Minería de Datos

Clustering

No dispongo de clases por lo que voy a agrupar los datos para inferir como se distribuyen los ejemplos del dataset.

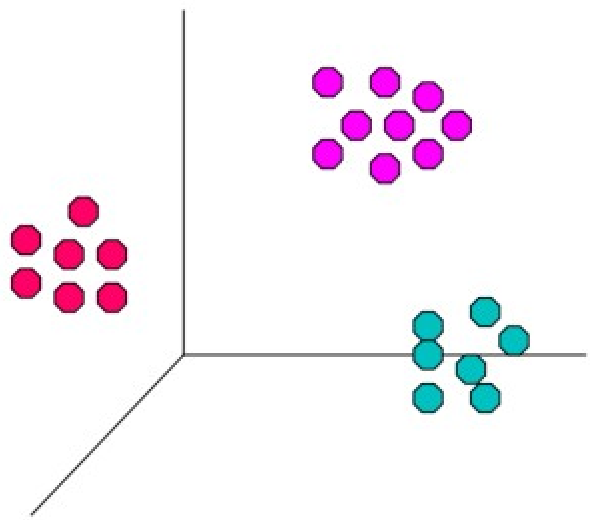
Clustering

No dispongo de clases por lo que voy a agrupar los datos para inferir como se distribuyen los ejemplos del dataset.

Clustering basado en distancias:

Minimizar distancias intracluster

Maximizar distancias intercluster



Clustering

Matriz de distancias entre objetos

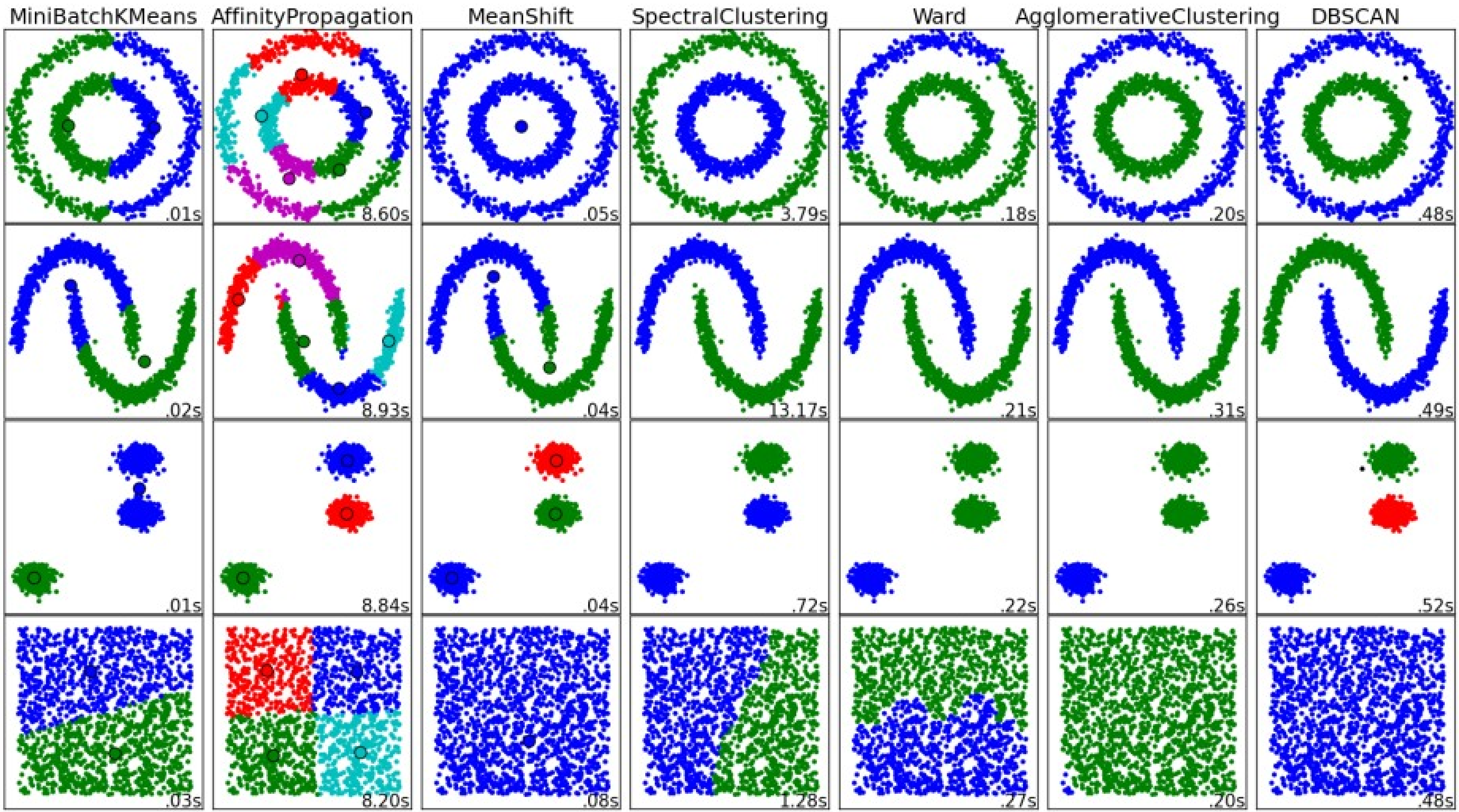
$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Clustering

