

Lecture 6

Clustering and k-mean

Instructor: Michela Taufer



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

BIG ORANGE. BIG IDEAS.®

Lesson Outline

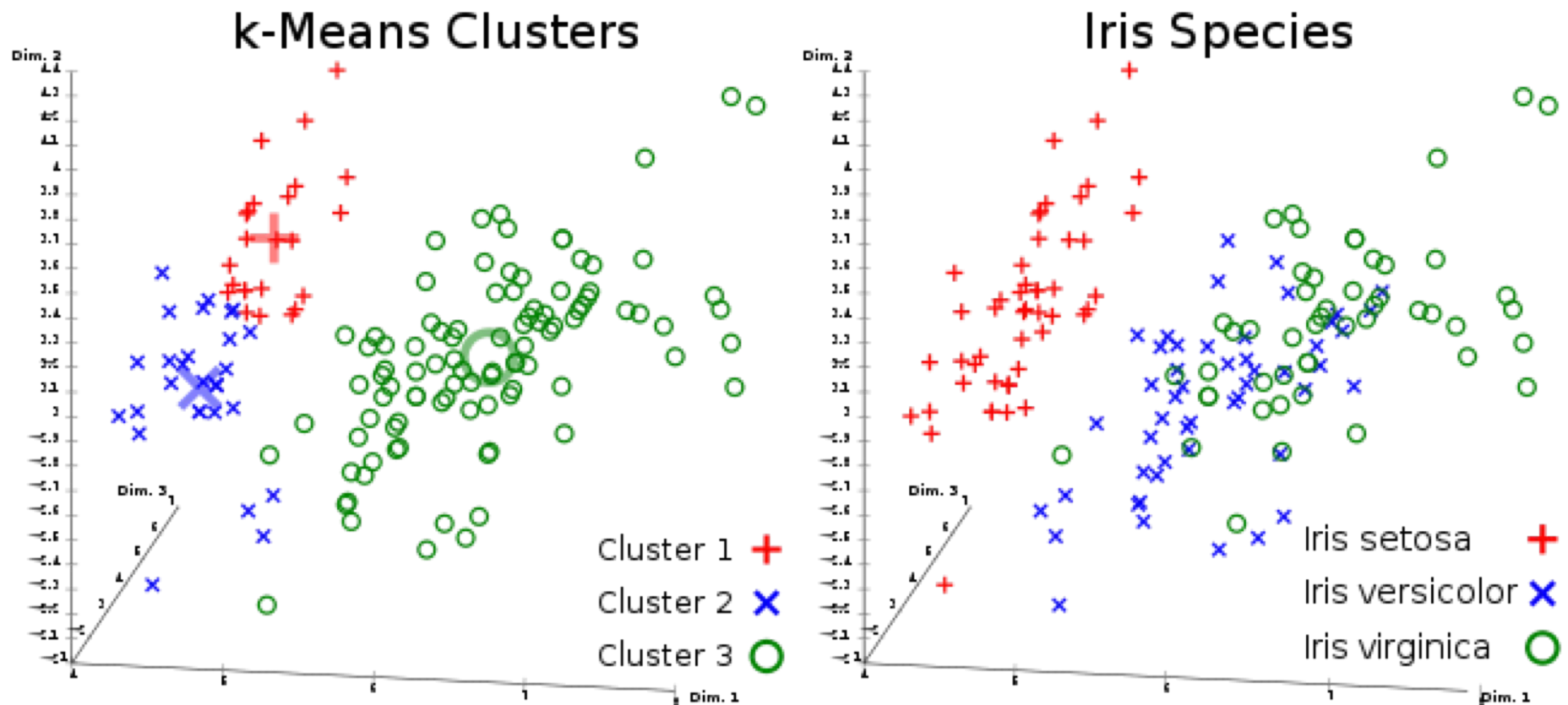
- Clustering and k-mean
- Linear regression
- Assignment 6
- Project (part 2)

Clustering

- Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other clusters
- Popular notions of clusters include **groups with small distances** between cluster members or **dense areas** of the data space
 - K-mean: partition n observations into k clusters
 - DBSCAN: density based clustering algorithm

Clustering

- Unsatisfactory k-means for clustering the Iris flower dataset



Distance Between Vectors

- Let $\mathbf{x} = (x_1, \dots, x_d)$ and $\mathbf{y} = (y_1, \dots, y_d)$ be vectors in \mathbb{R}^d
- **Question:** What is the distance from \mathbf{x} to \mathbf{y} ?
- **Answer:** It depends
 $\text{dist}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$
- There are many choices for the norm $\|\cdot\|$

▶ $\|\mathbf{x} - \mathbf{y}\|_1 := \sum_{i=1}^d |x_i - y_i|$ Manhattan distance

▶ $\|\mathbf{x} - \mathbf{y}\|_2 := \sqrt{\sum_{i=1}^d (x_i - y_i)^2}$ Euclidean distance

▶ $\|\mathbf{x} - \mathbf{y}\|_p := \left(\sum_{i=1}^d |x_i - y_i|^p \right)^{1/p}$ for fixed $1 < p < \infty$

▶ $\|\mathbf{x} - \mathbf{y}\|_\infty := \max_i |x_i - y_i|$

Estimators and Errors

- Let $x = (x_1, \dots, x_d)$ and $\mu = (\mu_1, \dots, \mu_d)$ be vectors in \mathbb{R}^d
- Suppose that μ is an estimator for x
- **Question:** How do we measure how good of an estimator μ is?
- **Answer:** $\|x - \mu\|$, the distance from x to μ
- **Question:** Is μ a good estimator for x if
 - ▶ $\|x - \mu\| < 1$?
 - ▶ $\|x - \mu\| < .1$?
 - ▶ $\|x - \mu\| < .001$?
- **Answer:** The specific application dictates what good means

Estimators and Error

- **BETTER Question:** If μ_1 and μ_2 are both estimators for x , which is better?
- **Answer:** μ_1 estimates x better than μ_2 if $\|x - \mu_1\| < \|x - \mu_2\|$

Estimators and Error

- Suppose that $S = \{x_1, \dots, x_d\}$ is a set of vectors and μ estimates S
- **Question:** How do we measure how well μ estimates S ?
- **Answer:** There are many ways we could do this

- ▶ $\text{error} = \sum_{i=1}^n \|\mathbf{x}_i - \mu\|$ (using any norm)
- ▶ $\text{SSE} = \sum_{i=1}^n \|\mathbf{x}_i - \mu\|_2^2$ (probably most common)
- ▶ $\text{error} = \max_i \|\mathbf{x}_i - \mu\|$ (also common)

SSE: sum of squared errors of prediction

K-mean clustering

- Given n data points, $\{x_1, \dots, x_n\}$ and an integer k
- Goal: Partition $\{x_1, \dots, x_n\}$ into k sets (clusters) S_1, \dots, S_k in such a way as to minimize the total **Within Cluster Sum of Squares** error:

$$WCSS = \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|_2^2$$

measure of the variability of the observations **within** each **cluster**

- where μ_i is the mean of the vectors in S_i , i.e.

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x.$$

Number of observations

- *The number of observations in each cluster in the final partition*
- Examine the number of observations in each cluster when you interpret the measures of variability, such as the average distance and the within-cluster sum of squares. The **variability** of a cluster may be **affected by its having a smaller or larger number of observations**. For example, the within-cluster sum of squares becomes larger as more observations are added.
- Examine clusters that have significantly fewer observations than other clusters. Clusters that **have very few observations may contain outliers or unusual observations with unique characteristics**.

Cluster centroid

- ***The middle of a cluster.*** A centroid is a vector that contains one number for each variable, where each number is the mean of a variable for the observations in that cluster. The centroid can be thought of as the ***multi-dimensional average of the cluster.***
- Use the cluster centroid as a general measure of cluster location and to help interpret each cluster
- Each centroid can be seen as **representing the "average observation" within a cluster** across all the variables in the analysis

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x.$$

Average distance from centroid

- *The average of the distances from observations to the centroid of each cluster*
- The average distance from observations to the cluster centroid is a **measure of the variability of the observations within each cluster**
- A cluster that has a **smaller average distance** is more compact than a cluster that has a larger average distance
- Clusters that have **higher values** exhibit greater variability of the observations within the cluster

Maximum distance from centroid

- *The maximum of the distances from observations to the centroid of each cluster.*
- The maximum distance from observations to the cluster centroid is a measure of the **variability of the observations within each cluster**
- A **higher maximum value**, especially in relation to the average distance, indicates an observation in the cluster that lies farther from the cluster centroid

Within cluster sum of squares

- *The sum of the squared deviations from each observation and the cluster centroid*
- The within-cluster sum of squares is a measure of the **variability of the observations within each cluster.**
 - A cluster that has a small sum of squares is more compact than a cluster that has a large sum of squares
 - Clusters that have higher values exhibit greater variability of the observations within the cluster

$$WCSS = \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|_2^2$$

Within cluster sum of squares

- The within-cluster sum of squares is ***influenced by the number of observations***
 - As the number of observations increases, the sum of squares becomes larger
- The within-cluster sum of squares is often **not directly comparable across clusters** with different numbers of observations
- To compare the within-cluster variability of different clusters, use the average distance from centroid instead

Distances between cluster centroids

- *The distances between cluster centroids measures how far apart the centroids of the clusters in the final partition are from one another.*
- The distance values are not very informative by themselves but **they are important when considered all together** to see how different the clusters are from each other
- A larger distance generally indicates a greater difference between the clusters

k-means clustering

We have data points x_1, \dots, x_n ,
initial clusters $S_1^{(0)}, \dots, S_k^{(0)}$,
and initial means of clusters $\mu_1^{(0)}, \dots, \mu_k^{(0)}$.

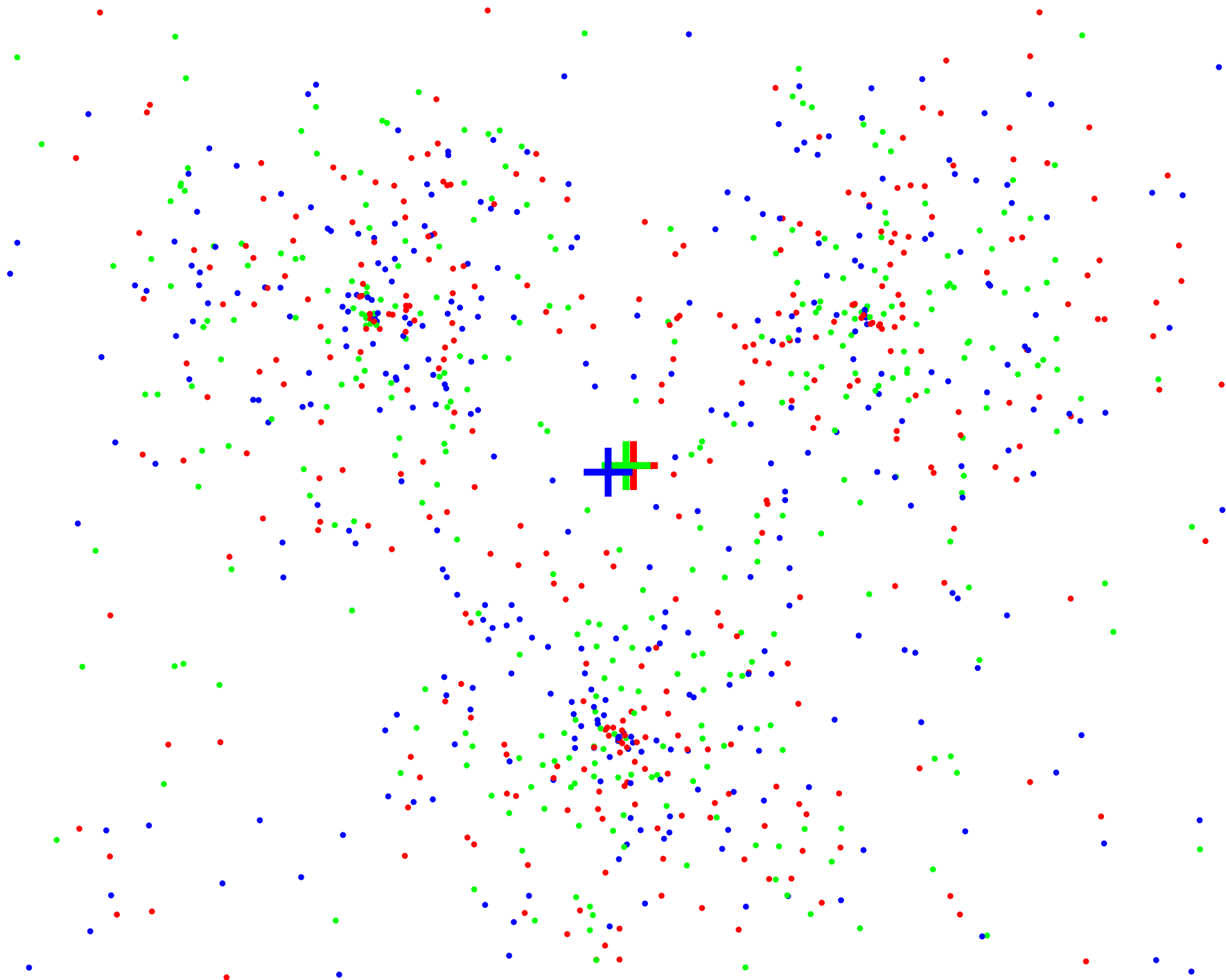
- ▶ **Step 1: Assignment.** (Iteration number t)
Put x_i in cluster $S_j^{(t)}$ if $\mu_j^{(t-1)}$ is the mean closest to x_i .
(Do this for every data point x_i .)
- ▶ **Step 2: Update the means.**
Compute the mean, $\mu_i^{(t)}$, of all the vectors in cluster $S_i^{(t)}$.
- ▶ **Repeat.**
Go back to Step 1 (iteration $t + 1$) unless none of the clusters changed.

