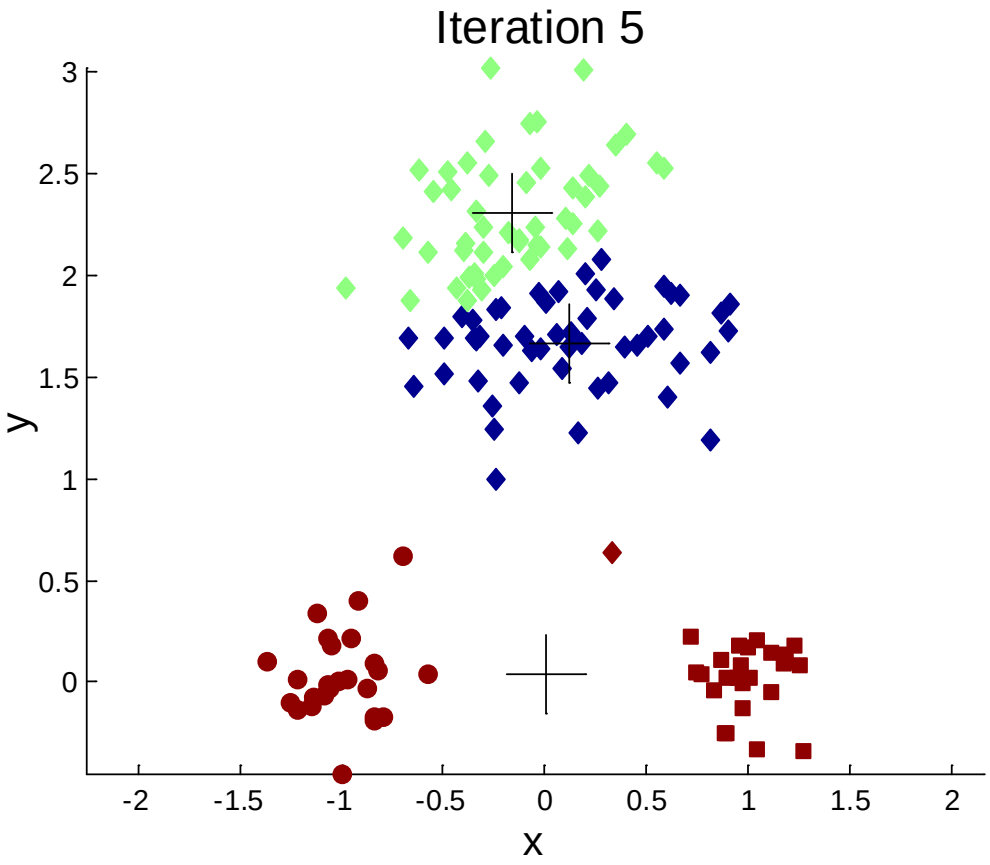




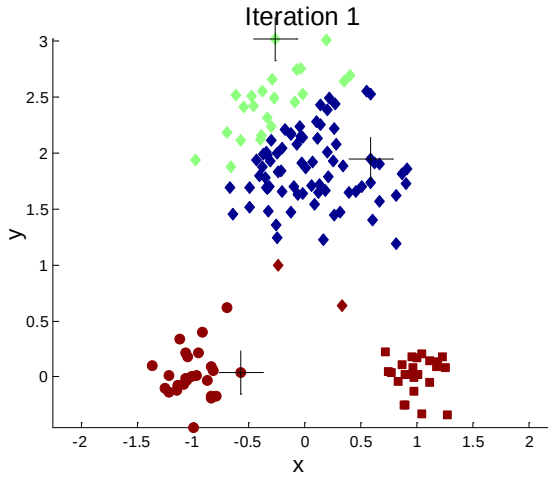


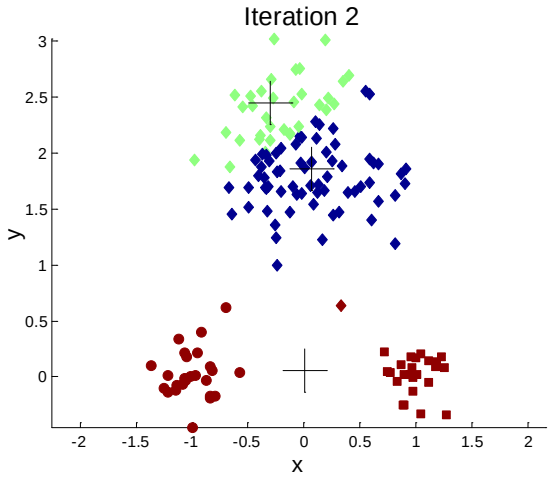


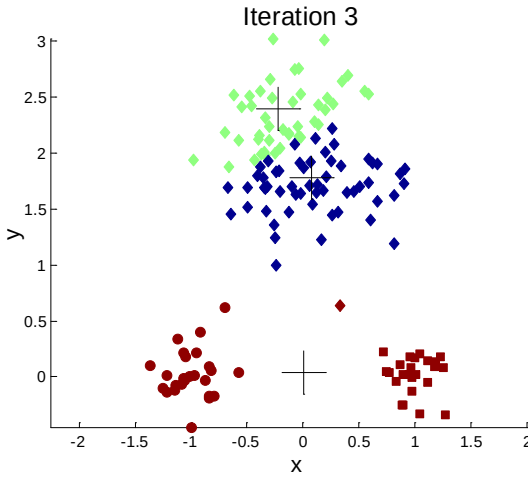
K-medias – Dependencia cluster inicial

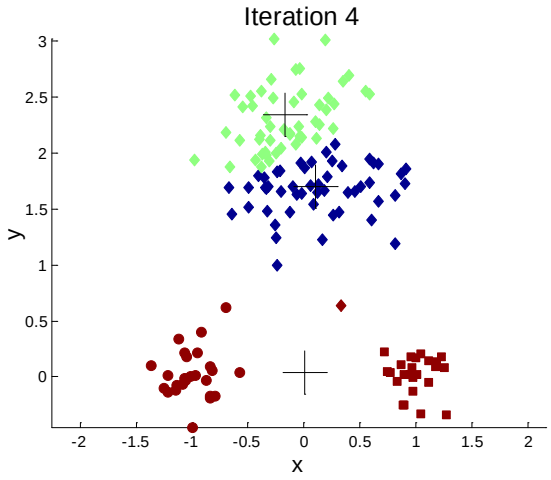


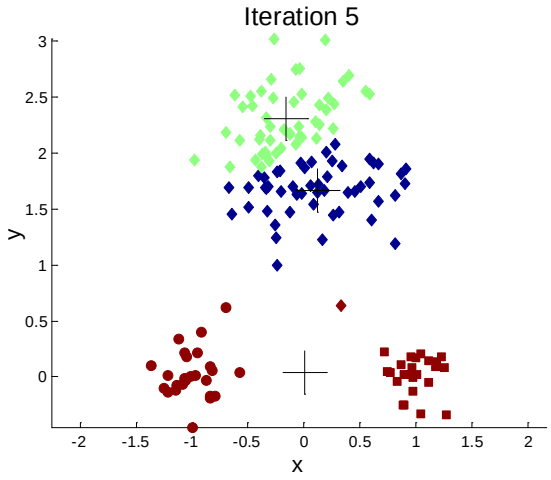
K-medias – Dependencia cluster inicial











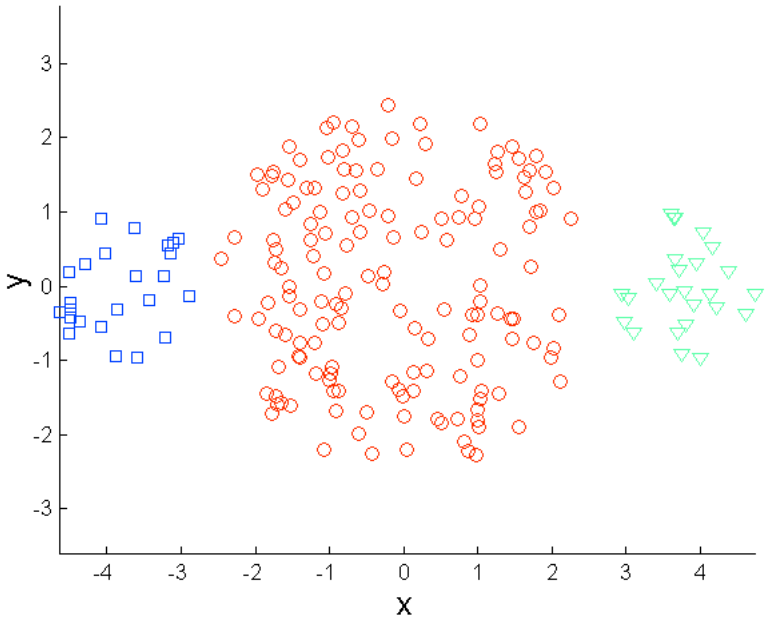
- Es un procedimiento relativamente eficiente y rápido.
- Determina la partición resultante en **O(tkn),** donde n es el número de objetos o muestras, k es el número de clusters y t es el número de iteraciones.