Matriz de vectores de objetos

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Clustering


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
.01s
8.60s
.05s
3.79s
.18s
.20s
.48s
.02s
8.93s
.04s
13.17s
.21s
.31s
.49s
.01s
8.84s
.04s
.72s
.22s
.26s
.52s
.03s
8.20s
.08s
1.28s
.27s
.20s
.48s

Clustering

Algoritmo	Parámetro	Escalabilidad	Caso de uso	Geometría
K-Means	Número de clusters	Escalable Mejora con modificación MiniBatch	Propósito general flat clustering K no muy grande	Distancia entre objetos
Affinity Propagation	Coeficiente de damping	No escalable	Non-flat clustering K grande	Grafo de distancias
Mean-shift	Ancho de banda	No escalable	Non-flat clustering K grande	Distancia entre objetos
Spectral clustering	Número de clusters	Escalabilidad media	Non-flat clustering K no muy grande	Grafo de distancias
Ward	Número de clusters	Escalable	K grande	Distancias entre objetos
Clustering aglomerativo	Número de clusters	Escalable	Distancias no Euclideanas K grande	Distancias entre objetos
DBSCAN	Tamaño del vecindario	Escalable	Clusters de tamaños distintos	Grafo de vecinos más cercanos
Mezcla de Gaussianas	Muchos	No escalable	Flat clustering Estimación de densidad	Distancias Mahalanobis a centroides

Clustering

Algoritmo	Parámetro	Escalabilidad	Caso de uso	Geometría
K-Means	Número de clusters	Escalable Mejora con modificación MiniBatch	Propósito general flat clustering K no muy grande	Distancia entre objetos
Affinity Propagation	Coeficiente de damping	No escalable	Non-flat clustering K grande	Grafo de distancias
Mean-shift	Ancho de banda	No escalable	Non-flat clustering K grande	Distancia entre objetos
Spectral clustering	Número de clusters	Escalabilidad media	Non-flat clustering K no muy grande	Grafo de distancias
Ward	Número de clusters	Escalable	K grande	Distancias entre objetos
Clustering aglomerativo	Número de clusters	Escalable	Distancias no Euclideanas K grande	Distancias entre objetos
DBSCAN	Tamaño del vecindario	Escalable	Clusters de tamaños distintos	Grafo de vecinos más cercanos
Mezcla de Gaussianas	Muchos	No escalable	Flat clustering Estimación de densidad	Distancias Mahalanobis a centroides

Clustering

Algoritmo	Parámetro	Escalabilidad	Caso de uso	Geometría
K-Means	Número de clusters	Escalable Mejora con modificación MiniBatch	Propósito general flat clustering K no muy grande	Distancia entre objetos
Affinity Propagation	Coeficiente de damping	No escalable	Non-flat clustering K grande	Grafo de distancias
Mean-shift	Ancho de banda	No escalable	Non-flat clustering K grande	Distancia entre objetos
Spectral clustering	Número de clusters	Escalabilidad media	Non-flat clustering K no muy grande	Grafo de distancias
Ward	Número de clusters	Escalable	K grande	Distancias entre objetos
Clustering aglomerativo	Número de clusters	Escalable	Distancias no Euclideanas K grande	Distancias entre objetos
DBSCAN	Tamaño del vecindario	Escalable	Clusters de tamaños distintos	Grafo de vecinos más cercanos
Mezcla de Gaussianas	Muchos	No escalable	Flat clustering Estimación de densidad	Distancias Mahalanobis a centroides

Clustering con k-means

- Cada cluster en K-means es definido por un centroide.
- Objetivo: optimizar alguna noción de distancia:
  1. Intra-cluster: (Minimizar) distancia entre objetos de un cluster a su centroide.
  2. Inter-cluster: (Maximizar) distancia entre objetos de clusters distintos.
- Centroide:

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x$$

  donde C_i denota un cluster.
- Idea del algoritmo:
  - Asignación inicial: `k` centroides al azar.
  - Reasignación: asignar cada objeto a su centroide más cercano (algoritmo avaro).
  - Recomputación: recalcular los centroides. □

