IIC 2433 Minería de Datos

https://github.com/marcelomendoza/IIC2433

- KMEANS -

Clustering permite entender como se agrupan los datos

Clustering con k-means

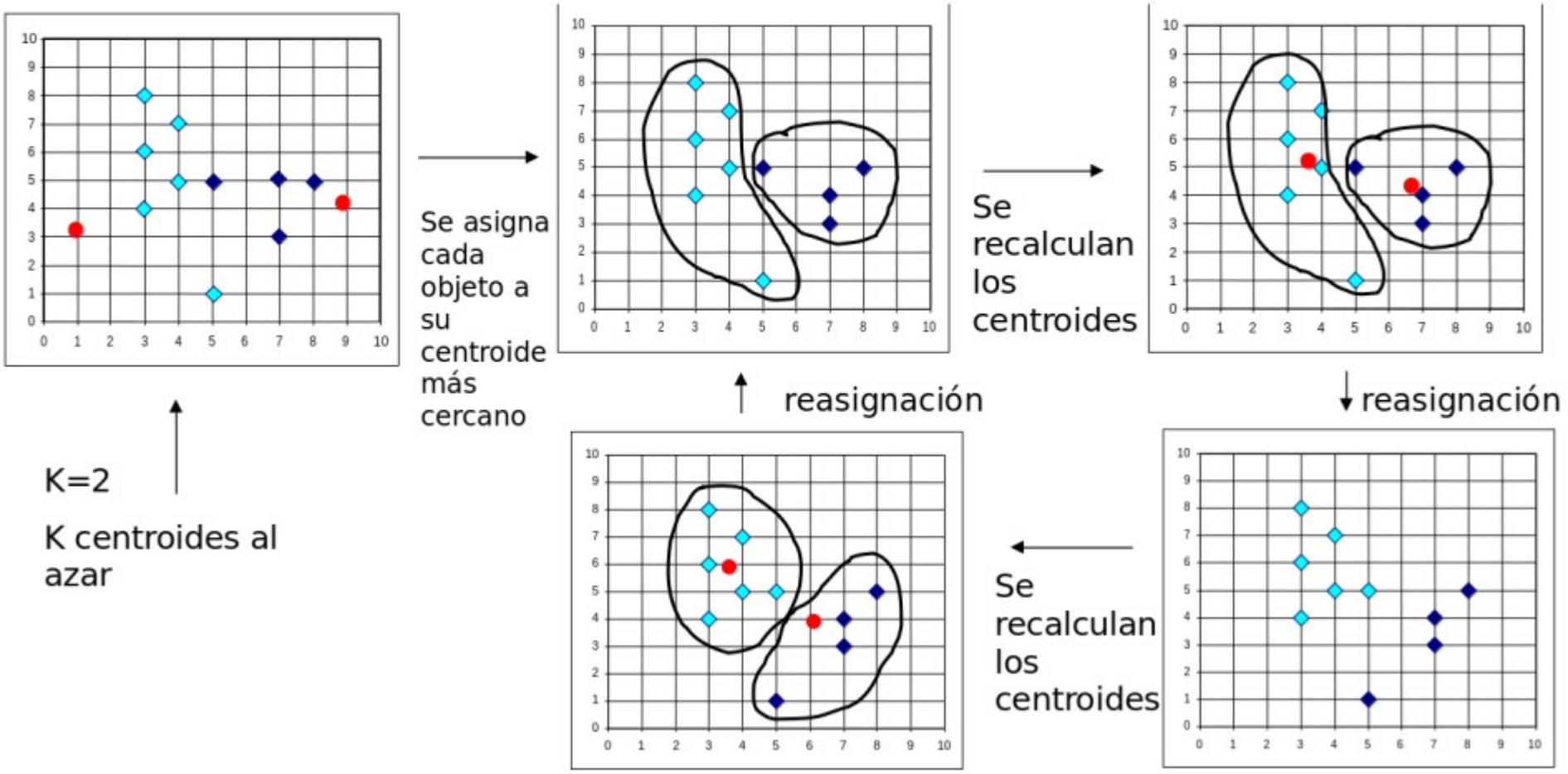
- Cada cluster en K-means es definido por un centroide.
- Objetivo: optimizar alguna noción de distancia:
  1. Intra-cluster: (Minimizar) distancia entre objetos de un cluster a su centroide.
  2. Inter-cluster: (Maximizar) distancia entre objetos de clusters distintos.
- Centroide:

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

  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

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: implícitamente

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

SSE: Suma de errores al ²

Clustering con k-means

k-means minimiza el SSE: implícitamente

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

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

SSE: Suma de errores al ²

Clustering con k-means

k-means minimiza el SSE: implícitamente

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

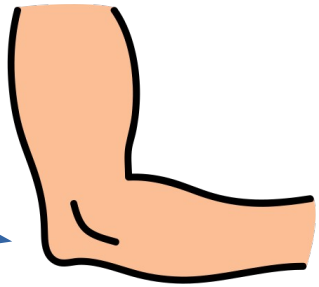
$$\begin{aligned} \frac{\partial}{\partial c_k}\mathrm{SSE} &= \frac{\partial}{\partial c_k} \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2 \\ &= \sum_{i=1}^{K} \sum_{x \in C_i} \frac{\partial}{\partial c_k} (c_i - x)^2 \\ &= \sum_{x \in C_k} 2 * (c_k - x_k) = 0 \end{aligned}$$

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

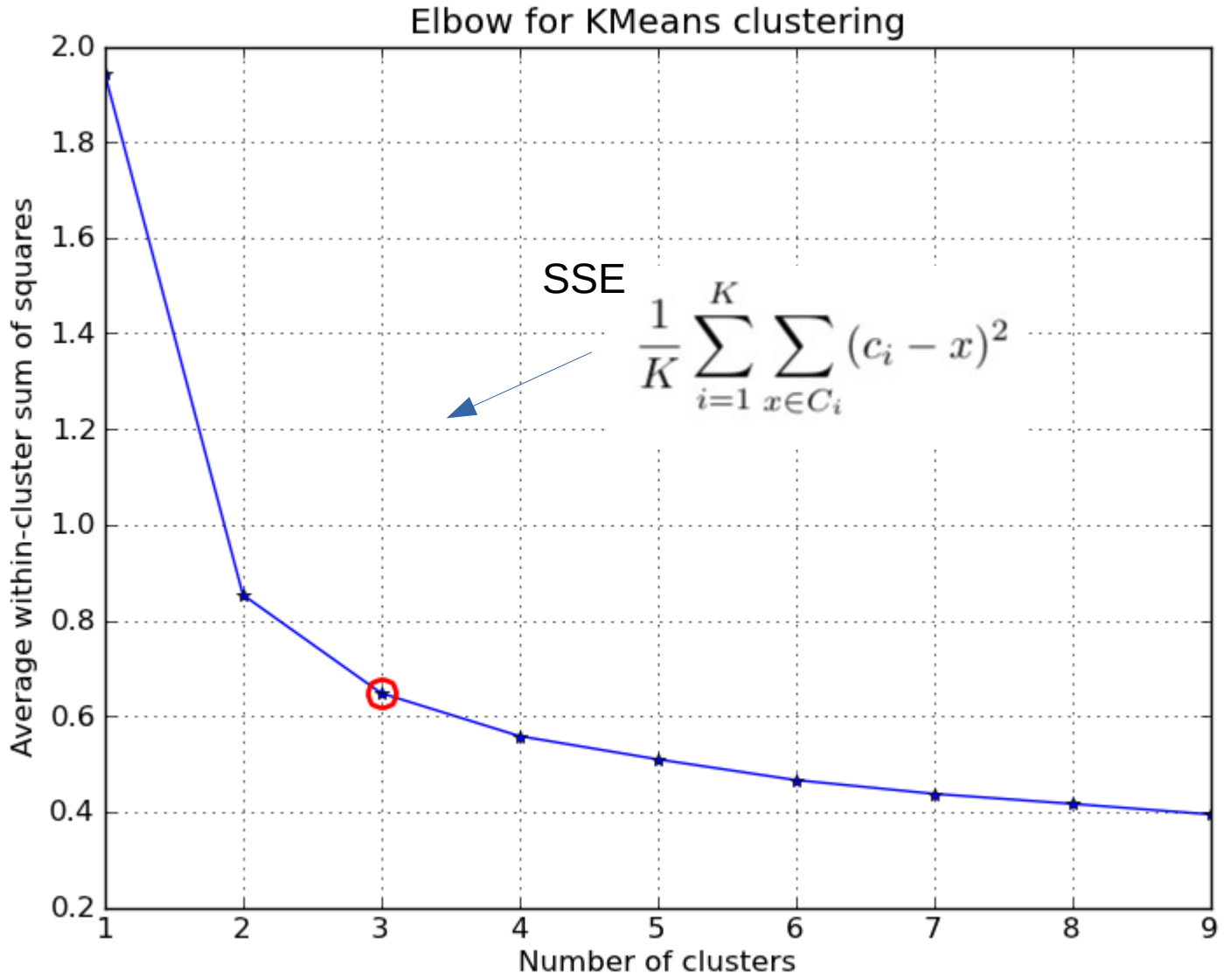
m_k: # elementos en el clúster

¿Cuántos prototipos usamos?