K-medias - Limitaciones

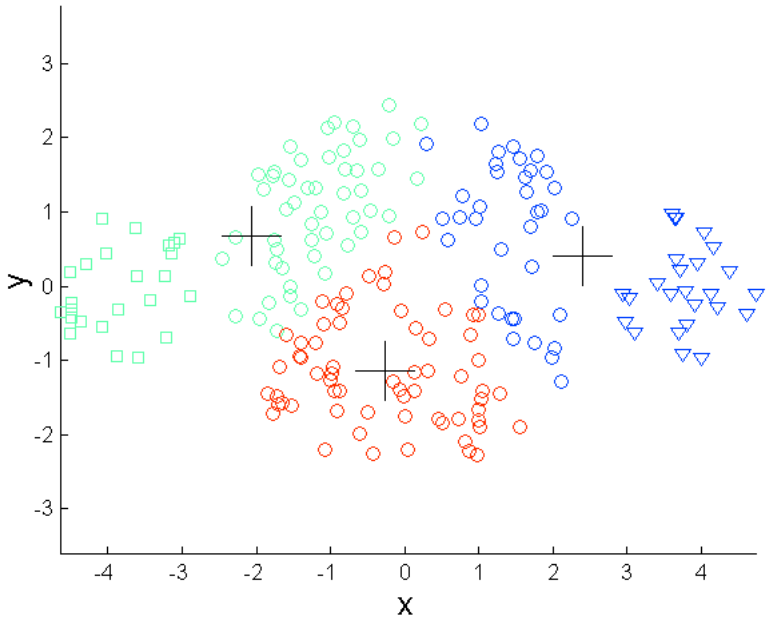
- K-medias no da buenos resultados cuando los clusters tienen:
  - Diferentes tamaños
  - Diferentes densidades
  - Formas no globulares
- Problema de las muestras fuera de rango (*outliers*)
- Una posible solución puede ser incrementar el valor de K para obtener más cluster $\Rightarrow$ dividir un único cluster en partes
  - Incorporar postproceso para agrupar los clusters

- Pueden quedar clusters vacios si se mueven varias muestras a la vez antes de actualizar