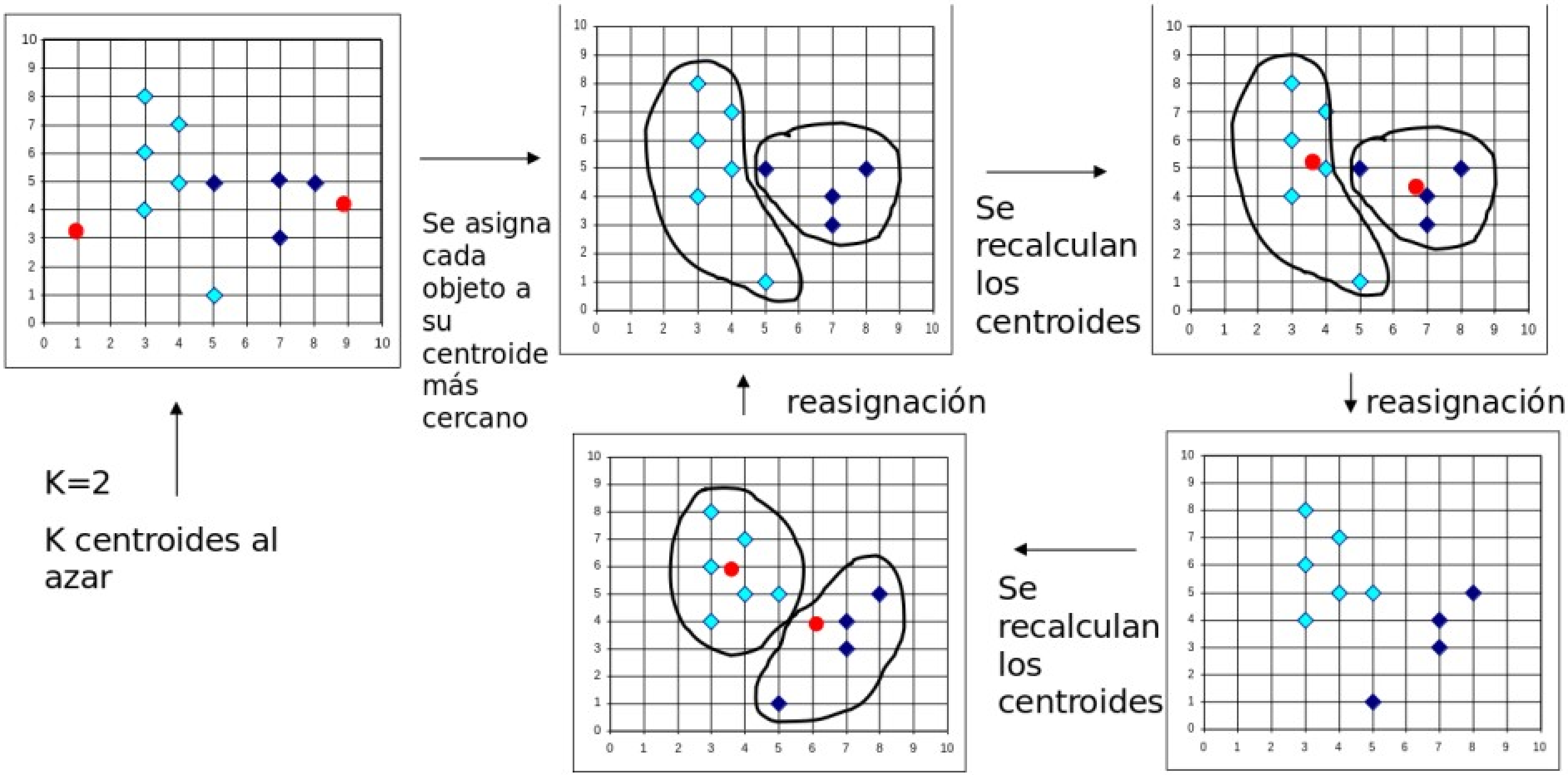
Clustering con k-means

$K\text{-MEANS}(\{\vec{x}_1, \ldots, \vec{x}_m\}, K)$

$(\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_K) \leftarrow \text{SELECTRANDOMSEEDS}(\{\vec{x}_1, \ldots, \vec{x}_m\}, K)$
for $k \leftarrow 1$ **to** K
do $\vec{c}_k \leftarrow \vec{s}_k$
while criterio convergencia no cumplido
do for $k \leftarrow 1$ **to** K
    do $C_k \leftarrow \{\}$
    for $i \leftarrow 1$ **to** m
    do $k \leftarrow Min_k \mid\mid \vec{c}_k - \vec{x}_i \mid\mid$ *(encontrar el centroide mas cercano)*
        $C_k \leftarrow C_k \cup \{\vec{x}_i\}$ *(agregar al cluster)*
    for $k \leftarrow 1$ **to** K
    do $\vec{c}_k \leftarrow \frac{1}{m_k} \sum_{\vec{x} \in C_k} \vec{x}$ *(recomputacion de centroides)*
return $\{C_1, \ldots, C_K\}$

Clustering con k-means

Ejemplo



Clustering con k-means

Hechos importantes:

- K-means converge. (McQueen, 67)
- Criterios de parada
  1. Iteraciones: (Máximo) número de iteraciones.
  2. Error tolerado: (Optimizar) alguna noción de distancia entre objetos.
- Complejidad:
  1. K-means es $\mathtt{NP-hard}$ en cualquier espacio d-dimensional con distancia Euclideana o coseno.
  2. K-means es $\mathtt{NP-hard}$ para cualquier valor de k.

SSE: Suma de errores al ²

Clustering con k-means

k-means minimiza el SSE: $$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2$$

SSE: Suma de errores al ²

Clustering con k-means

k-means minimiza el SSE: $$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2$$

$$\begin{aligned}
\frac{\partial}{\partial c_k}\mathrm{SSE} &= \frac{\partial}{\partial c_k}\sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2 \\
&= \sum_{i=1}^{K} \sum_{x \in C_i} \frac{\partial}{\partial c_k}(c_i - x)^2 \\
&= \sum_{x \in C_k} 2 * (c_k - x_k) = 0
\end{aligned}$$

SSE: Suma de errores al ²

Clustering con k-means

k-means minimiza el SSE: $$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2$$

$$\begin{aligned}
\frac{\partial}{\partial c_k}\mathrm{SSE} &= \frac{\partial}{\partial c_k}\sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2 \\
&= \sum_{i=1}^{K} \sum_{x \in C_i} \frac{\partial}{\partial c_k}(c_i - x)^2 \\
&= \sum_{x \in C_k} 2 * (c_k - x_k) = 0
\end{aligned}$$

$$\sum_{x \in C_k} \cancel{2} * (c_k - x_k) = 0 \Rightarrow m_k c_k = \sum_{x \in C_k} x_k \Rightarrow c_k = \frac{1}{m_k} \sum_{x \in C_k} x_k$$

elementos en el clúster

SAE: Suma de errores absolutos

Clustering con k-means

Variante: $$\mathrm{SAE} = \sum_{i=1}^{K} \sum_{x \in C_i} dist_{\mathrm{L}_1}(c_i, x)$$