Variar k buscando el codo



ELBOW (codo):



SSE

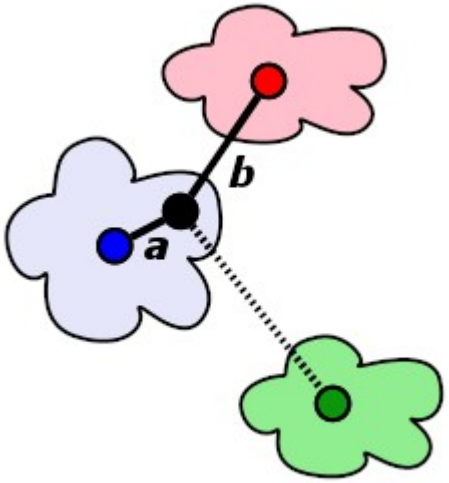
$$\frac{1}{K}\sum_{i=1}^{K}\sum_{x \in C_i}(c_i - x)^2$$

¿Cuántos prototipos usamos?

Silhouette:

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

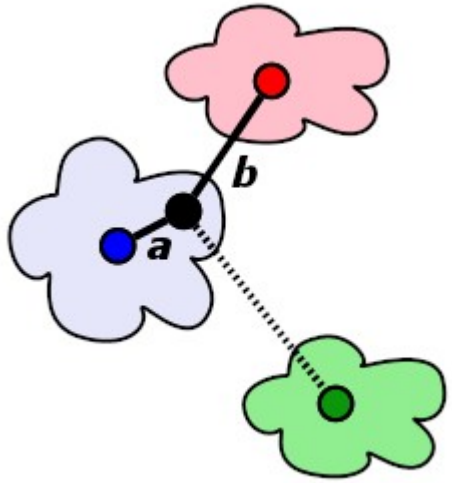
Congruencia de x_i a otros clusters: $$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$



¿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: $$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

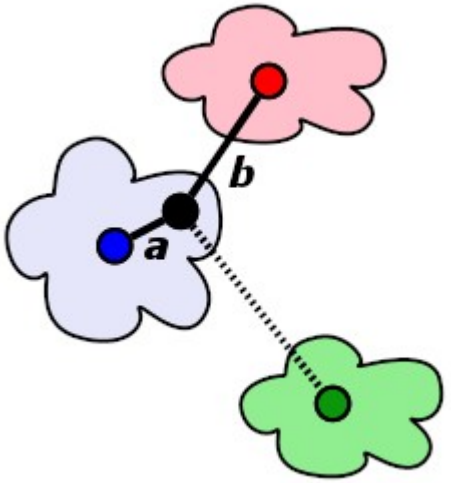
Silhouette Coef.: $$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad \text{si} \quad |C_i| > 1,$$

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

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

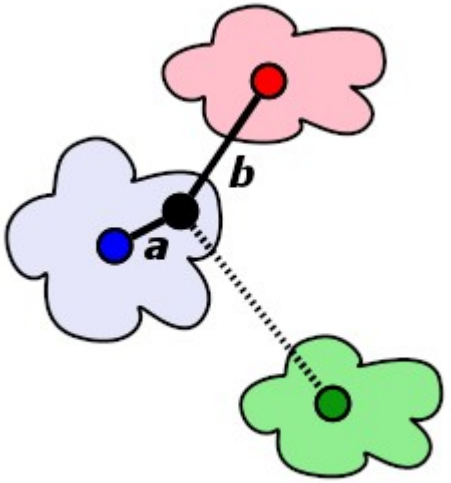
$$s(i) = 0, \quad \text{si} \quad |C_i| = 1.$$

¿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]

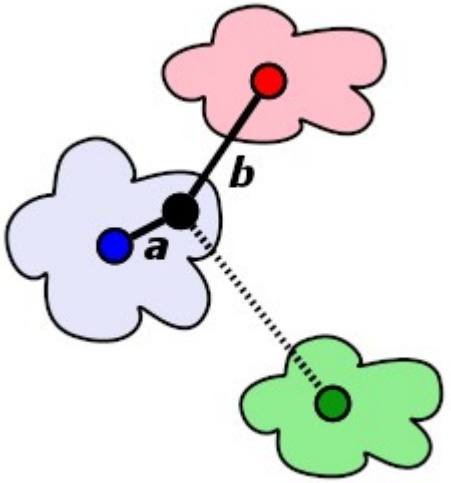
¿Cuántos prototipos usamos?

