ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

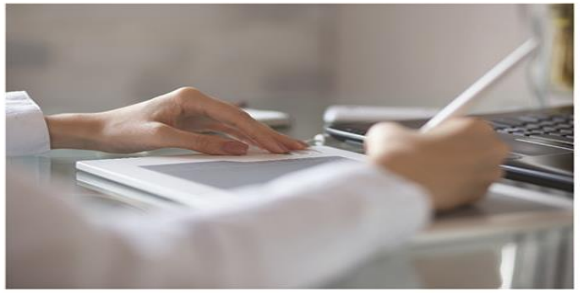
Diplomado en Big Data y Ciencias de Datos

Minería de Datos

Análisis de Clústeres

Educación Profesional - Escuela de Ingeniería UC

Sebastián Raveau




Análisis de clústeres

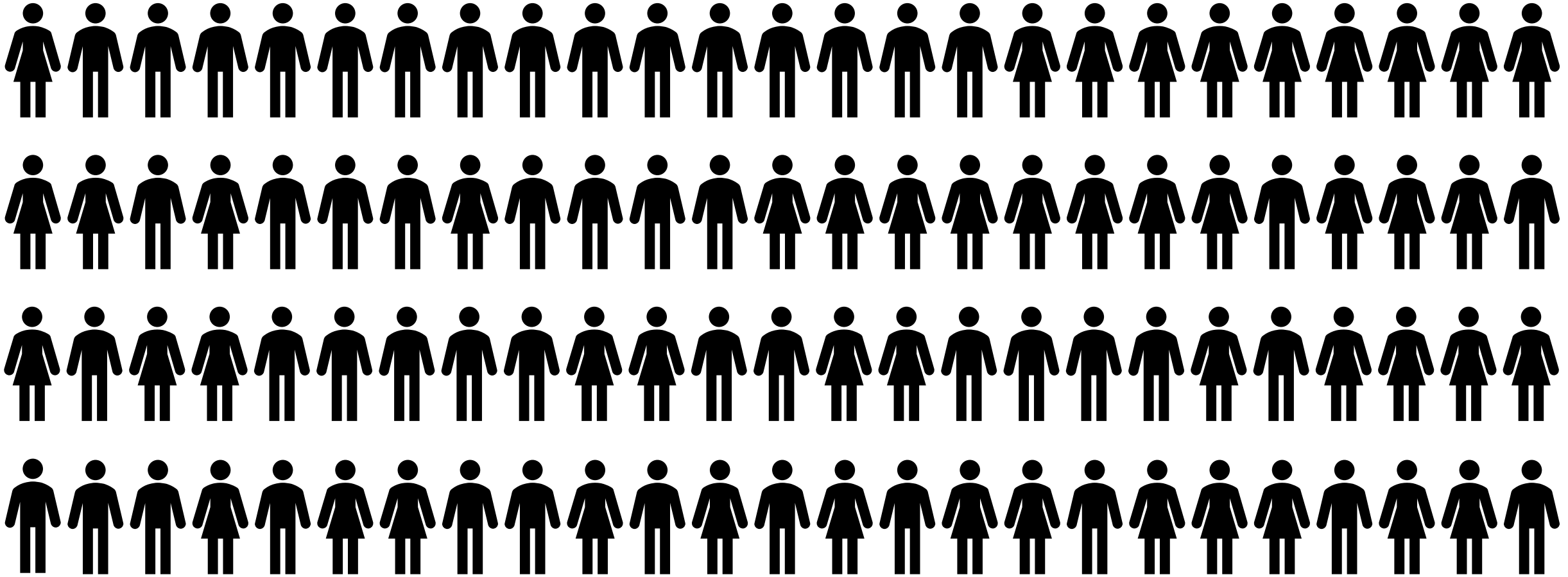
El análisis de clústeres corresponde a una técnica de aprendizaje no supervisado, ya que no tenemos una “clasificación real” contra la cual comparar el resultado

El objetivo es analizar y entender la estructura de los datos, agrupando “observaciones similares”

No es posible comparar los resultados de distintos métodos mediante éxito predictivo (i.e. no hay clasificaciones erróneas)

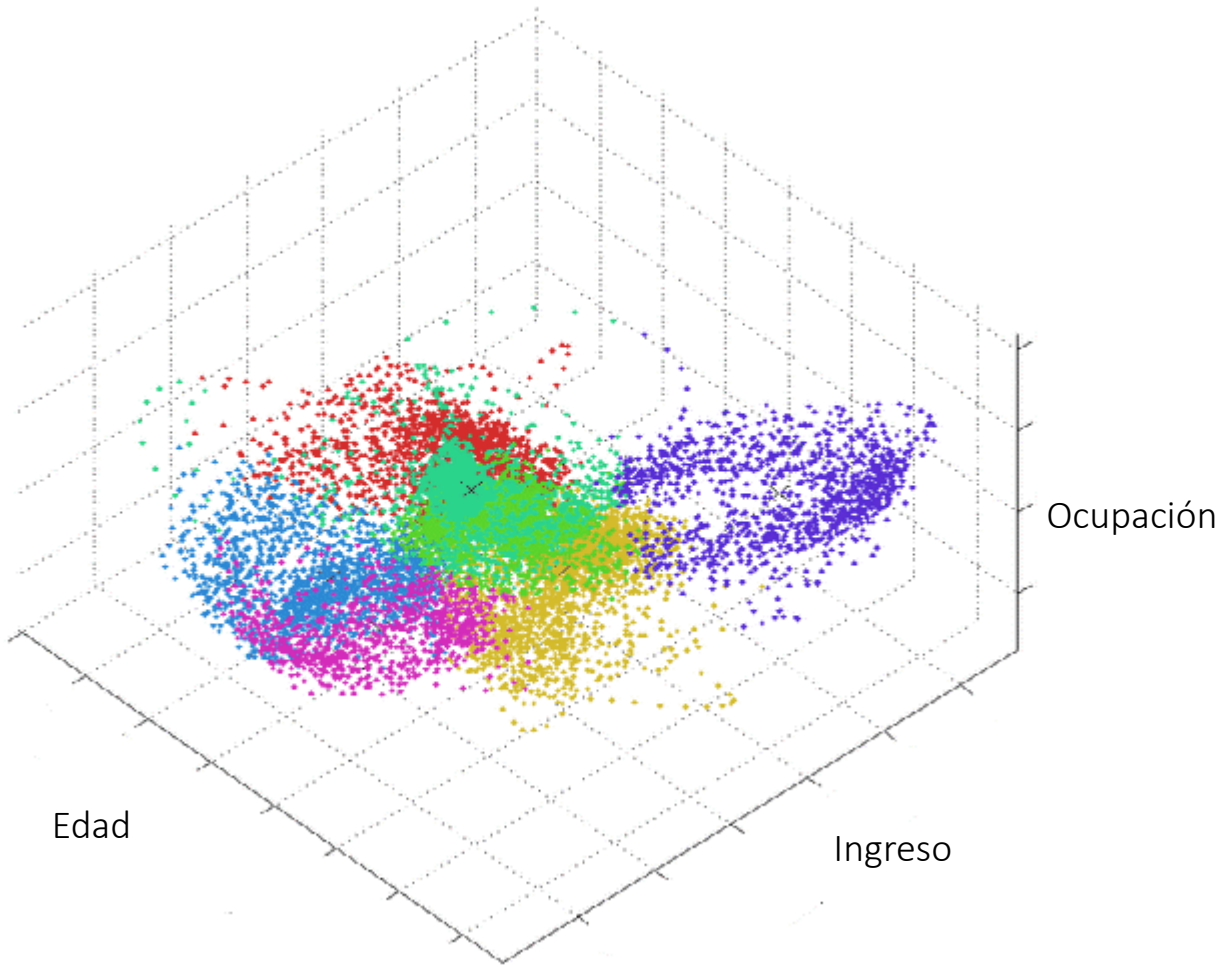
Análisis de clústeres

Supongamos que queremos identificar perfiles de clientes



Análisis de clústeres

Supongamos que queremos identificar perfiles de clientes



Arriesgados

Hípsters

Cautos

Aventureros

Despreocupados

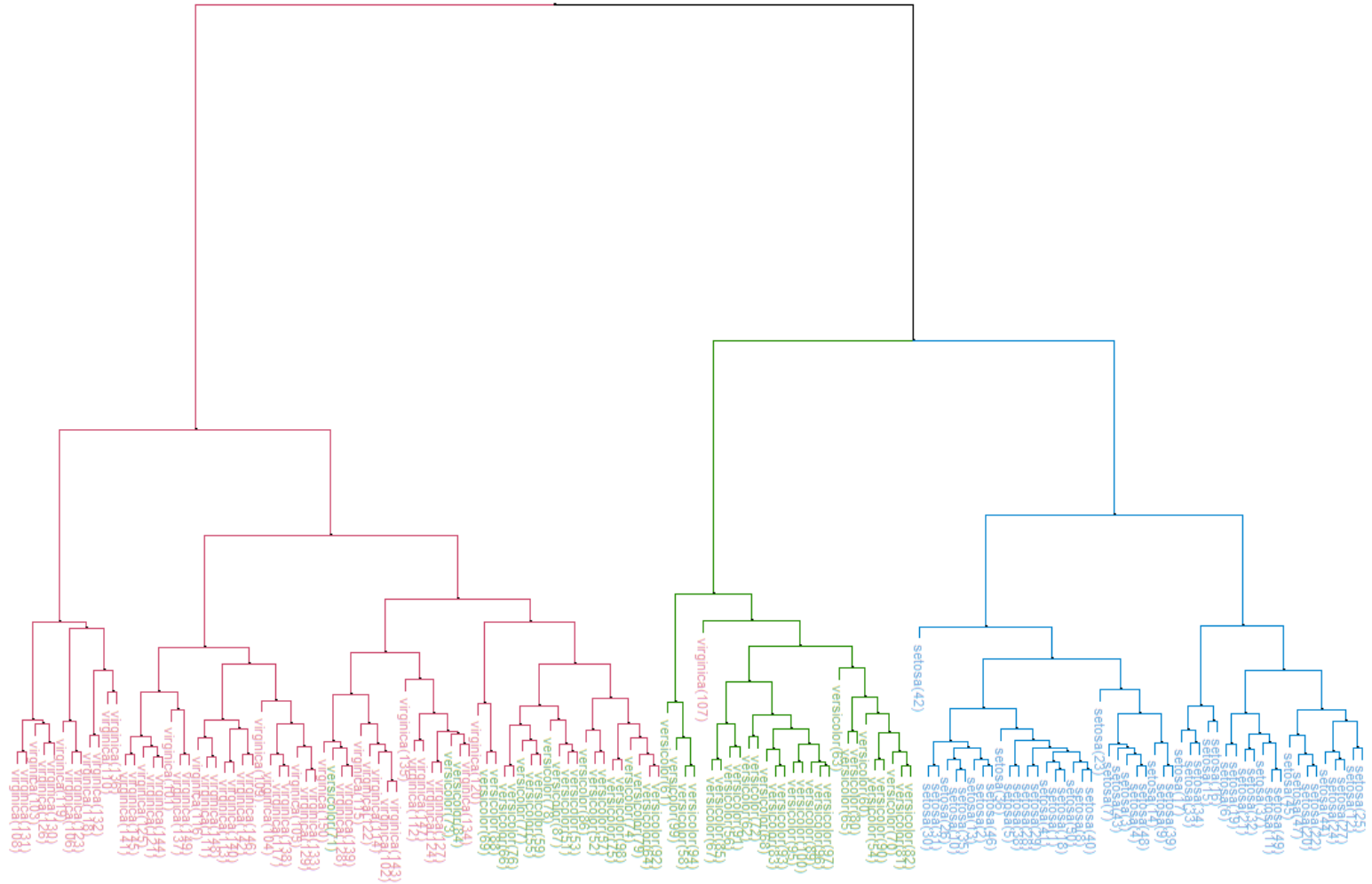
Racionales

Análisis de clústeres

Supongamos que queremos identificar perfiles de clientes

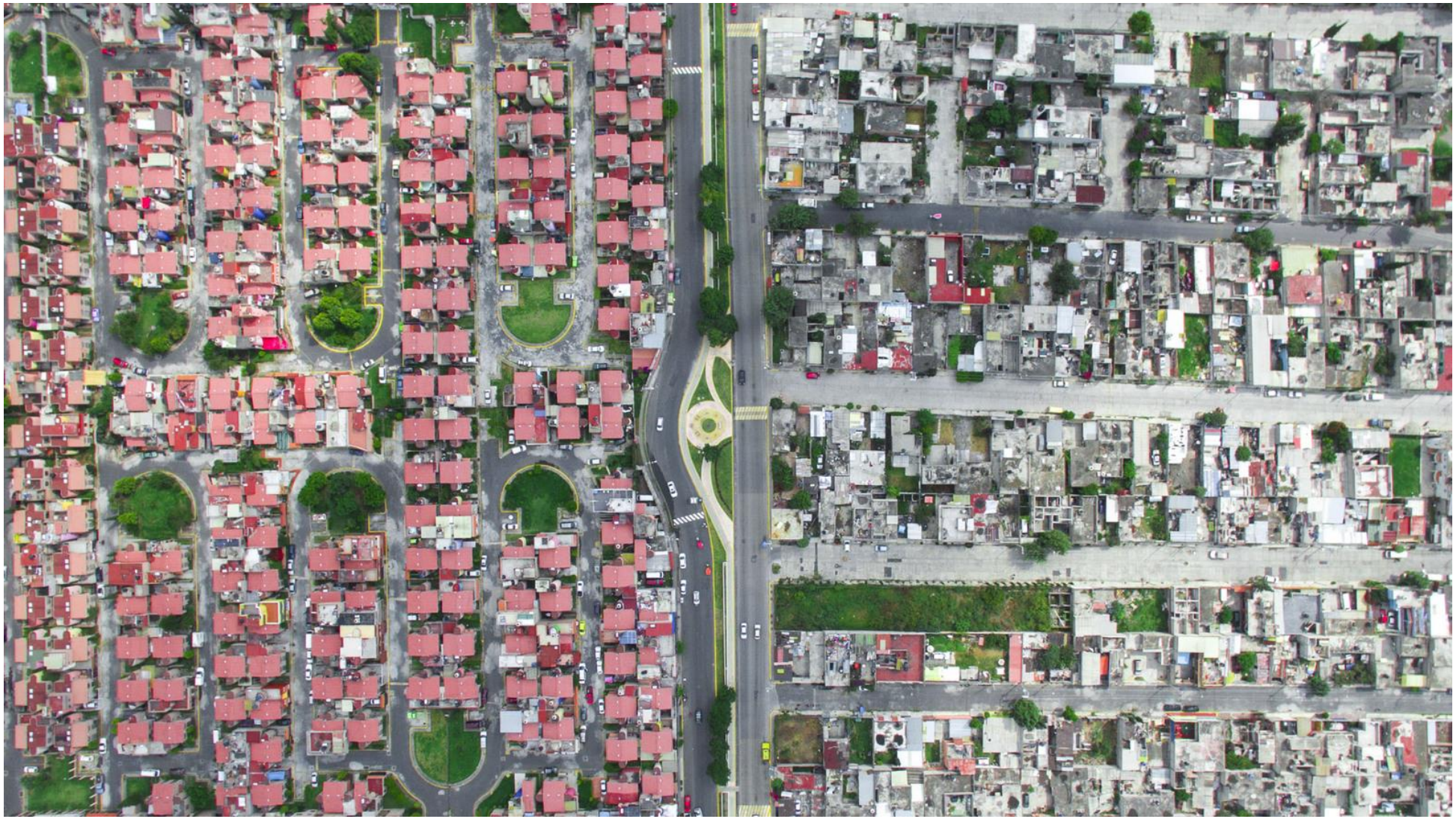
Algunas aplicaciones habituales

Clasificación de plantas y animales



Segmentación de clientes

Mercado inmobiliario y clasificación espacial



Detección de fraudes

Compresión de imágenes

K Means

K Means

Este algoritmo busca segmentar los datos en K clústeres excluyentes

Se buscan dos objetivos:

1. Que los datos que pertenecen a cada clúster sean similares

   (i.e. poca variabilidad intra-clúster)

2. Que los clústeres sean diferentes entre sí

   (i.e. alta variabilidad inter-clúster)

K Means

Dos observaciones son “similares” si se encuentran a poca distancia

Aspectos a considerar:

- Métrica de distancia (e.g. Euclidiana o Manhattan)
- Tratamiento de variables categóricas u ordinales
- Normalización de datos con distintas escalas

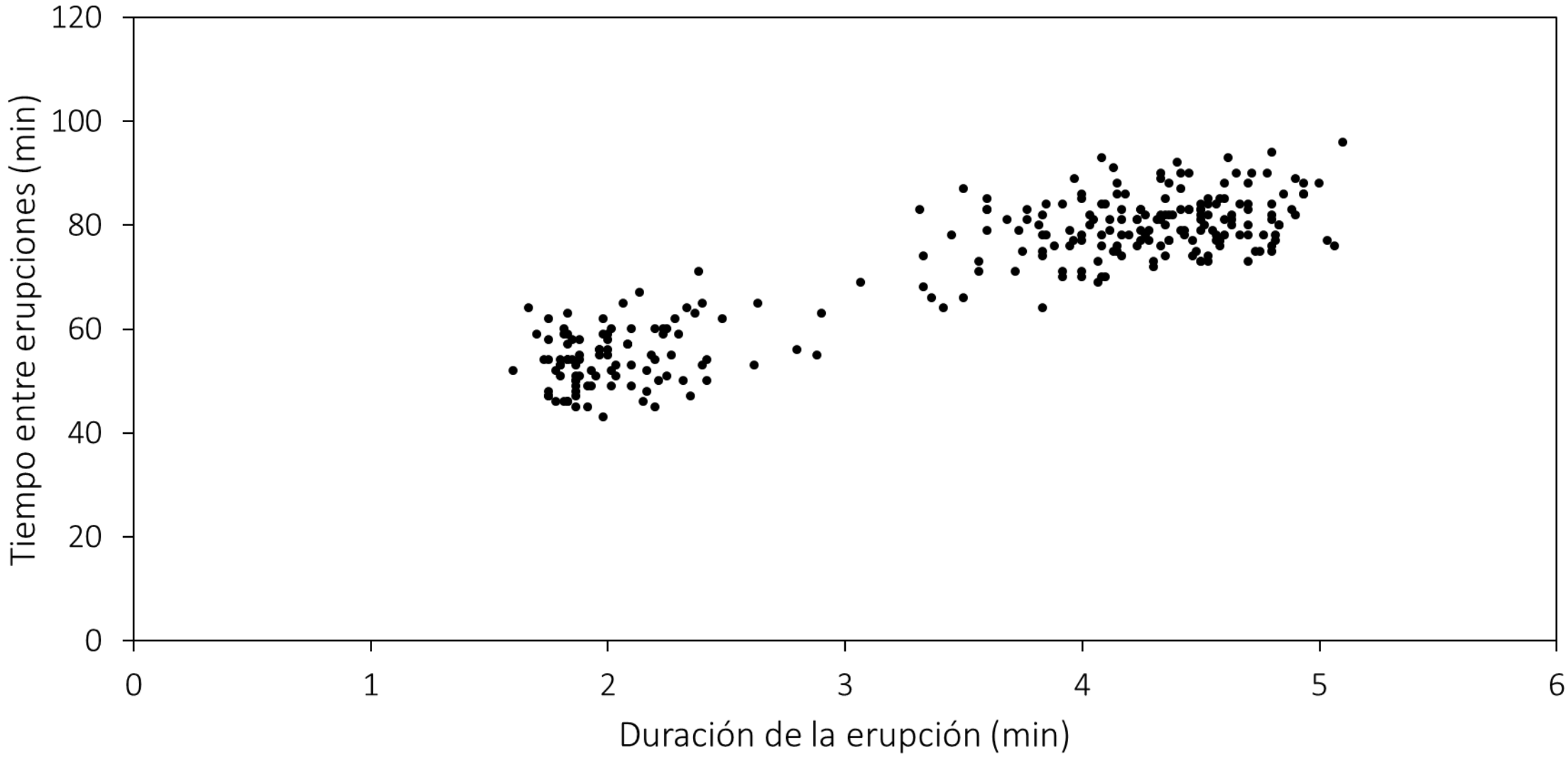
K Means

El algoritmo opera en forma iterativa:

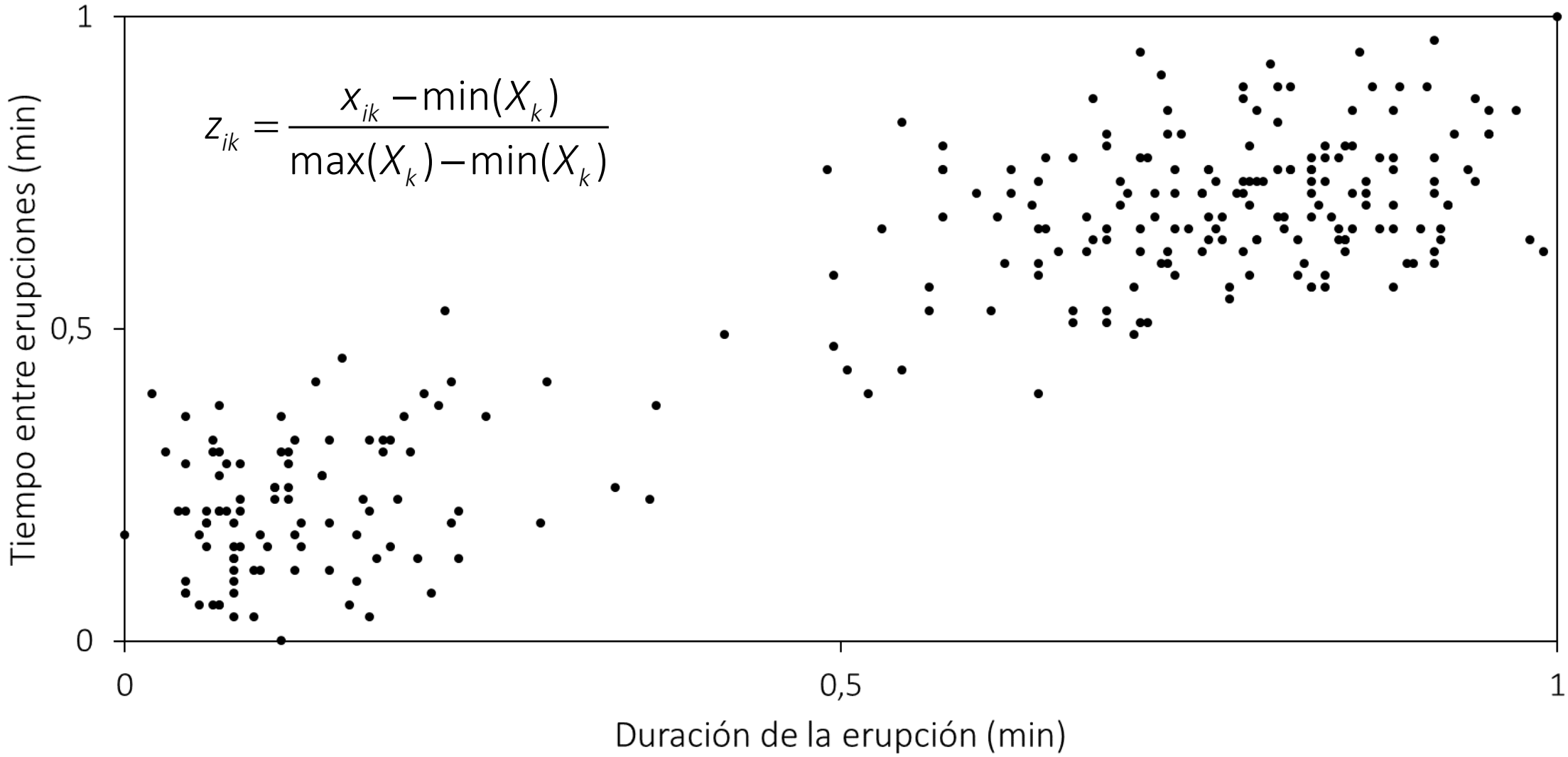
1. Seleccionar K observaciones al azar, a ser utilizadas como centros iniciales
2. Asignar las observaciones al clúster definido por el centro más cercano
3. Actualizar los centros de cada clúster a partir del promedio de las observaciones que lo componen
4. Volver al punto 2 e iterar hasta converger

Ejemplo: el geiser Old Faithful (Wyoming, EEUU)

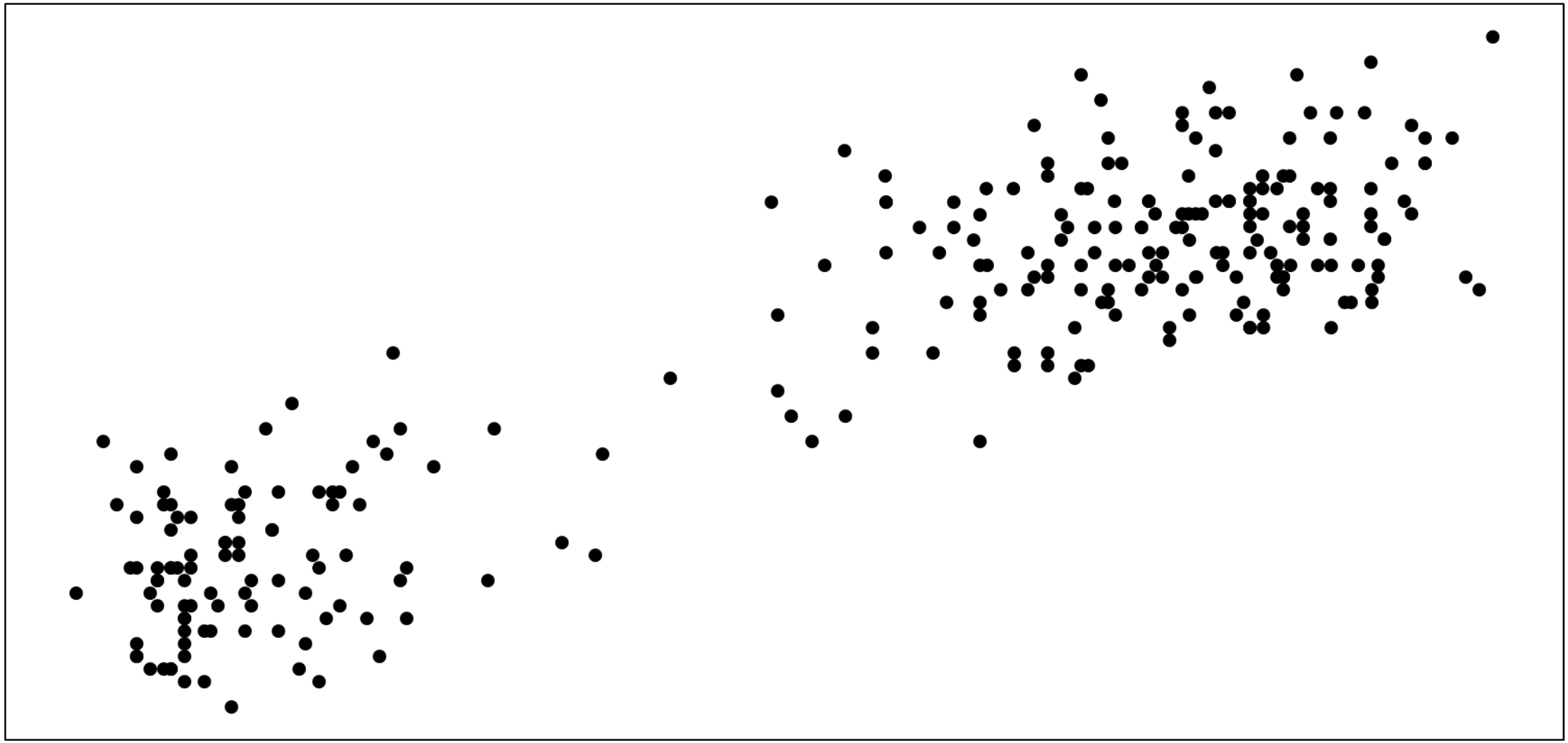
Los datos


120
100
80
60
40
20
0
Tiempo entre erupciones (min)
0
1
2
3
4
5
6
Duración de la erupción (min)

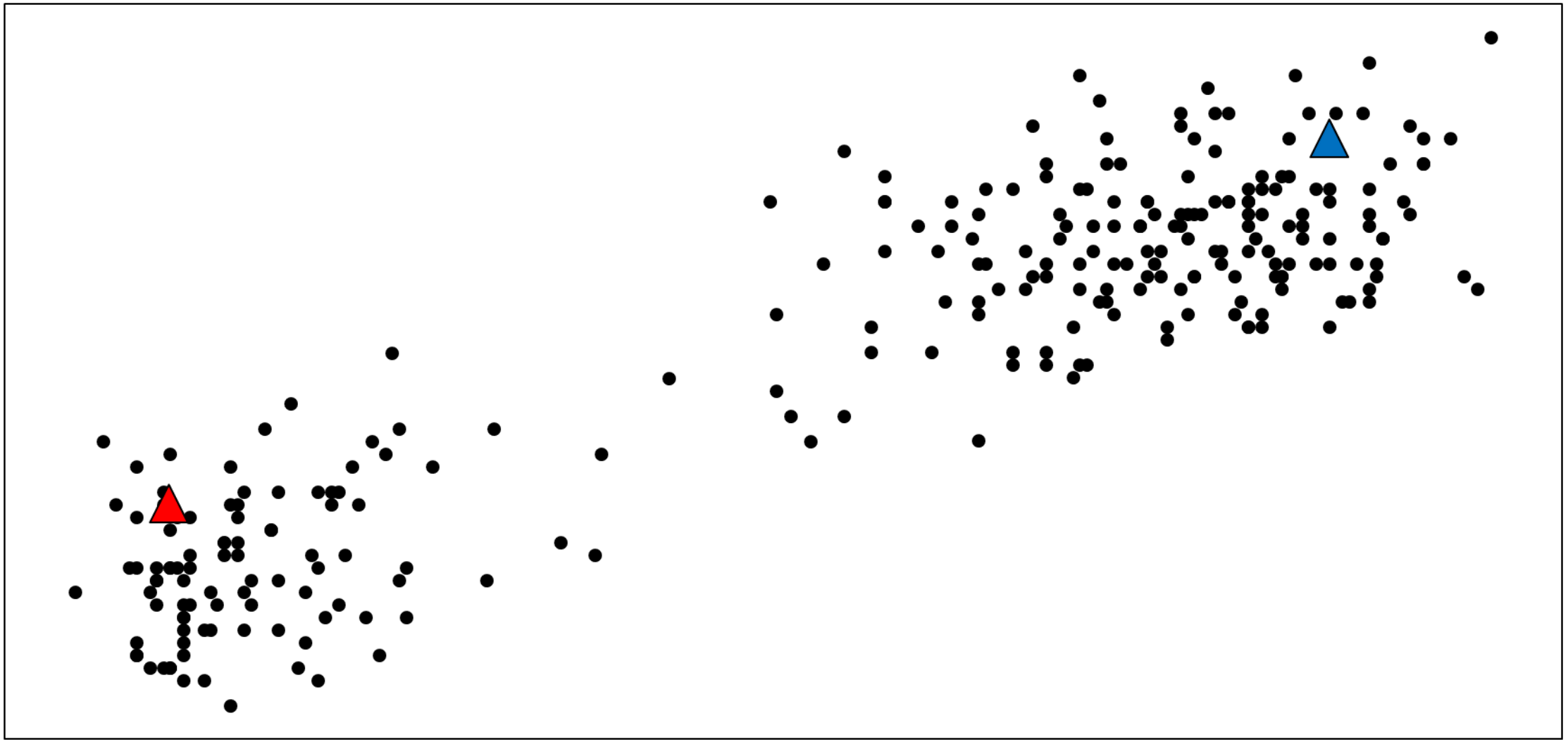
Trabajaremos con datos normalizados



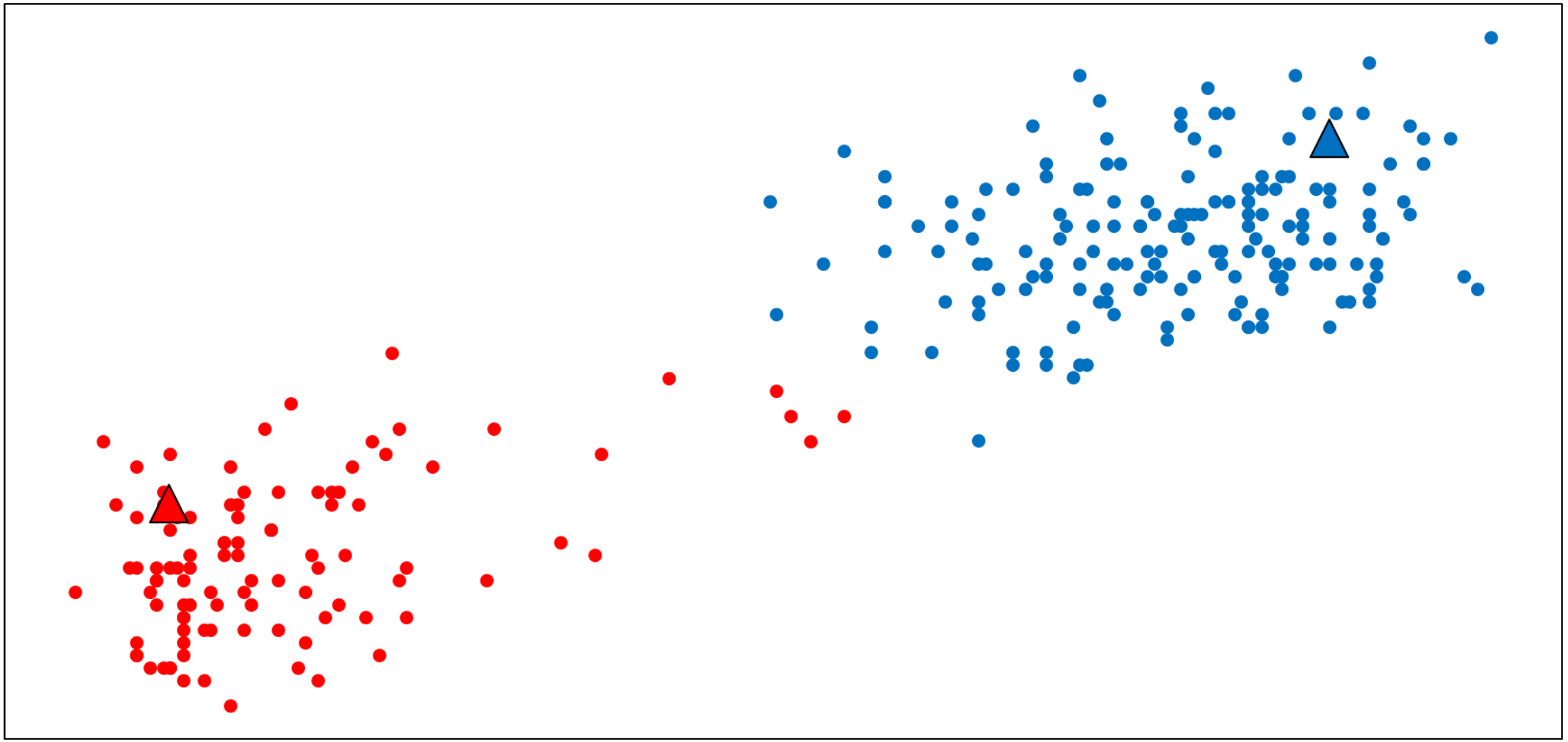
Apliquemos el algoritmo con K = 2



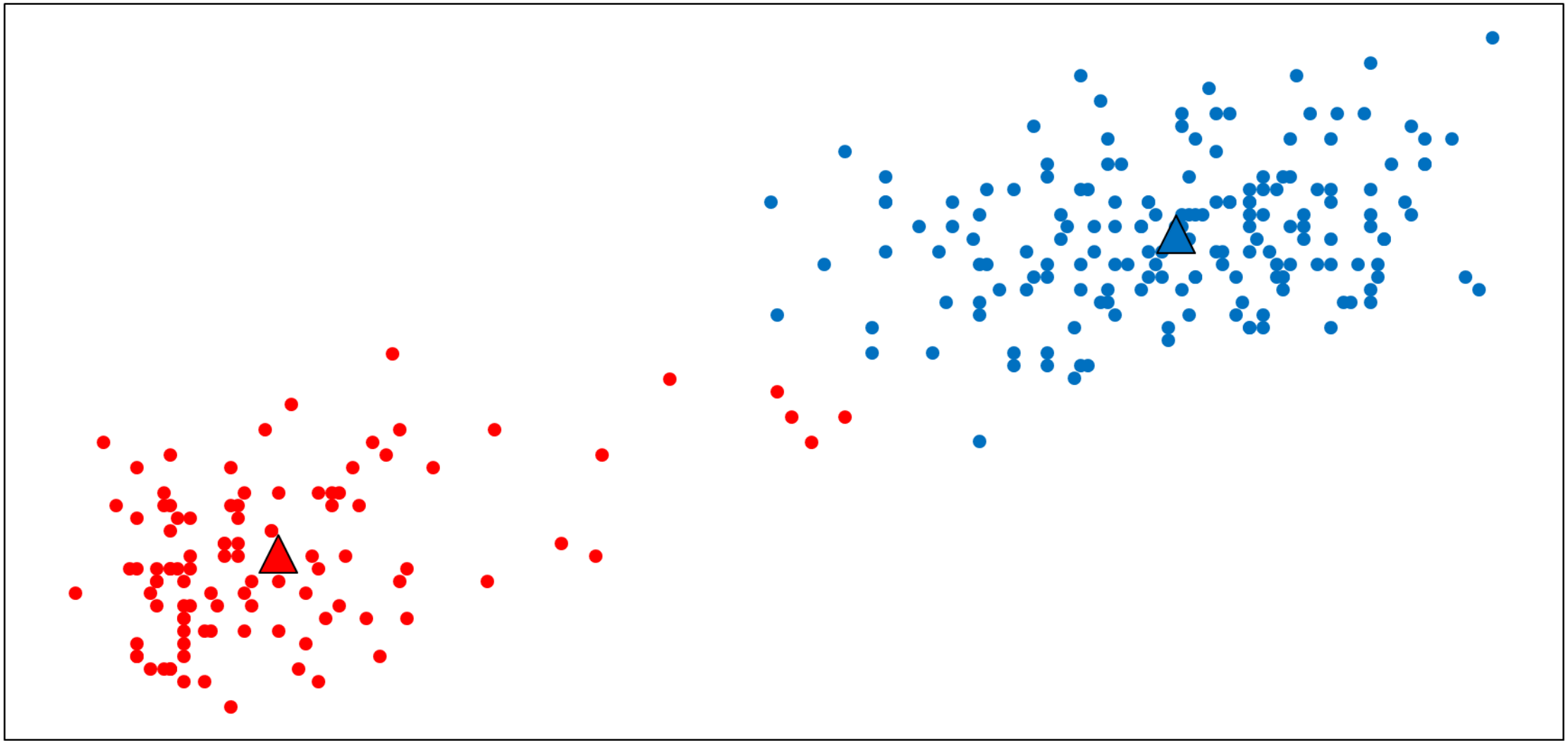
Iteración 0: seleccionamos dos centros al azar