K-medias – Diferentes tamaños de clusters

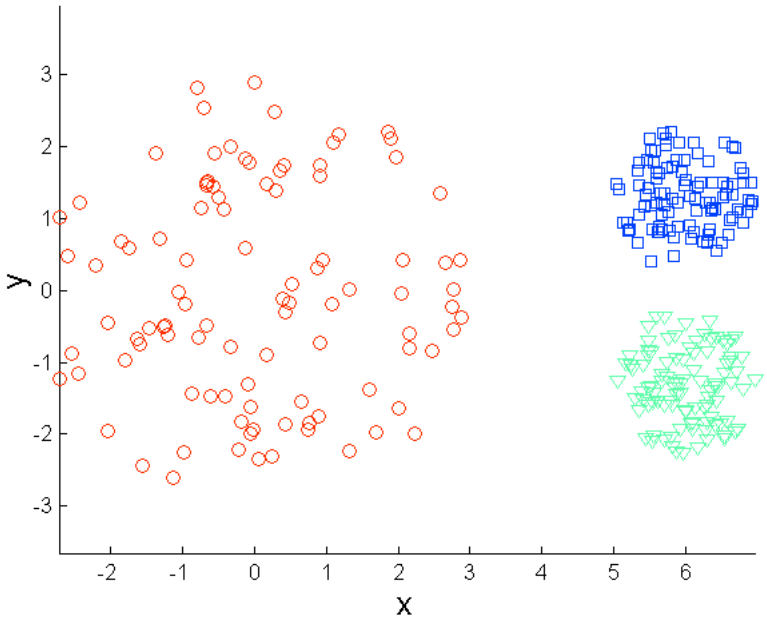


Muestras

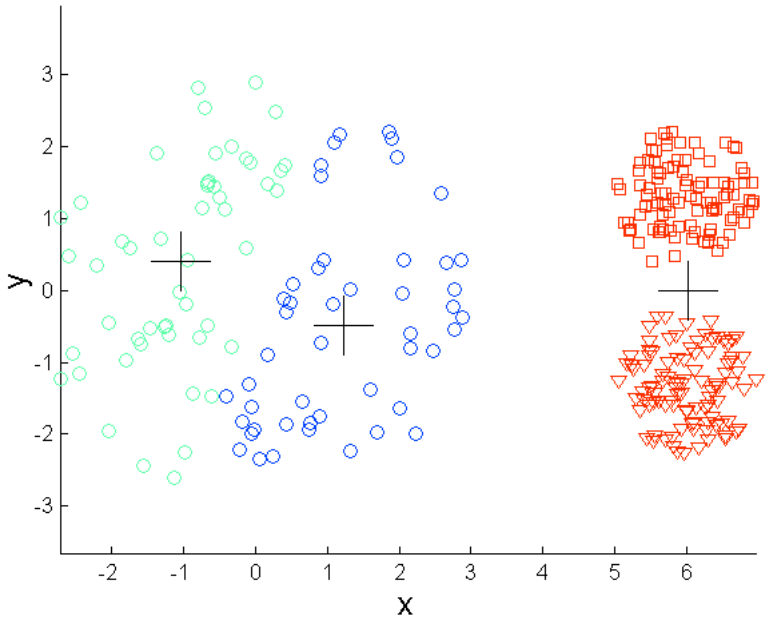


Resultado K-media

K-medias – Diferentes densidades de clusters

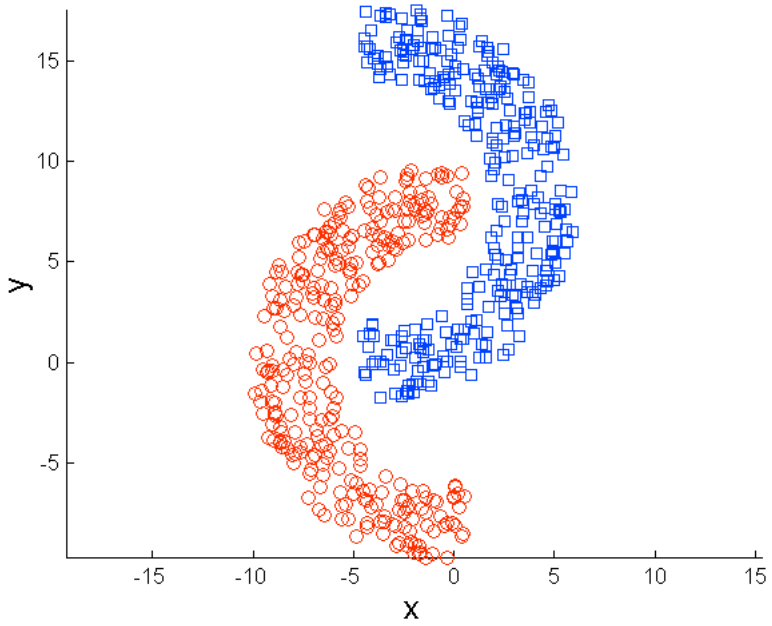


Muestras

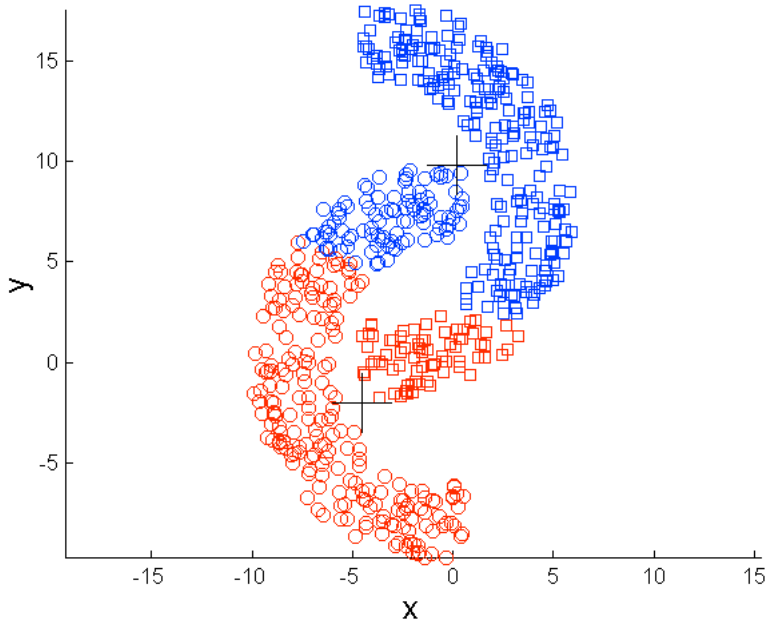


Resultado K-media

K-medias – Clusters no globulares

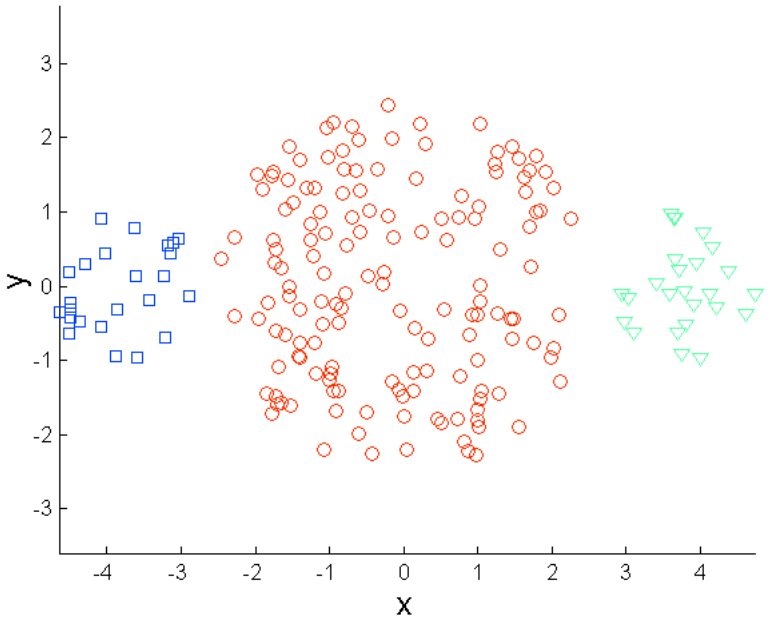


Muestras

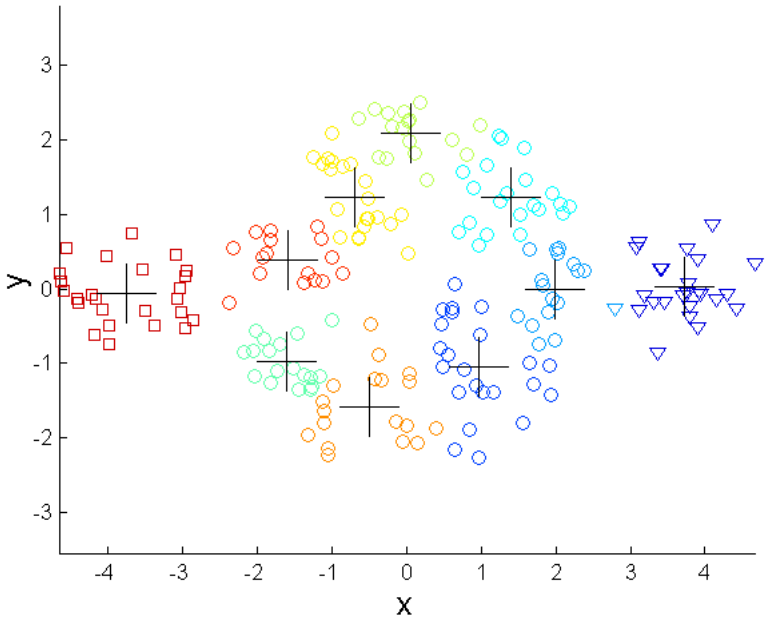


Resultado K-media

K-medias – Aumentar K

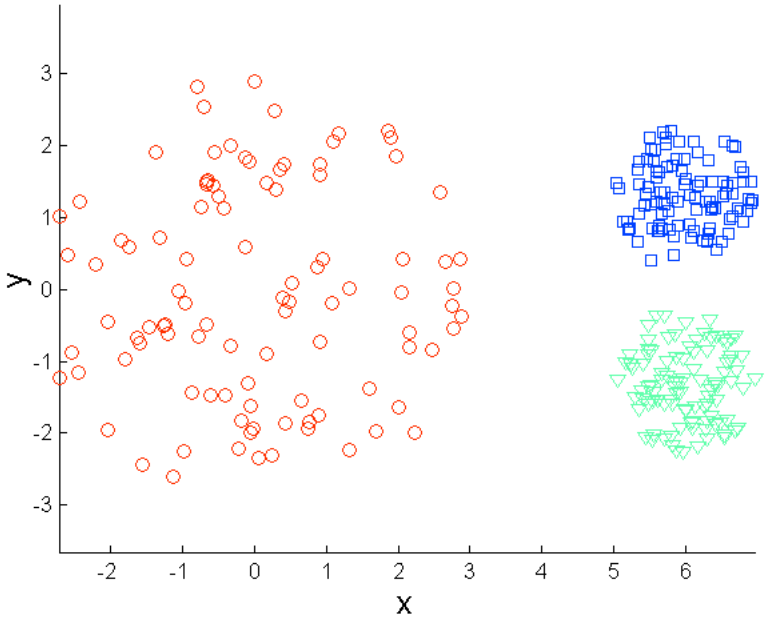


Muestras

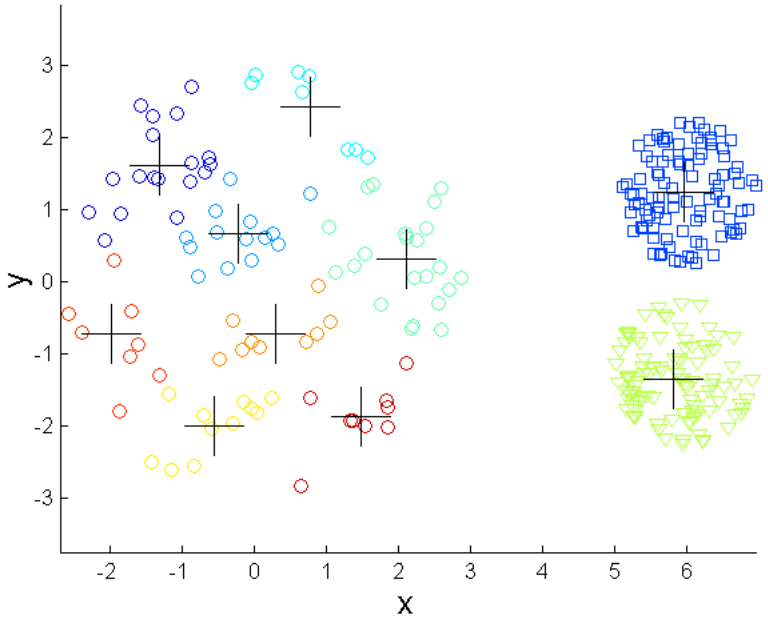


Resultado K-media

K-medias – Aumentar K

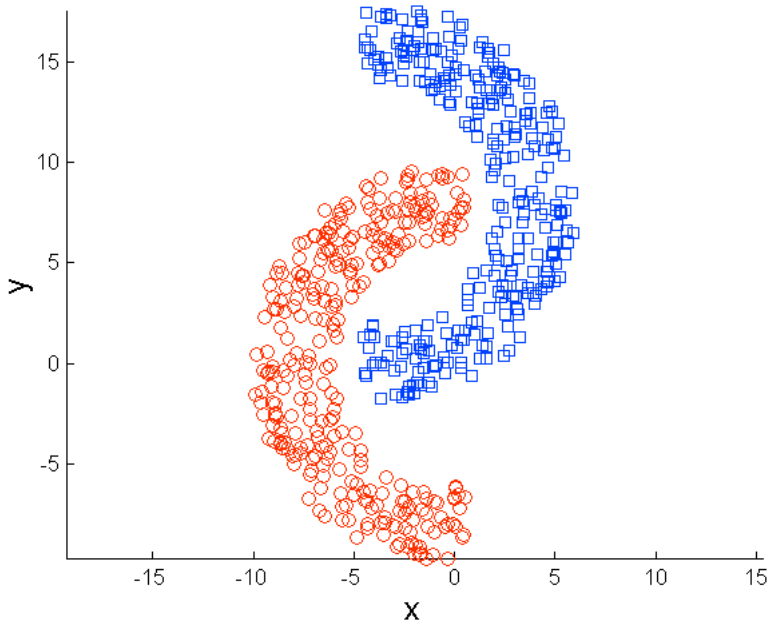


Muestras

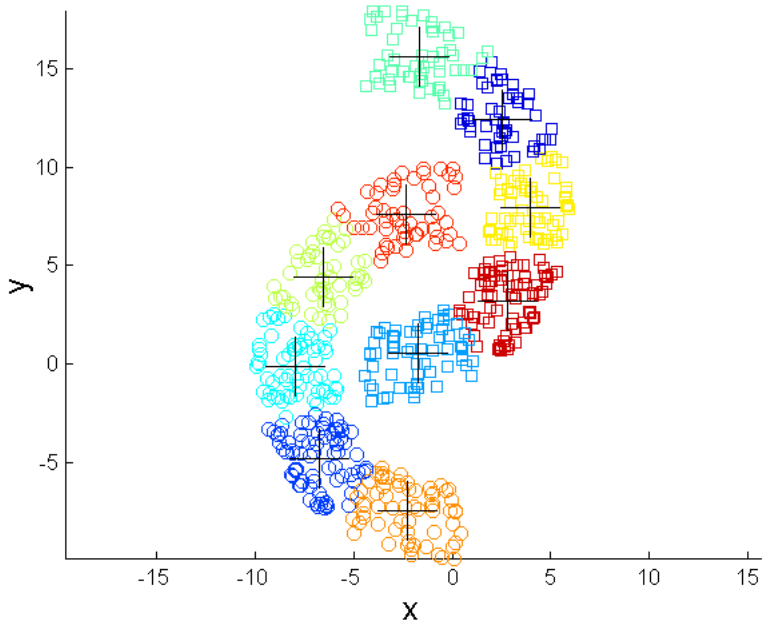