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]

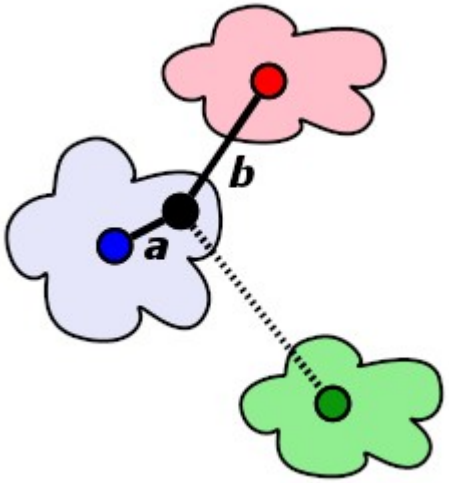
¿Cuántos prototipos usamos?

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.:

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

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

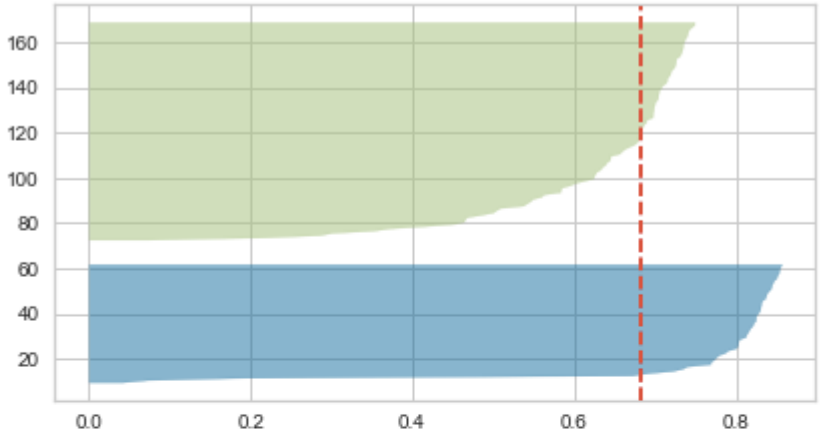
Un valor alto indica alta congruencia

[-1, 1]

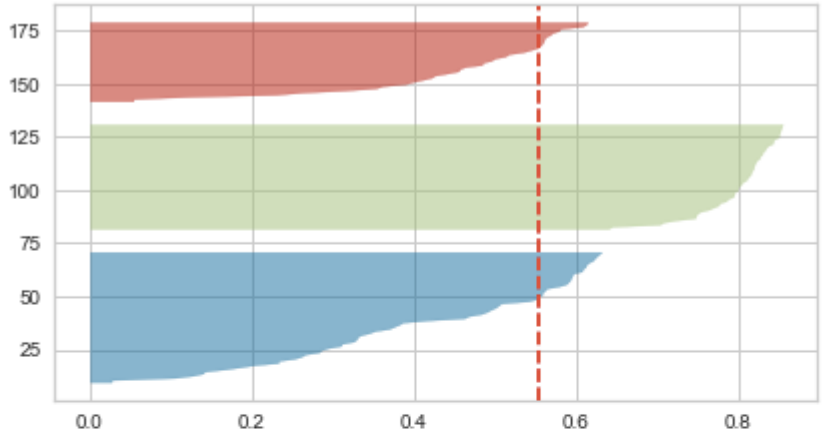
¿Cuántos prototipos usamos?