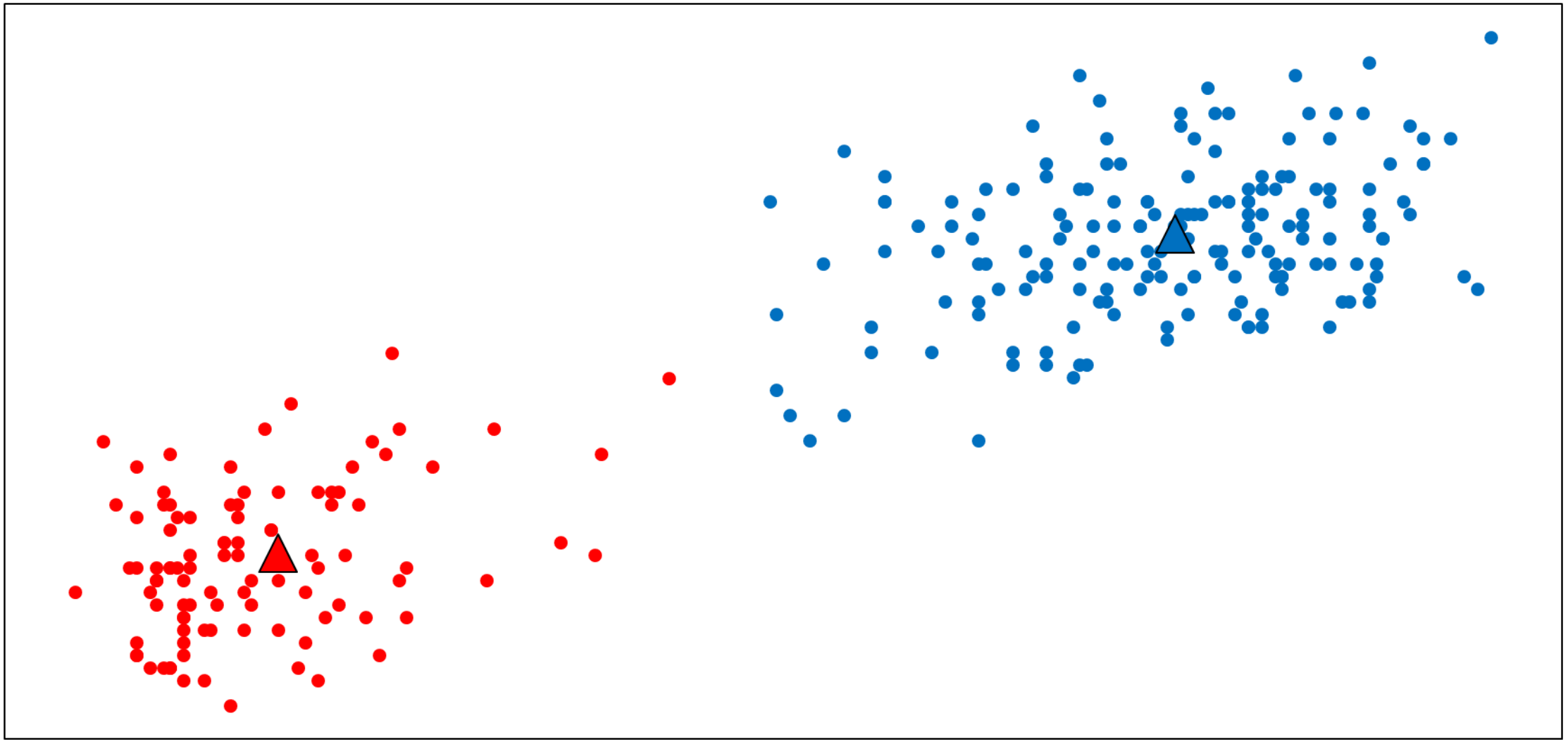
Iteración 1: Asignamos los datos a cada clúster



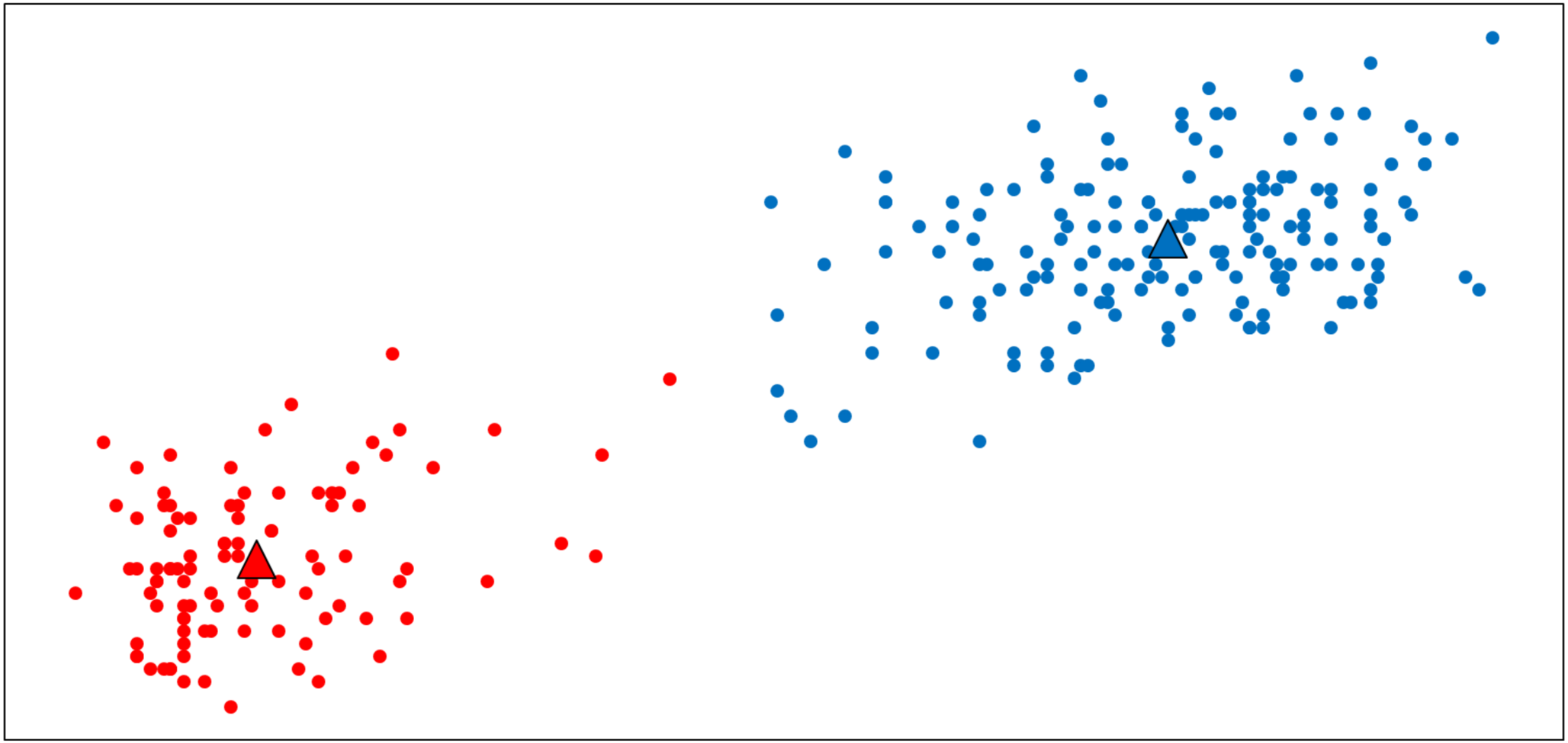
Iteración 1: Actualizamos los centros



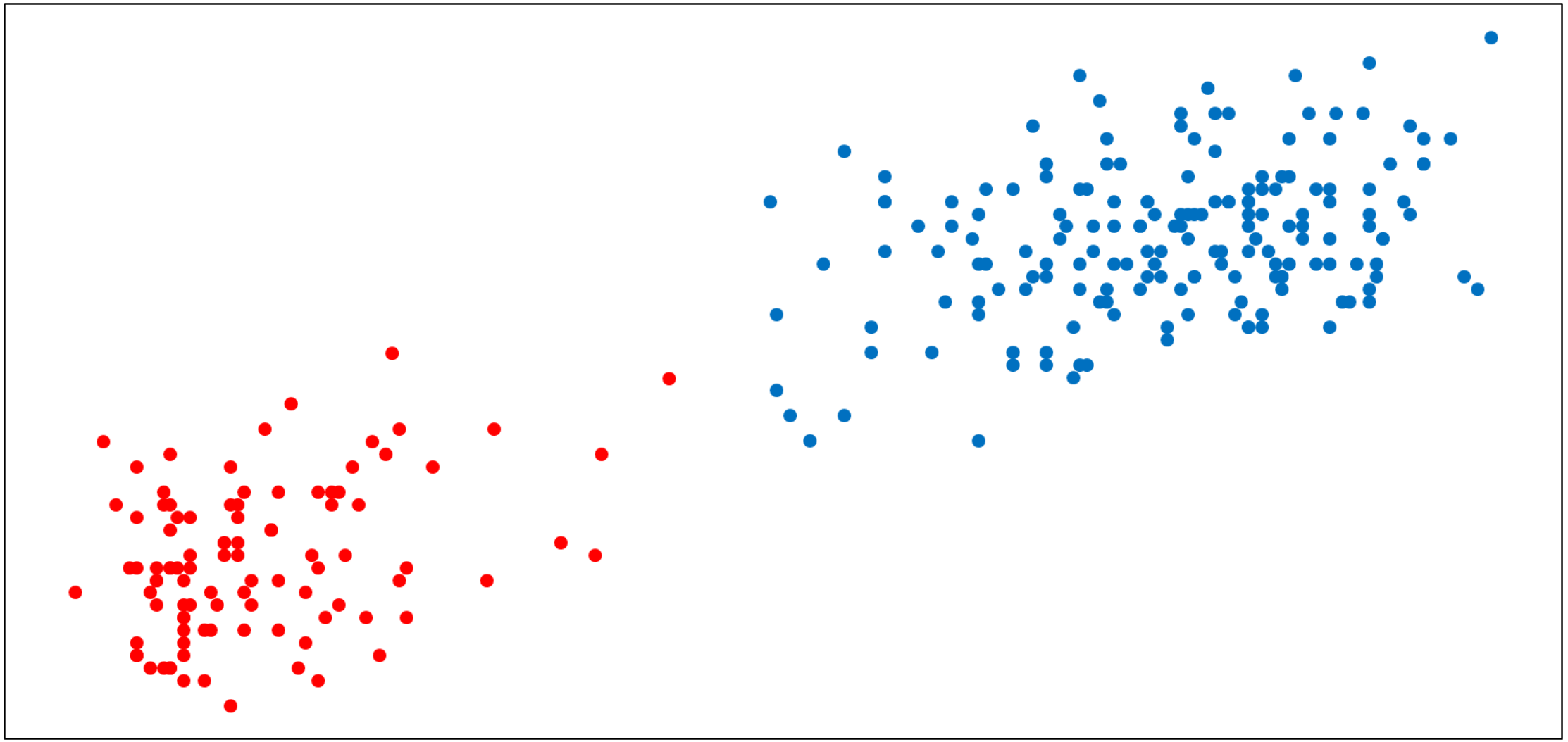
Iteración 2: Asignamos los datos a cada clúster



Iteración 2: Actualizamos los centros



La asignación de datos a cada clúster no cambia, convergimos



K Means

Cada observación ha sido clasificada en uno de los dos clústeres

Clúster Rojo 98 observaciones

centro = { 0,128 ; 0,220 }

Clúster Azul 174 observaciones

centro = { 0,771 ; 0,699 }

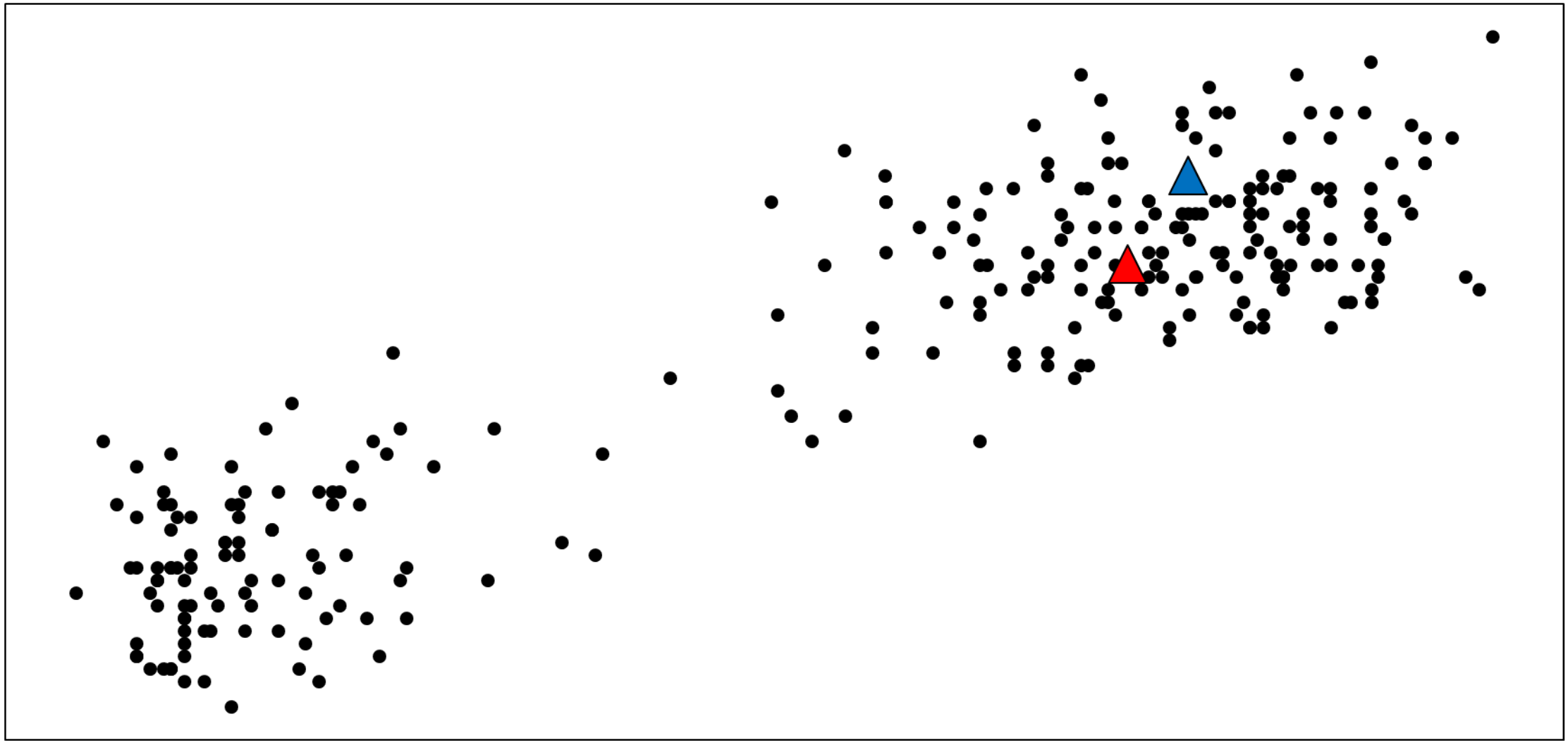
K Means

Una métrica que utilizaremos para evaluar la calidad del resultado es la suma cuadrática de las distancias de cada observación al centro de su clúster

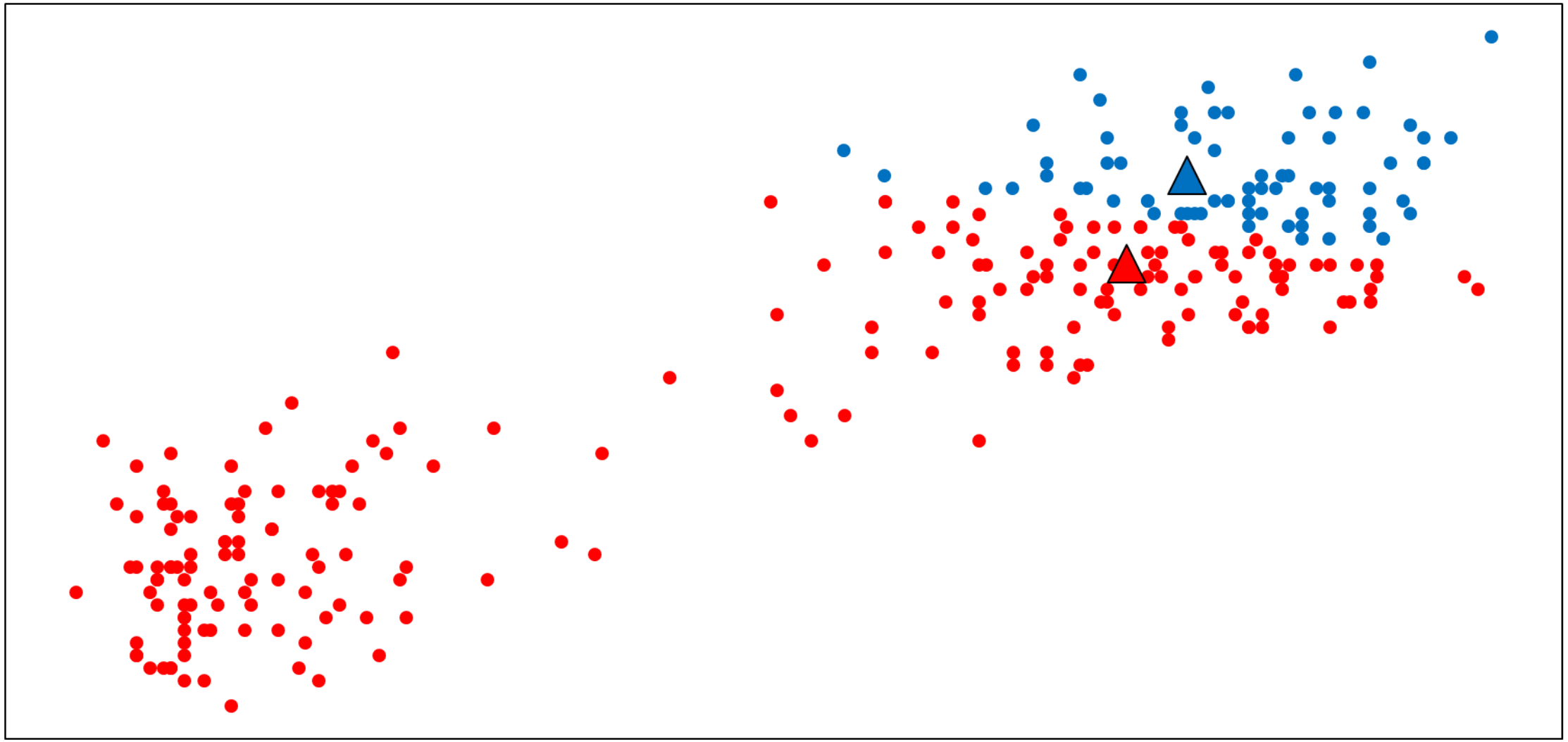
$$SSD_i = \sum_{j \in C_i} \left(x_j - \overline{x}_i \right)^2$$

Clúster Rojo	*SSD* = 1,883	(98 obs.	promedio = 0,019)
Clúster Azul	*SSD* = 4,457	(174 obs.	promedio = 0,026)

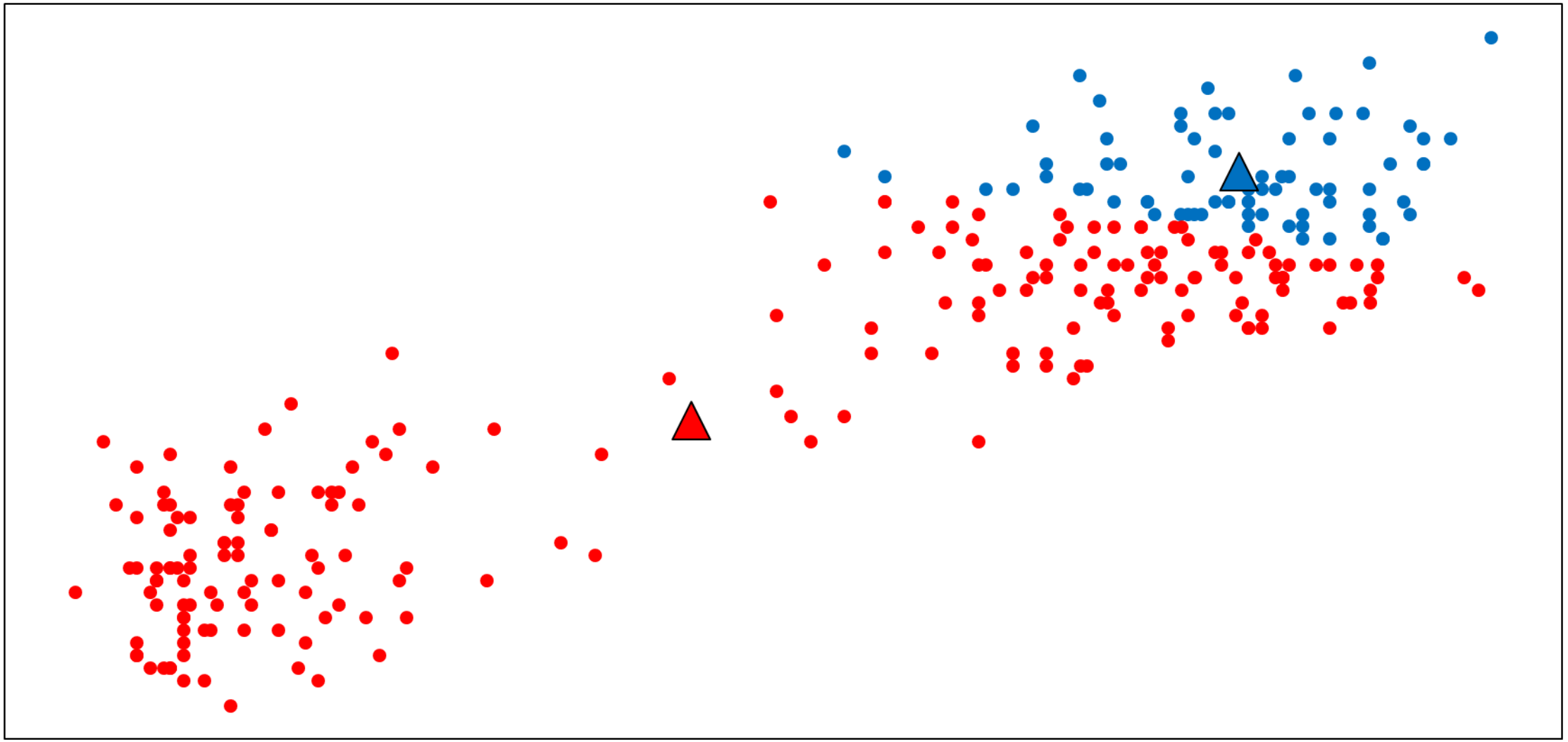
¿Qué pasa si partimos de “peores” centros iniciales?



