Lecture 6:
Clustering and k-mean

COSC 526: Introduction to Data Mining
Spring 2020

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE
KNOXVILLE
BIG ORANGE. BIG IDEAS.®

Today Outline

- Assignment 6
- Clustering and k-mean
- Linear regression
- Live Chat

Assignment 6

3

Assignment 6

- During the last lecture, we learned about the **Apache Spark** implementation of the **MapReduce** programming model

In this assignment

- We will use **PySpark** (the Spark Python API) to perform one of the text parsing problems that we solved in the last assignment *with the power of parallel processing*
- In the previous assignment, we defined three sequential methods:
  - mapSequential
  - reduceSequential
  - reduceByKeySequential
- In this assignment, we will be using PySpark's parallel version of these functions

Assignment 6: Test PySpark

- Run the cell below to verify that your Java, Spark, and PySpark installations are successful

In [ ]:
```
from pyspark import SparkContext
sc = SparkContext.getOrCreate()
data = sc.parallelize(range(1,10))
print(data.reduce(lambda x,y: x+y))
sc.stop()
```

Assignment 6: Problem 0

- Now that we are in Jetstream, clone Assignment05 in Jetstream, open your completed Assignment05, and rerun it
  - Note that you are running your job on a remote 6-core node and not on your laptop
- Executing the same code on different machines is a valuable test of the portability of your code

Assignment 6: Problem 1

- **We redo the problems from Assignment05 using Apache Spark**
  - We move from *sequential* text processing in Python (i.e., Assignments 4 and 5) to the *parallel* implementation in *Apache Spark*
- *Note that the code you wrote for the sequential version should work with the parallel version*
- *You will only need to adapt the code to use **Spark's parallelized data structure, the RDD***

Assignment 6: Problem 2

- We analyze the text for letter frequency
  - If you've taken a crypto course and/or have seen substitution ciphers then you are probably aware that 'e' is the most common letter used in the English language
- **Use the pre-processed text words to count the frequency of each letter in the text using the parallel MapReduce methods of Spark**

Assignment 6: Problem 3

- If we really wanted to crack a substitution cipher (or win on "Wheel of Fortune") then we should be aware that, although 'e' is the most common letter used in English, it may not be the most common first letter in a word
- **Count the positional frequencies of each letter using the parallel MapReduce methods of Spark**
  - Count the number of times each letter appears as the first letter in a word, as the last letter in a word, and as an interior letter in a word (i.e. a letter that is neither first nor last)

Assignment 6: Problem 4

- As you did the previous assingnments, use matplotlib to create histograms for Problems 1-3