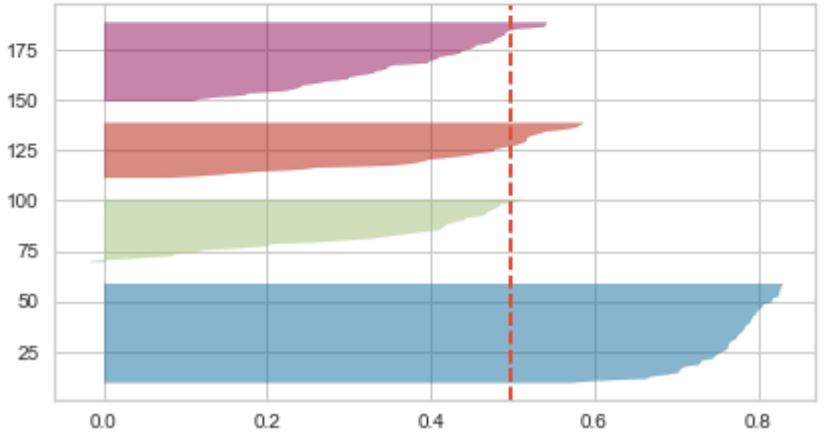
Silhouette promedio:

k=2

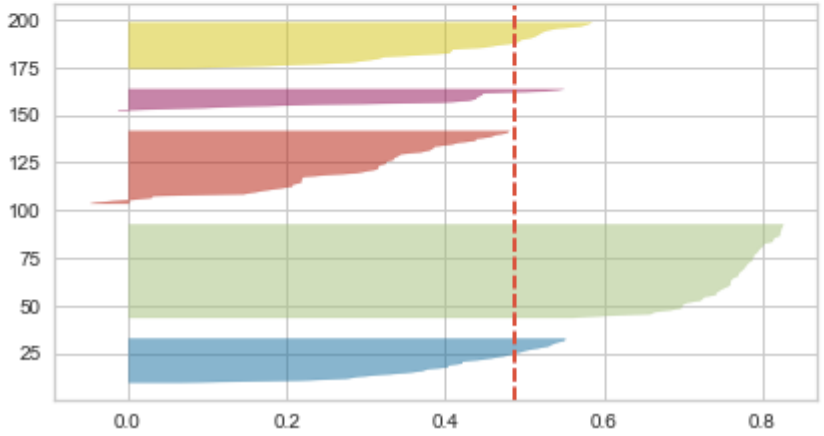


k=3





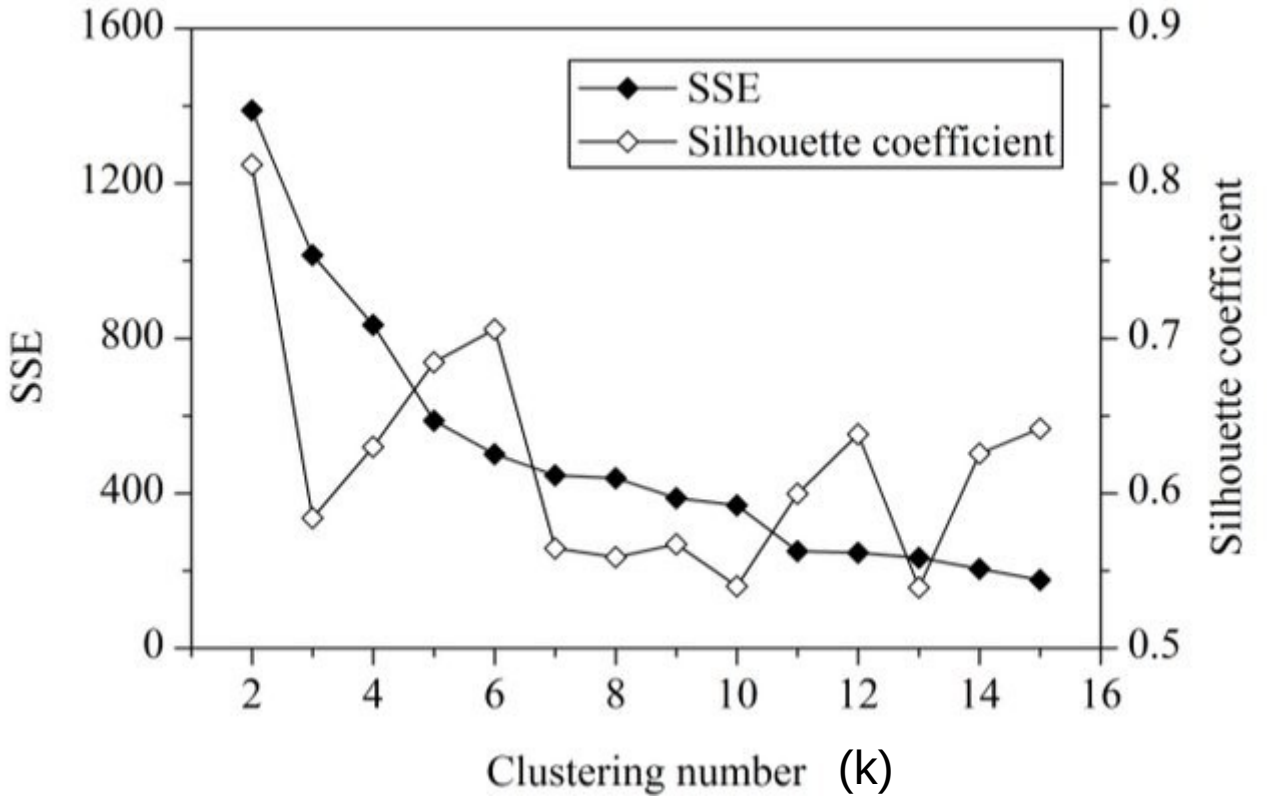
k=4



k=5

¿Cuántos prototipos usamos?