SAE: Suma de errores absolutos

Clustering con k-means

Variante: $$\mathrm{SAE} = \sum_{i=1}^{K} \sum_{x \in C_i} dist_{\mathrm{L}_1}(c_i, x)$$

$$\begin{aligned}
\frac{\partial}{\partial c_k}\mathrm{SAE} &= \frac{\partial}{\partial c_k}\sum_{i=1}^{K}\sum_{x \in C_i} |c_i - x| \\
&= \sum_{i=1}^{K}\sum_{x \in C_i} \frac{\partial}{\partial c_k}|c_i - x| \\
&= \sum_{x \in C_k} \frac{\partial}{\partial c_k}|c_k - x| = 0
\end{aligned}$$

SAE: Suma de errores absolutos

Clustering con k-means

Variante: $$\text{SAE} = \sum_{i=1}^{K} \sum_{x \in C_i} dist_{\text{L}_1}(c_i, x)$$

$$\begin{aligned}
\frac{\partial}{\partial c_k}\text{SAE} &= \frac{\partial}{\partial c_k} \sum_{i=1}^{K} \sum_{x \in C_i} |c_i - x| \\
&= \sum_{i=1}^{K} \sum_{x \in C_i} \frac{\partial}{\partial c_k} |c_i - x| \\
&= \sum_{x \in C_k} \frac{\partial}{\partial c_k} |c_k - x| = 0
\end{aligned}$$

$$\sum_{x \in C_k} \frac{\partial}{\partial c_k} |c_k - x| = 0 \Rightarrow \sum_{x \in C_k} sign(x - c_k) = 0$$

$$c_k = median\{x \in C_k\}$$

SAE: Suma de errores absolutos

Clustering con k-means

Variante: $$\mathrm{SAE} = \sum_{i=1}^{K} \sum_{x \in C_i} dist_{\mathrm{L}_1}(c_i, x)$$

$$\begin{aligned}
\frac{\partial}{\partial c_k}\mathrm{SAE} &= \frac{\partial}{\partial c_k} \sum_{i=1}^{K} \sum_{x \in C_i} |c_i - x| \\
&= \sum_{i=1}^{K} \sum_{x \in C_i} \frac{\partial}{\partial c_k} |c_i - x| \\
&= \sum_{x \in C_k} \frac{\partial}{\partial c_k} |c_k - x| = 0
\end{aligned}$$

$$\sum_{x \in C_k} \frac{\partial}{\partial c_k} |c_k - x| = 0 \Rightarrow \sum_{x \in C_k} sign(x - c_k) = 0$$

$$c_k = median\{x \in C_k\}$$

k-medians

SAE: Suma de errores absolutos

Clustering con k-means

Variante: $$\mathrm{SAE} = \sum_{i=1}^{K} \sum_{x \in C_i} dist_{\mathrm{L}_1}(c_i, x)$$

$$\begin{aligned} \frac{\partial}{\partial c_k}\mathrm{SAE} &= \frac{\partial}{\partial c_k} \sum_{i=1}^{K} \sum_{x \in C_i} |c_i - x| \\ &= \sum_{i=1}^{K} \sum_{x \in C_i} \frac{\partial}{\partial c_k} |c_i - x| \\ &= \sum_{x \in C_k} \frac{\partial}{\partial c_k} |c_k - x| = 0 \end{aligned}$$

$$\sum_{x \in C_k} \frac{\partial}{\partial c_k} |c_k - x| = 0 \Rightarrow \sum_{x \in C_k} sign(x - c_k) = 0$$

$$c_k = median\{x \in C_k\}$$

k-medians

K-medoids: datapoints como centroides

SAE: Suma de errores absolutos

Clustering con k-means

Variante: $$\mathrm{SAE} = \sum_{i=1}^{K} \sum_{x \in C_i} dist_{\mathrm{L}_1}(c_i, x)$$

$$\begin{aligned}
\frac{\partial}{\partial c_k}\mathrm{SAE} &= \frac{\partial}{\partial c_k}\sum_{i=1}^{K}\sum_{x \in C_i} |c_i - x| \\
&= \sum_{i=1}^{K}\sum_{x \in C_i} \frac{\partial}{\partial c_k}|c_i - x| \\
&= \sum_{x \in C_k} \frac{\partial}{\partial c_k}|c_k - x| = 0
\end{aligned}$$

$$\sum_{x \in C_k} \frac{\partial}{\partial c_k}|c_k - x| = 0 \Rightarrow \sum_{x \in C_k} sign(x - c_k) = 0$$