Resultado K-media

K-medias – Aumentar K



Muestras



Resultado K-media

Determinación del número k de clúster para k-medias

¿Cómo fijar k?

- La principal limitación del k-medias es que necesita como argumento el número k de clases en las que hay que partir los datos.
- Una manera de intentar definir el número k de clusters óptimo es a través de visualización de los datos
- La visualización tiene una primera limitación que es la dimensión de los datos, que complica la tarea
- Hay soluciones alternativas, por ejemplo:
  - **Método del Codo** (Elbow method)
  - **Método de la Silueta** (Silhouette method)

Método del Codo (Elbow Method)

- Sea que definimos la **Inercia** o Suma de Errores Cuadrático (Sum of Squared Error, SSE), también denominada dispersión intraclase (Within Class Dispersion, WCD) de todas muestras a los centroides de la clase a la que están asignados:

$$J_{SSE} = J_{WCD} = \sum_{i=1}^{k} \sum_{j=1}^{m} w_{ij} \left\| \boldsymbol{X}_j - \boldsymbol{Z}_i \right\|^2$$

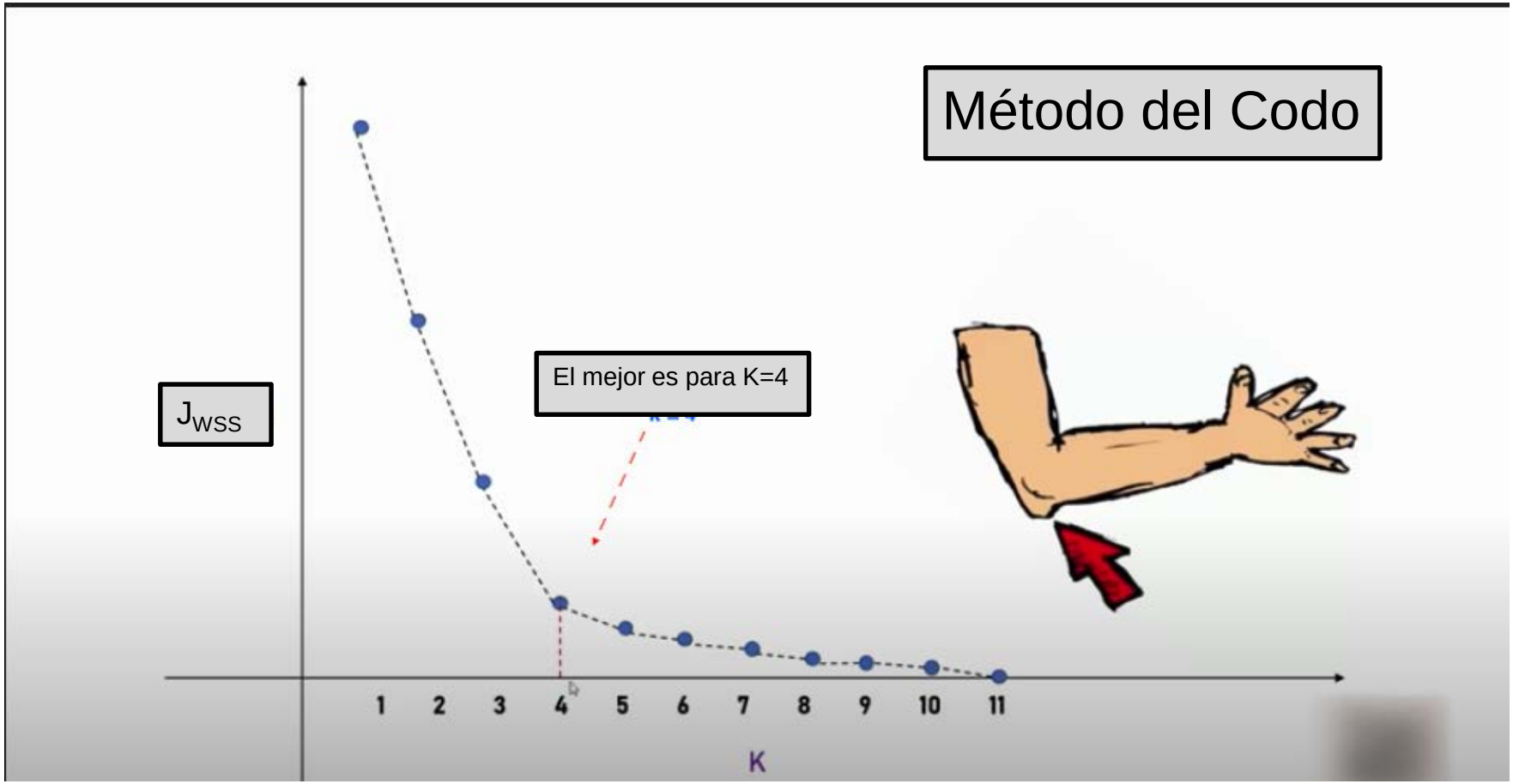
Método del Codo

- Para un K dado, a menor SSE, mejor es el modelo (más compacto)
- Si aumentamos K, la dispersión SSE disminuirá. Esto se debe a que las muestras estarán más cercanas a los centroides de las clases a las que son asignadas.
- Podemos representar gráficamente los resultados de SSE sobre los conjuntos de clases resultantes de dividir las muestras utilizando, por ejemplo K-medias según diferentes valores de K