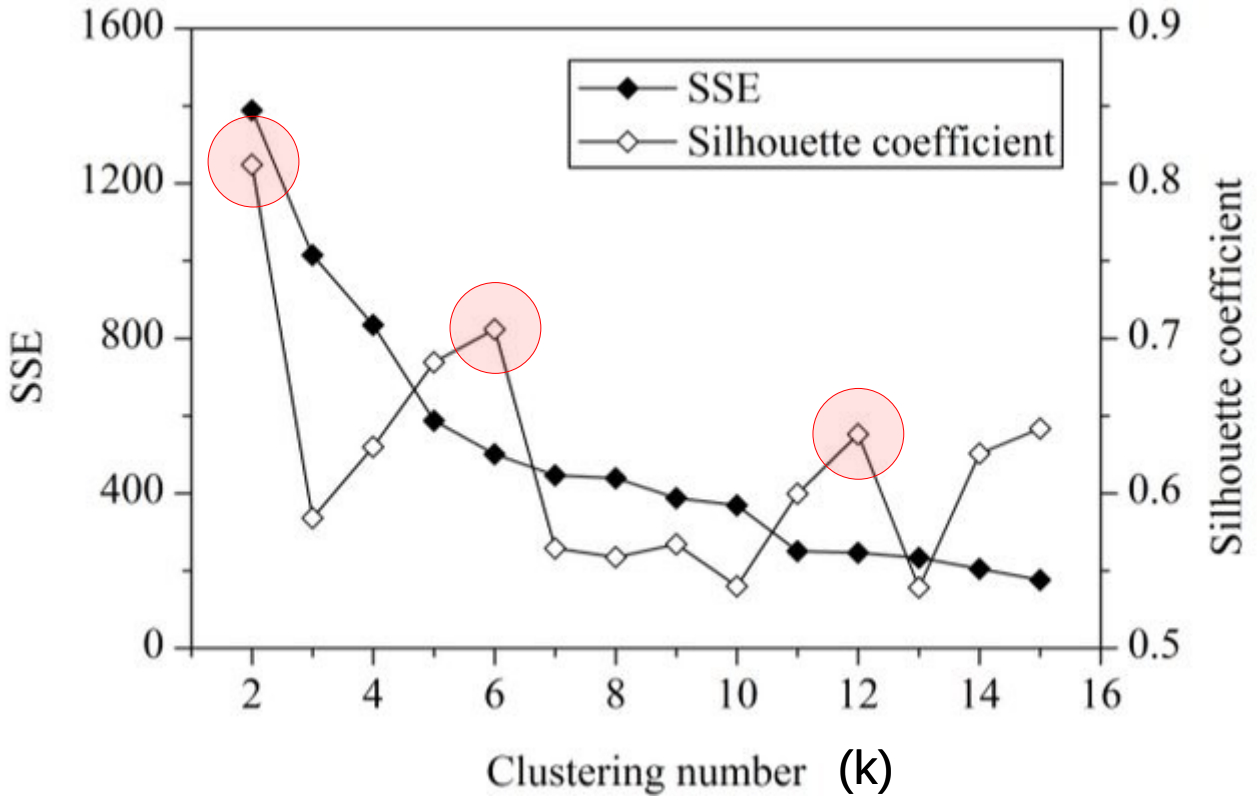
Silhouette v/s ELBOW:



¿Con cuál **k** se quedan?

¿Cuántos prototipos usamos?

Silhouette v/s ELBOW:



¿Con cuál **k** se quedan?