$$c_k = median\{x \in C_k\}$$

k-medians

K-medoids: datapoints como centroides

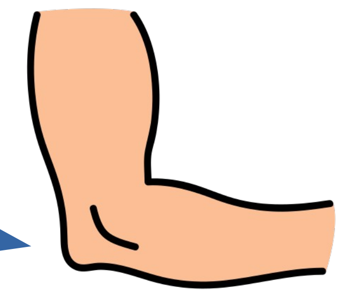
K-means++: grid tune for seeds

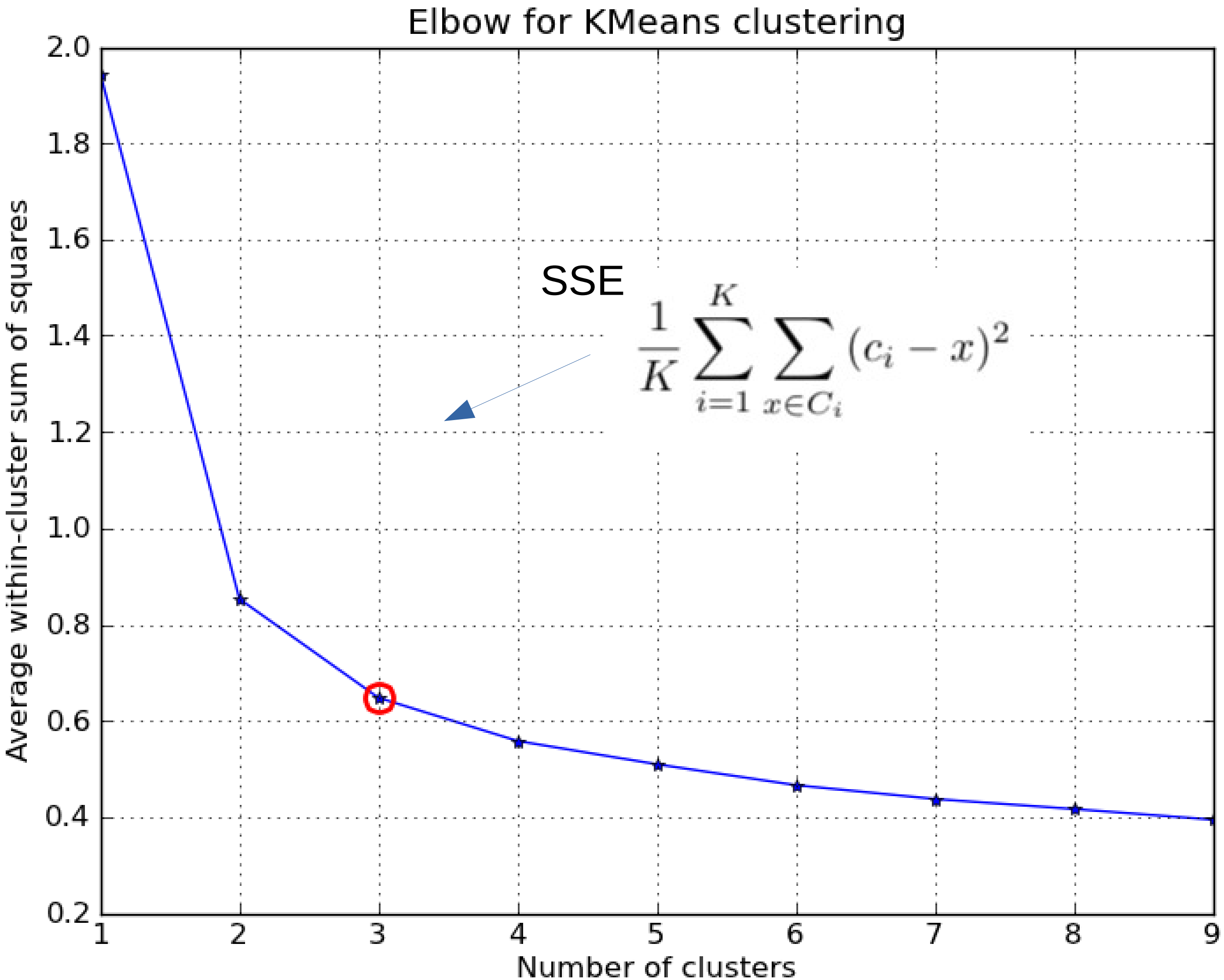
¿Cuántos prototipos usamos?

¿Cuántos prototipos usamos?

ELBOW (codo):

Variar k buscando el codo





SSE

$$\frac{1}{K}\sum_{i=1}^{K}\sum_{x\in C_i}(c_i - x)^2$$

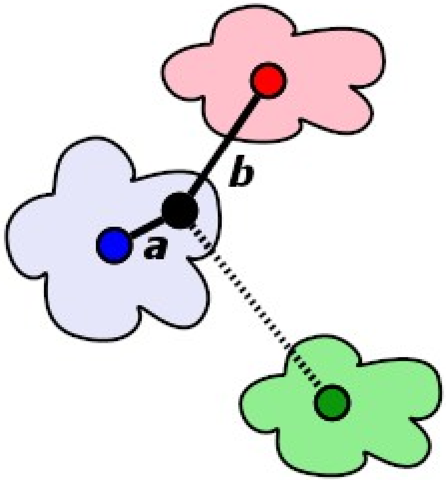
¿Cuántos prototipos usamos?

Silhouette:

Congruencia de x_i a C_i: $$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

¿Cuántos prototipos usamos?