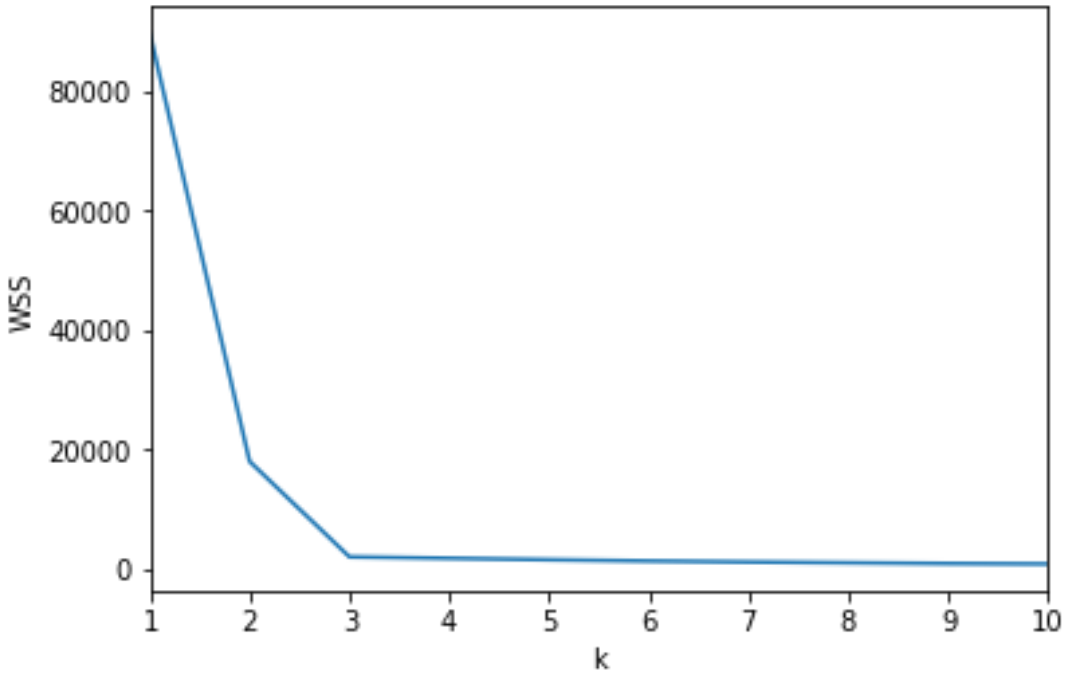
Método del Codo

- Calcular la Suma de Errores Cuadráticos Intra-Cluster (Within-Cluster-Sum of Squared Errors, WSS) para diferentes valores de k, y escojer el k para el cual la JWSS reduzca la pendiente hacia la asíntota. En el gráfico de JWSS-versus-k, es visible como si fuse un codo en la curva.

Método del Codo



Método del Codo



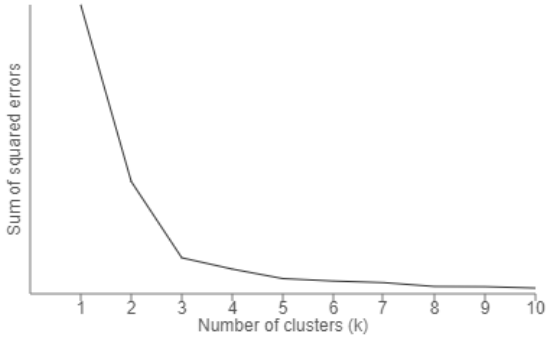
Método del Codo

- El valor del punto del codo es el que nos fija el valor de K
- Es un punto en el que la pendiente se incrementa relativamente respecto a los anteriores
- Después del punto del codo, la curva suele comportarse más o menos asintóticamente (tender a una llanura o “plateau”)

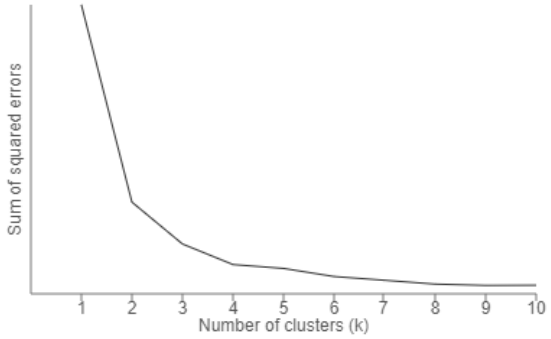
Método del Codo

- El método es muy simple para obtener el K óptimo pero a veces la gráfica resulta vaga o ambigua para realizar la toma de decisiones.

Dataset A



Dataset B



- Para el Dataset A, el codo está claramente en k = 3. Sin embargo esta elección resulta ambigua para el Dataset B, donde se podría escoger 3 or 4.
- En estos casos resulta mejor utilizar un método intrínseco de puntuación como el que veremos a continuación

Método de la Silueta (Silhouette Method)

- Para la interpretación y validación de la coherencia en análisis de agrupamientos
- **Silueta:** figura de mérito que mide la cohesión de cada muestra (objeto) a su propio cúmulo (grupo, clase, cluster) en comparación con otros cúmulos
- Es una medida (s) cuyo valor se encuentra en el rango $-1 \leq s \leq 1$
  - $s=+1 \Rightarrow$ el objeto esta perfectamente emparejado con su cúmulo y mal con los vecinos. Si la mayoría de los objetos tienen un valor alto, la configuración del cúmulo es apropiada
  - Si muchos objetos tienen un valor bajo (s por debajo de 0) la configuración es inapropiada. Mas inapropiada cuando más se acerca a -1
- Sea que el conjunto de muestras de aprendizaje no supervisadas han sido agrupadas por k-media en un cierto número k de grupos, clusters o cúmulos