k-means clustering

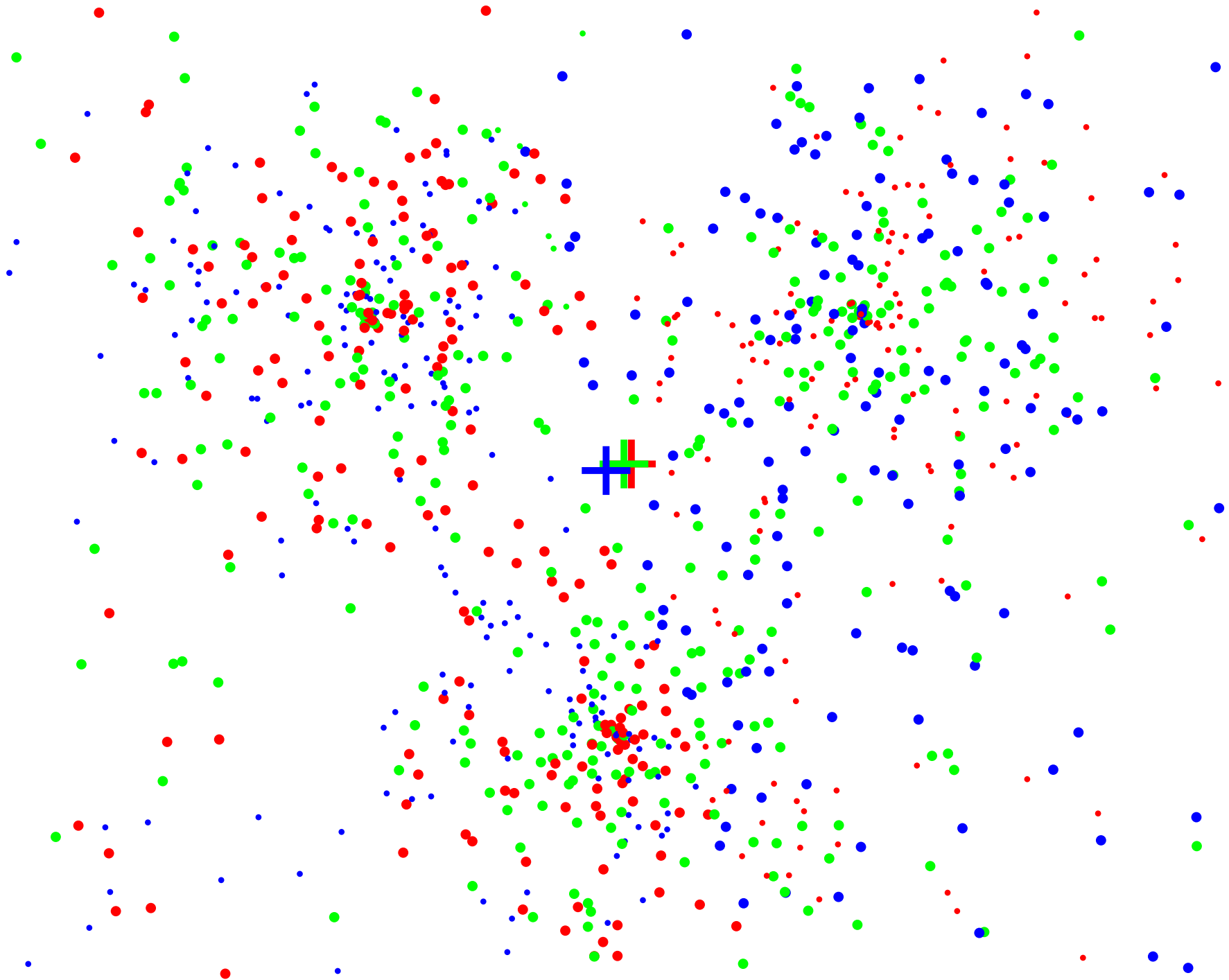
Initialization:

There are two standard ways to start the algorithm.

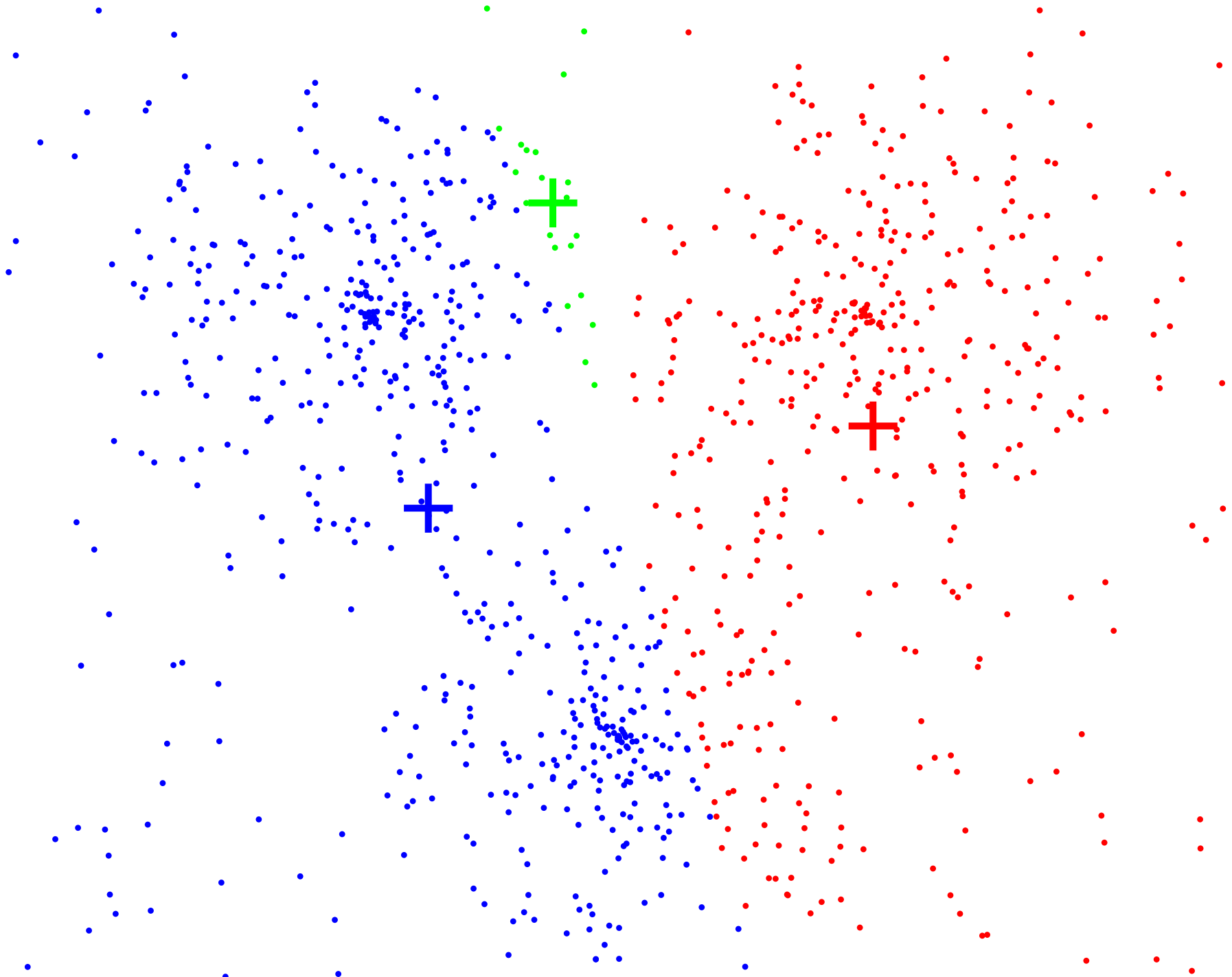
- ▶ Randomly assign each x_i to a cluster $S_j^{(0)}$ and then compute the means of each cluster.
- ▶ Set each initial mean, $\mu_j^{(0)}$, to a (distinct) random data point.



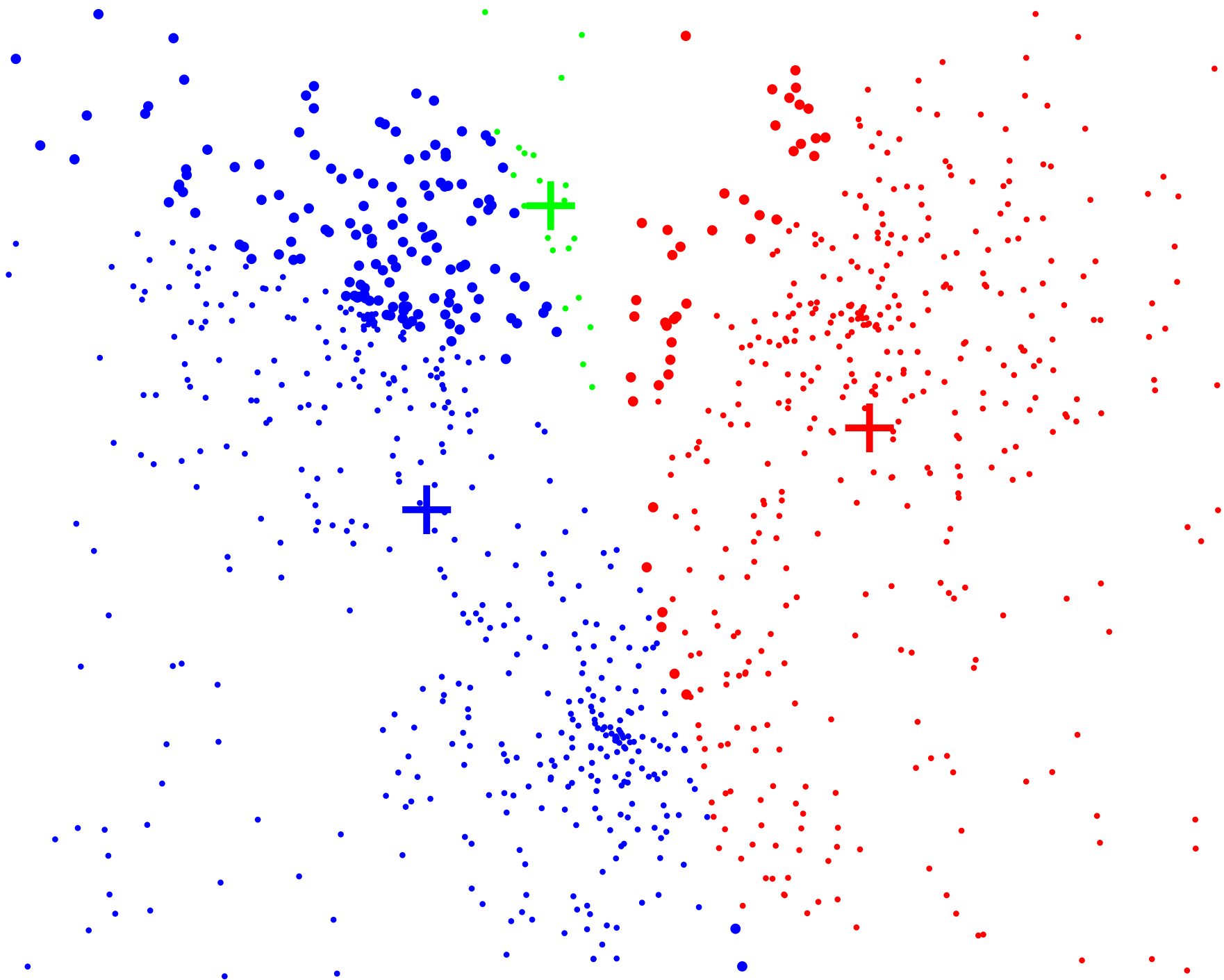
k-Clustering with $k = 3$, initial assignment