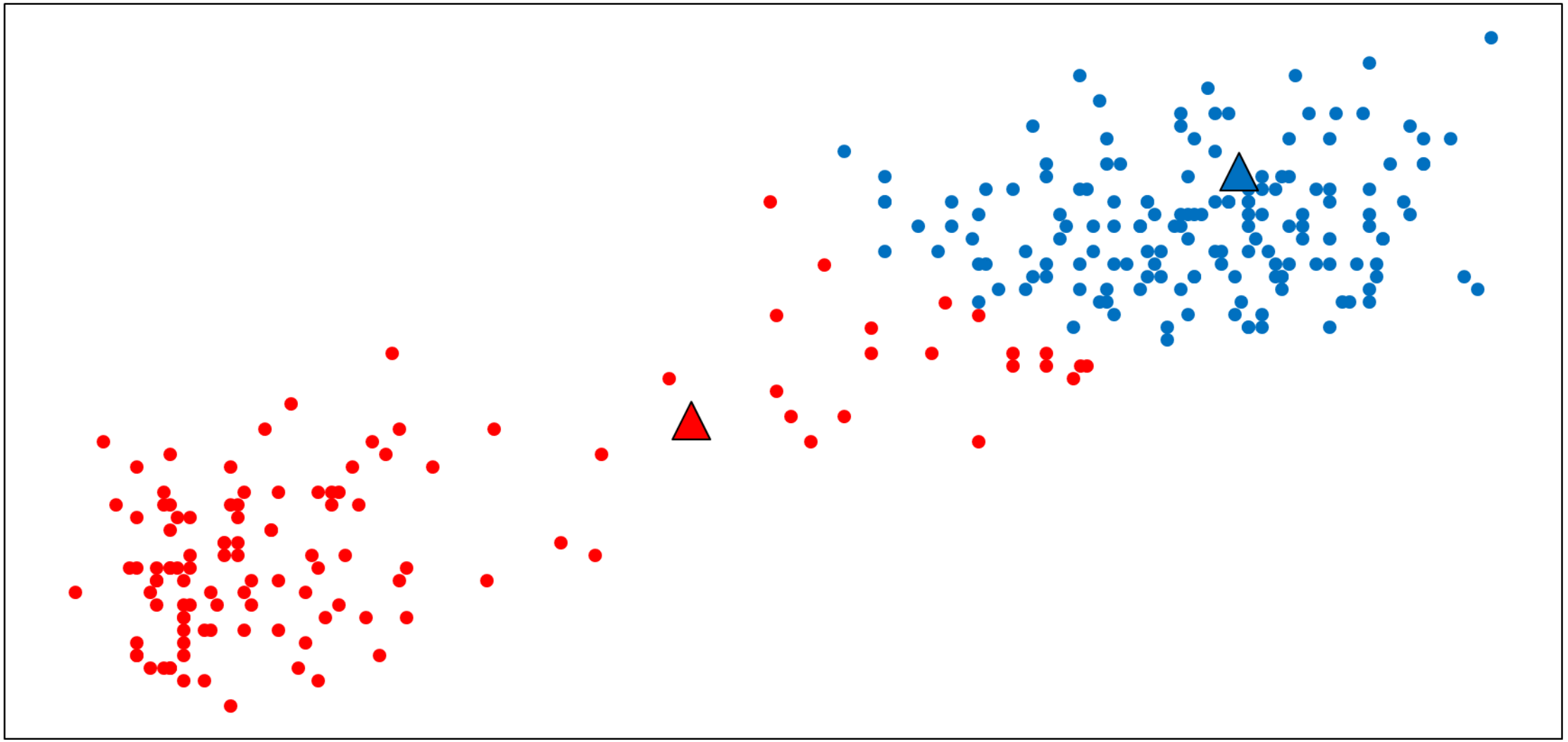
Iteración 1: Asignamos los datos a cada clúster



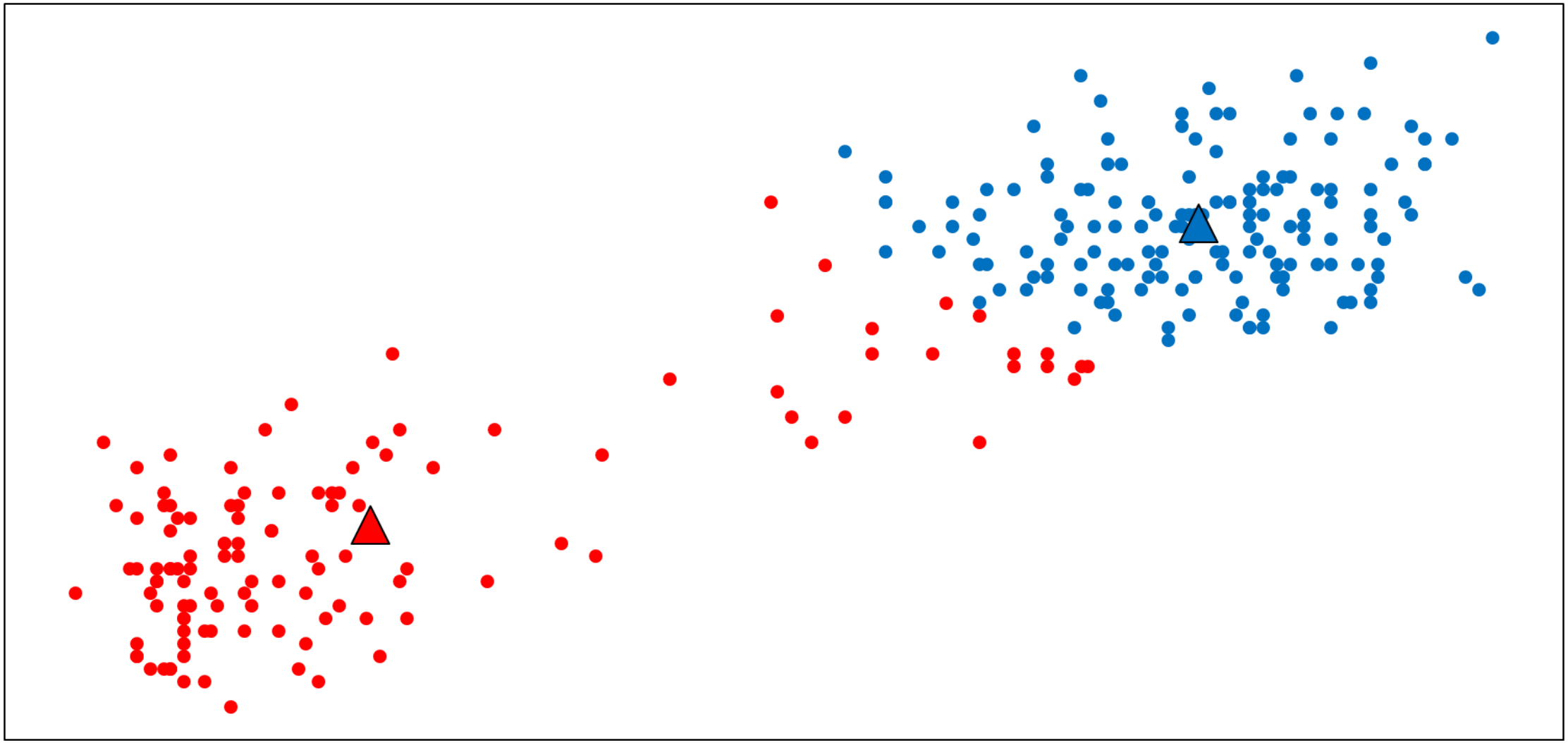
Iteración 1: Actualizamos los centros



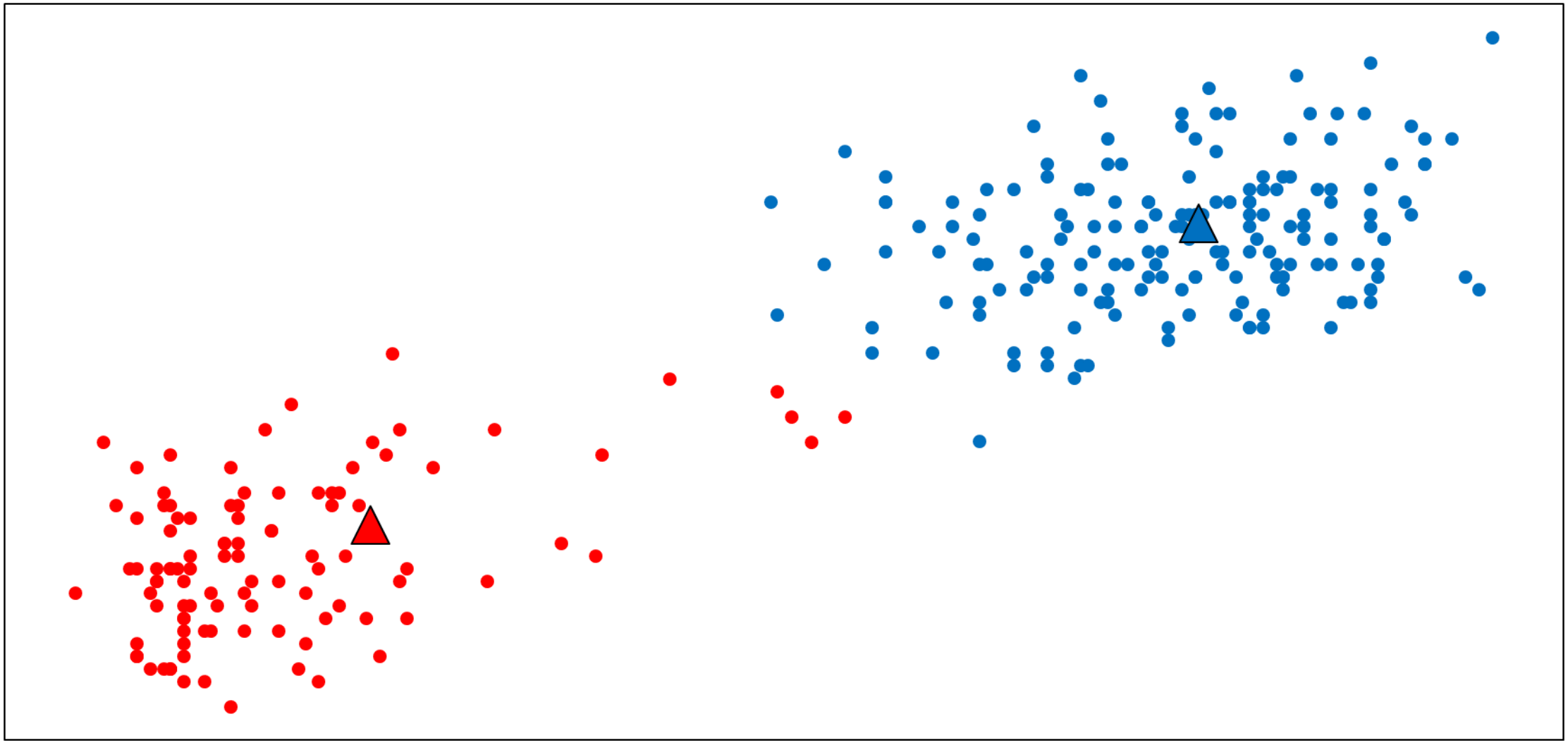
Iteración 2: Asignamos los datos a cada clúster



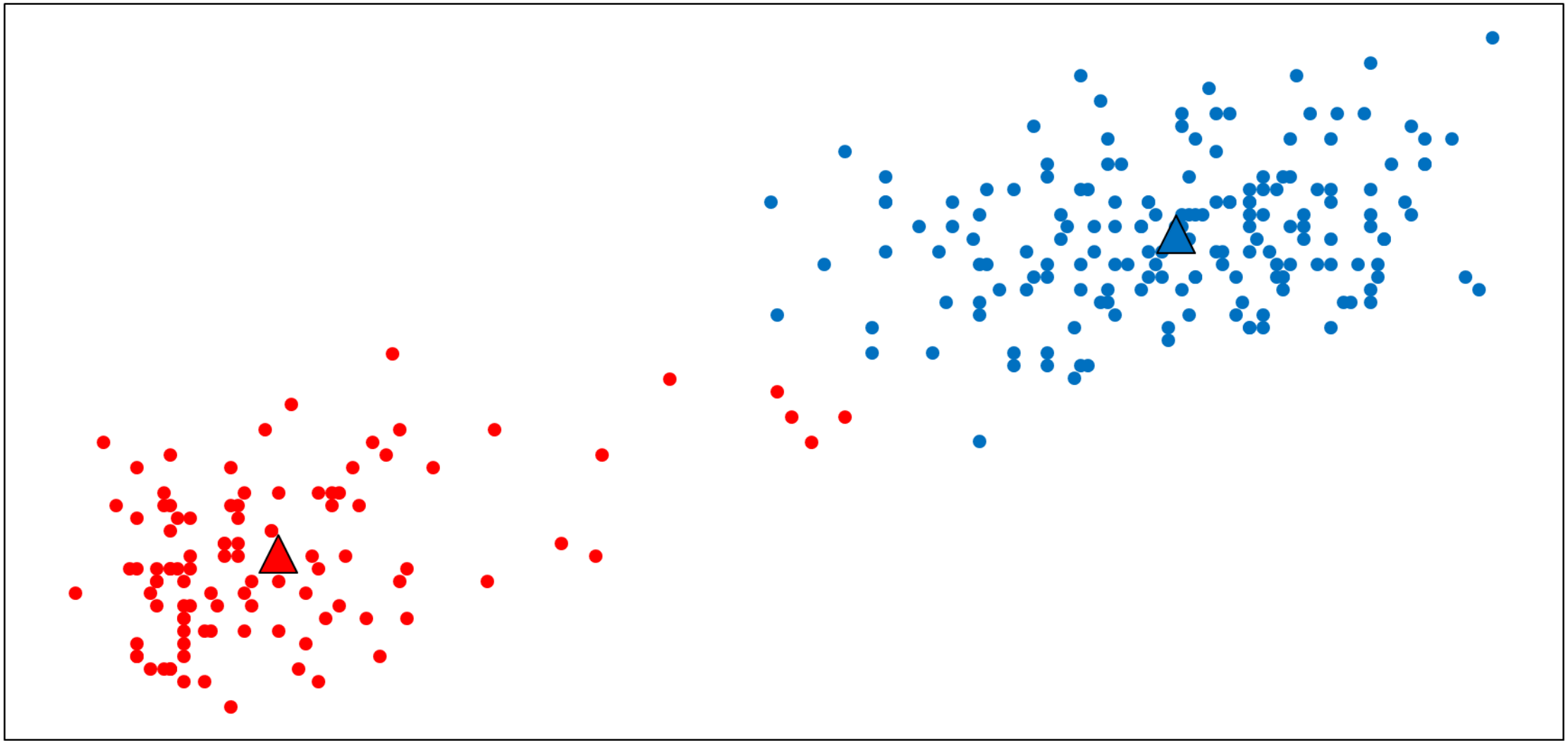
Iteración 2: Actualizamos los centros



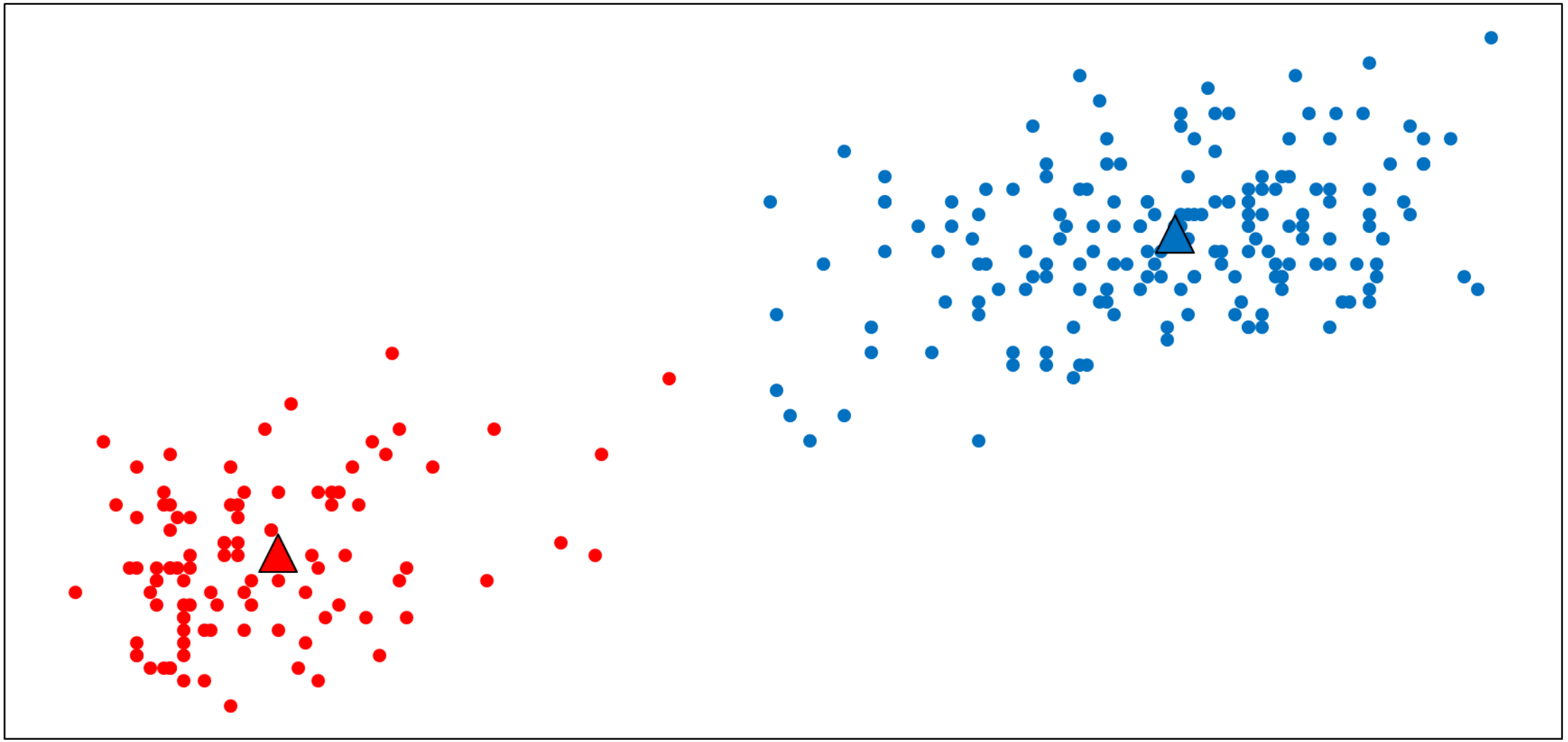
Iteración 3: Asignamos los datos a cada clúster