Clustering

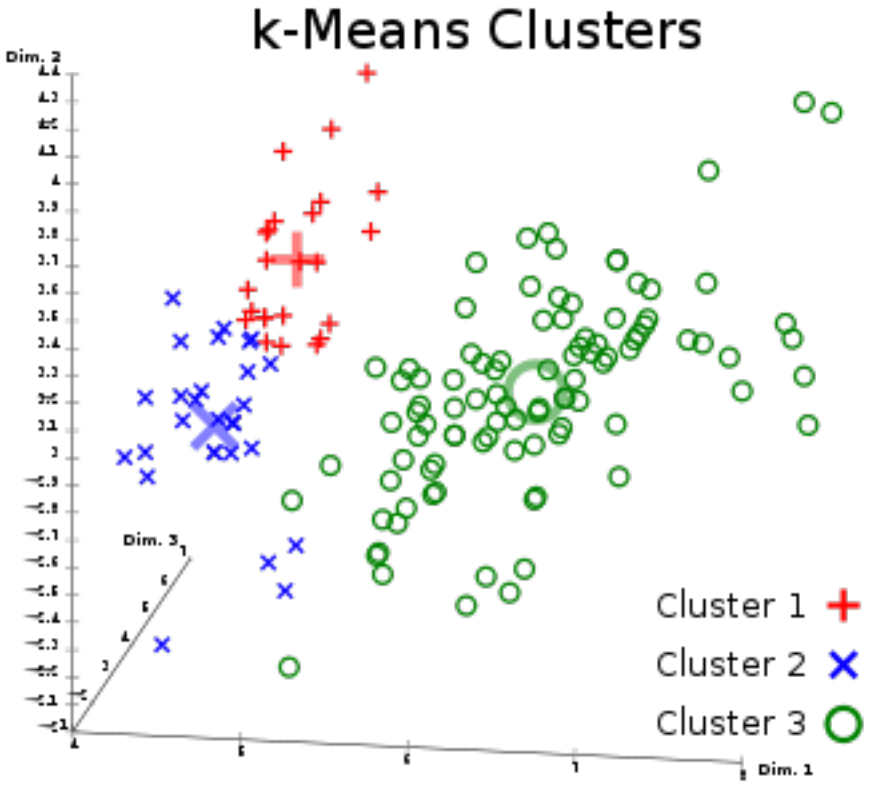
11

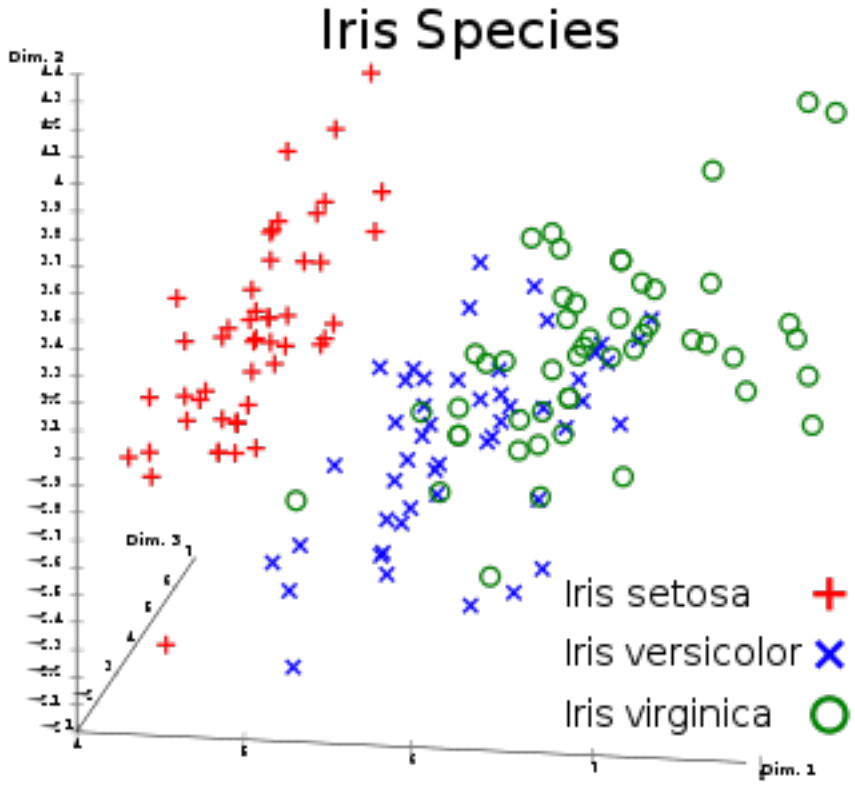
Clustering

- 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





https://en.wikipedia.org/wiki/Iris_flower_data_set

Distance Between Vectors

- Let x = (x1, … , xd) and y = (y1 , … , yd) be vectors in R
- **Question:** What is the distance from x to y?
- **Answer:** It depends

  dist(x, y) = $\|x - y\|$
- There are many choices for the norm $\|\blacksquare\|$

  - $\|\mathbf{x} - \mathbf{y}\|_1 := \sum_{i=1}^{d} |x_i - y_i|$ 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 = (x1, …, xd) and $\mu = (\mu 1, \ldots., \mu d)$ be vectors in $\mathbb{R}$
- 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
  - $\|\mathbf{x} - \boldsymbol{\mu}\| < 1$?
  - $\|\mathbf{x} - \boldsymbol{\mu}\| < .1$?
  - $\|\mathbf{x} - \boldsymbol{\mu}\| < .001$?
- **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 = {x1, …, xd } 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

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

SSE: sum of squared errors of prediction

K-mean clustering

- Given n data points, {x1, …, xn } and an integer k
- Goal: Partition {x1, …, xn } into k sets (clusters) S1 , …, Sk in such a way as to minimize the total **Within Cluster Sum of Squares** error:

$$\mathrm{WCSS} = \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2$$

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

- where μ i is the mean of the vectors in Si , i.e.,

$$\boldsymbol{\mu}_i = \frac{1}{|S_i|} \sum_{\mathbf{x} \in S_i} \mathbf{x}.$$

https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-k-means/interpret-the-results/all-statistics-and-graphs/

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**.

https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-k-means/interpret-the-results/all-statistics-and-graphs/

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

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

https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-k-means/interpret-the-results/all-statistics-and-graphs/

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

https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-k-means/interpret-the-results/all-statistics-and-graphs/

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

https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-k-means/interpret-the-results/all-statistics-and-graphs/

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

$$\mathrm{WCSS} = \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2$$

https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-k-means/interpret-the-results/all-statistics-and-graphs/

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

https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-k-means/interpret-the-results/all-statistics-and-graphs/

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 $x_1, \ldots, x_n$,
initial clusters $S_1^{(0)}, \ldots, S_k^{(0)}$,
and initial means of clusters $\mu_1^{(0)}, \ldots, \mu_k^{(0)}$.