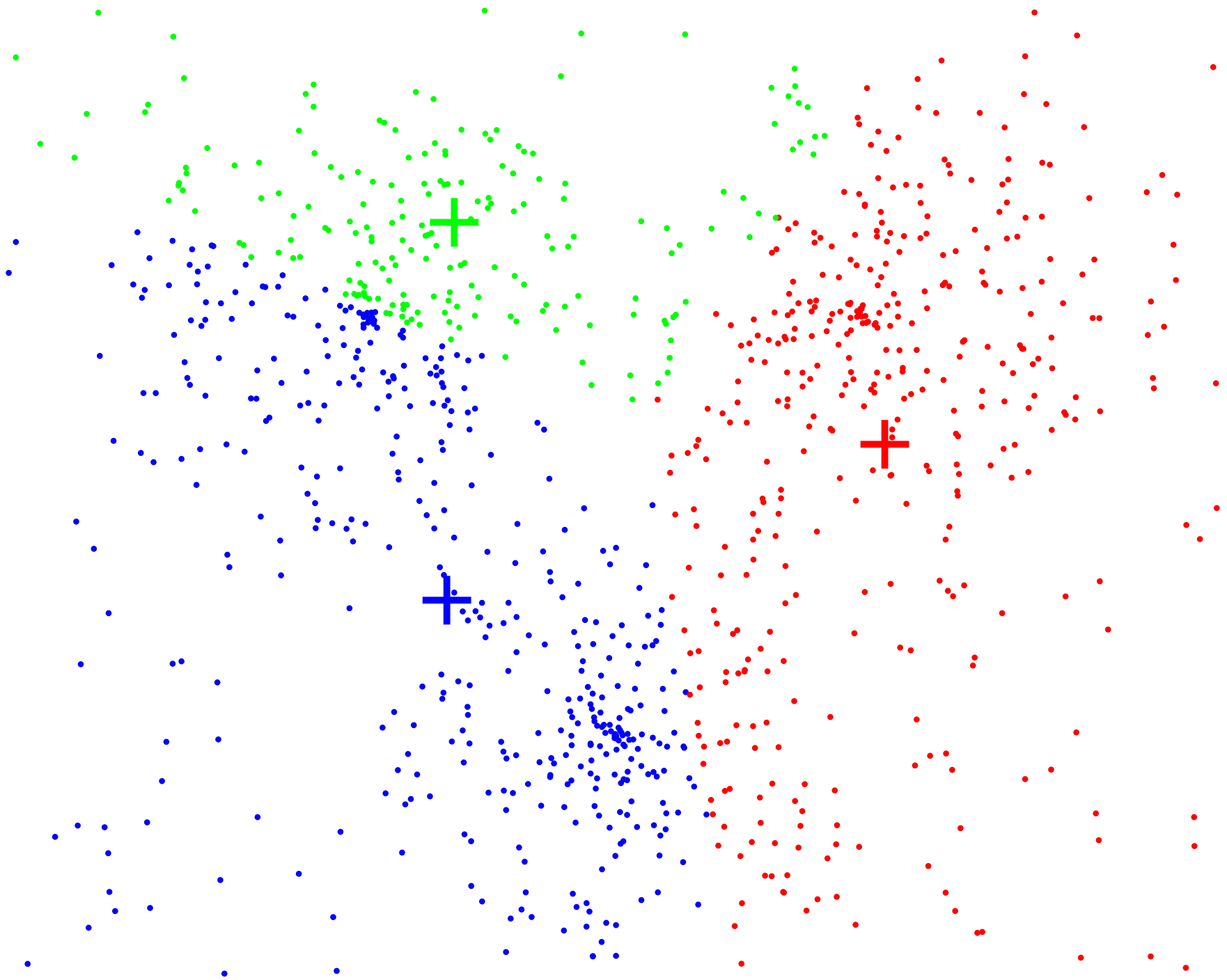
k-Clustering with $k = 3$, points changing cluster



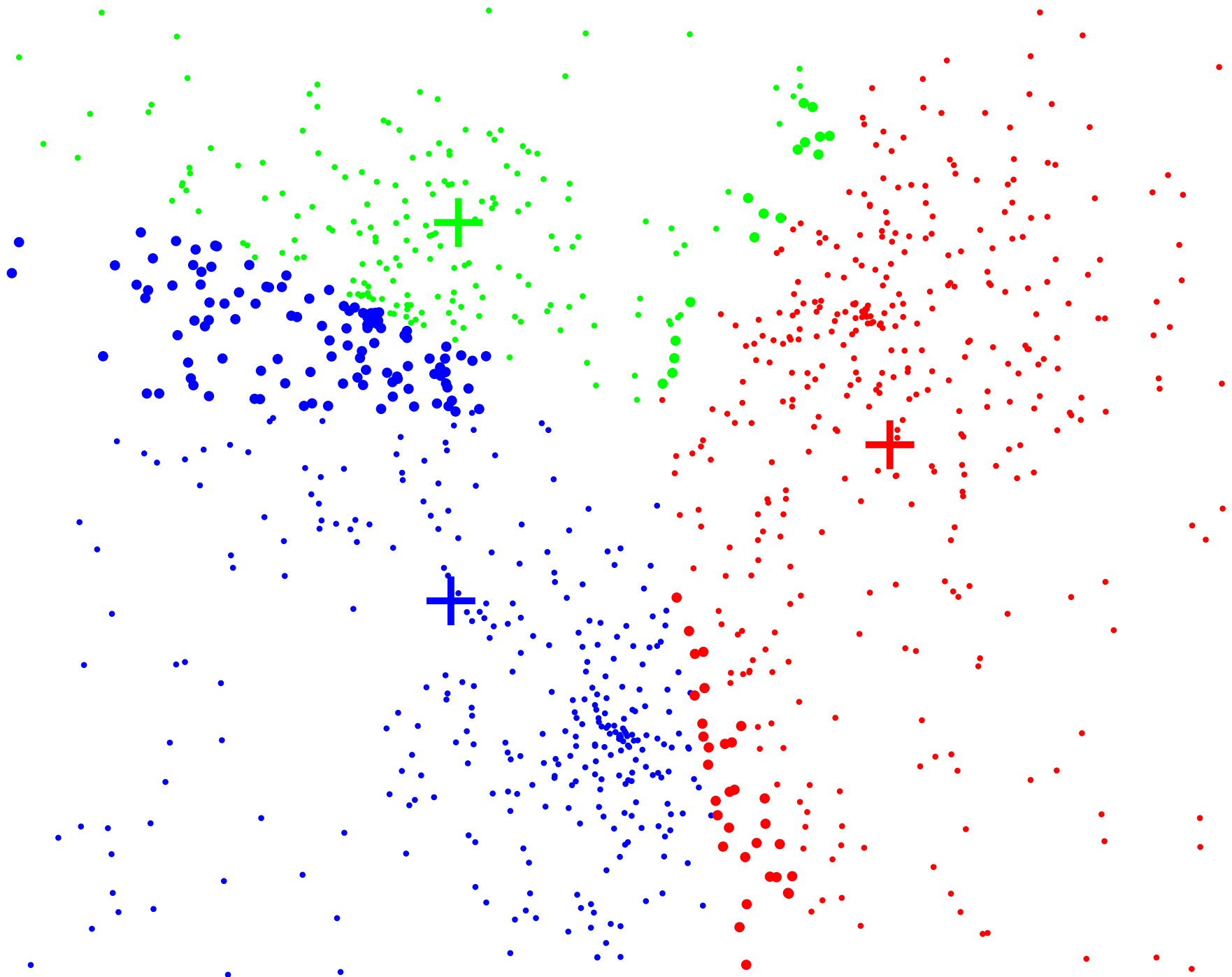
k-Clustering with $k = 3$, after rst iteration



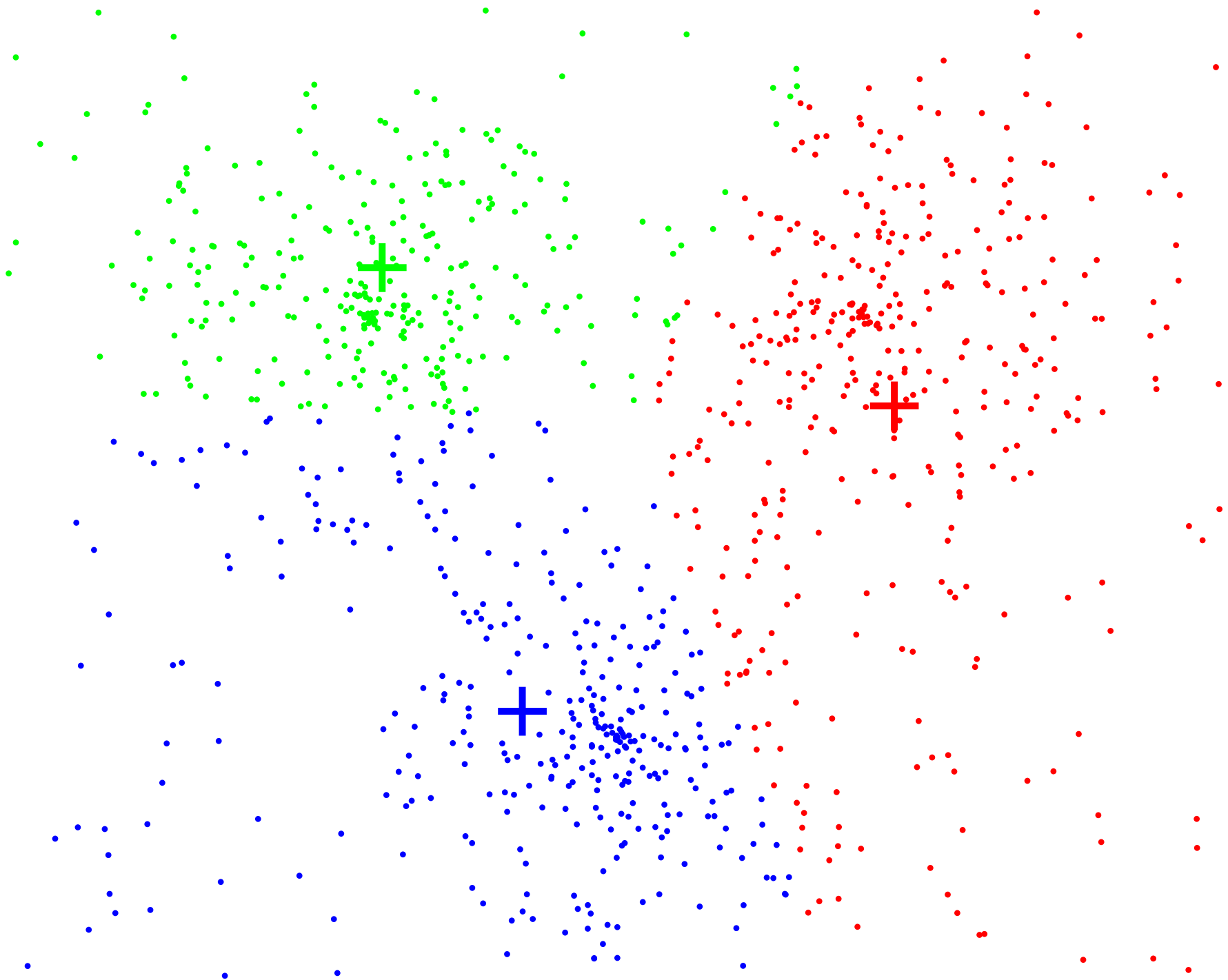
k-Clustering with $k = 3$, points changing cluster