Silhouette:

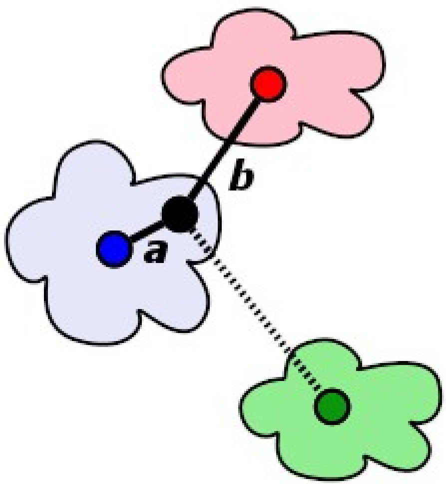


Congruencia de x_i a C_i: $$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

Congruencia de x_i a otros clusters: $$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

¿Cuántos prototipos usamos?

Silhouette:



Congruencia de x_i a C_i: $$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

Congruencia de x_i a otros clusters: $$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

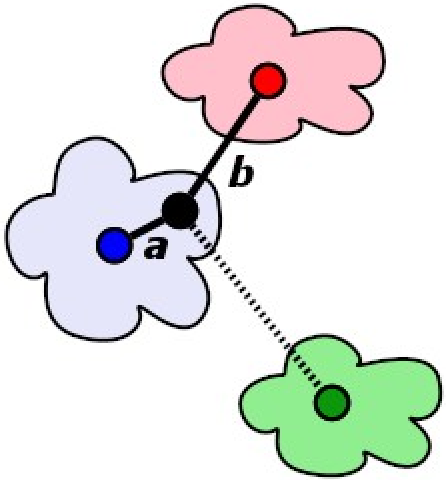
Silhouette Coef.: $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad \text{si} \quad |C_i| > 1,$$

$$s(i) = 0, \quad \text{si} \quad |C_i| = 1.$$

¿Cuántos prototipos usamos?

Silhouette:

Congruencia de x_i a C_i: $$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$



Congruencia de x_i a otros clusters: $$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

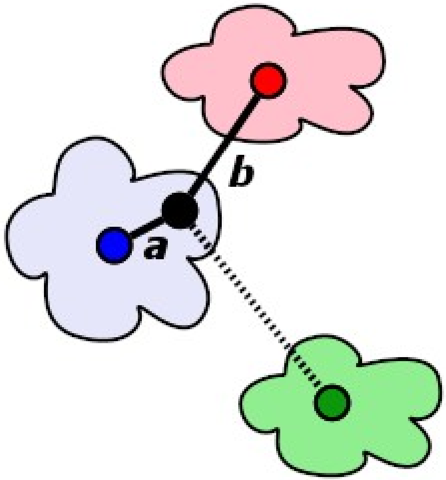
Silhouette Coef.: $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad \text{si} \quad |C_i| > 1,$$

$$s(i) = 0, \quad \text{si} \quad |C_i| = 1.$$

¿Intervalo?

¿Cuántos prototipos usamos?

Silhouette:



Congruencia de x_i a C_i: $$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

Congruencia de x_i a otros clusters: $$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

Silhouette Coef.: $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad \text{si} \quad |C_i| > 1,$$

$$s(i) = 0, \quad \text{si} \quad |C_i| = 1.$$

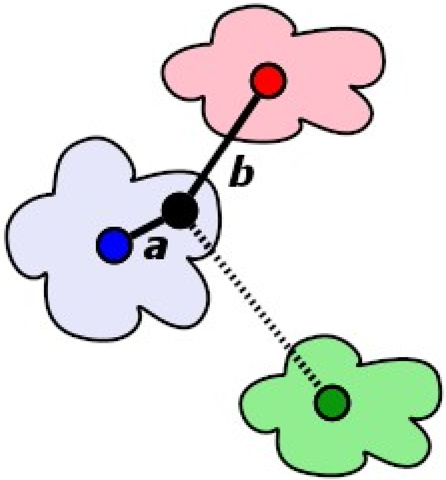
[-1, 1]

¿Cuántos prototipos usamos?

Silhouette:

Un valor alto indica poca congruencia

Congruencia de x_i a C_i: $$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$



Congruencia de x_i a otros clusters: $$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

Silhouette Coef.: $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad \text{si} \quad |C_i| > 1,$$

$$s(i) = 0, \quad \text{si} \quad |C_i| = 1.$$

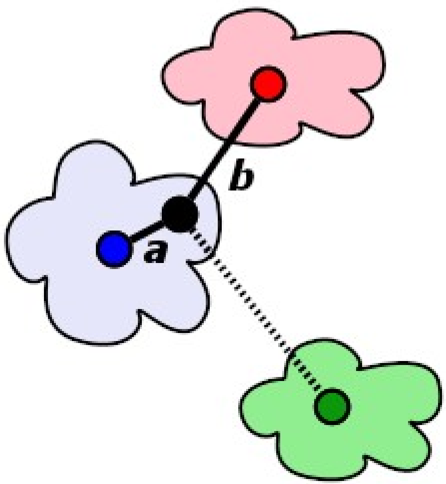
[-1, 1]

¿Cuántos prototipos usamos?