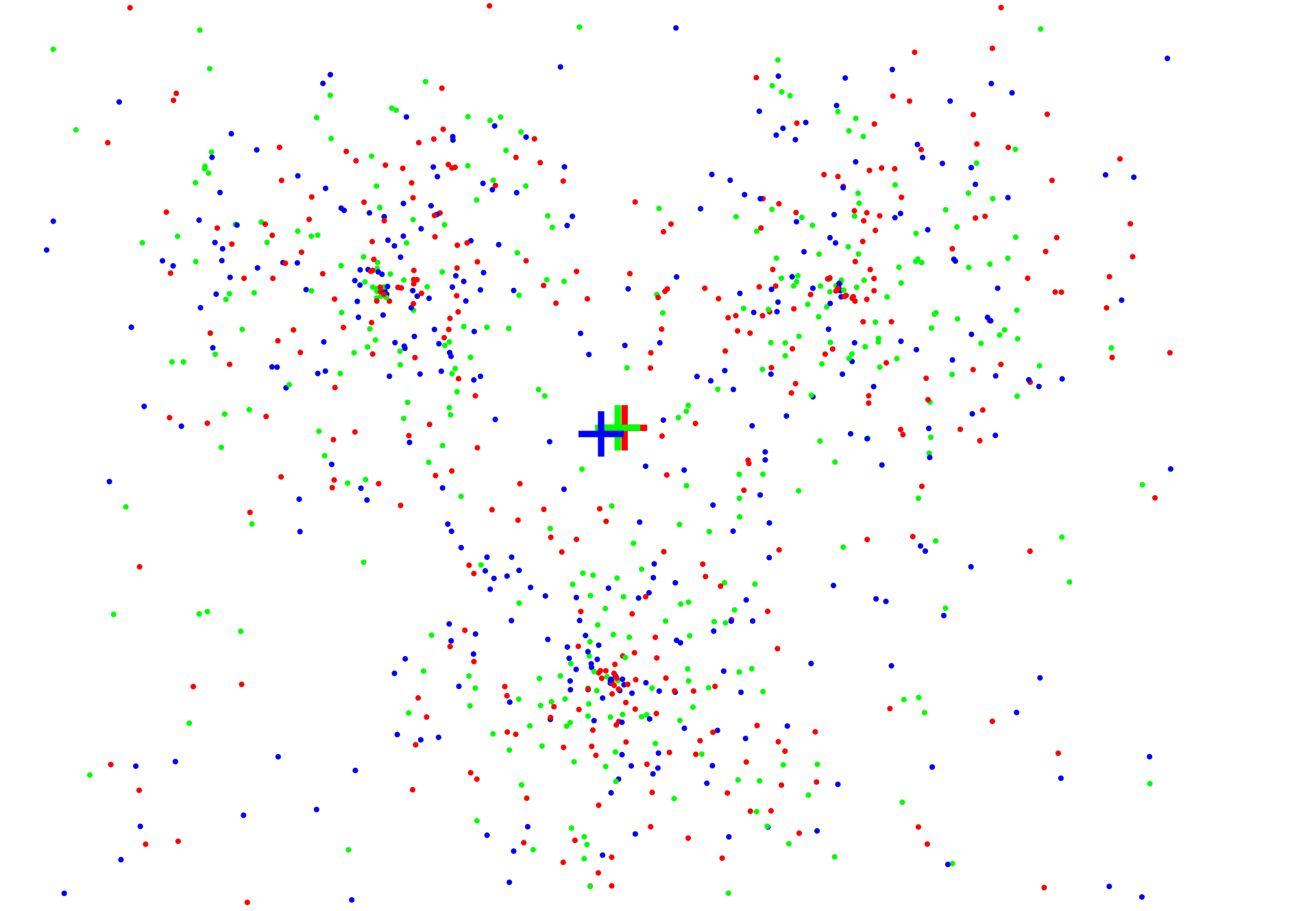
- **Step 1: Assignment.** (Iteration number t)
  Put x_i in cluster $S_j^{(t)}$ if $\mu_j^{(t-1)}$ is the mean closest to x_i.
  (Do this for every data point x_i.)
- **Step 2: Update the means.**
  Compute the mean, $\mu_i^{(t)}$, of all the vectors in cluster $S_i^{(t)}$.
- **Repeat.**
  Go back to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

k-means clustering

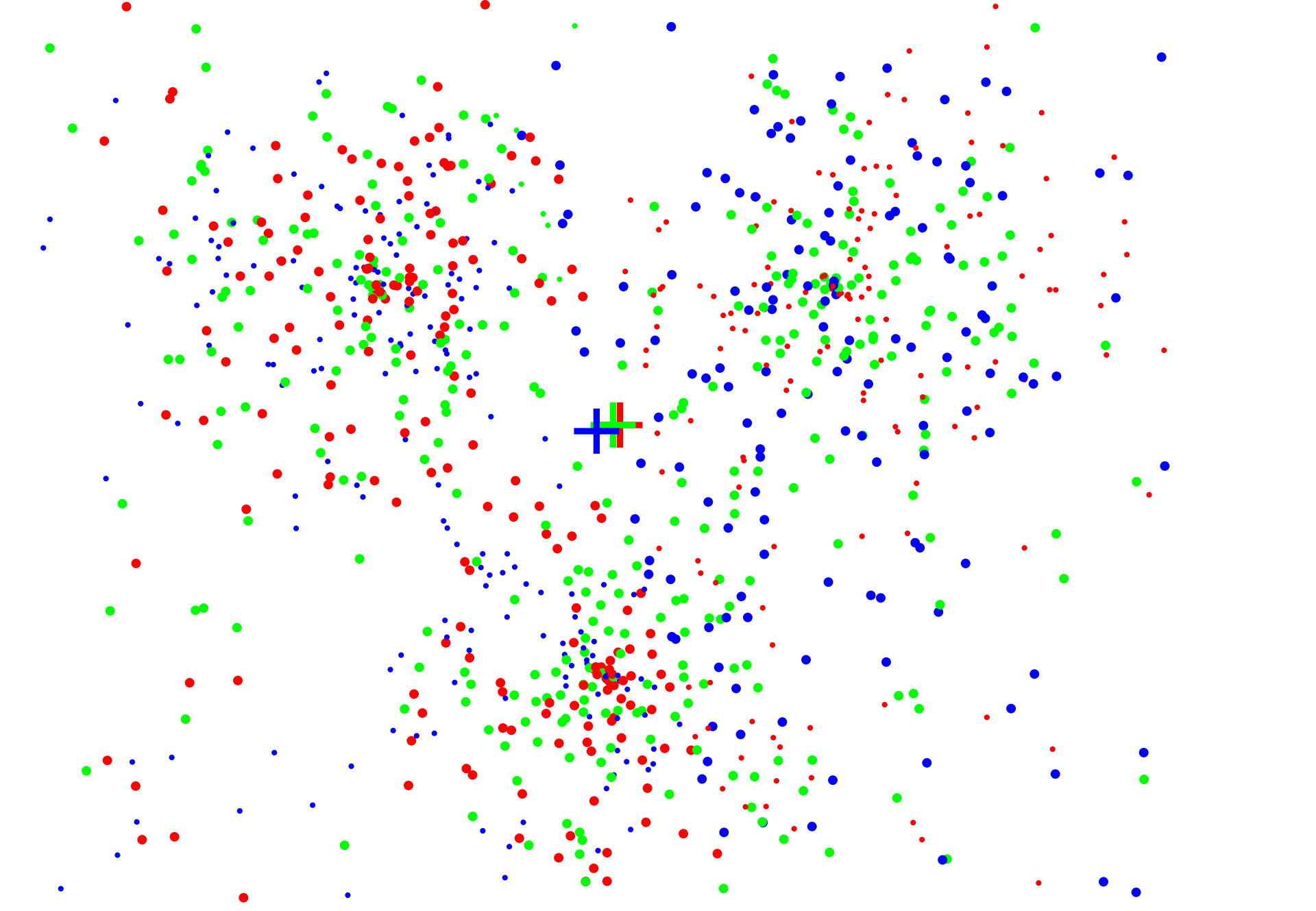
Initialization:

There are two standard ways to start the algorithm.