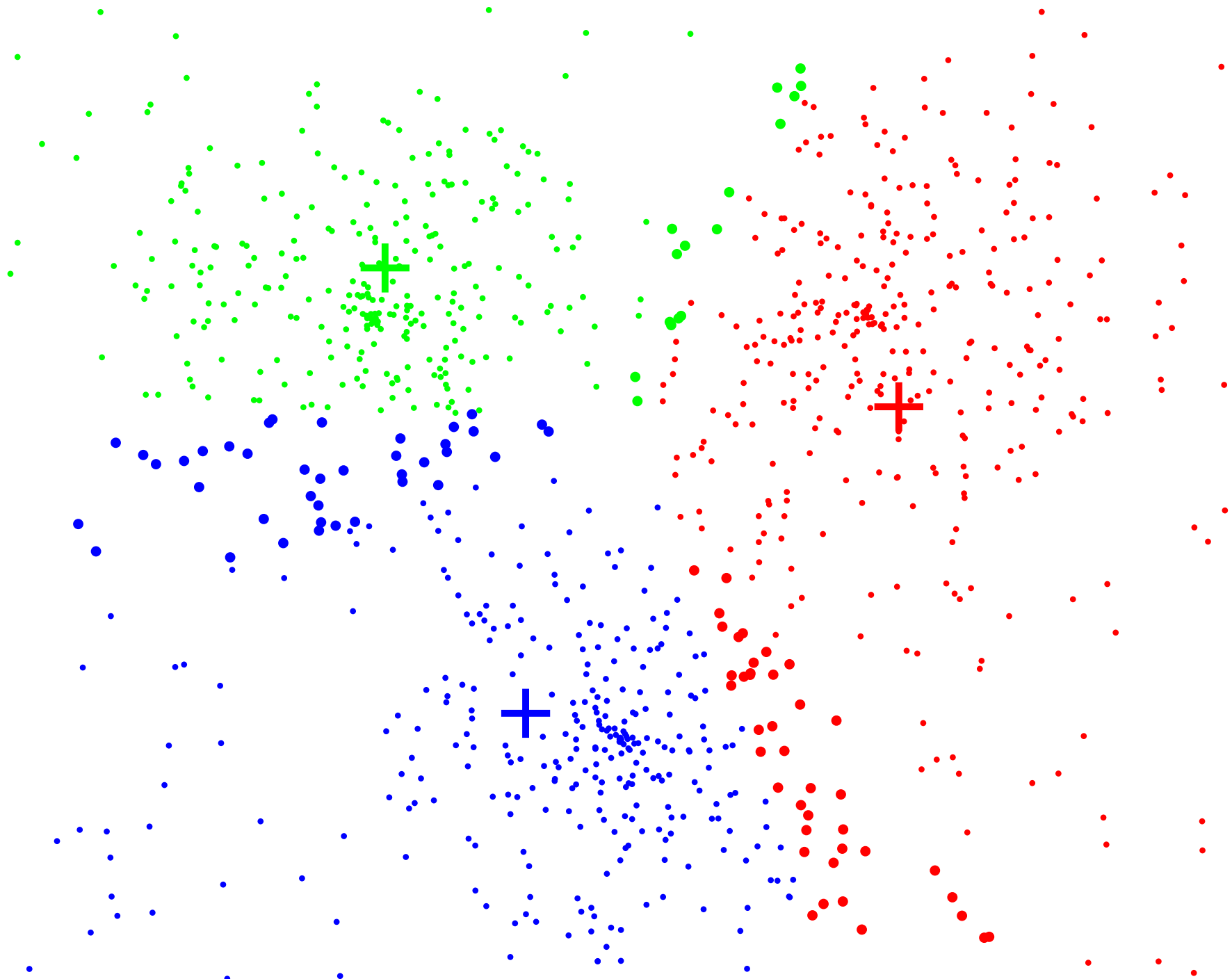
k-Clustering with $k = 3$, after second iteration



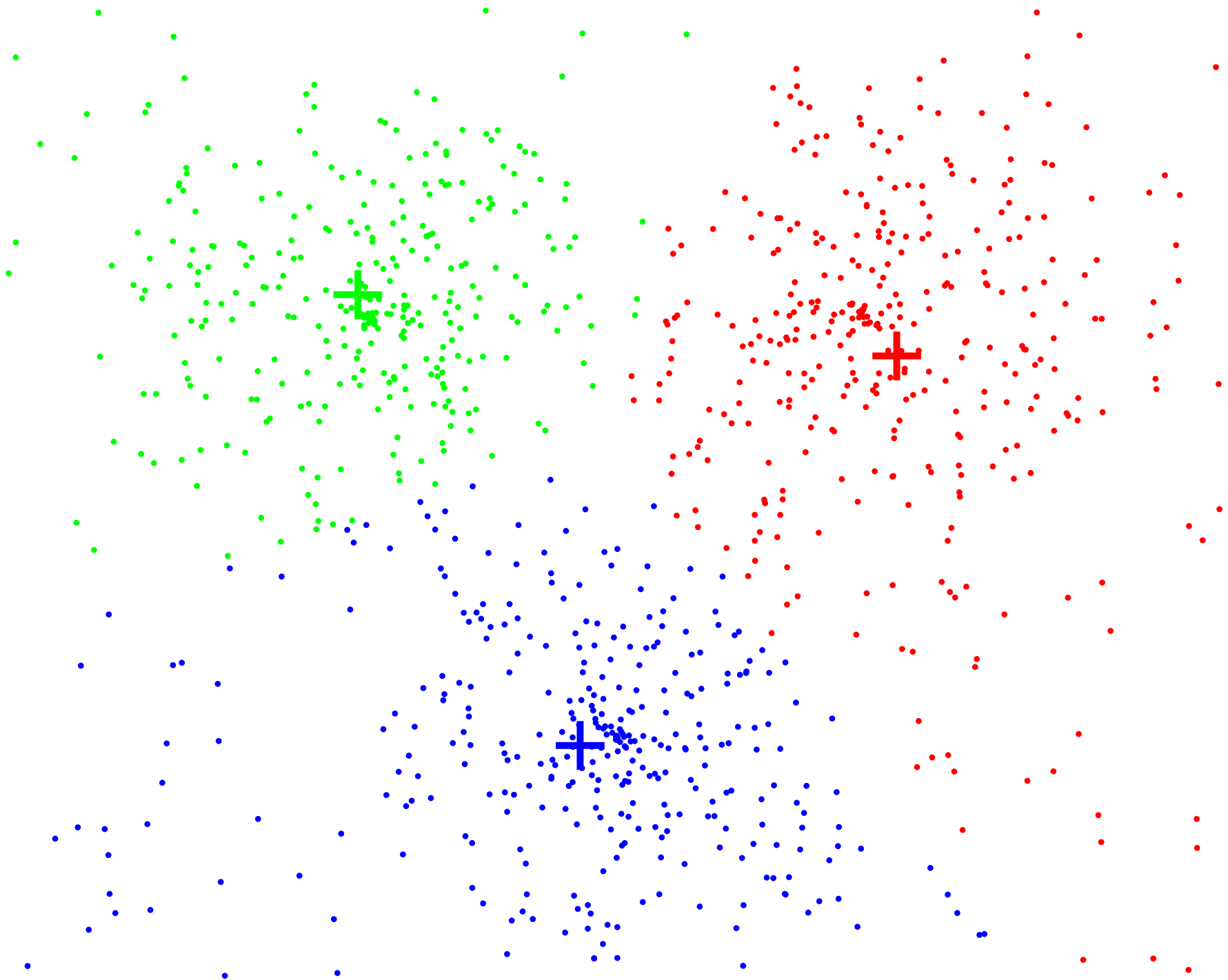
k-Clustering with $k = 3$, points changing cluster



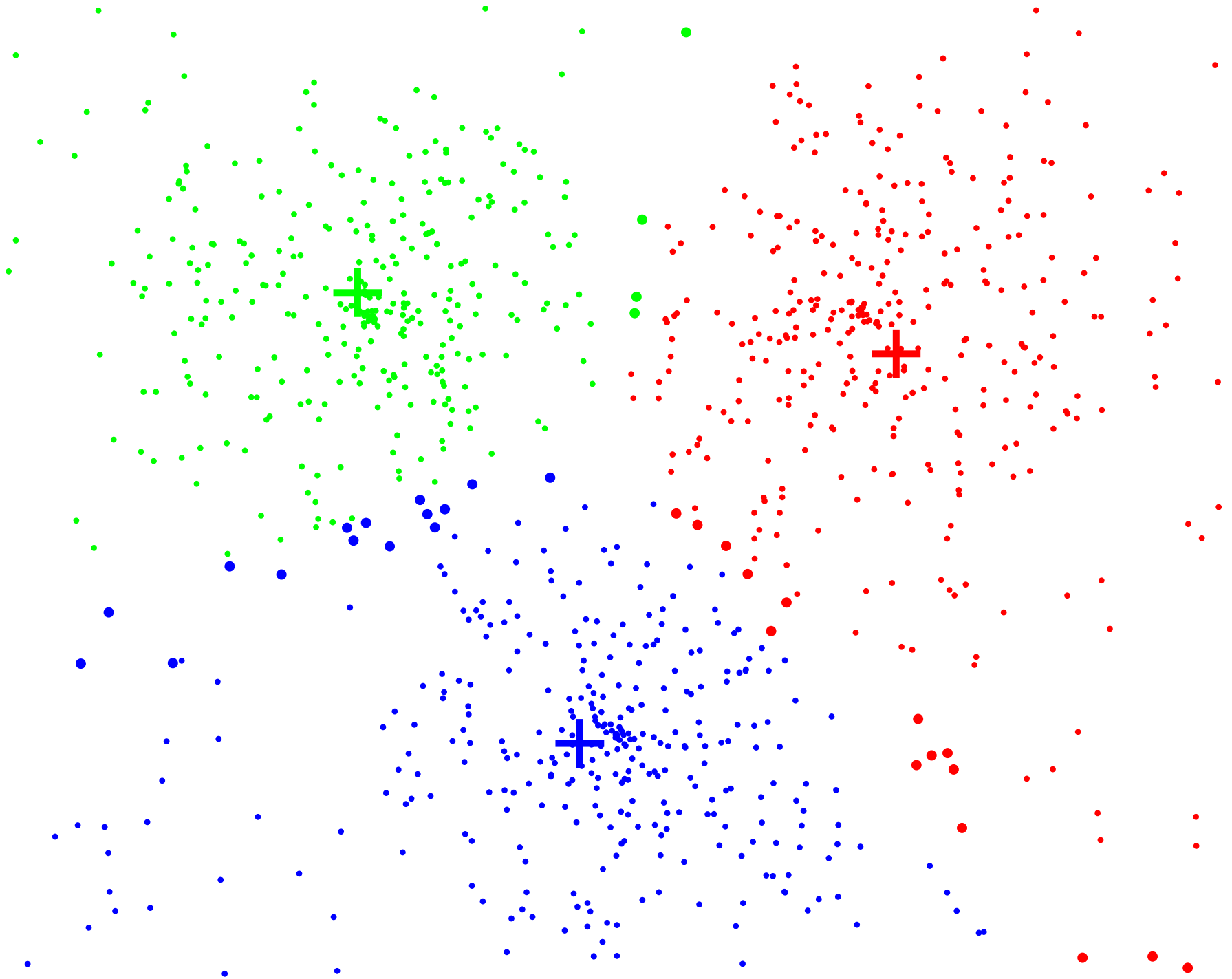
k-Clustering with $k = 3$, after third iteration



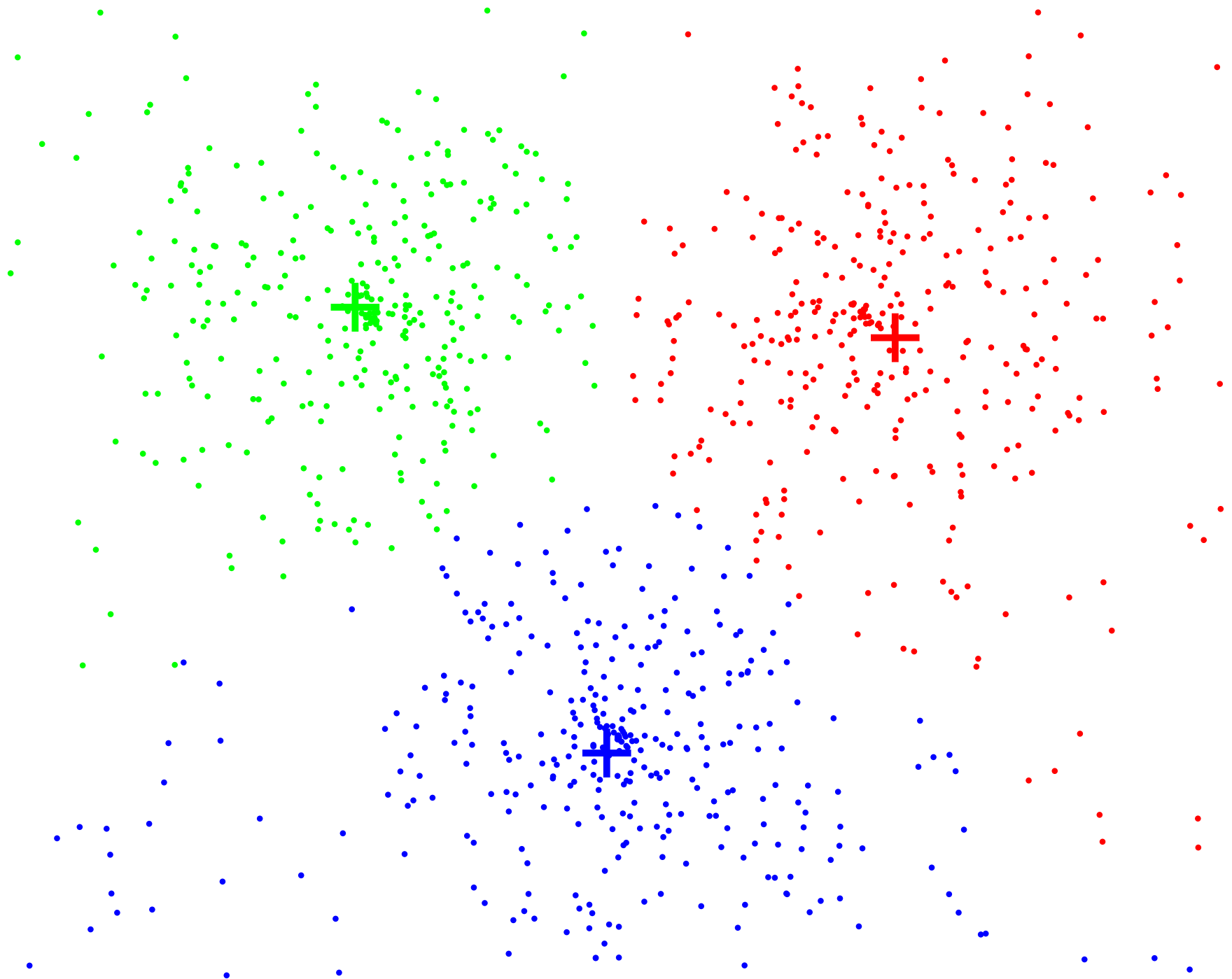
k-Clustering with $k = 3$, points changing cluster