Silhouette:

Un valor alto indica poca congruencia

Congruencia de x_i a C_i: $$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$



Congruencia de x_i a otros clusters: $$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

Silhouette Coef.: $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad \text{si} \quad |C_i| > 1,$$

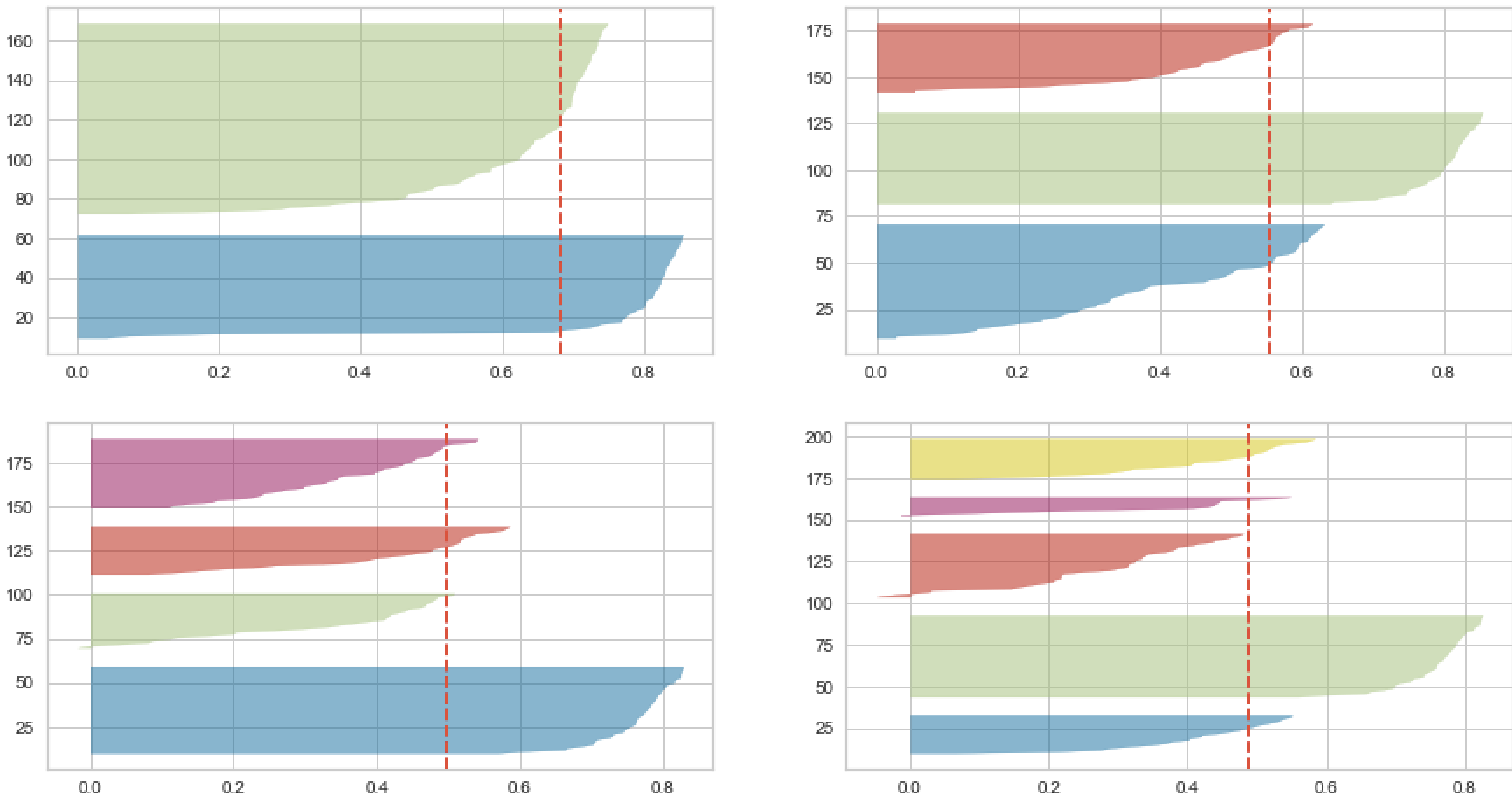
$$s(i) = 0, \quad \text{si} \quad |C_i| = 1.$$

Un valor alto indica alta congruencia

[-1, 1]