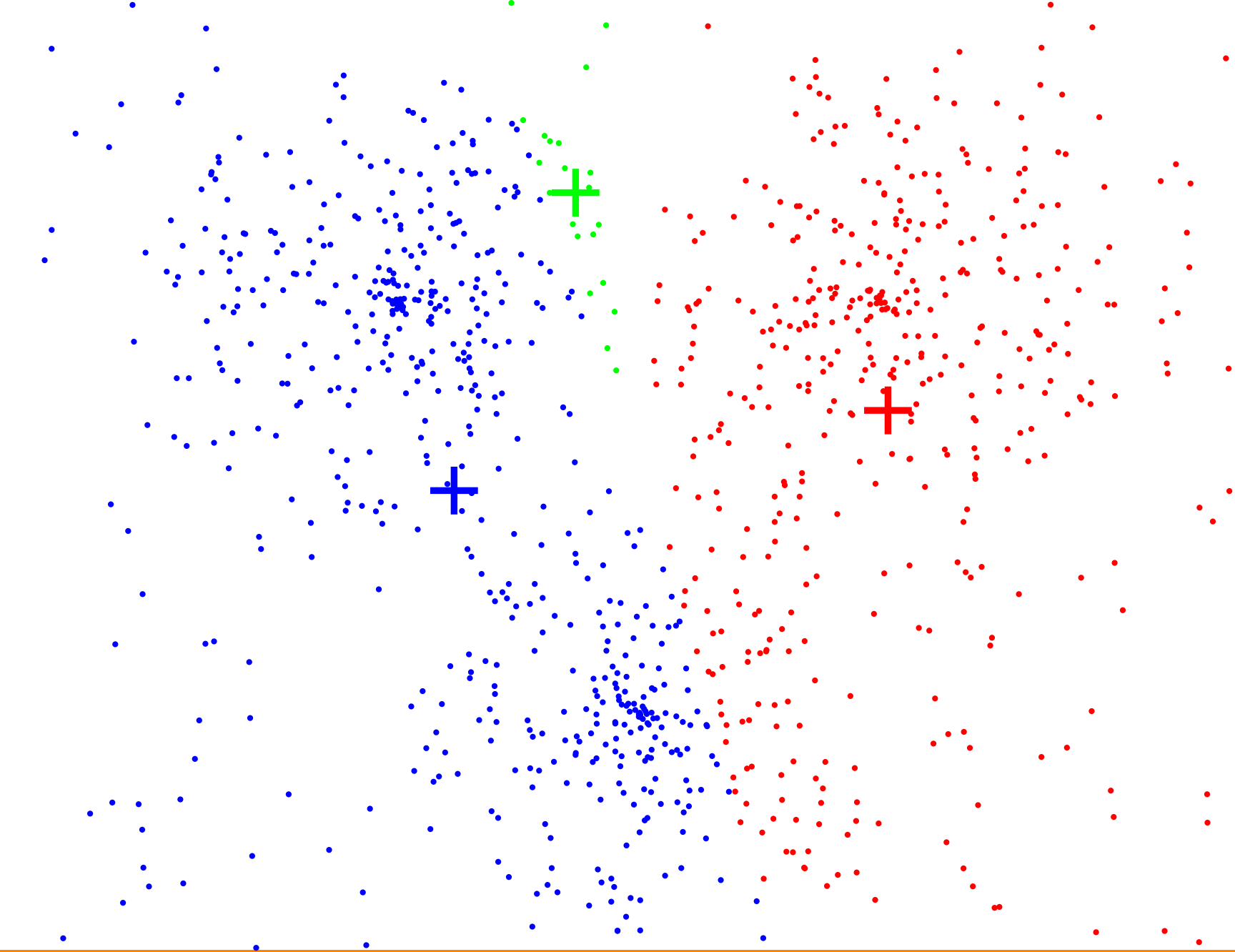
- Randomly assign each x_i to a cluster $S_j^{(0)}$ and then compute the means of each cluster.
- Set each initial mean, $\mu_i^{(0)}$, to a (distinct) random data point.