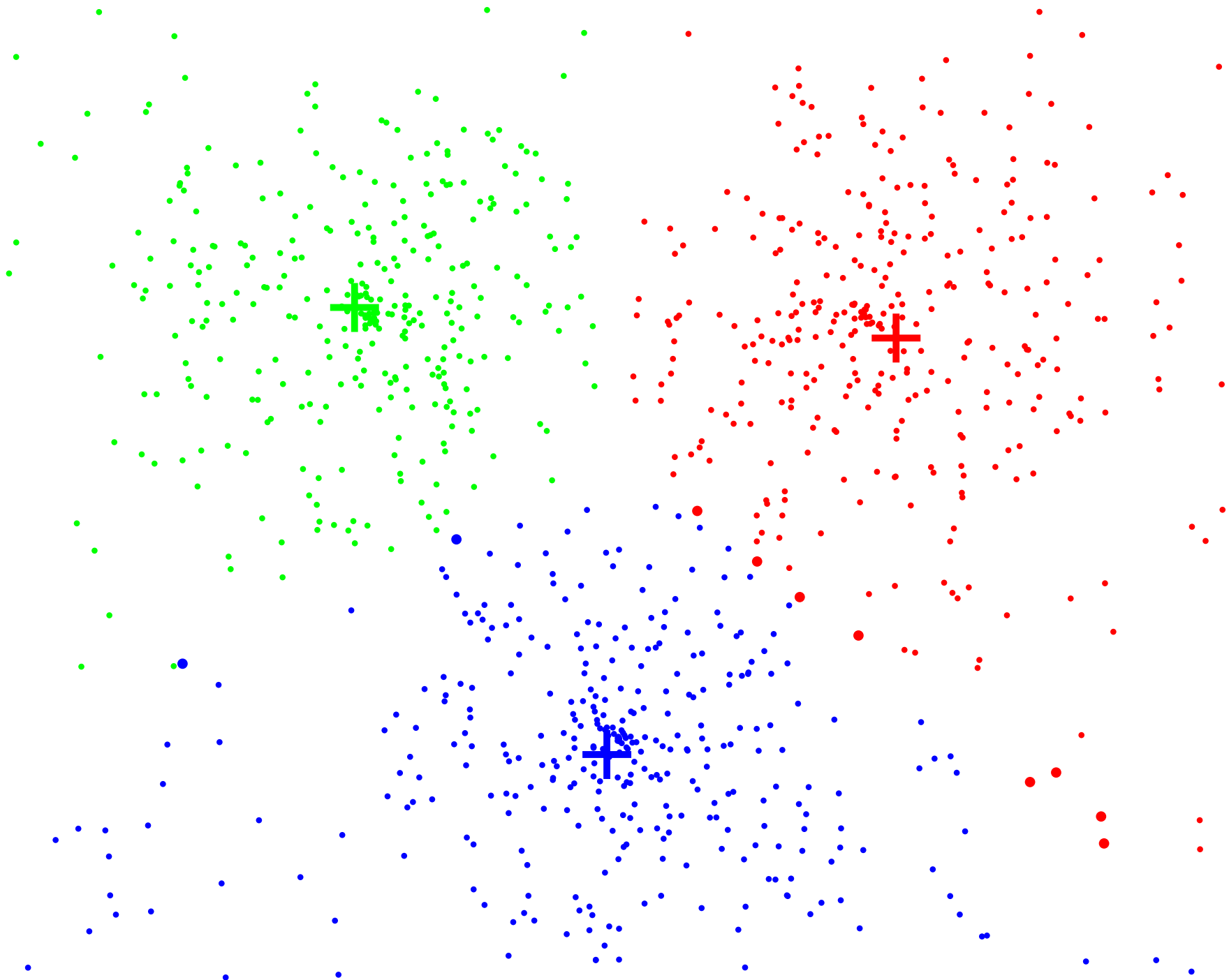
k-Clustering with $k = 3$, after fourth iteration



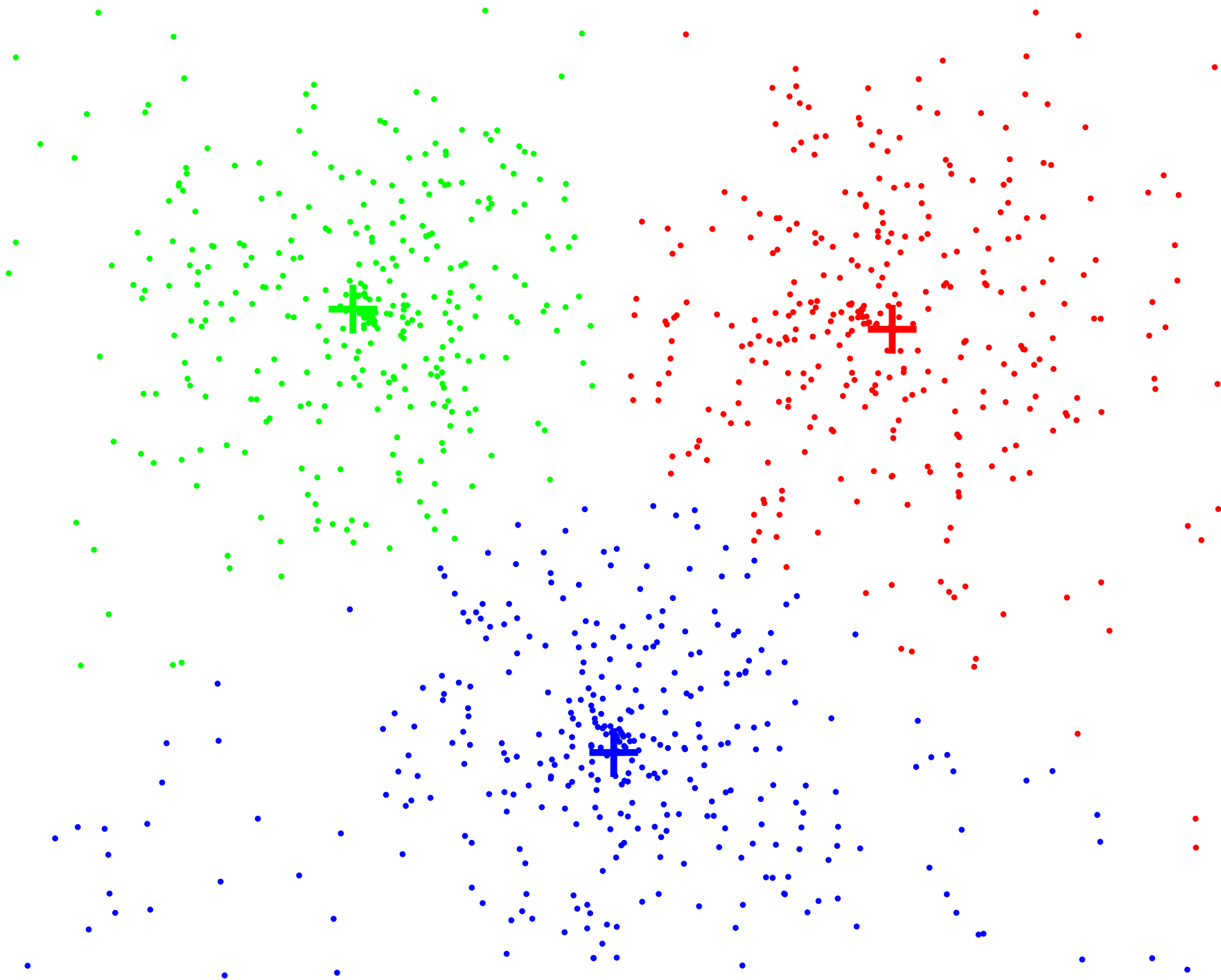
k-Clustering with $k = 3$, points changing cluster



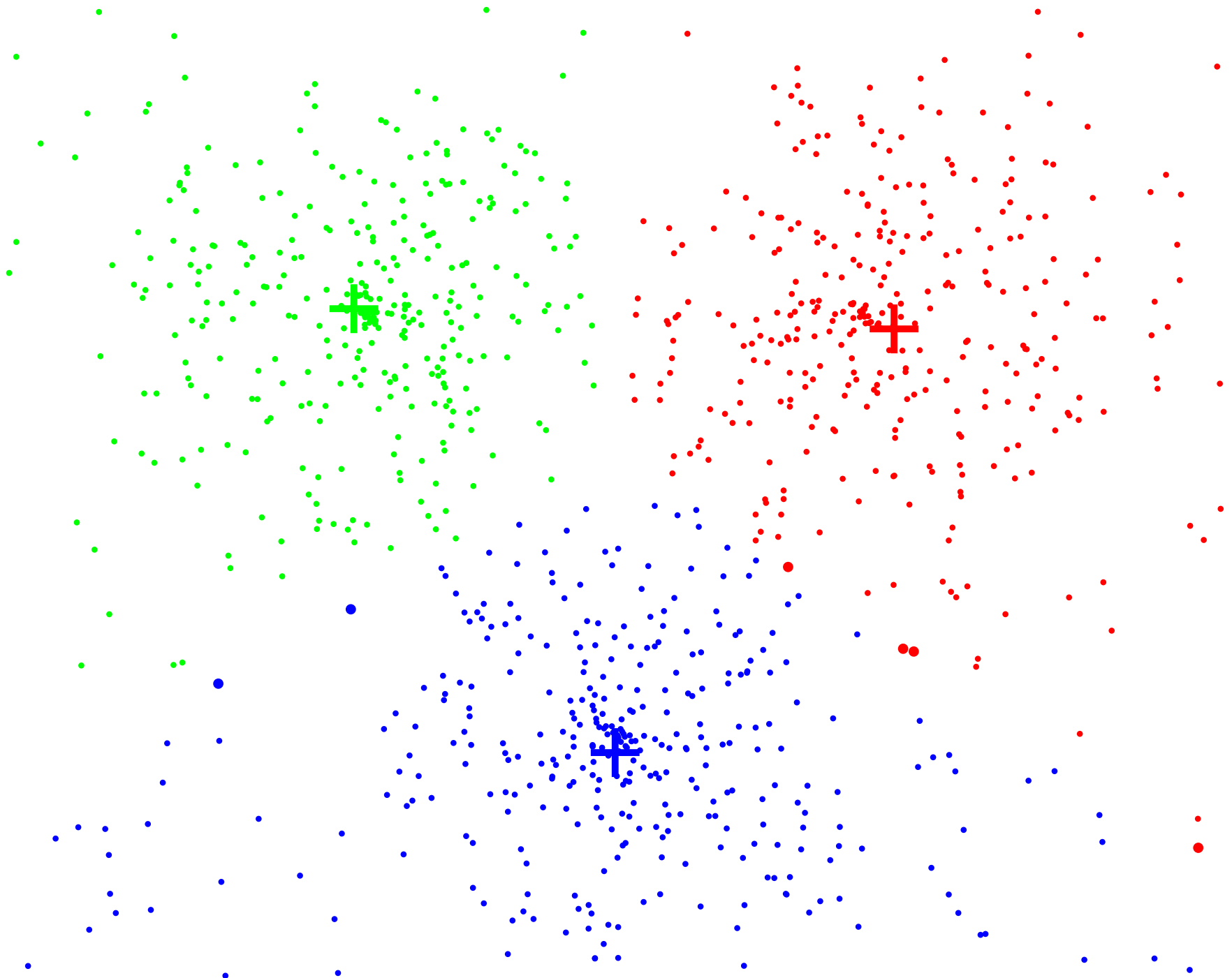
k-Clustering with $k = 3$, after f th iteration