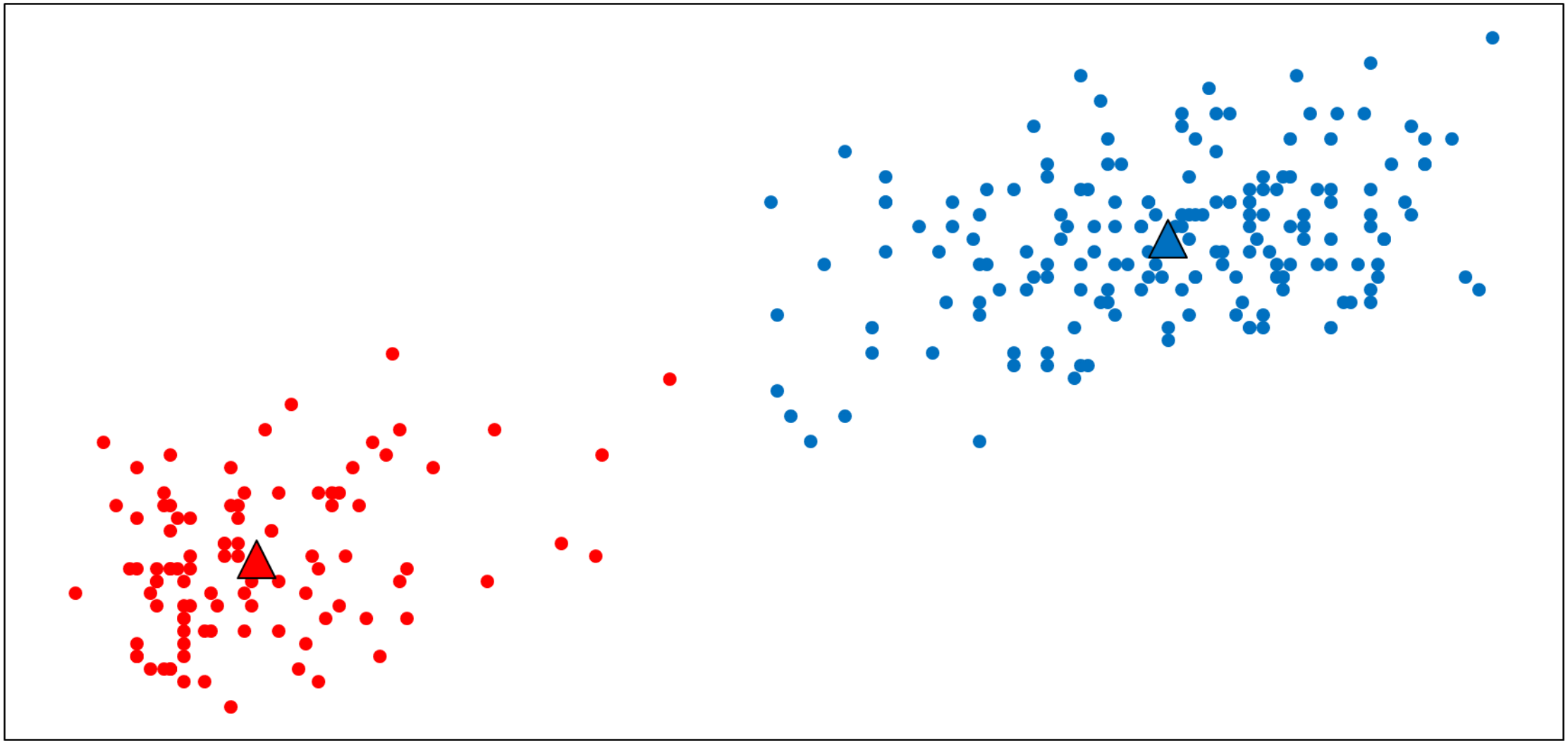
Iteración 3: Actualizamos los centros



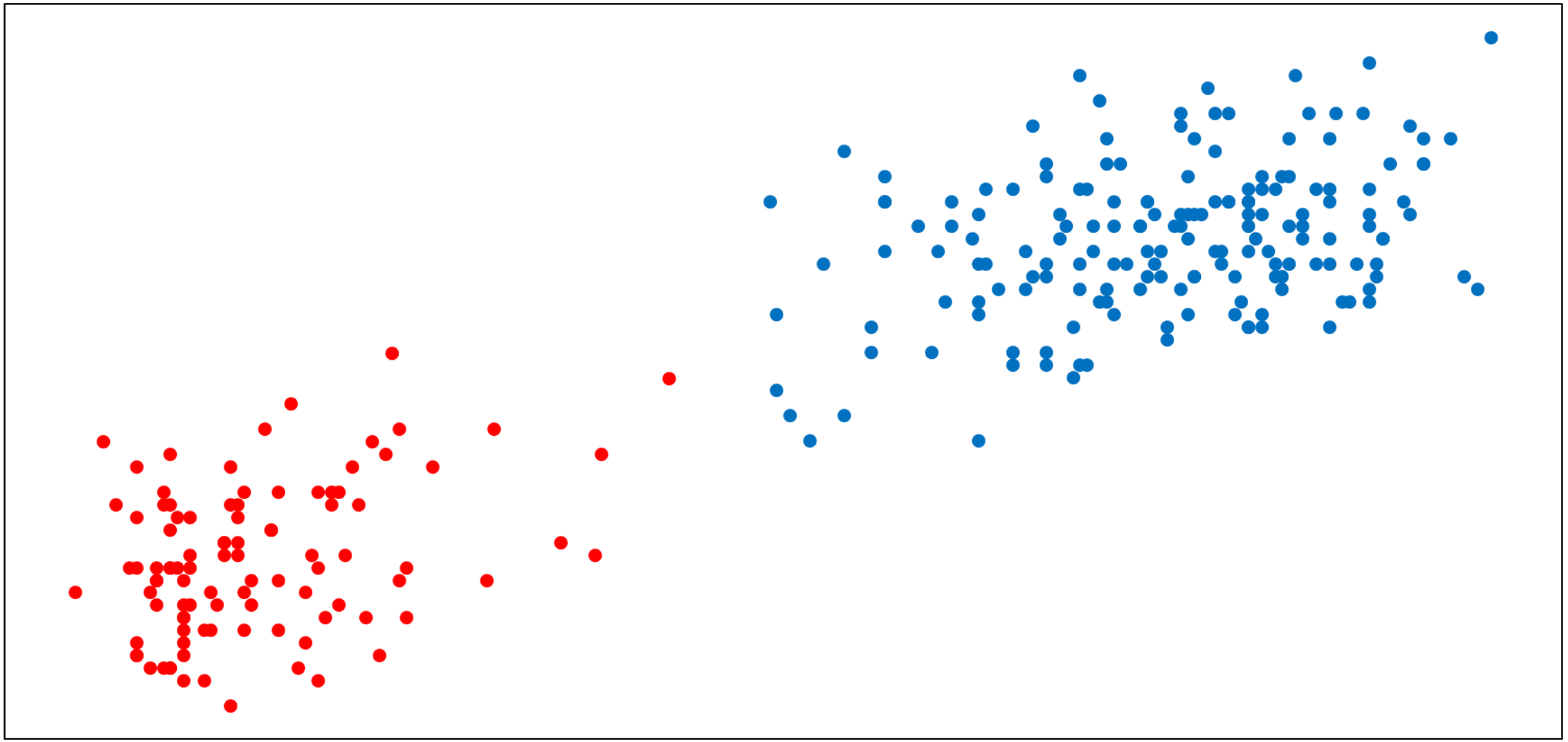
Iteración 4: Asignamos los datos a cada clúster



Iteración 4: Actualizamos los centros



La asignación de datos a cada clúster no cambia, convergimos



Algunos comentarios sobre K Means

1. El algoritmo siempre converge

2. Como la selección de centros iniciales es aleatoria, el algoritmo puede no converger siempre al mismo resultado

   Podemos aplicar el algoritmo varias veces y elegir el resultado con menor SSD total

Algunos comentarios sobre K Means

3. El algoritmo es escalable a muchas observaciones

4. El algoritmo no es tan escalable a muchas variables

5. Se puede establecer un criterio de parada anticipada (e.g. si SSD no varía demasiado en dos iteraciones consecutivas)

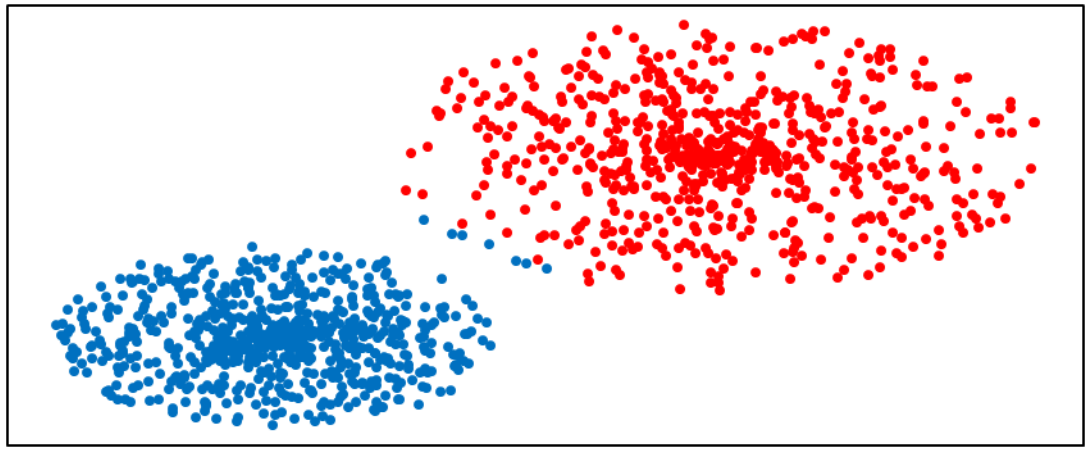
Algunos comentarios sobre K Means

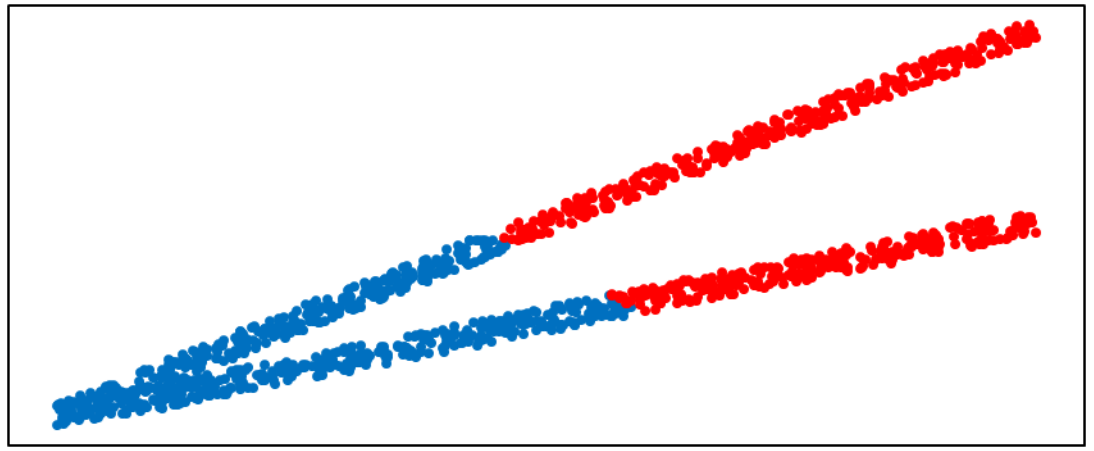
6. Interpretar los resultados (e.g. “ponerle nombre” a los clústeres) puede no ser fácil

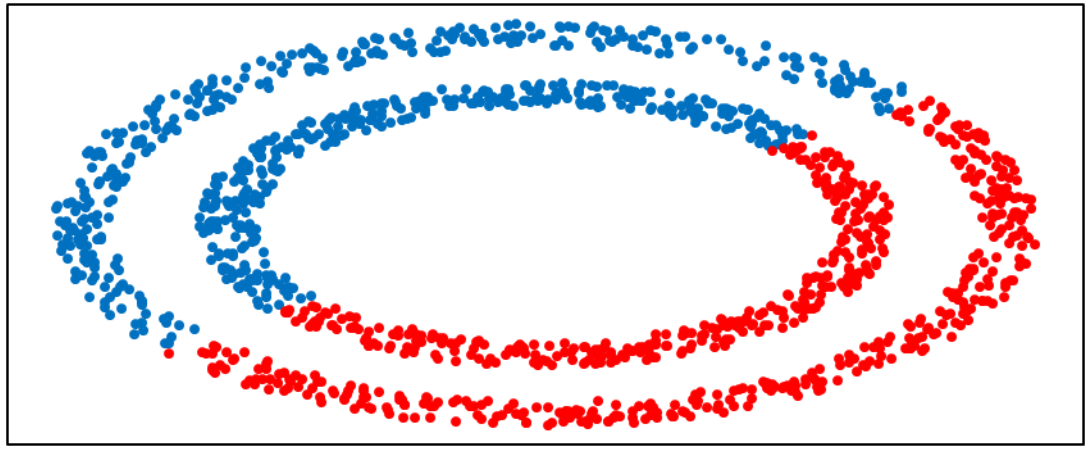
7. La presencia de datos atípicos (i.e. *outliers*) puede afectar fuertemente el resultado del algoritmo

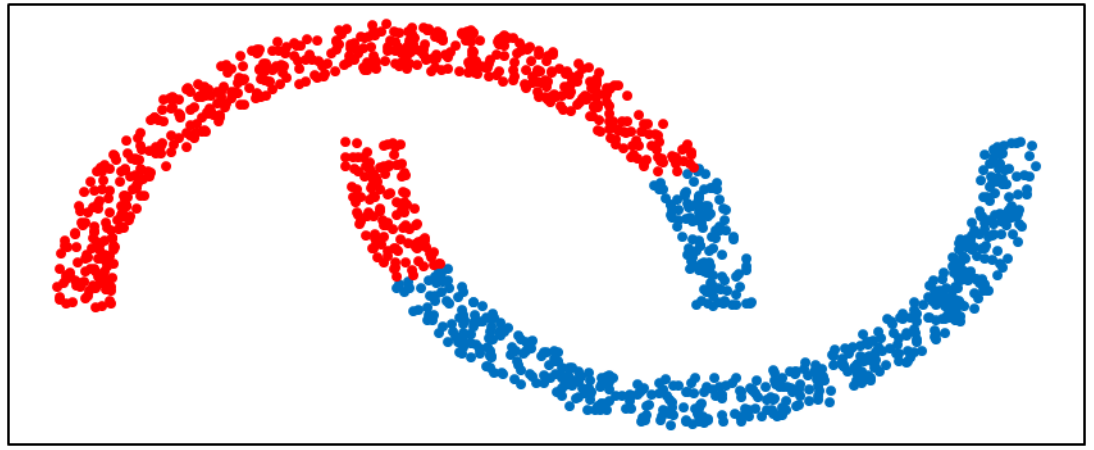
8. El algoritmo puede fallar ante relaciones no-esféricas de datos