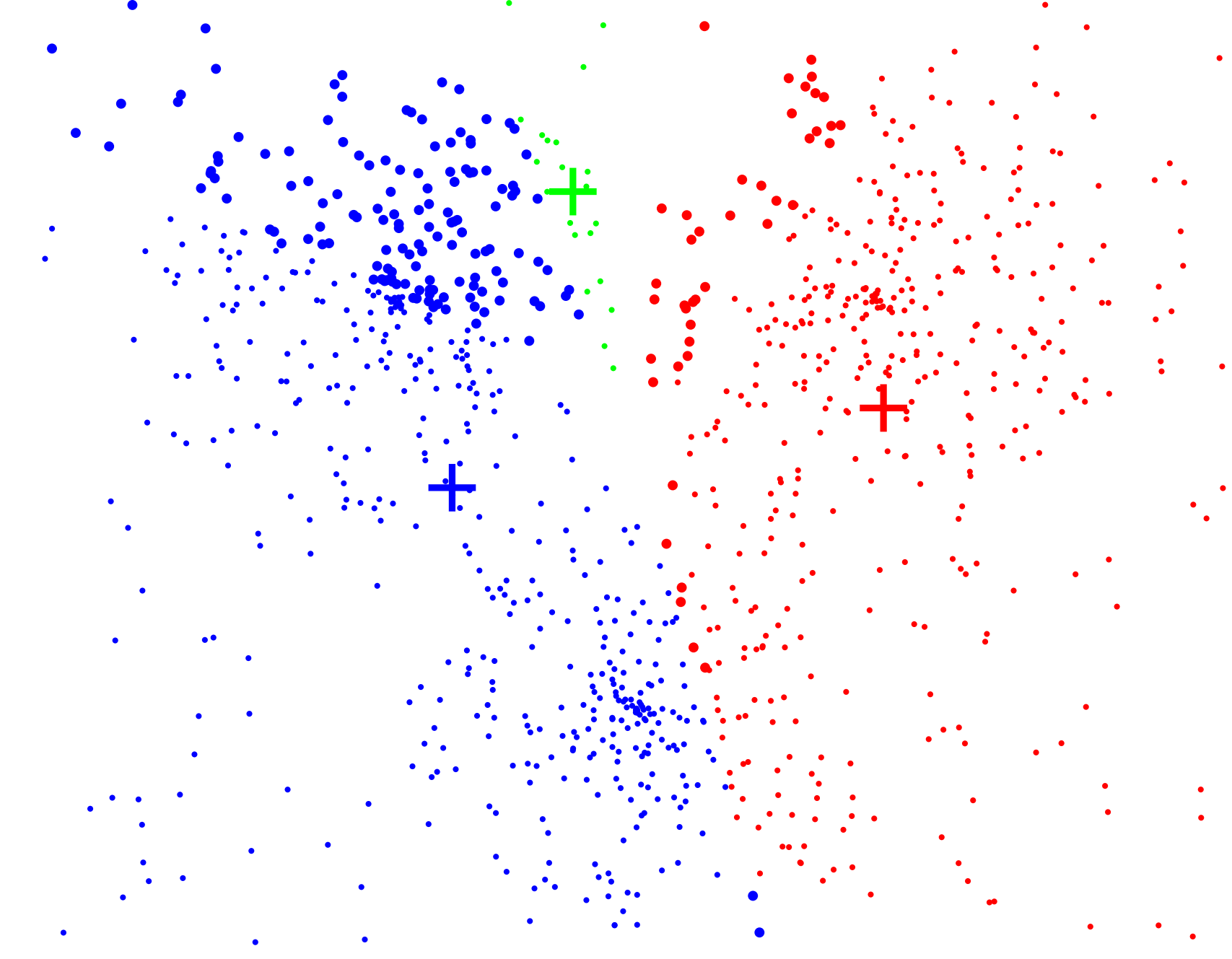
k-Clustering with k = 3, initial assignment



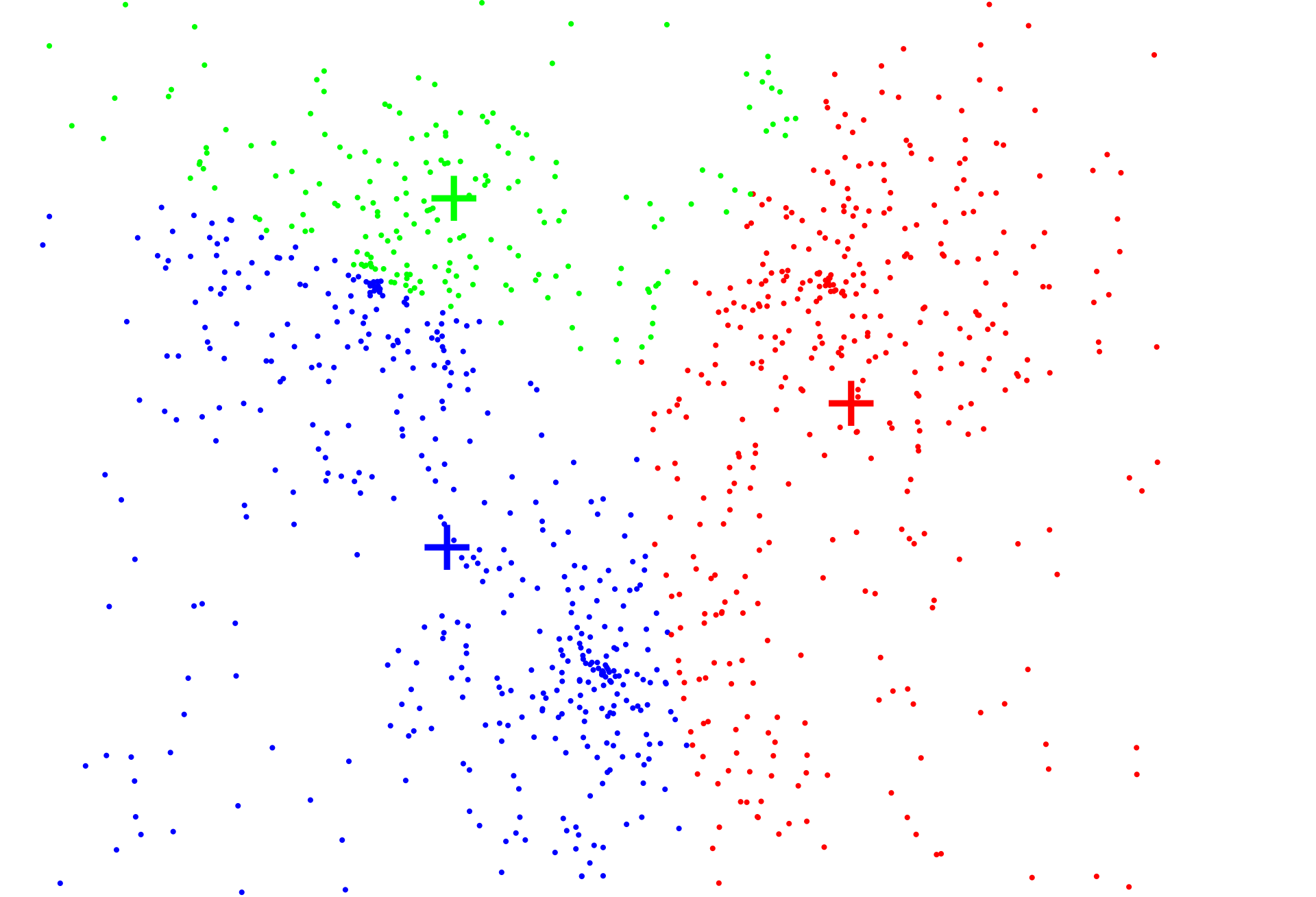
k-Clustering with k = 3, points changing cluster