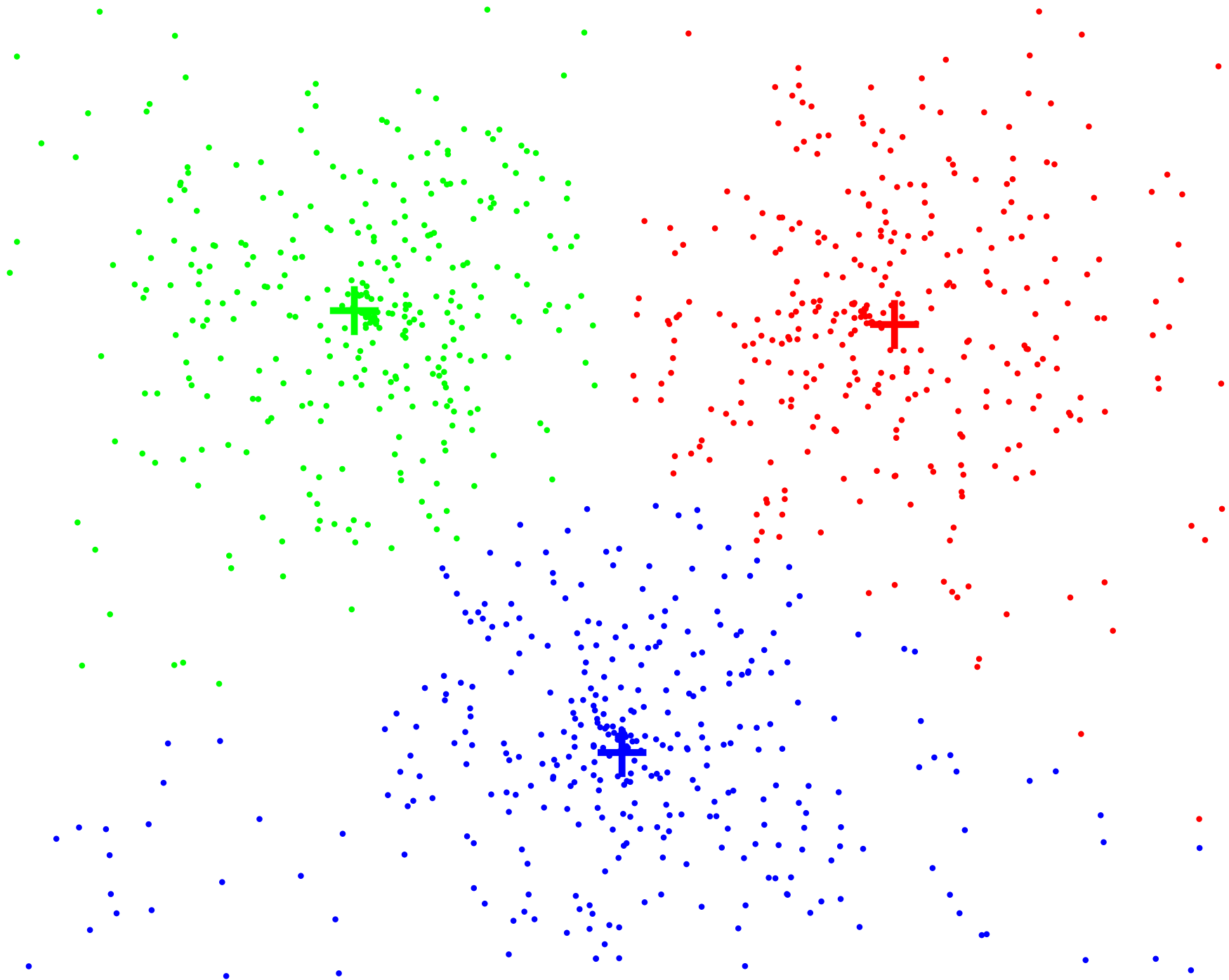
k-Clustering with $k = 3$, points changing cluster



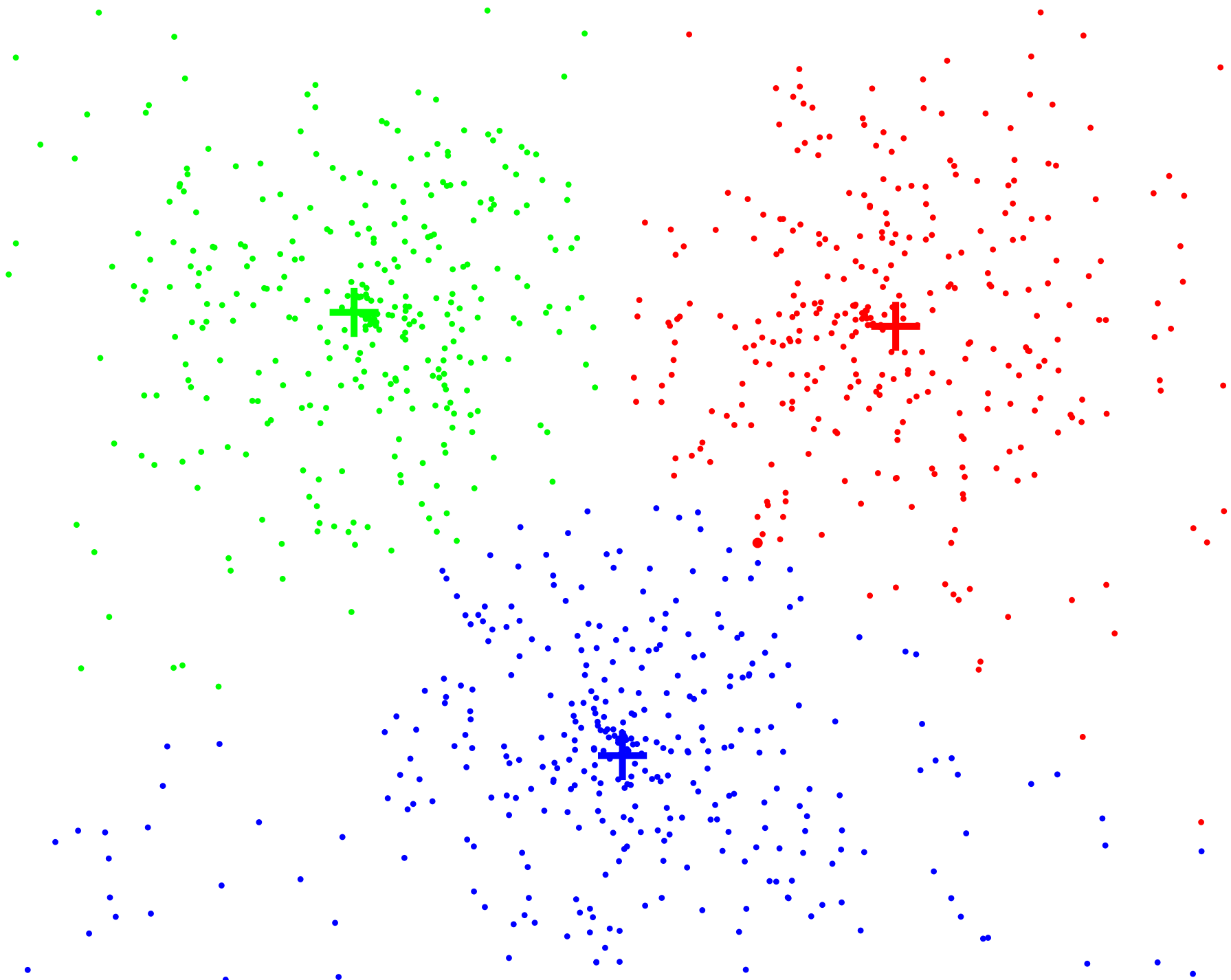
k-Clustering with $k = 3$, after sixth iteration



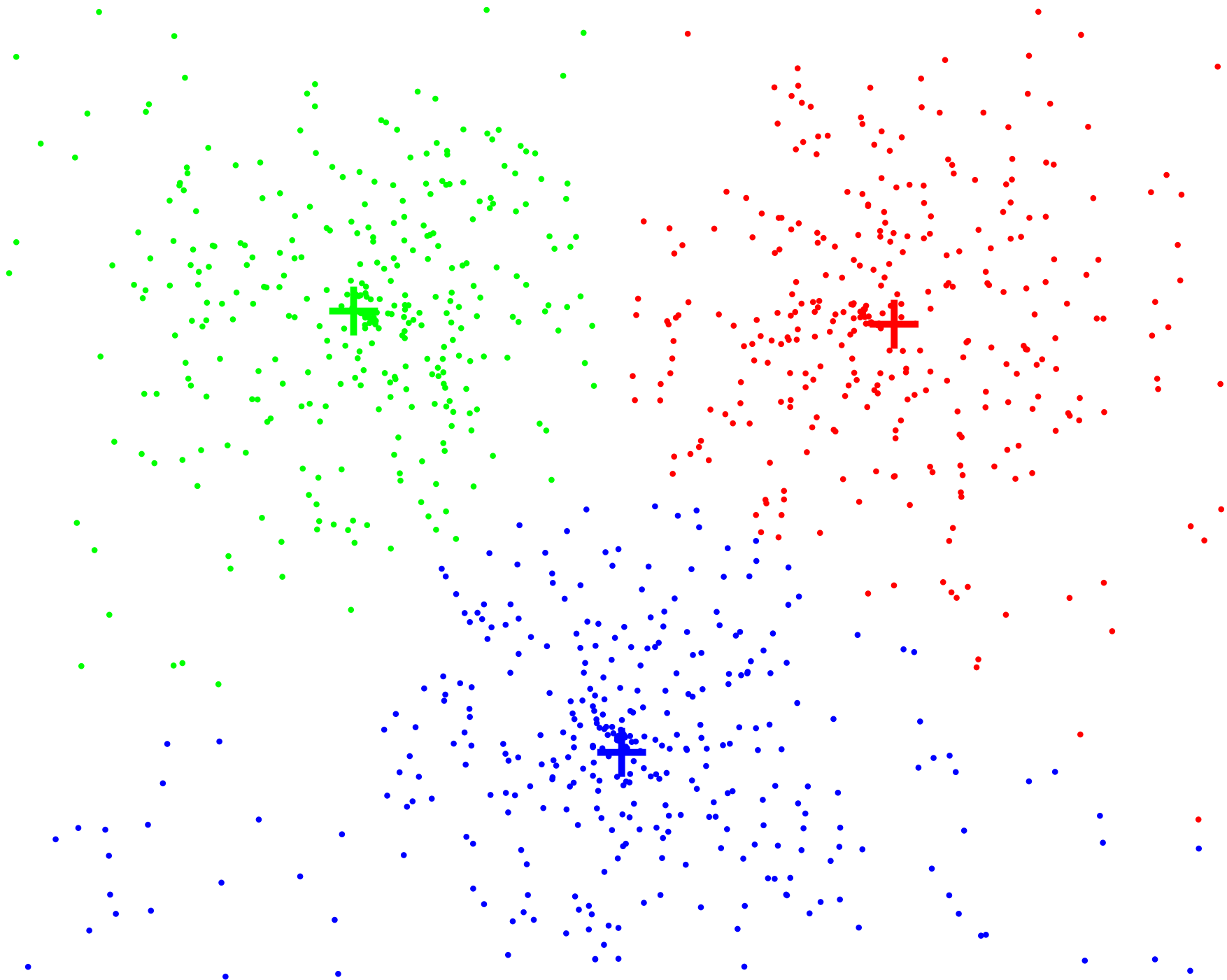
k-Clustering with $k = 3$, points changing cluster