Relaciones esféricas y no-esféricas de datos

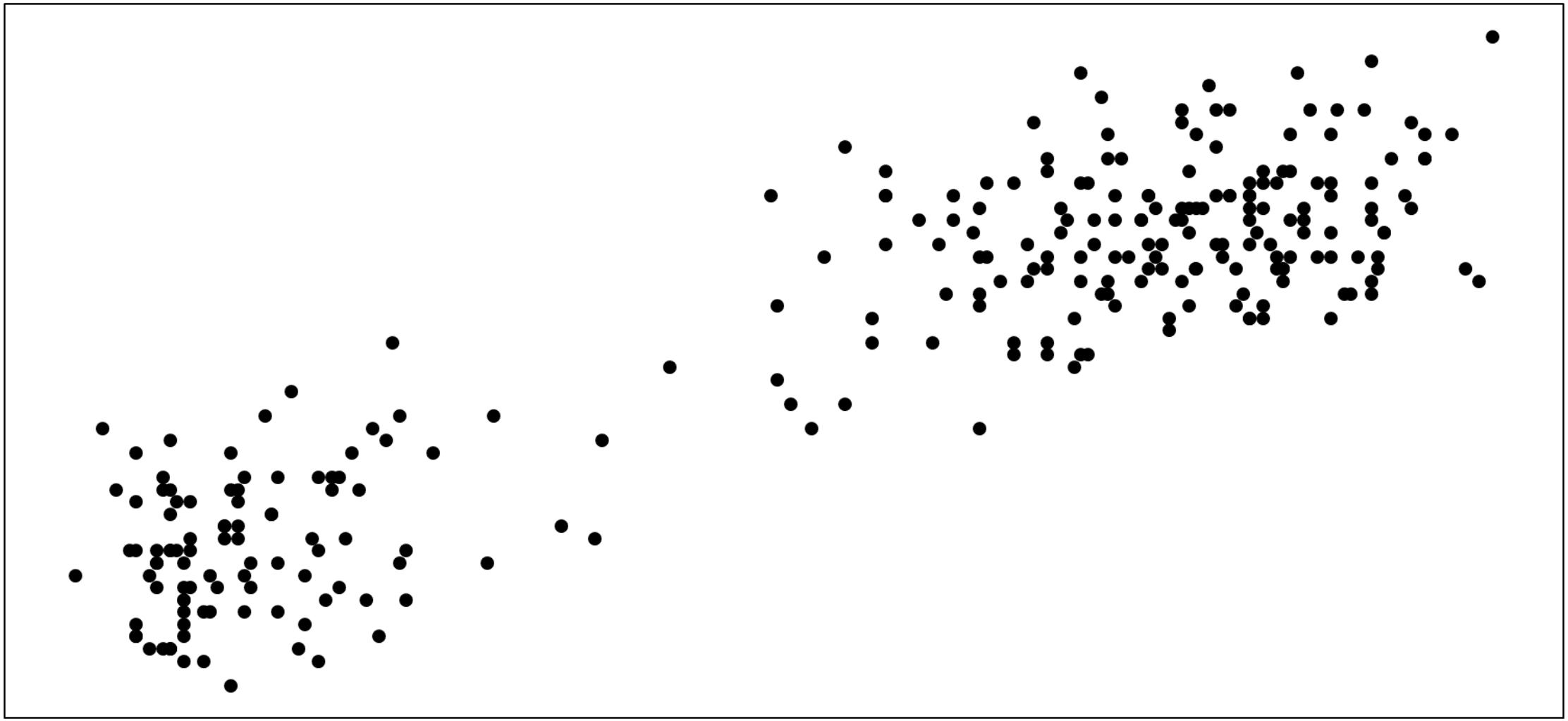




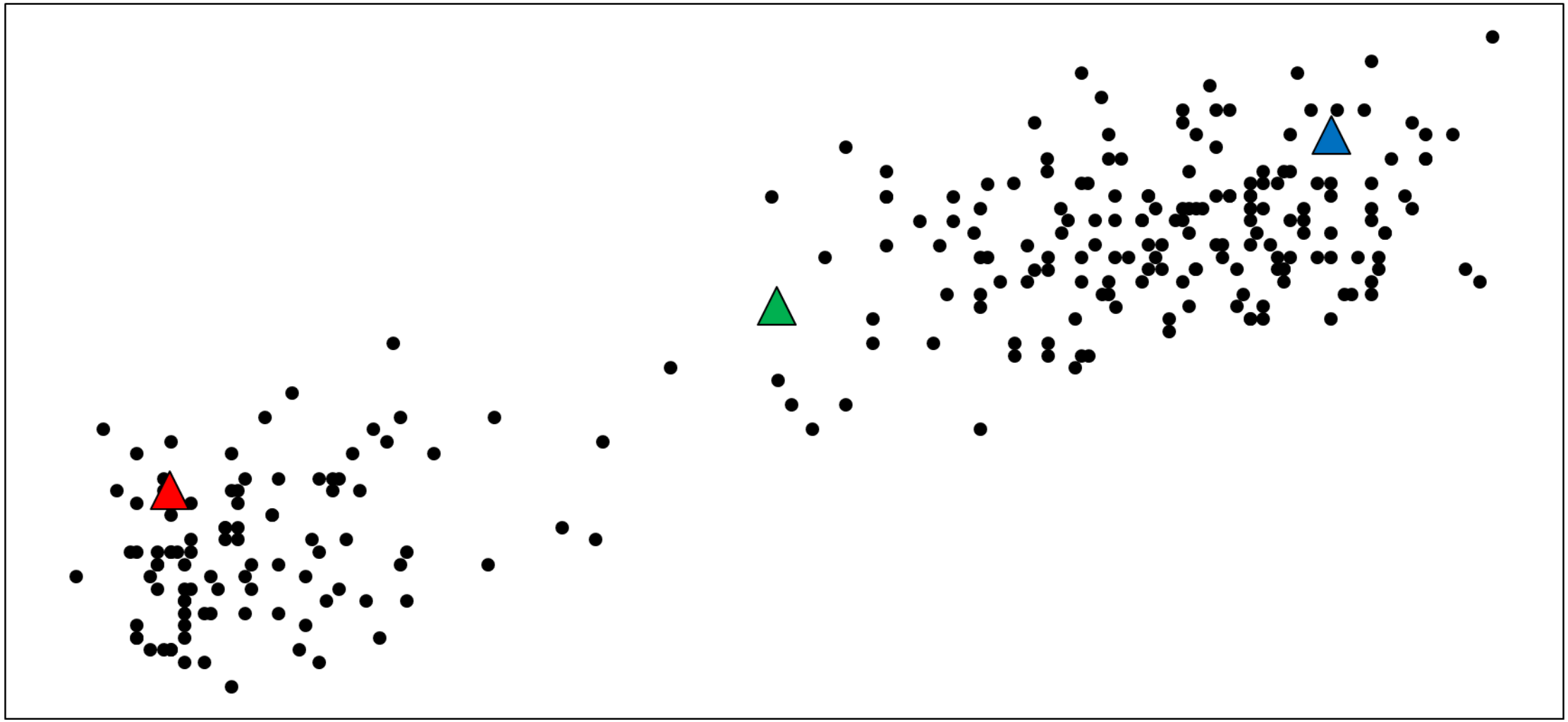




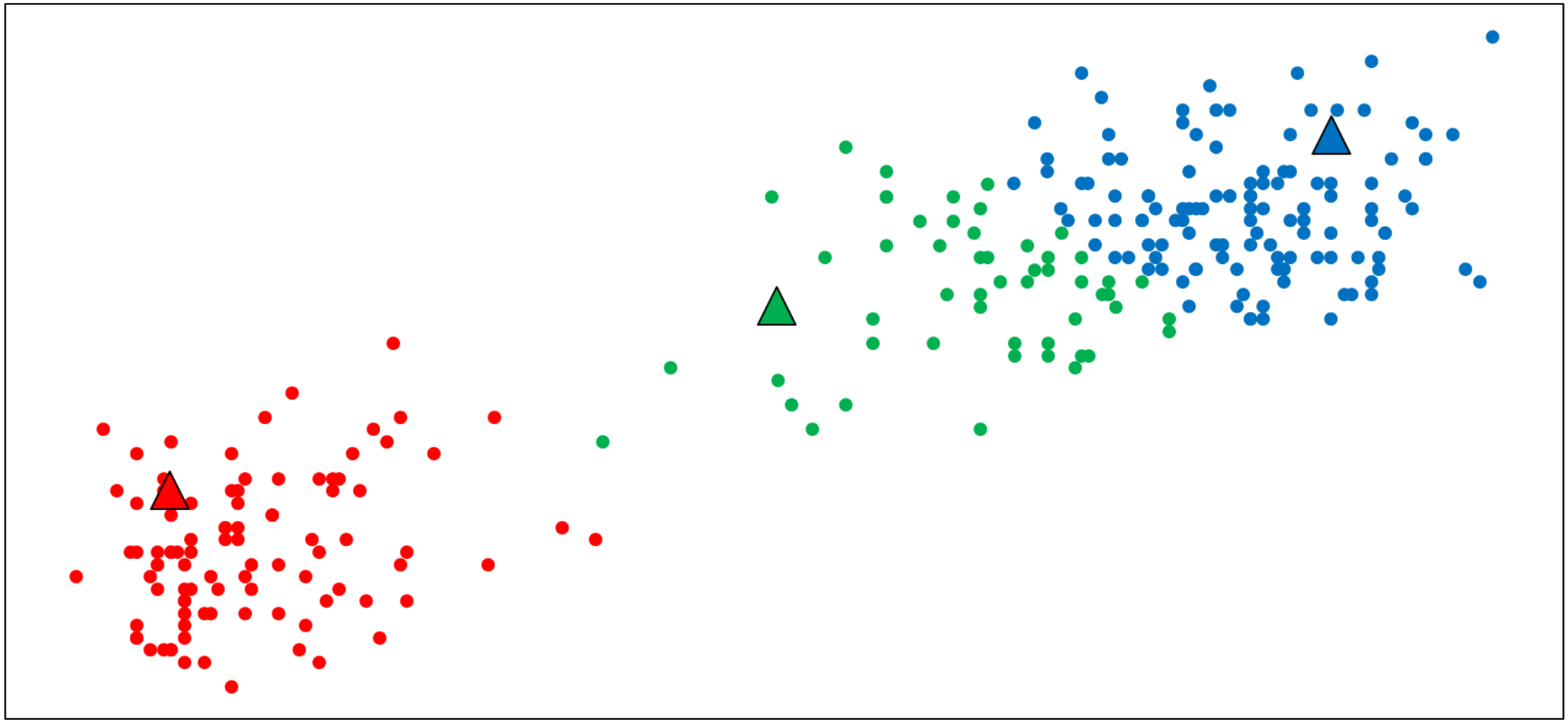
Apliquemos el algoritmo con K = 3



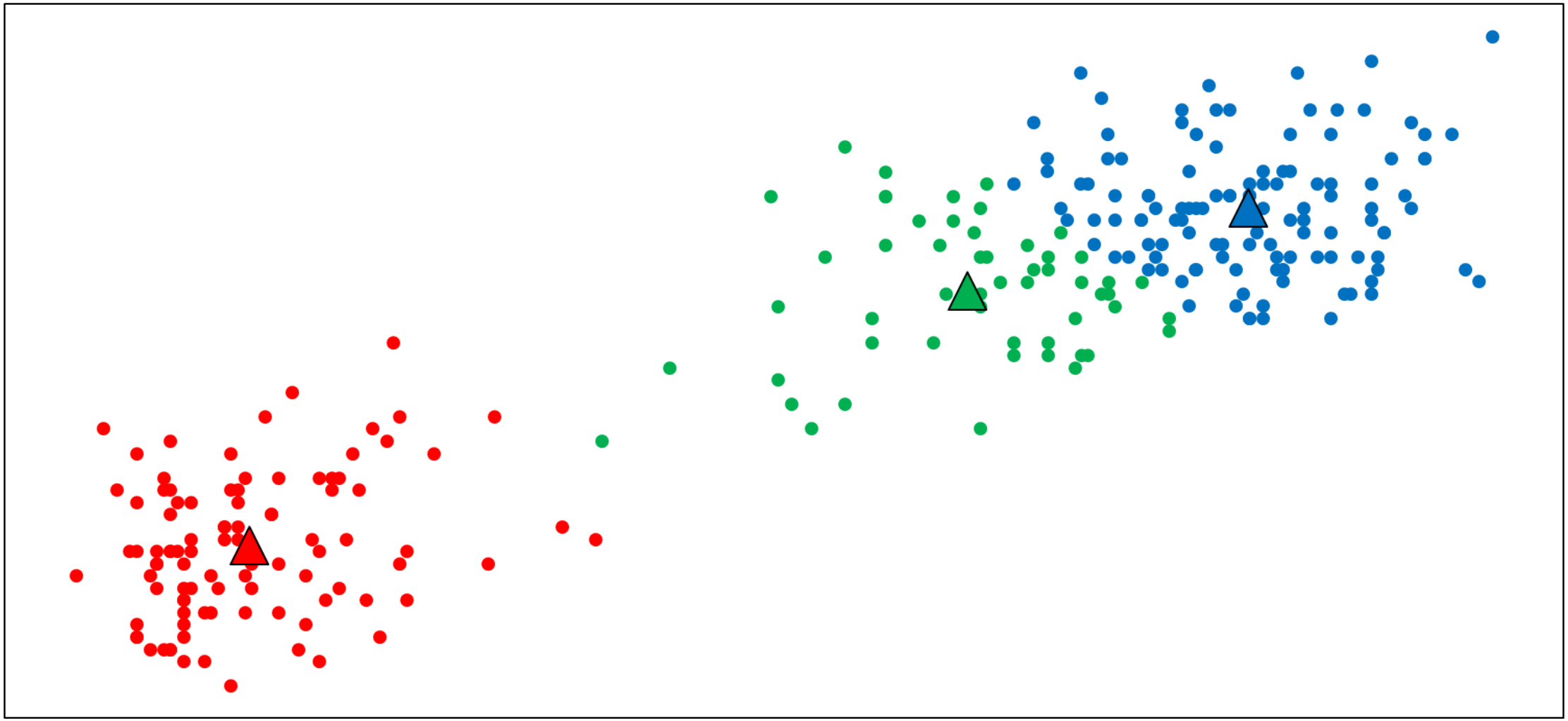
Iteración 0: seleccionamos tres centros al azar



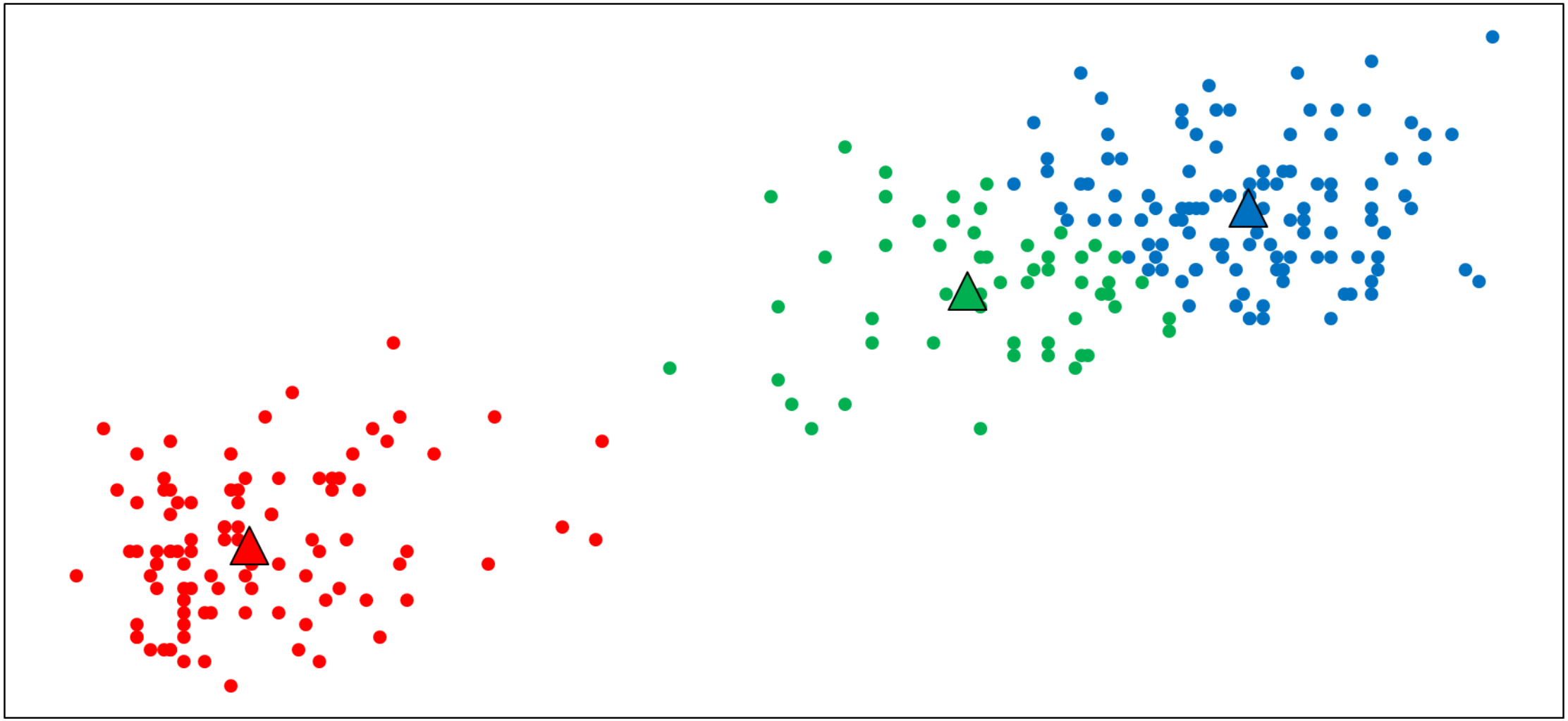
Iteración 1: Asignamos los datos a cada clúster



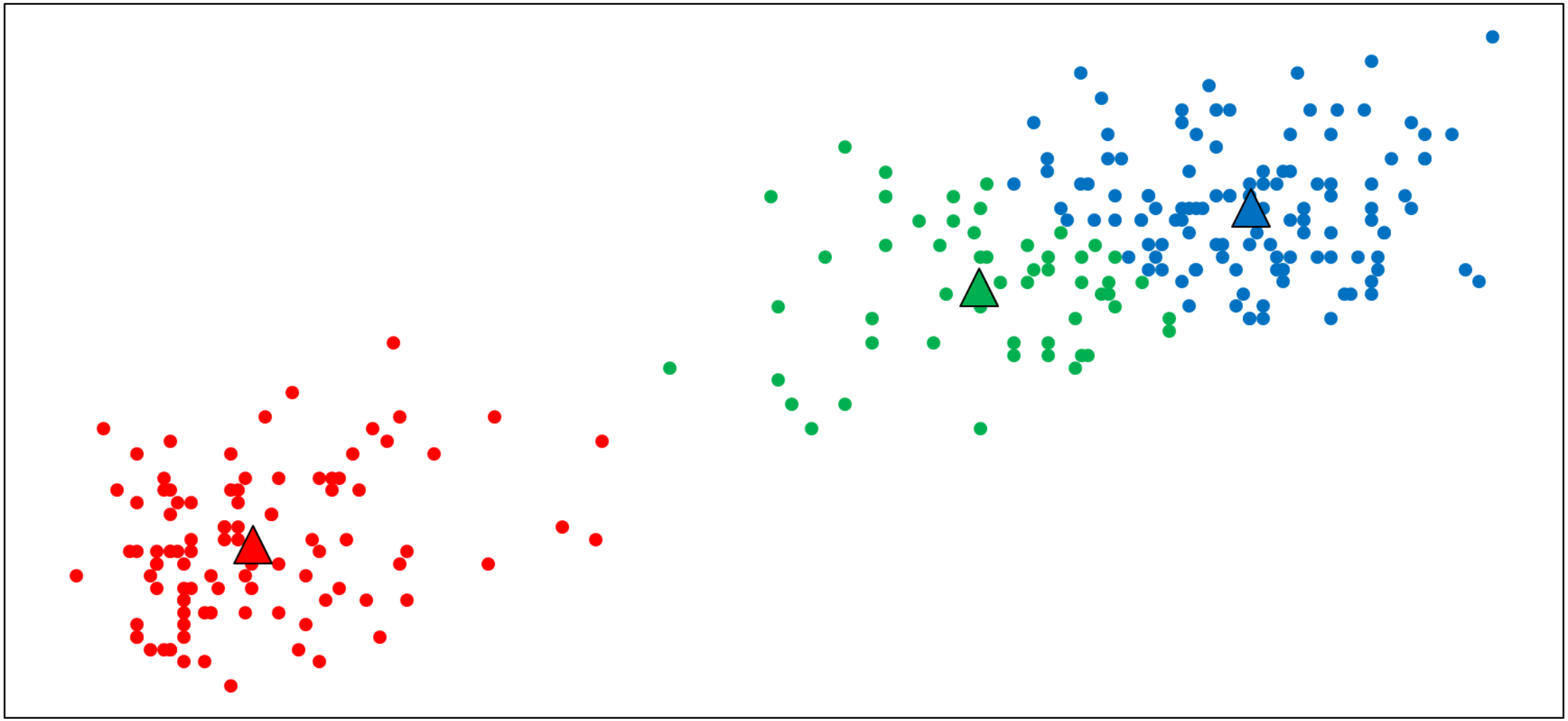
Iteración 1: Actualizamos los centros



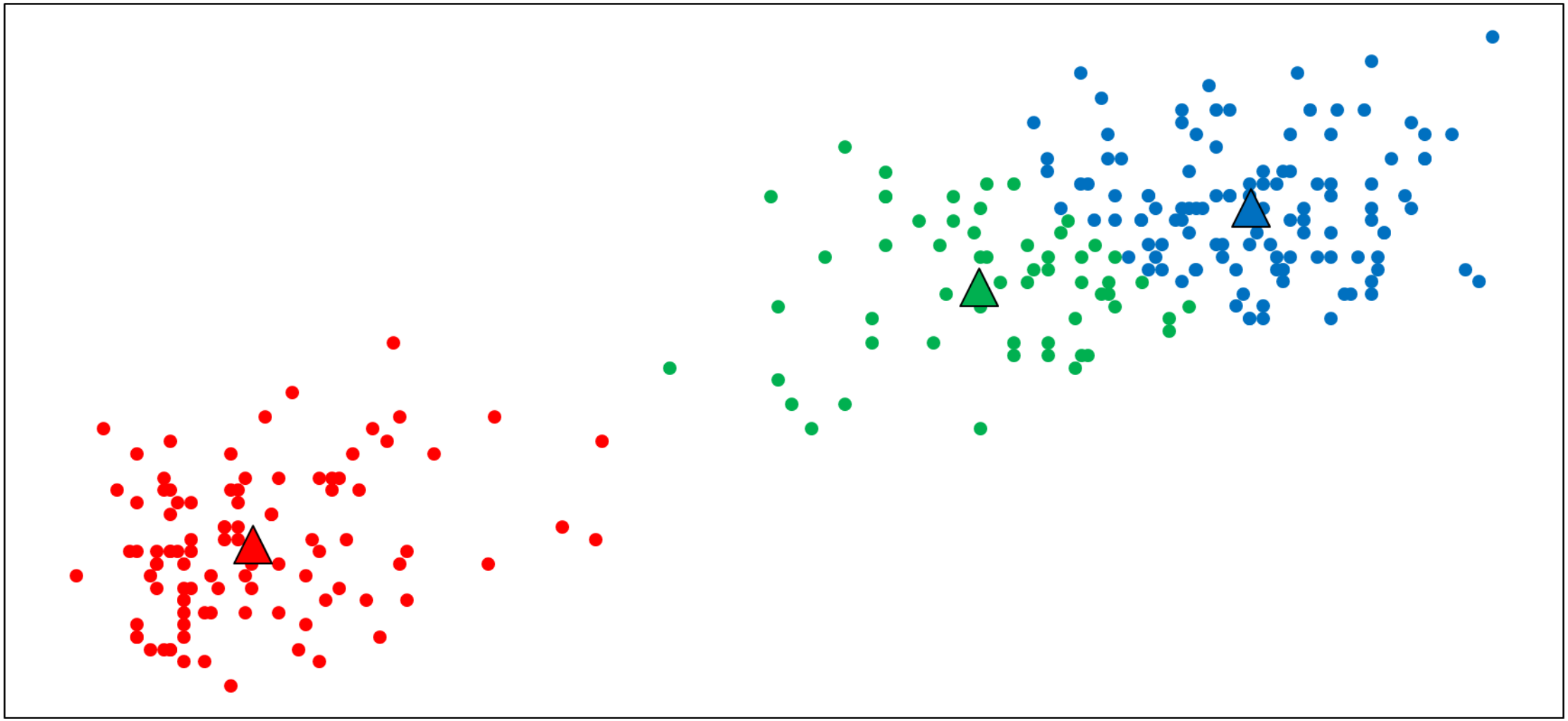
Iteración 2: Asignamos los datos a cada clúster