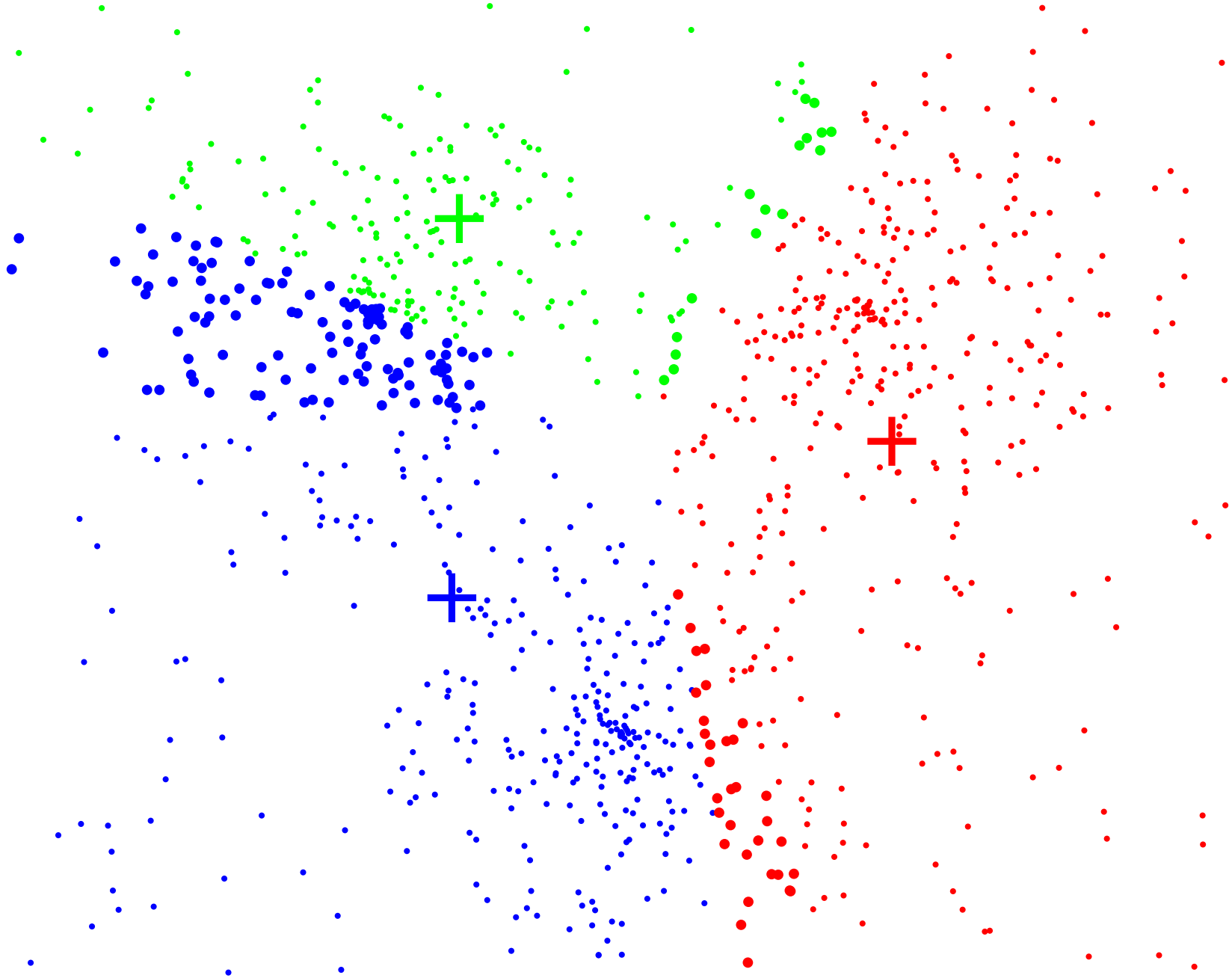
k-Clustering with k = 3, after rst iteration