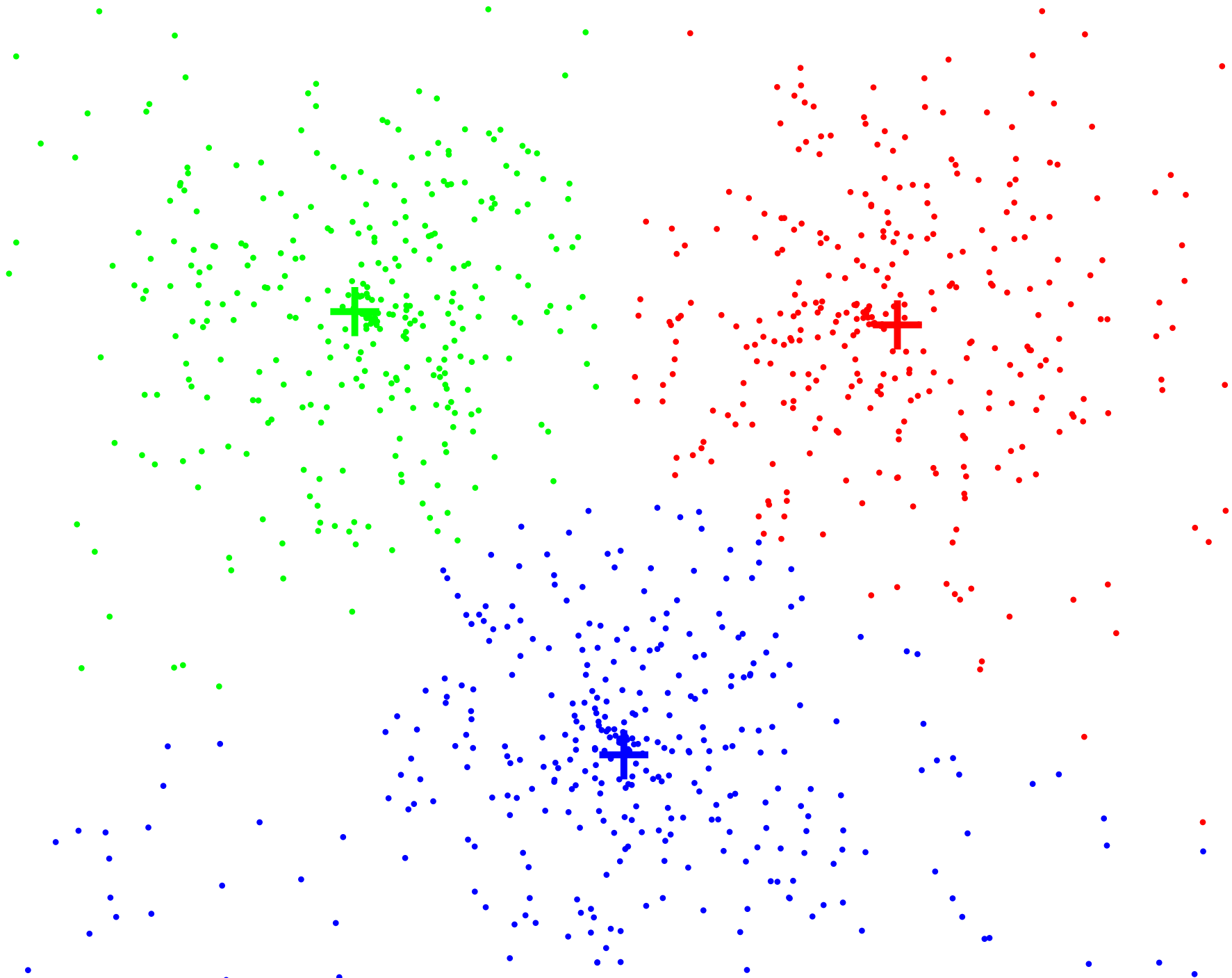
k-Clustering with $k = 3$, after seventh iteration



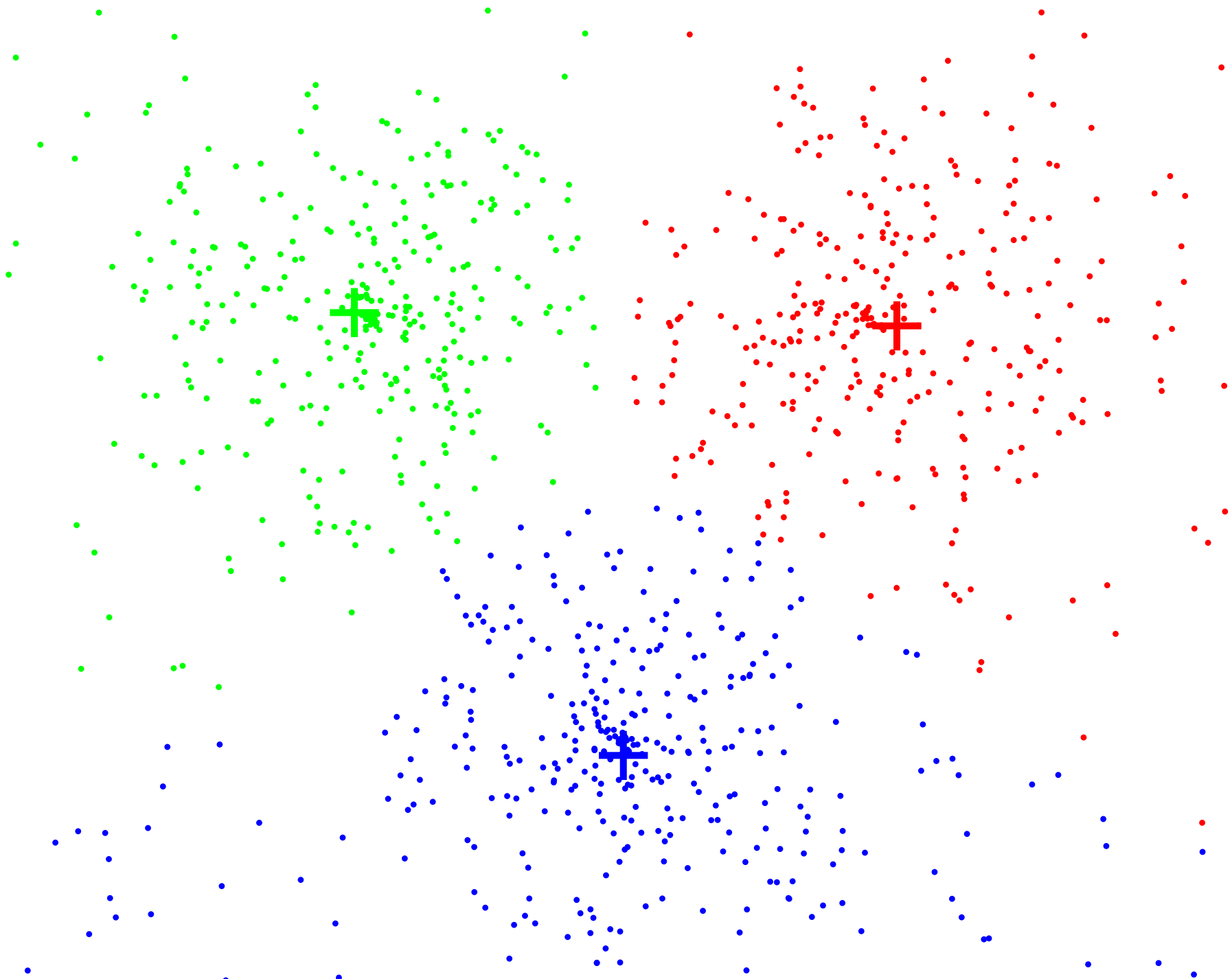
k-Clustering with $k = 3$, points changing cluster



k-Clustering with $k = 3$, after eighth iteration



k-Clustering with $k = 3$, points changing cluster? (No.)



k-Clustering with $k = 3$, nished.

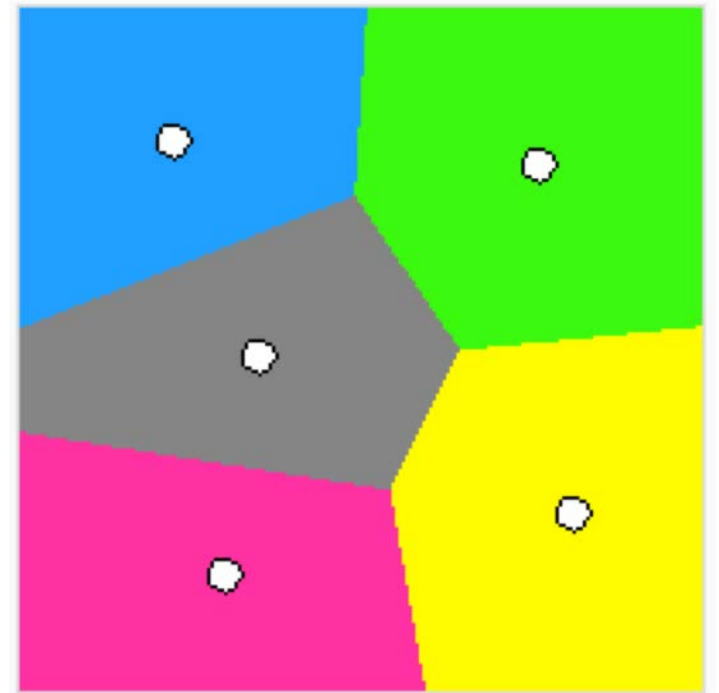
k-means clustering

Advantages:

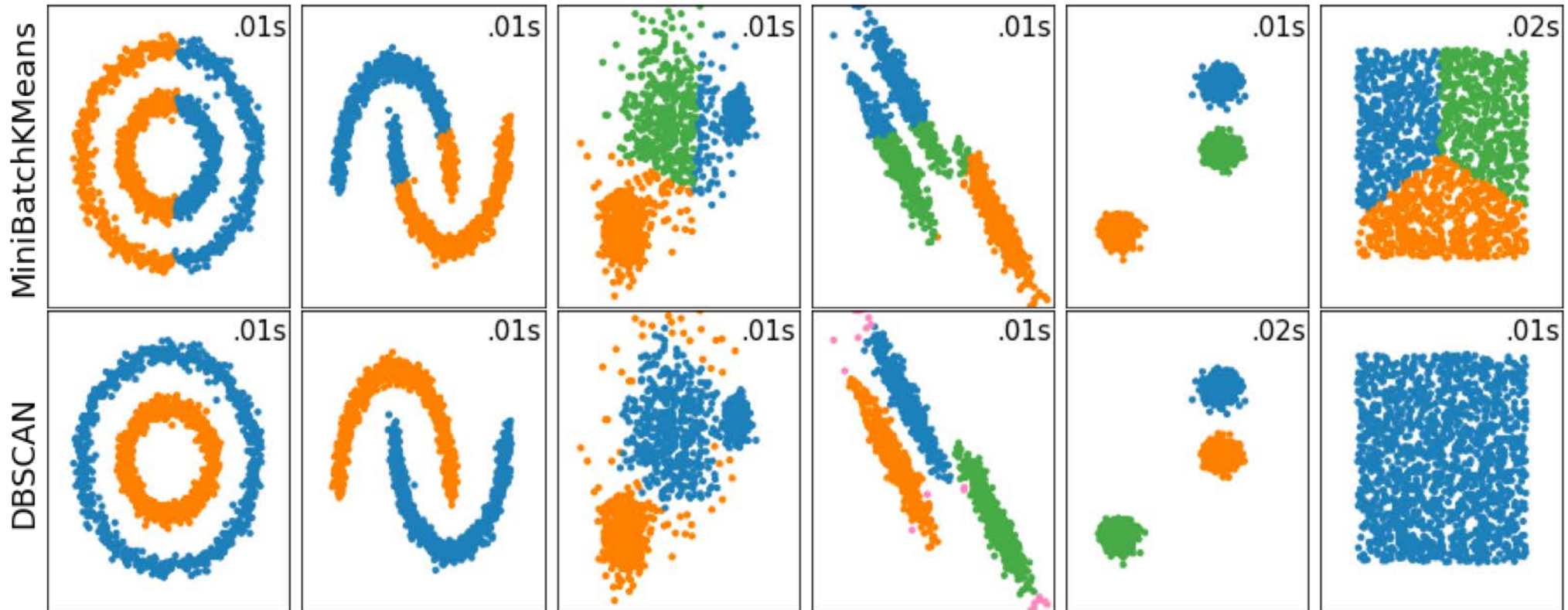
- I tends to converge very fast, $O(n)$

Disadvantages:

- finds a locally optimal set of clusters
- requires the user to specify the number of clusters
- tends to produce clusters that are similarly sized Voronoi cells



k-means clustering versus DBSCAN



Finding k

Different techniques:

- Elbow method --> open Jupyter code
code_elbow_method.ipynb
- Dendrogram
- Silhouette score Analysis

Elbow method

- Method of interpretation and validation of consistency within cluster analysis:
 - help finding the appropriate number of clusters in a dataset
- Look at the percentage of variance explained as a function of the number of clusters
- Choose a number of clusters so that adding another cluster does not give much better modeling of the data