Definición de a(i)

- Es una medida de lo bien que la muestra i está asignada al cluster k
- Cuanto más pequeño es su valor, mejor es la asignación
- Para cada muestra (objeto) i asignada por k-media al cluster k C_k se define:

$$a(i) = \frac{1}{n_k - 1} \sum_{j \in C_k,\, j \neq i} d(i,j)$$

- Donde $d(i,j)$ es la distancia (por ejemplo euclídea o Manhattan) del punto i a cada punto j del cluster C_k , excluido aquella al propio punto i $d(i,i)$ (que será nula)

Definición de b(i)

- Sea ahora cada una de las distancias medias de la muestra i a todas las muestras j pertenecientes a cada cluster *l D(i;l)* excepto al cluster C_k al que pertenece, es decir:

$$D(i;l) = \frac{1}{n_l} \sum_{j \in C_k, k \neq i} d(i,j)$$

- Donde n_l es el número de muestras del cluster l en cuestión
- b(i) se define como el mínimo de las distancias promedio *D(i;l),* es decir, la **distancia promedio de i al cúmulo vecino más cercano**, o también se puede decir como la distancia de i al siguiente cúmulo que mejor se ajusta a la muestra i

$$b(i) = \min_l D(i;l)$$

Definición de silueta s(i)

- Se define para un objeto o muestra i como:

$$\mathrm{s(i)} = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \text{ si } n_k > 1$$

y: $\mathrm{s(i)} = 0$ si $n_k = 1$

- Es decir se asigna puntuación nula a las muestras pertenecientes a grupos de tamaño $n_k = 1$
- Esta última restricción se incluye para que el número de clusters no aumente significativamente

Definición de silueta s(i)

- También la podemos expresar como:

$$s(i)=\begin{cases} 1-\frac{a(i)}{b(i)} & si\ a(i)<b(i) \\ 0 & si\ a(i)=b(i) \\ \frac{b(i)}{a(i)}-1 & si\ a(i)>b(i) \end{cases}$$

Discusión de la silueta s(i)

- Como adelantamos:

$$-1 \leq s \leq 1$$

- Como a(i) es una medida de diferencia de la muestra i con su propio grupo, si a(i) es pequeño entonces i está bien emparejado
- Además, si b(i) es grande, i está mal emparejado con su cúmulo vecino más cercano
- Por tanto, si s(i) es cercano a 1, el dato está apropiadamente agrupado
- Si s(i) es próximo a -1, sería más adecuado asignar i al cúmulo más cercano al actualmente asignado
- Si s(i) es 0, la muestra está al borde de dos cúmulos.

Media de s(i) de todos los puntos de un clúster

- Medida de cuán estrechamente agrupados están todos los puntos de un cúmulo o clúster
- Si hay demasiados o muy pocos cúmulos a causa de una mala elección de k en el k-media, habrá cúmulos que mostrarán siluetas más estrechas que el resto
- Las gráficas de las siluetas y sus medias por clúster pueden utilizarse para determinar el número natural de cúmulos dentro de un conjunto de muestras de aprendizaje

Coeficiente de Silueta de Kaufman

- Es el valor máximo del promedio para todos los datos de un conjunto de datos.
- Así si $\bar{s}(k)$ es el **SCORE** o media de las siluetas s(i) de todas las n muestras del conjunto de datos para una partición en k clusters:

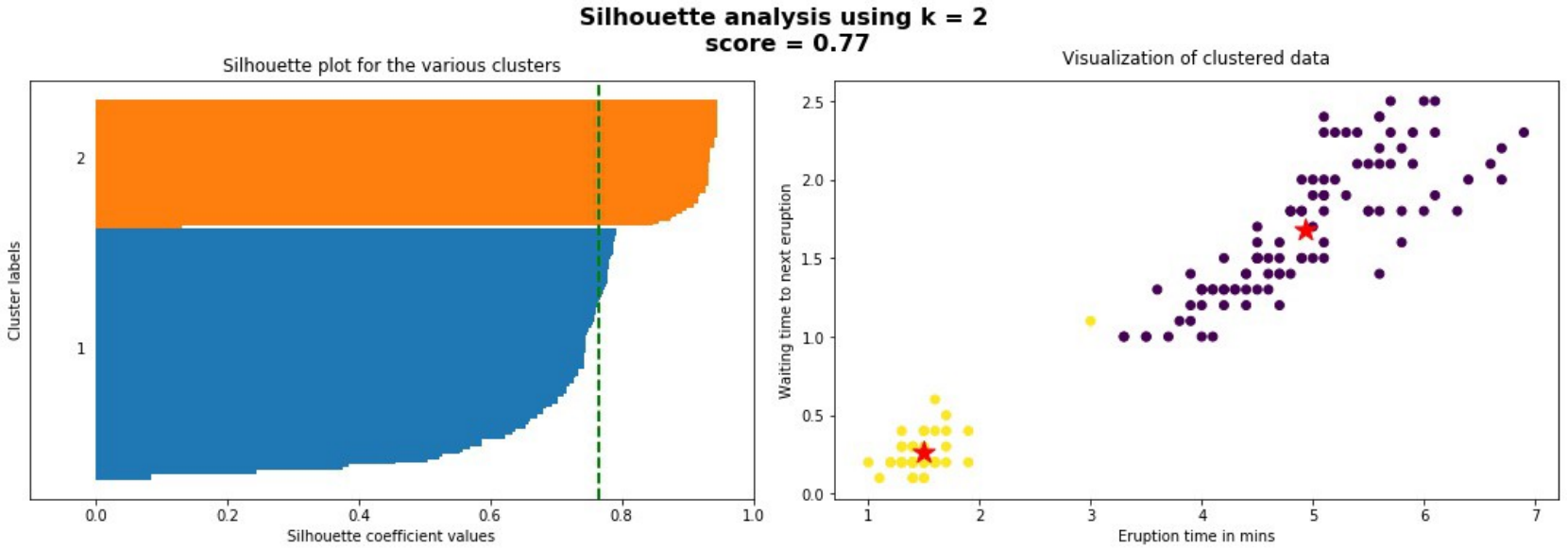
$$\bar{s}(k) = \frac{1}{n}\sum_{i=1}^{n} s(i)$$

- El **Coeficiente de Silueta de Kaufman (SC)** se define para todas las particiones en k diferentes grupos con k-media como:

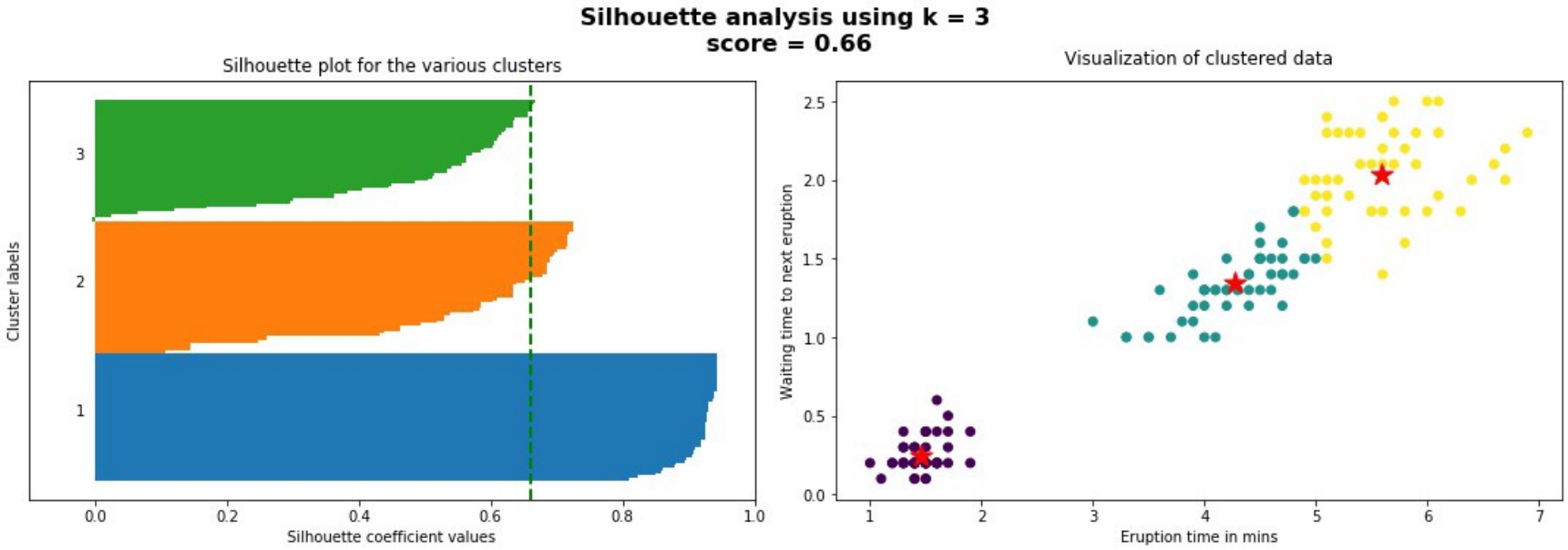
$$\mathbf{SC} = \max_{k} \bar{s}(k)$$

- Ese valor que da el máximo es el del número de clusters óptimo para cada problema dado:

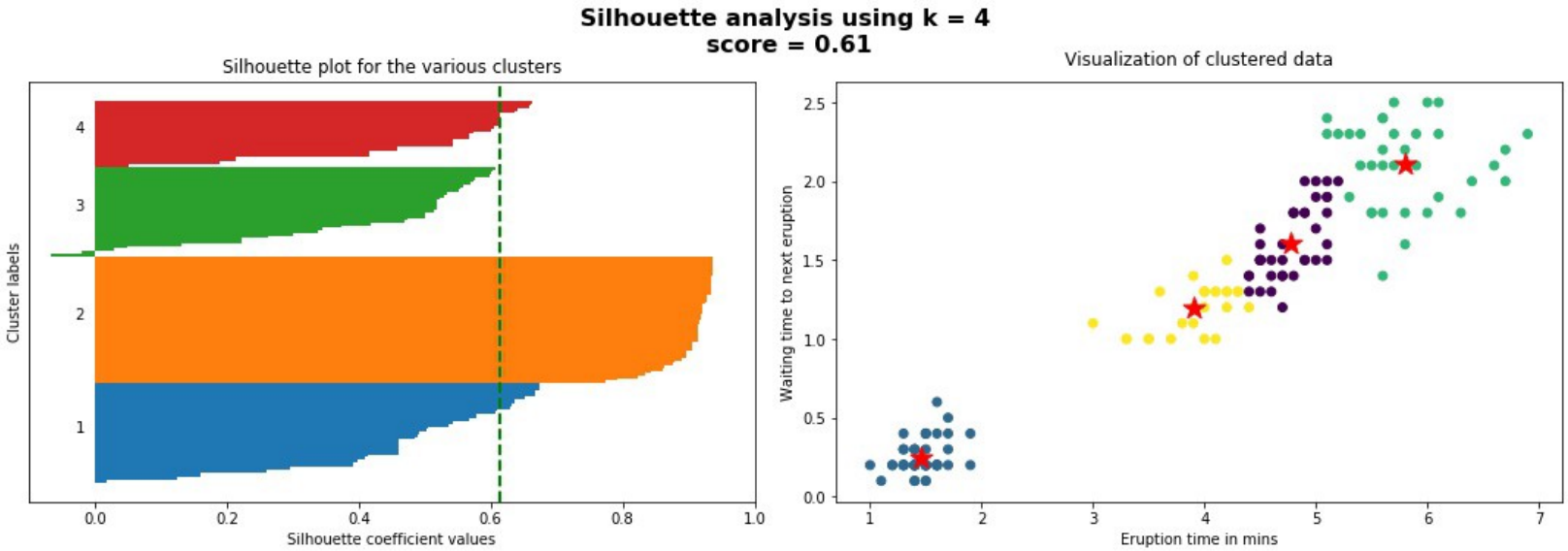
Ejemplo I


Silhouette analysis using k = 2
score = 0.77
Silhouette plot for the various clusters
Cluster labels
2
1
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette coefficient values
Visualization of clustered data
Waiting time to next eruption
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
Eruption time in mins

Ejemplo II


Silhouette analysis using k = 3
score = 0.66
Silhouette plot for the various clusters
Cluster labels
3
2
1
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette coefficient values
Visualization of clustered data
Waiting time to next eruption
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
Eruption time in mins

Ejemplo III


Silhouette analysis using k = 4
score = 0.61
Silhouette plot for the various clusters
Cluster labels
4
3
2
1
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette coefficient values
Visualization of clustered data
Waiting time to next eruption
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
Eruption time in mins

Fin Técnicas de Reagrupamiento