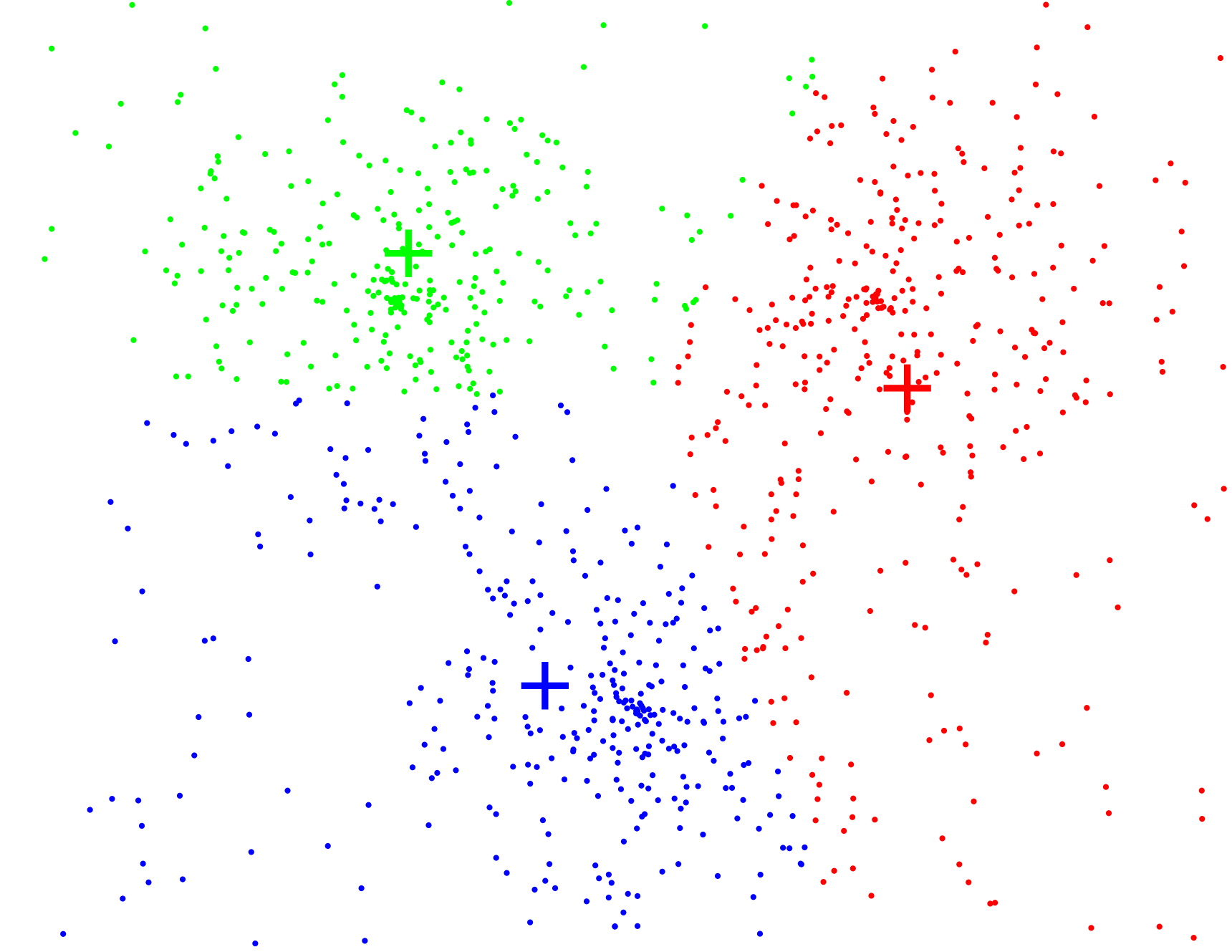
k-Clustering with k = 3, points changing cluster



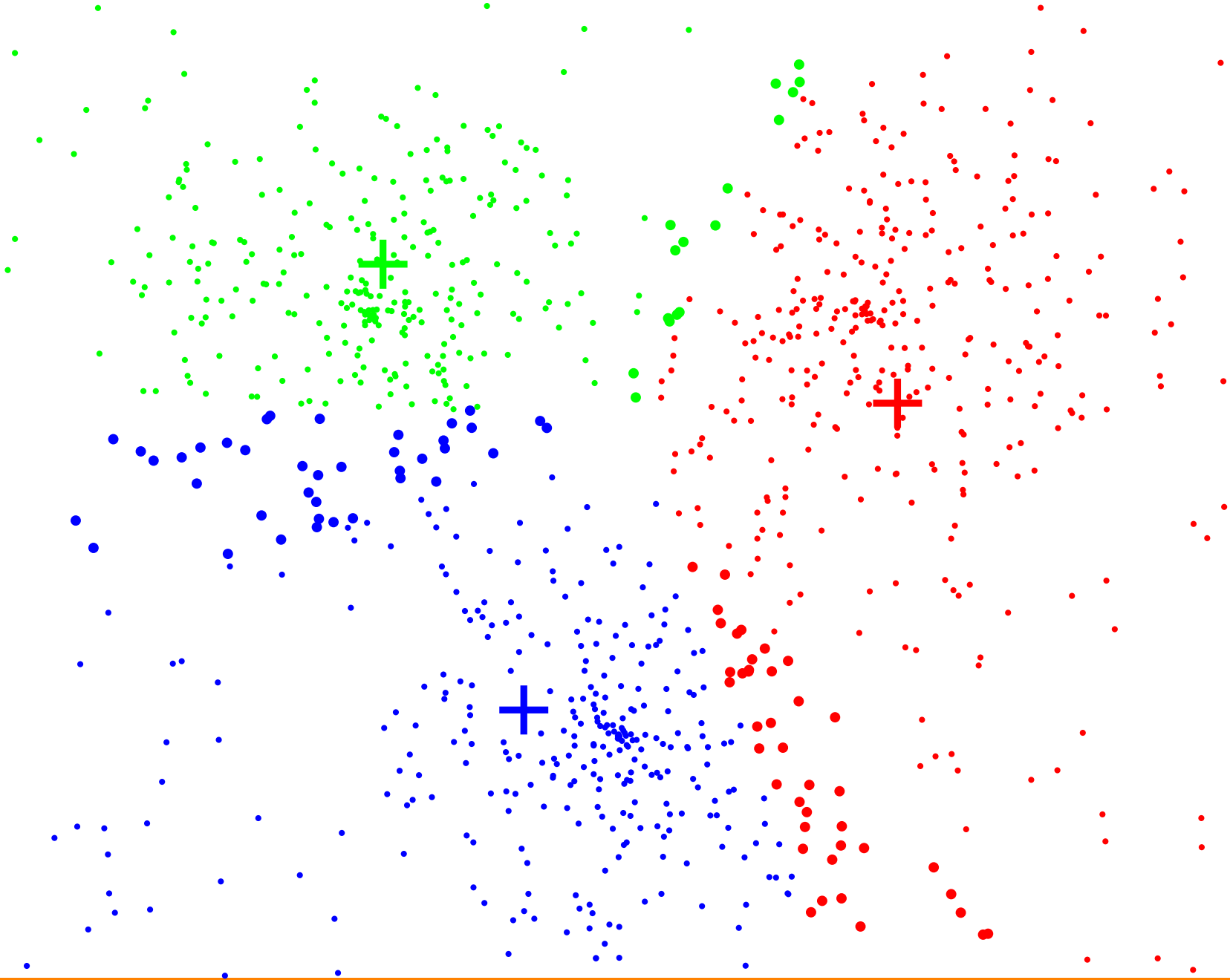
k-Clustering with k = 3, after second iteration