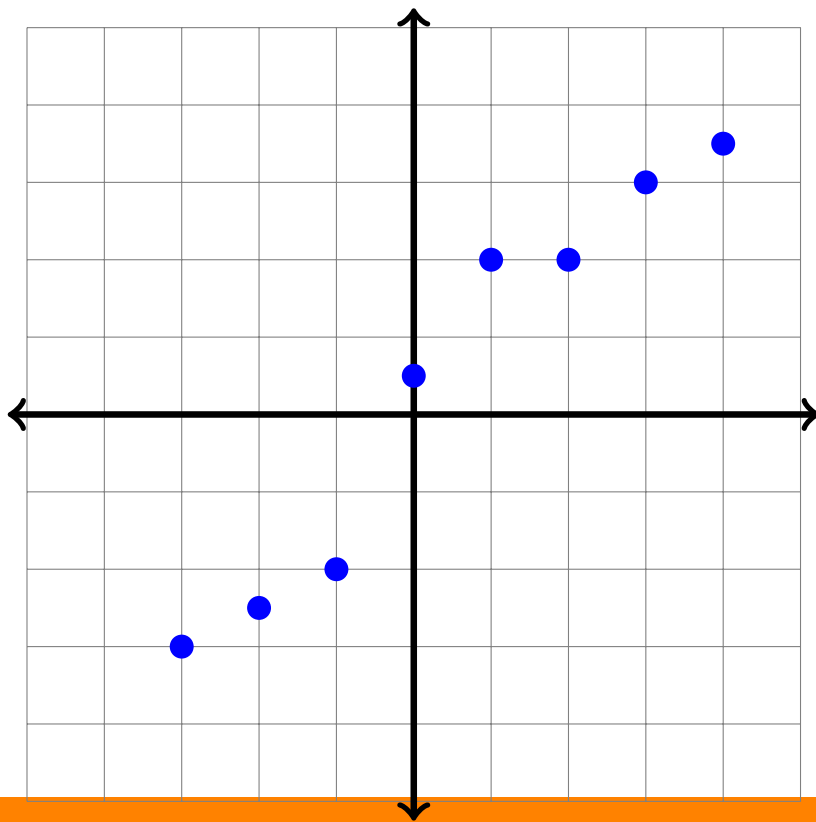
Elbow method

- Compute clustering algorithm (e.g., k-means clustering) for different values of k
 - By varying k from 1 to 10 clusters
- For each k , calculate the total within-cluster sum of square (wss)
- Plot the curve of wss according to the number of clusters k
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters

<http://www.sthda.com/english/articles/29-cluster-validation-essentials/96-determining-the-optimal-number-of-clusters-3-must-know-methods/#elbow-method>

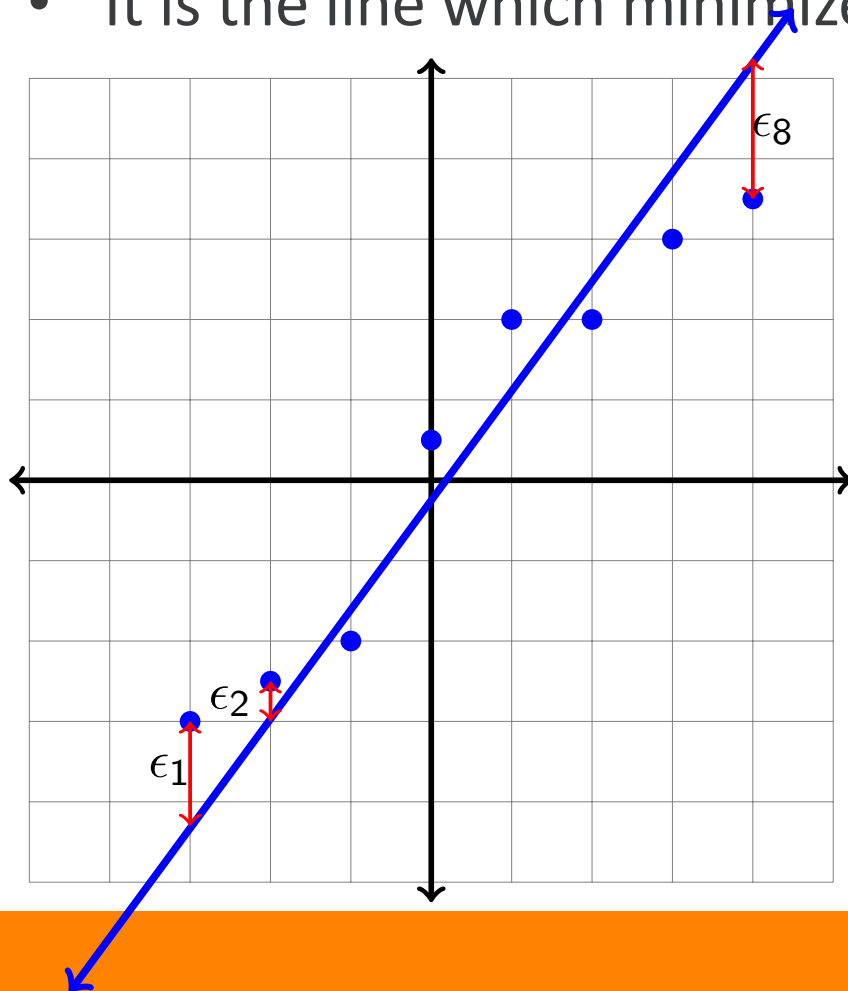
Linear regression

- Given a set of data points in the plane, $(x_1; y_1); \dots; (x_n; y_n)$, find an equation of the line $y = mx + b$ that best describes the points



Linear regression

- What does the best mean? (In terms of a best t line.)
- It is the line which minimizes the sum-of-squares error:



$$\sum_{i=1}^n \epsilon_i^2.$$

Linear regression

Given n data points, $(x_1, y_1), \dots, (x_n, y_n)$,

$$\text{let } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

The **y-intercept** of the *best fit line* is such that it passes through the center of mass (\bar{x}, \bar{y}) of the data points:

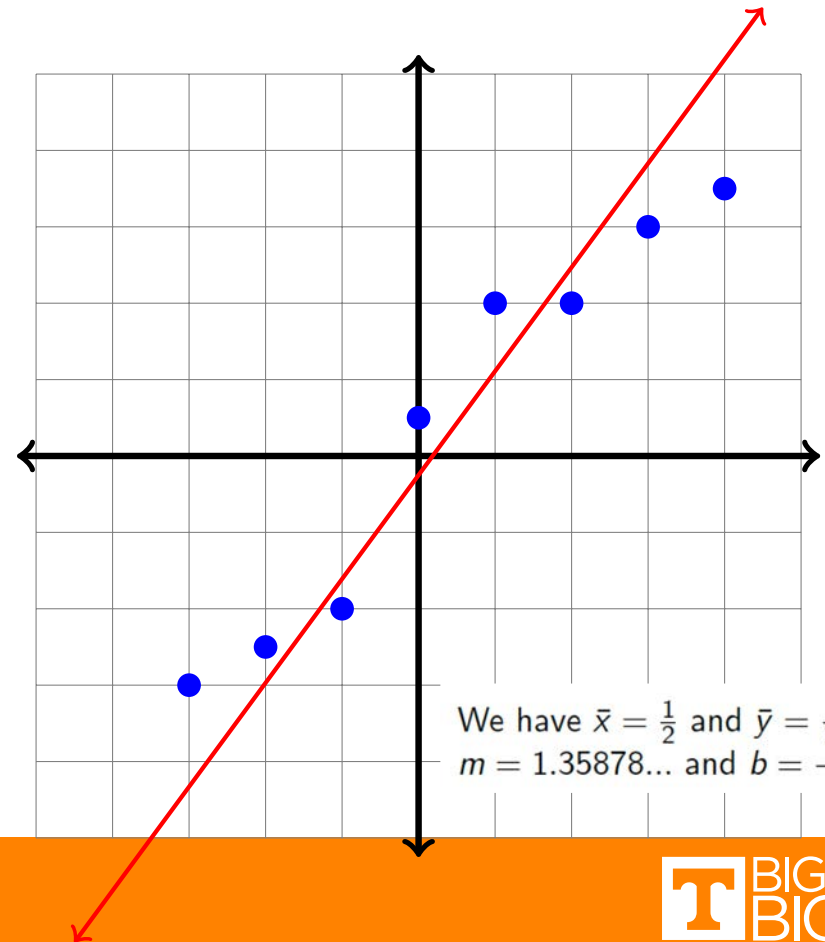
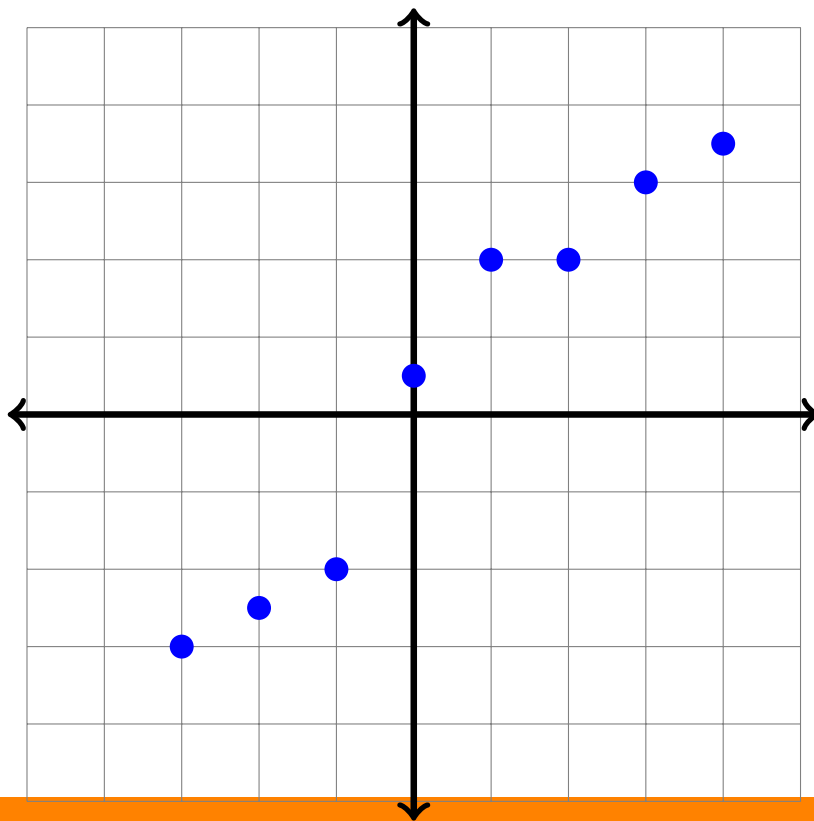
$$b = \bar{y} - m\bar{x}.$$

The **slope** of the *best fit line* is:

$$m = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Linear regression

- Given a set of data points in the plane, $(x_1; y_1); \dots; (x_n; y_n)$, find an equation of the line $y = mx + b$ that best describes the points



Assignment 6

- During today's lecture, we learned about kk-means clusterings.
- In the previous assignment, we learned to use PySpark's parallel versions of the map and reduce functions.
- In this assignment, we will be implementing the kk-means algorithm in parallel with PySpark.

DEVELOP THE CODE BY YOURSELF, DO NOT SEARCH FOR AN EXSITING LIBRARY

Project (I)

- Step 0: search for datasets (due today)
 - Discussion in class of datasets identified

NHANES Dataset

- National Health and Nutrition Examination Survey (NHANES) is a cross-sectional survey that is conducted every two years in the United States
- Individuals are asked to complete
 - demographics questionnaire
 - 2-day, 24 hour dietary recall data for all individuals
- Sampled population in NHANES for the years 1999 thru 2016.
 - In the early years (1999 and 2001) only 1 day of dietary information was collected
 - For years 2003 thru 2016, 2 days of dietary data were collected

NHANES Dataset

- For each year, individuals have a unique identifier called a sequence number
- All files for that year can be linked by the sequence number (variable name: SEQN)
- Every year, different participants are sampled, so you cannot track the same person over time
 - You can track the average population intakes over time.

NHANES Dataset

- What do the nutrient intake profiles of individuals look like? For example, do people who consume more saturated fats consume less fiber?
- Do nutrient intake profile patterns differ by gender, age, race, education, or poverty level?
- What do time trends in nutrient patterns look like?
- How consistent are individuals in their consumption over two days? For example if they are consuming high amounts of fiber on day one, are they also consuming high amounts on day 2?

Medicaid-Vital Statistics Data

- The Medicaid and Medicare Administration in the State of Delaware (DMMA) examines medical usage and health outcomes of their clients on a regular basis
- A report came out in 2014 that stated that individuals with mental illness were twice as likely to die and that they die at a much earlier age (almost 20 years earlier) compared to those without mental illness in the United States
- DMMA wanted to know if this was true of their Medicaid population
- The DSAMH_Medicaid dataset is a subset of about 6000 individuals who have Medicaid as their primary insurance and have at least one instance of mental illness.
- These data were then linked to vital statistics to determine mortality
 - There were approximately 200 deaths

Medicaid-Vital Statistics Data

- What cause of death codes are most commonly reported in the DSAMH_Medicaid dataset?
- Do the cause of death codes cluster into groups by gender? Or by disability status?
- Are there differences in the patterns of medical care (i.e., time spent in Medicaid, number of medical claims, number of hospital claims, number of emergency department claims, total billed amounts) reported by those that die versus those that didn't die?

Project (II)

- Step 1: Address (and answer) these questions
 - Do you work alone or in team? If in team, indicate who is your collaborator and what are the skills / expertise he/she is bringing to the project
 - What is the dataset you are considering?
 - What are the possible key question(s) you want to answer? Are the questions too general? Are the questions too narrow?



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

BIG ORANGE. BIG IDEAS.®