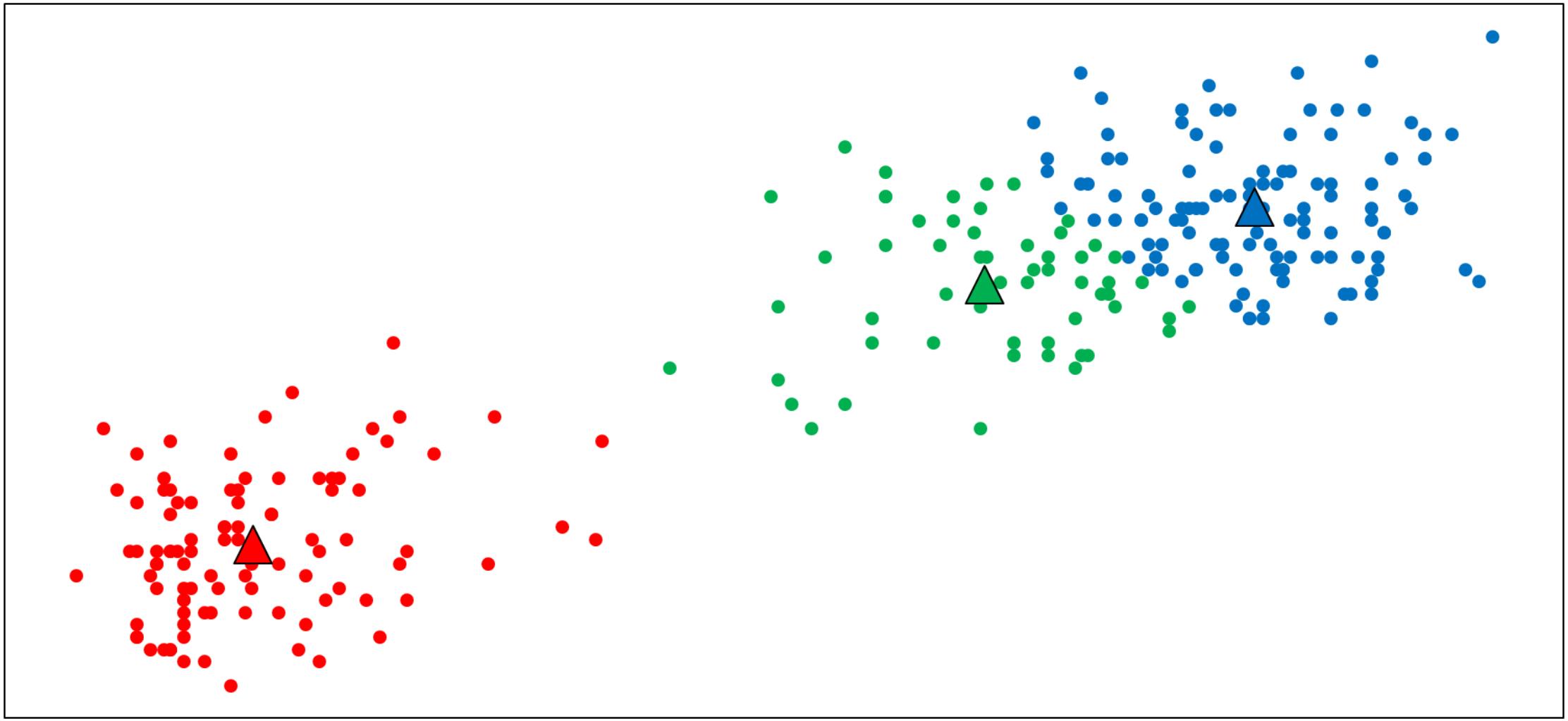
Iteración 2: Actualizamos los centros



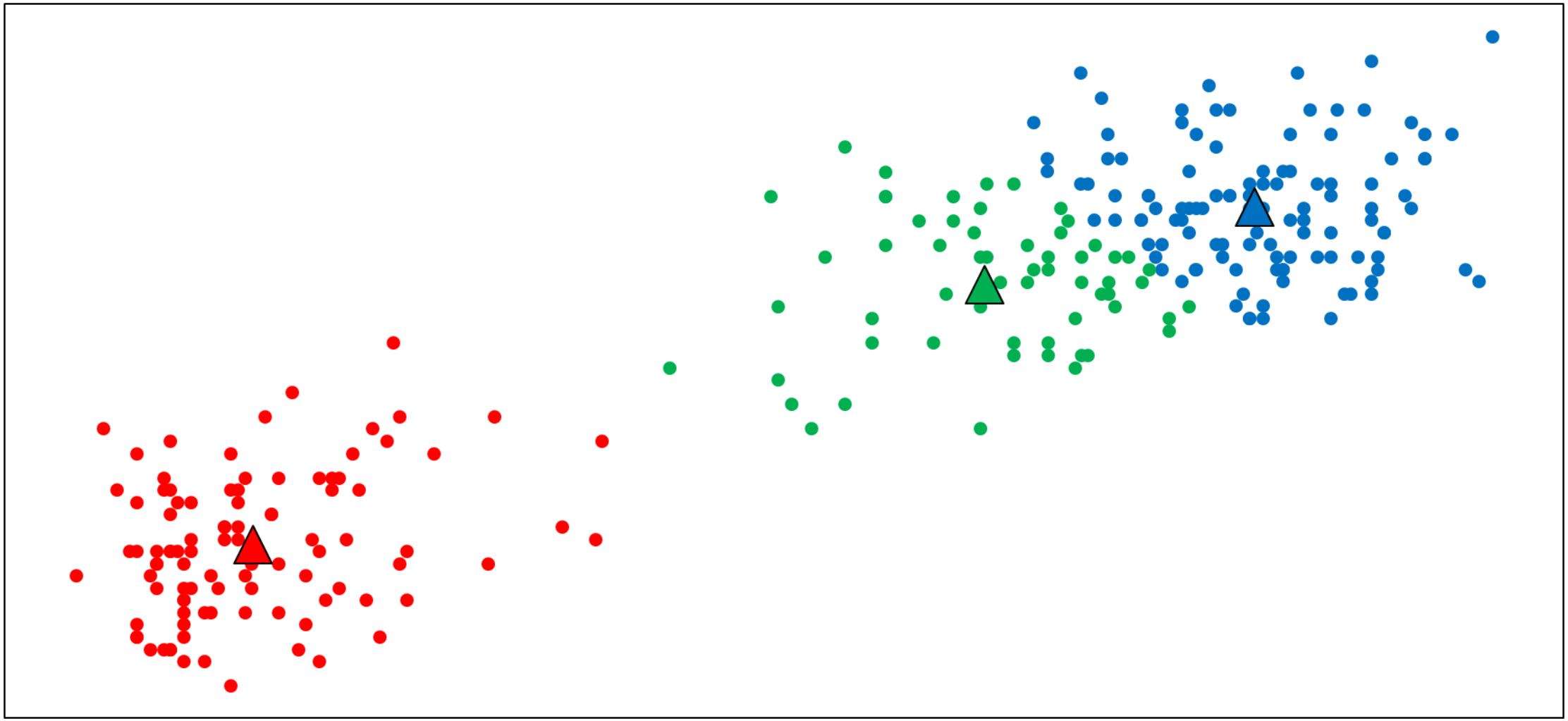
Iteración 3: Asignamos los datos a cada clúster



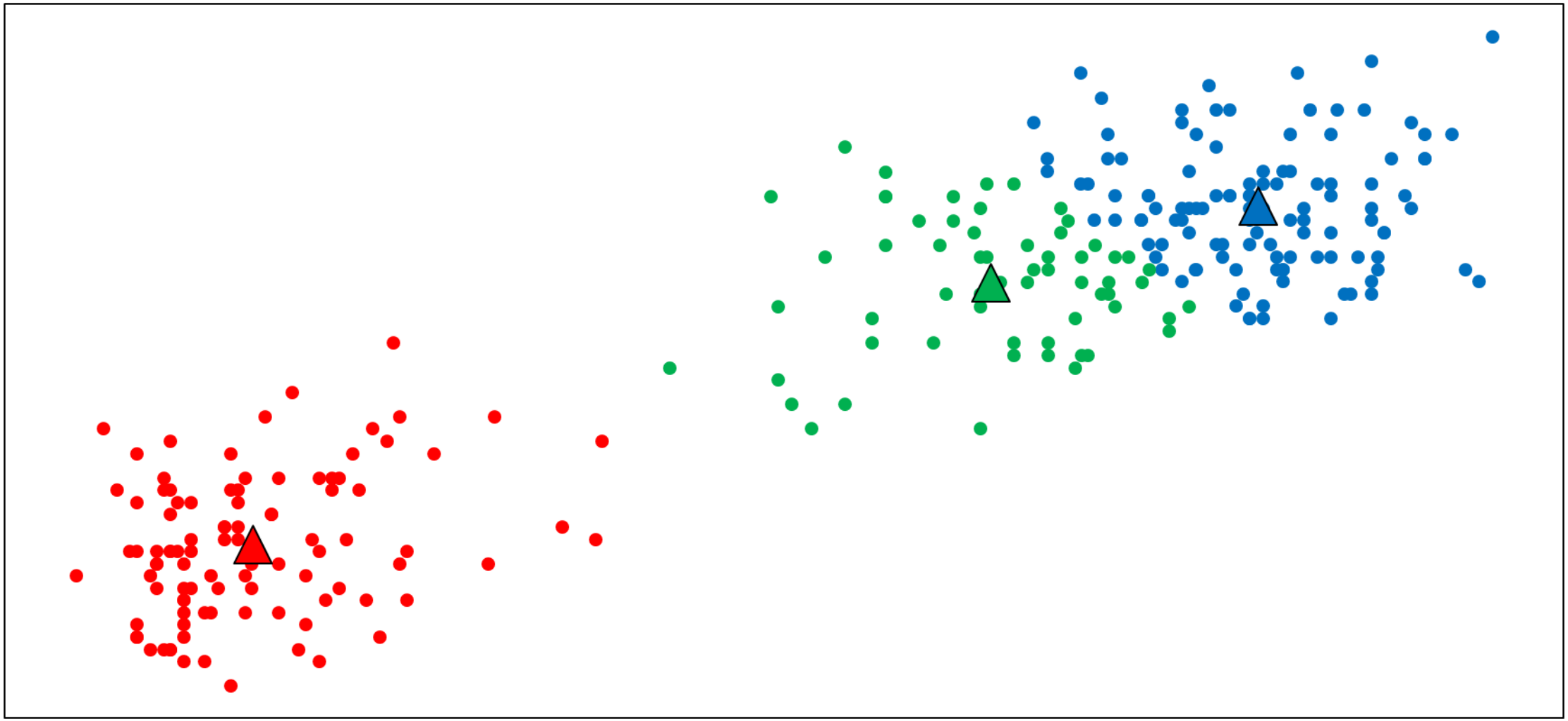
Iteración 3: Actualizamos los centros



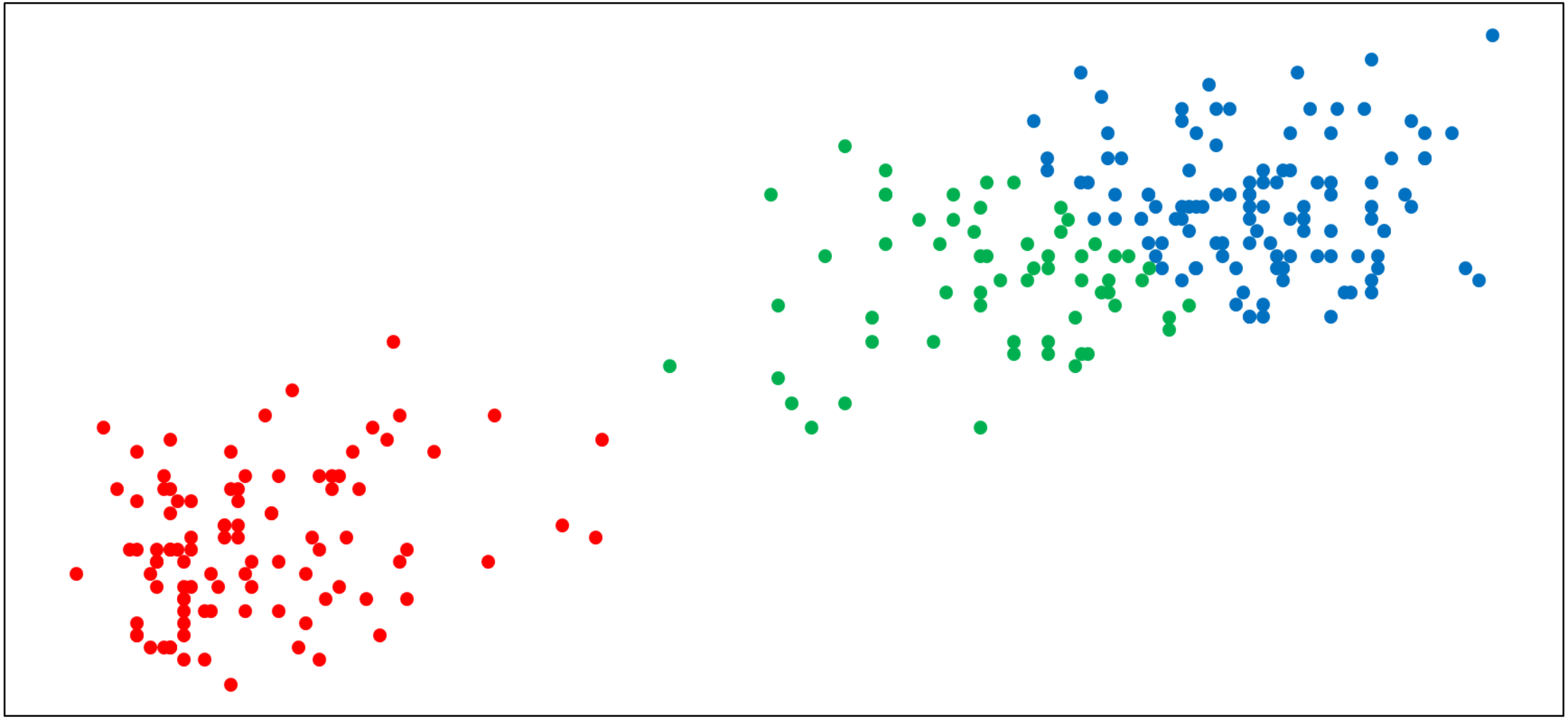
Iteración 4: Asignamos los datos a cada clúster



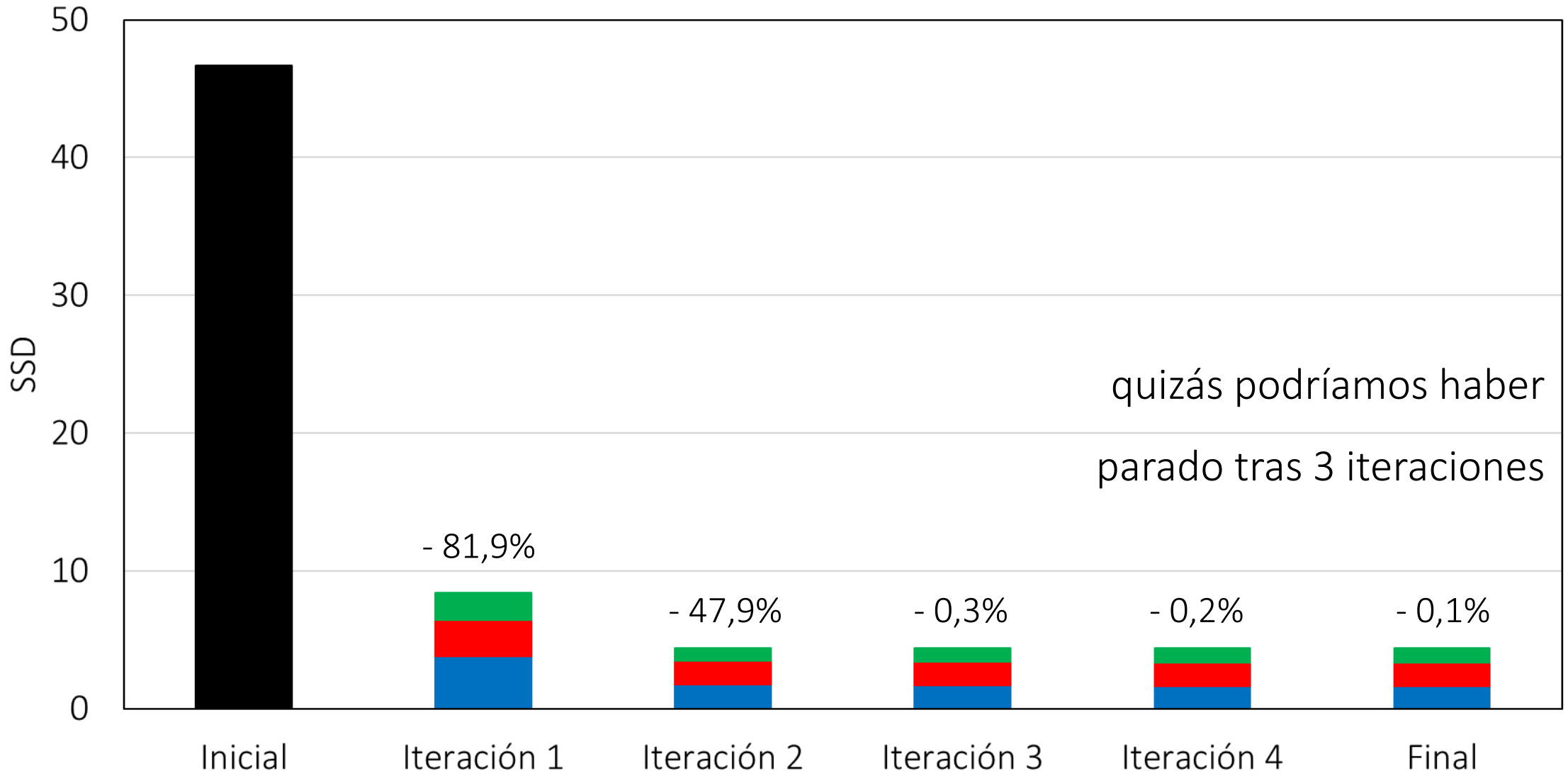
Iteración 4: Actualizamos los centros



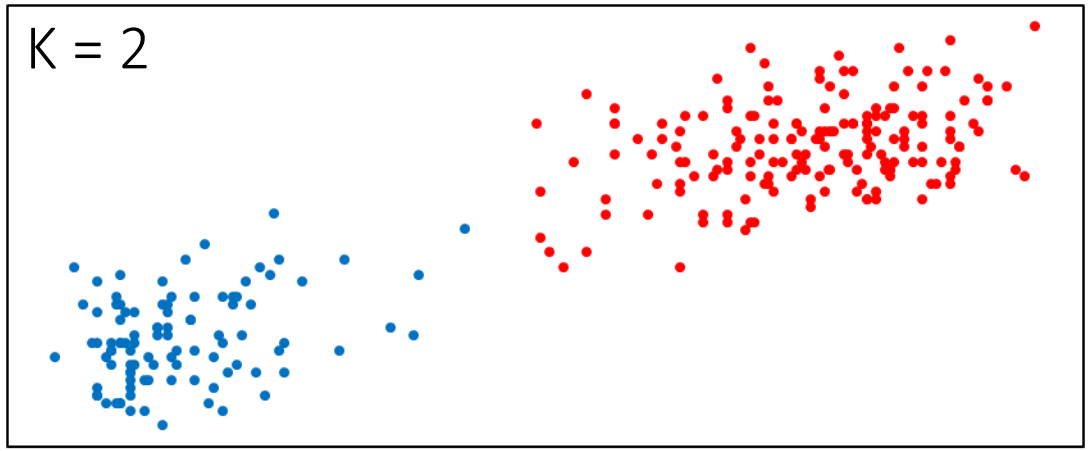
La asignación de datos a cada clúster no cambia, convergimos

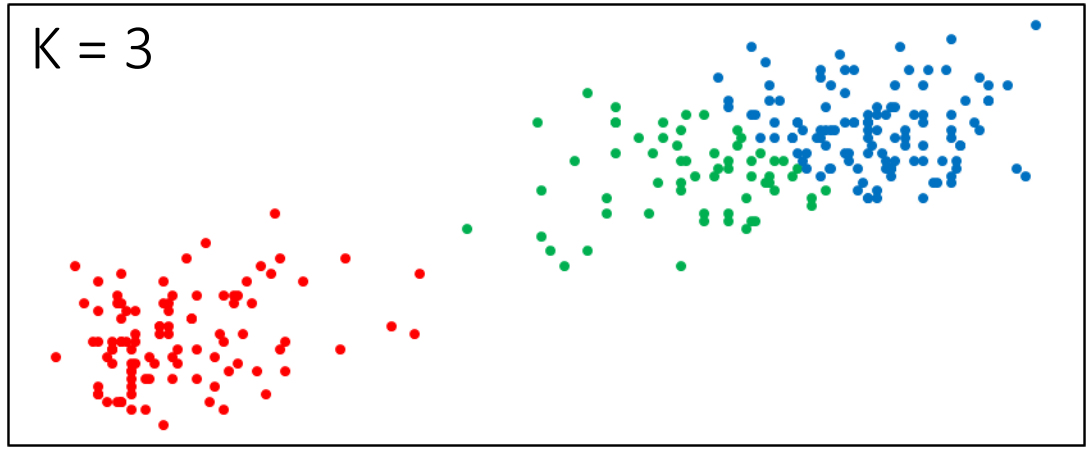


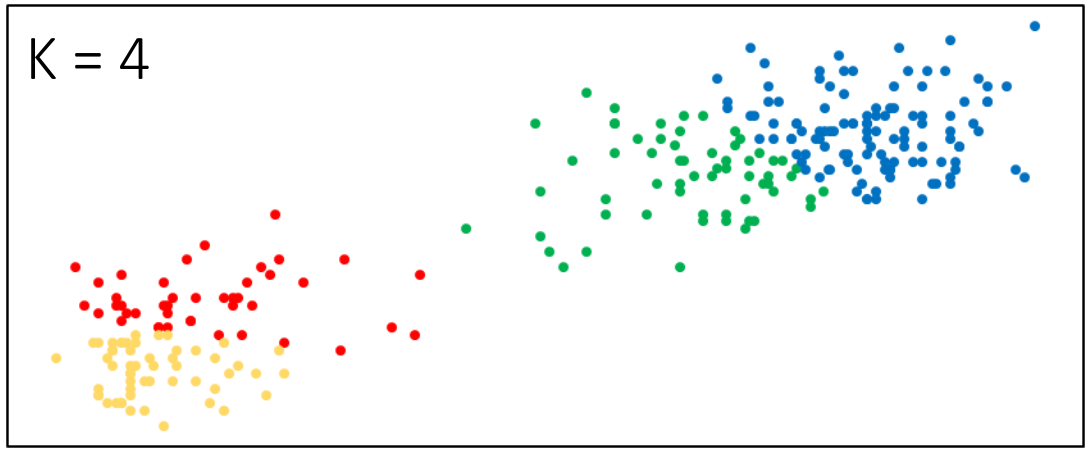
Suma cuadrática de distancias en cada iteración

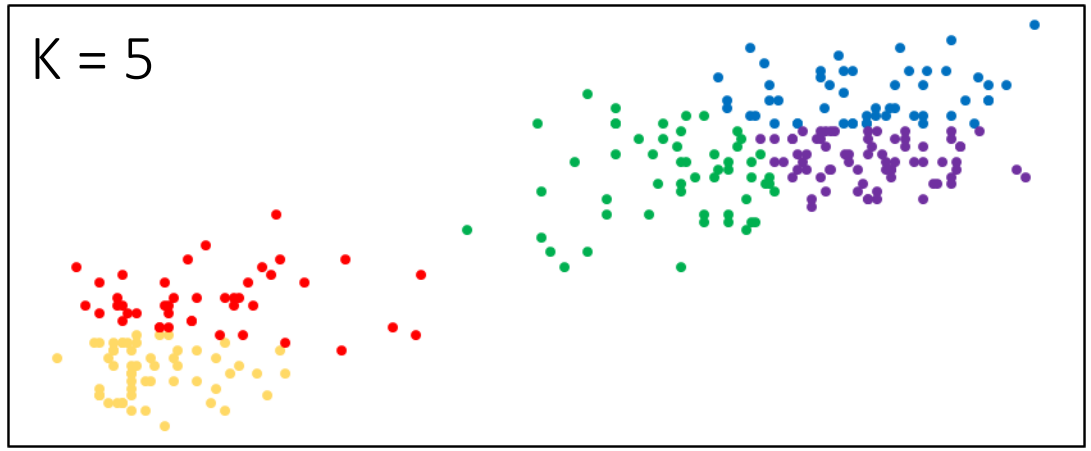


¿Cuánto cambia el resultado según K?