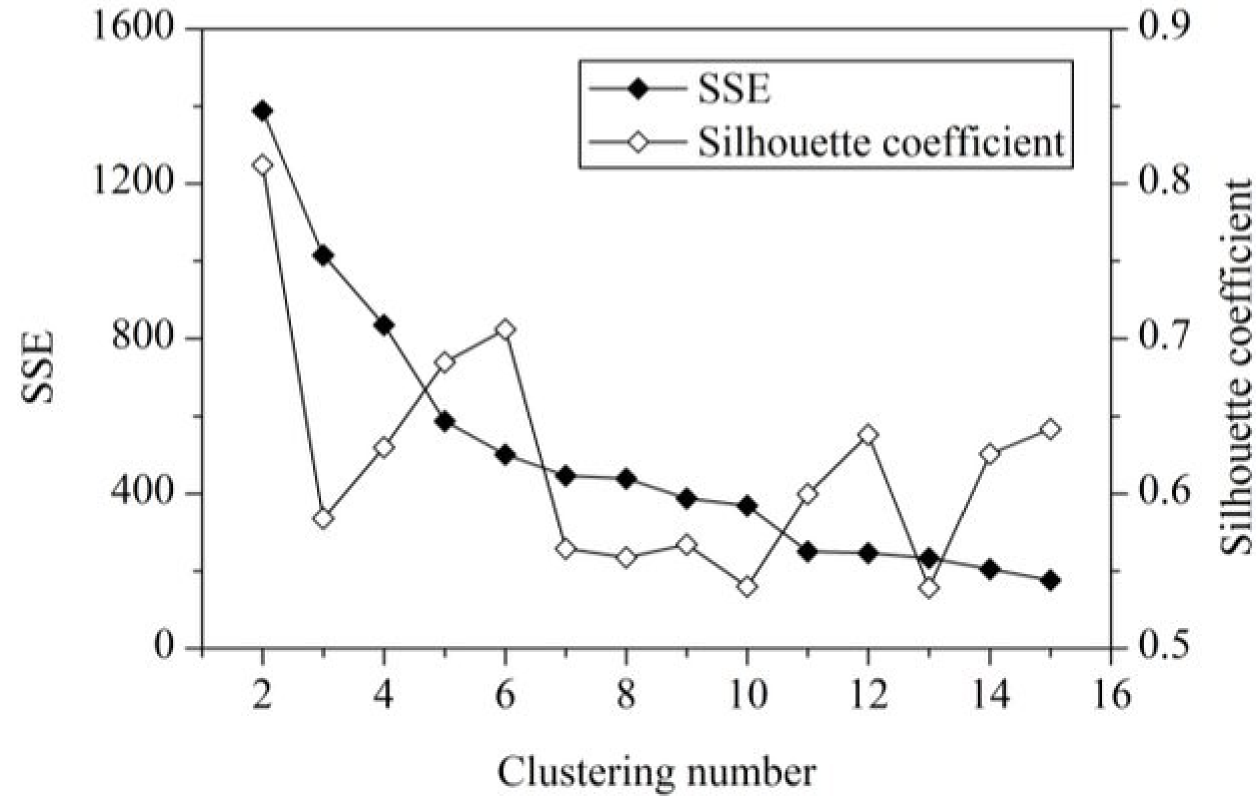
¿Cuántos prototipos usamos?

Silhouette promedio:



¿Cuántos prototipos usamos?

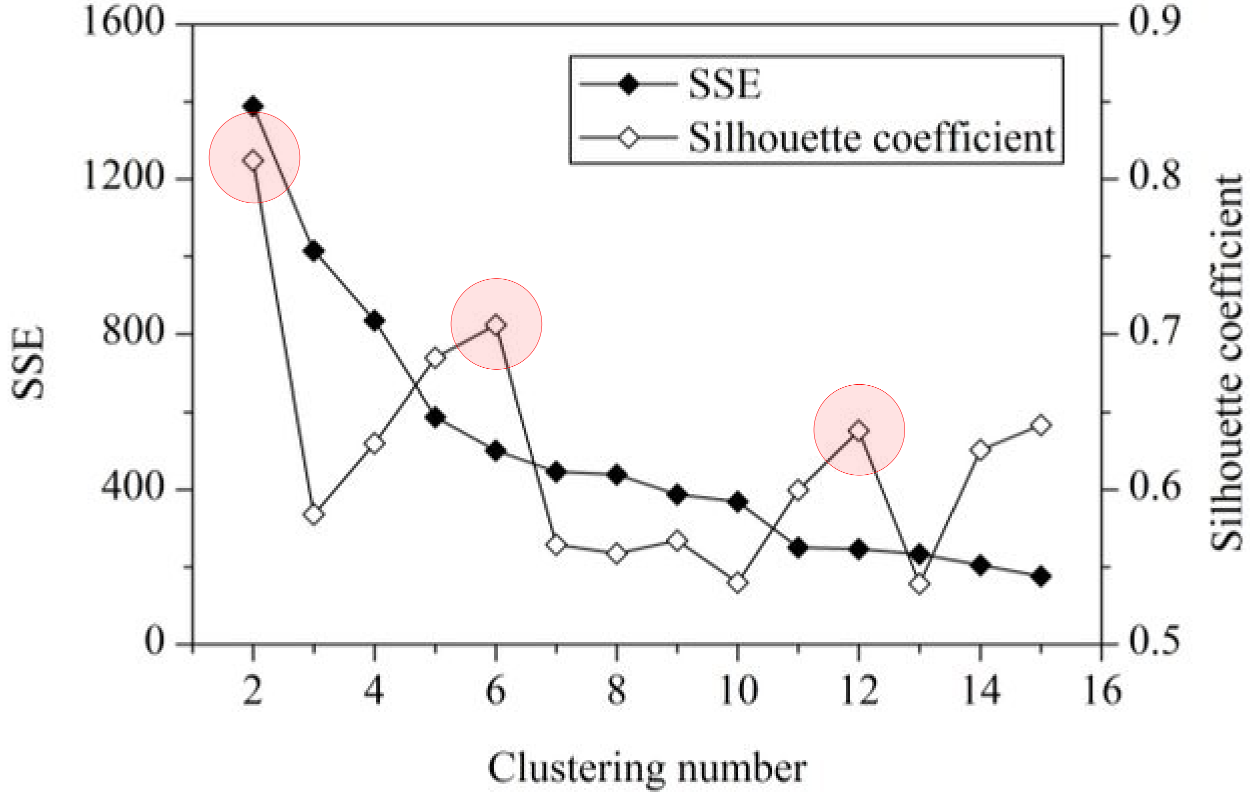
Silhouette v/s ELBOW:



¿Con cuál se quedan?

¿Cuántos prototipos usamos?

Silhouette v/s ELBOW:



¿Con cuál se quedan?