¿Cómo elegir K?

Existen distintos enfoques

Interpretabilidad de los resultados

Según la suma cuadrática de las distancias (i.e. el codo)

Según la cohesión y separación de cada clúster (i.e. la silueta)

El codo

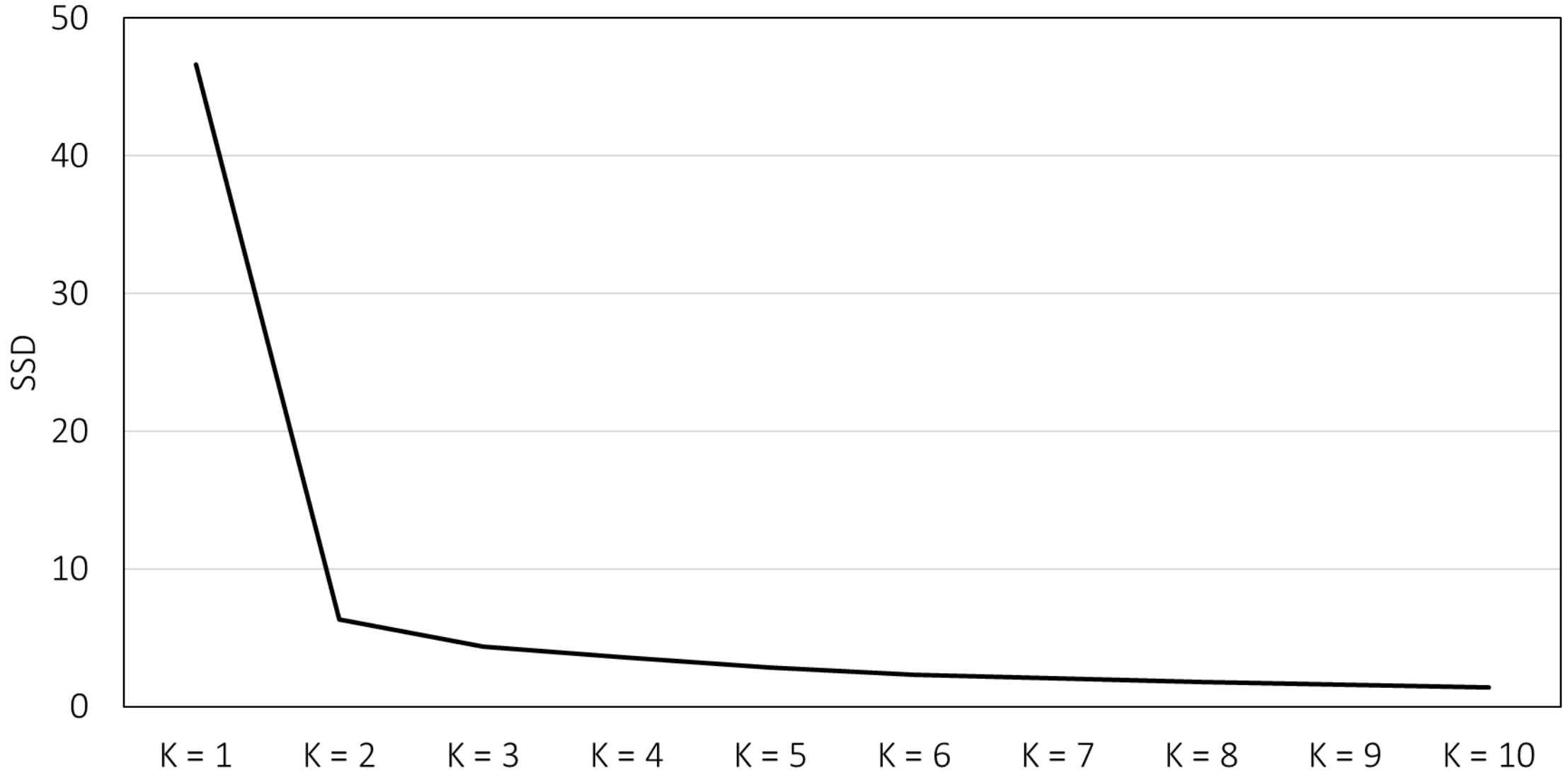
A medida que se aumenta K, la suma cuadrática de distancias decrece

Las ganancias en términos de SSD son marginalmente decrecientes

(i.e. cada clúster adicional genera ganancias menores que el anterior)

Seleccionamos K tal que las ganancias compensen la complejidad adicional

El codo



La silueta

Este indicador mide cuán similar es cada observación a su clúster (cohesión) y cuán distinta es de los otros clústeres (separación)

El rango de la silueta va de -1 a 1, mientras mayor es el valor:

- La observación se empareja adecuadamente con su clúster
- La observación se empareja pobremente con los otros clústeres

Un bajo valor de la silueta se puede deber a muchos o pocos clústeres

La silueta

Supongamos que tenemos K clústeres, para cada observación *i* que pertenece al clúster C_i calculamos la distancia promedio a las otras observaciones del clúster C_i

$$a(i) = \frac{1}{|C_i| - 1} \cdot \sum_{\substack{j \in C_i \\ j \neq i}} d(i,j)$$

Mientras menor sea $a(i)$, mejor se empareja la observación *i* con su clúster

La silueta

De igual manera, para cada observación *i* que pertenece al clúster C_i calculamos la menor distancia promedio a las observaciones de los otros clúster C_k

$$b(i) = \underset{k}{\text{Min}} \left\{ \frac{1}{|C_k|} \cdot \sum_{j \in C_k} d(i,j) \right\}$$

Mientras mayor sea $b(i)$, peor se empareja la observación *i* con su clúster vecino (i.e. aquel más cercano)

La silueta

Finalmente, para cada observación *i* calculamos su valor de silueta

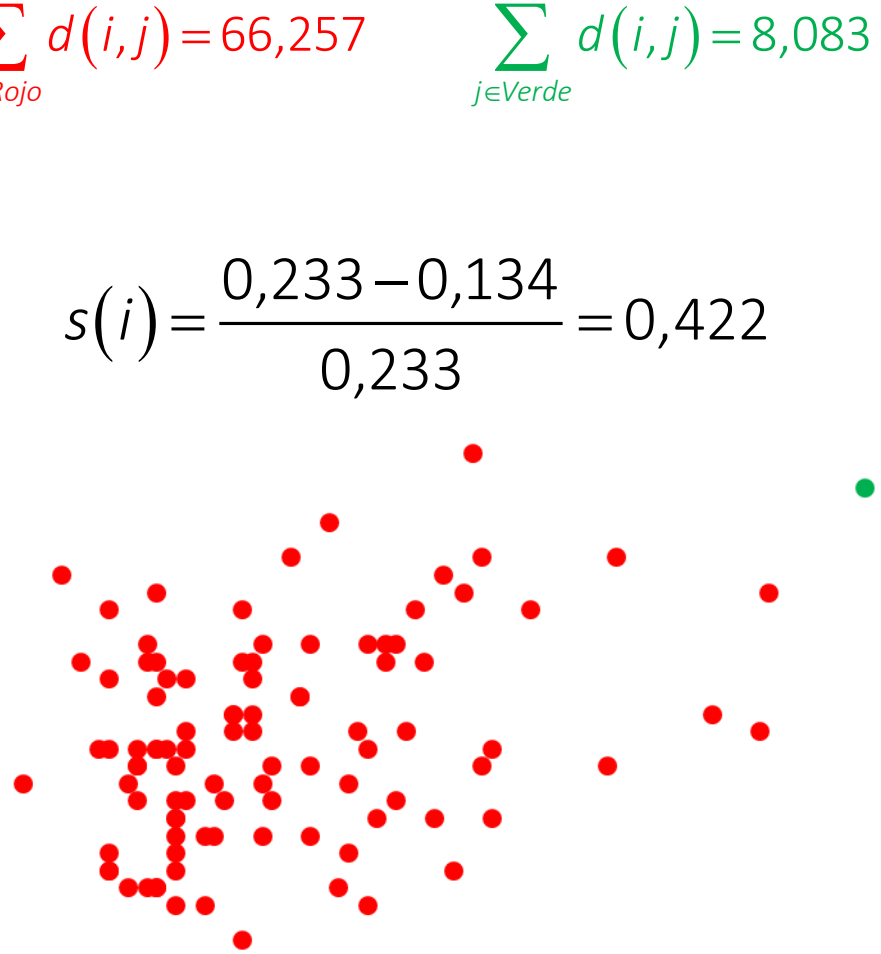
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

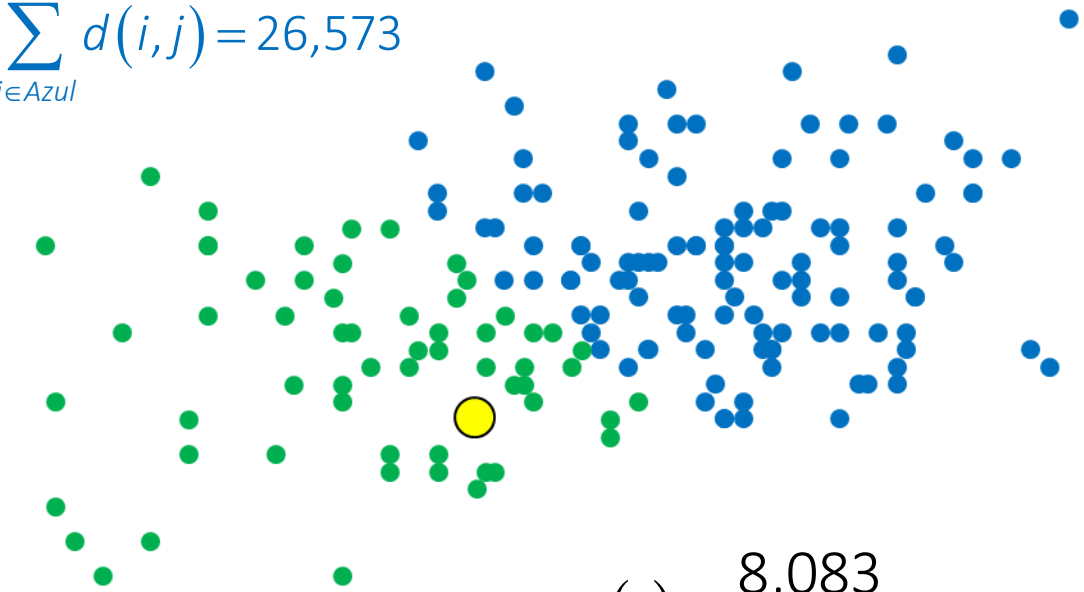
Se cumple que $-1 \leq s(i) \leq 1$

Podemos analizar el valor promedio o la distribución de valores de silueta

Analicemos una observación

$$\sum_{j \in Rojo} d(i,j) = 66{,}257 \qquad \sum_{j \in Verde} d(i,j) = 8{,}083 \qquad \sum_{j \in Azul} d(i,j) = 26{,}573$$



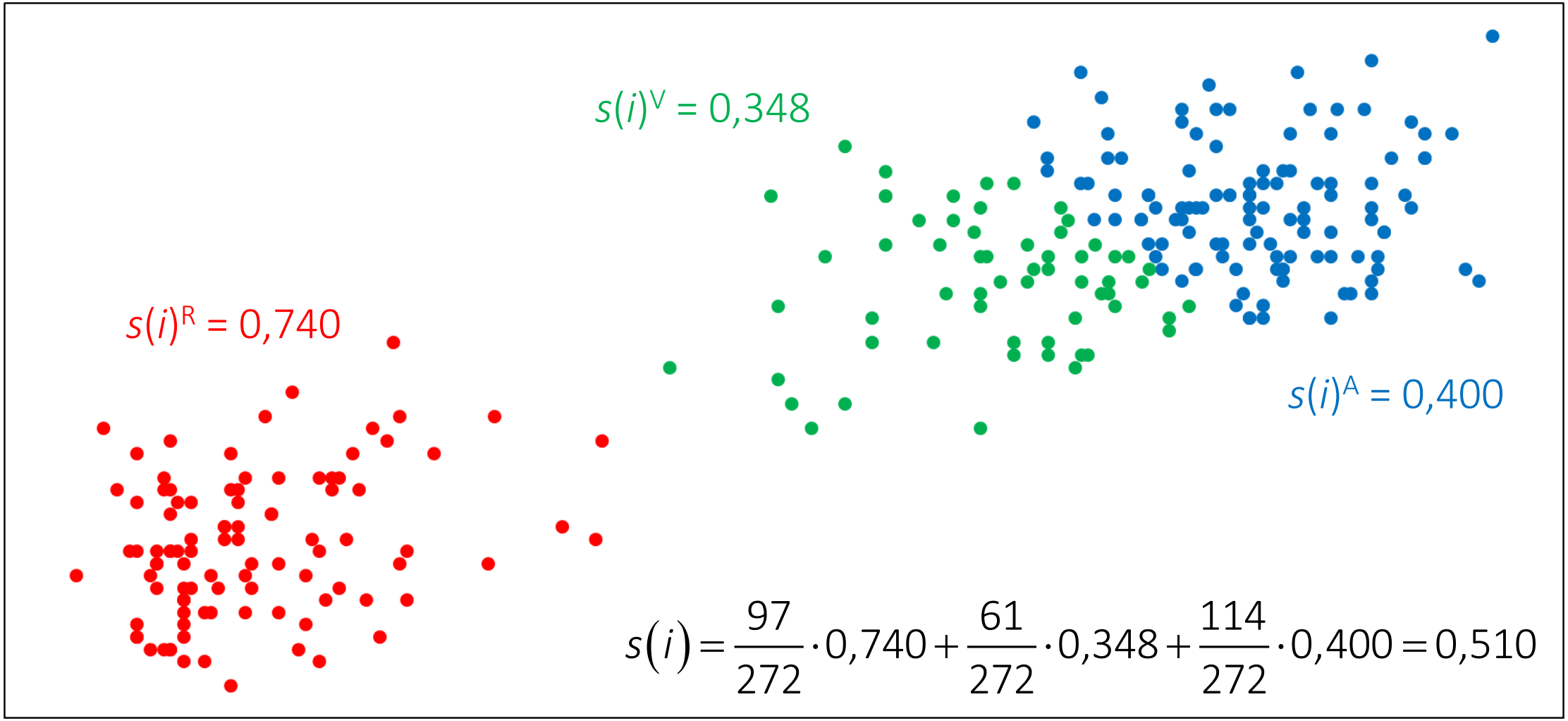


$$s(i) = \frac{0{,}233 - 0{,}134}{0{,}233} = 0{,}422$$

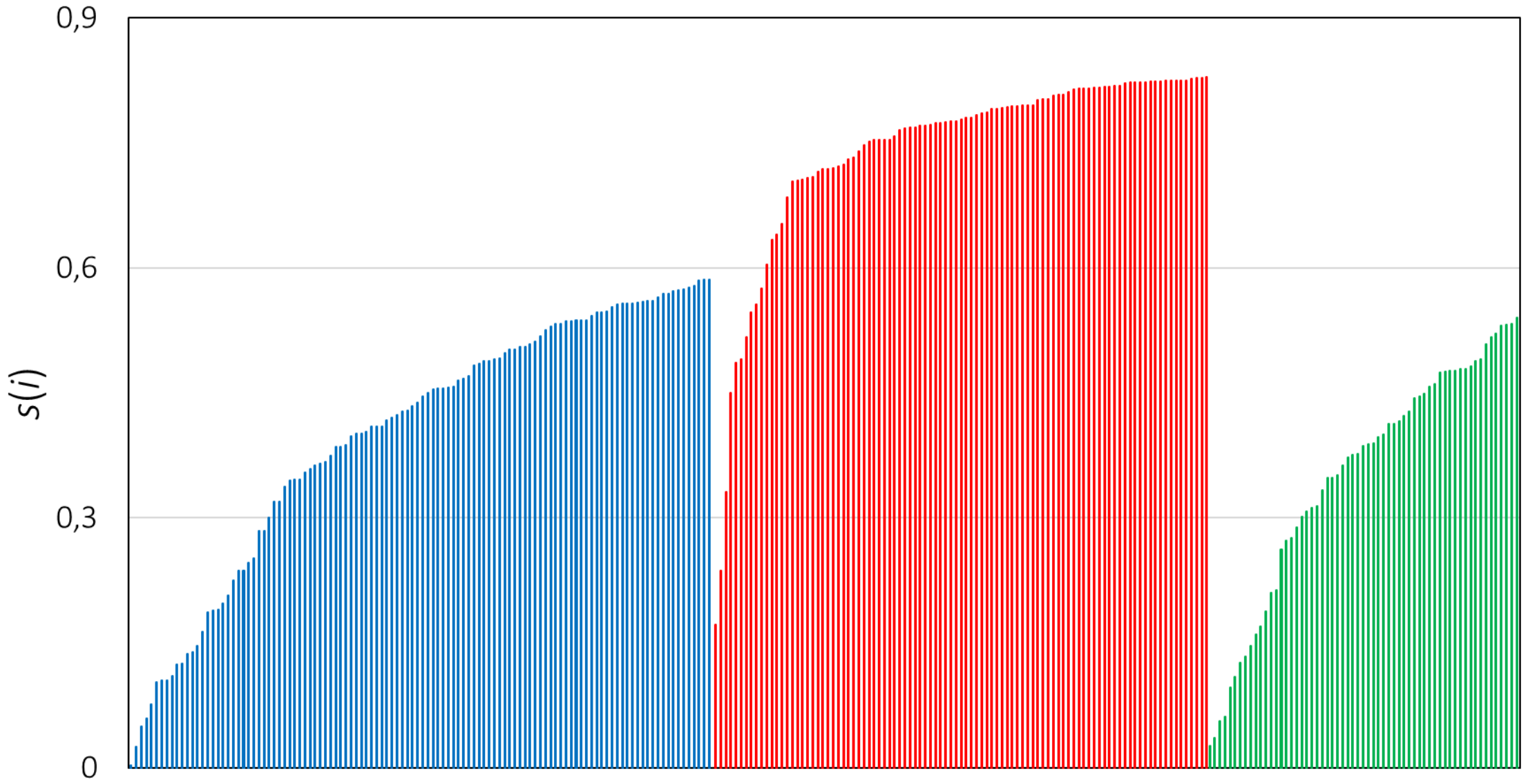
$$a(i) = \frac{8{,}083}{61-1} = 0{,}134$$

$$b(i) = \text{Min}\left\{\frac{26{,}573}{114}; \frac{66{,}257}{97}\right\} = 0{,}233$$

Valores promedio de siluetas con K = 3



Distribución de siluetas con K = 3



La silueta