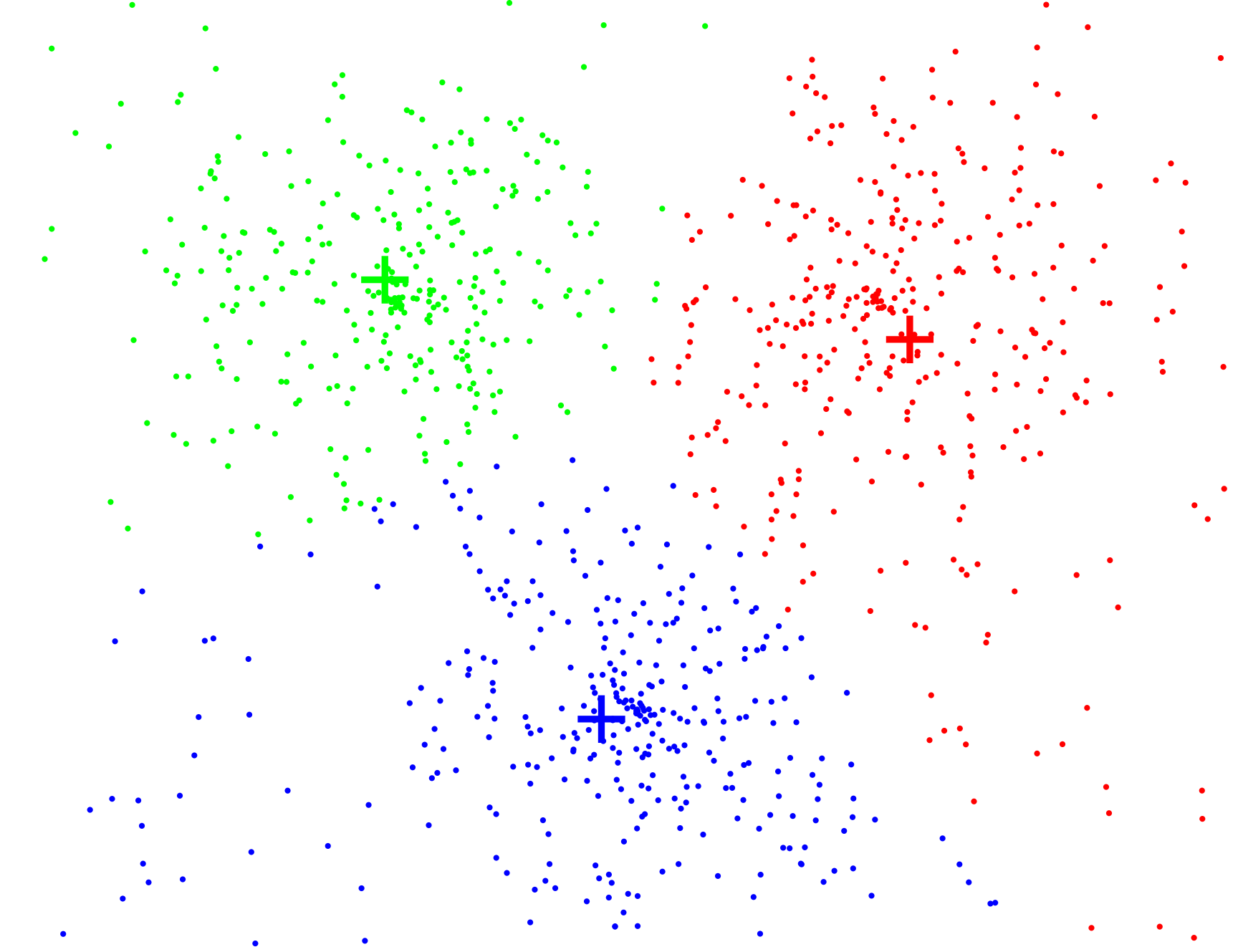
k-Clustering with k = 3, points changing cluster



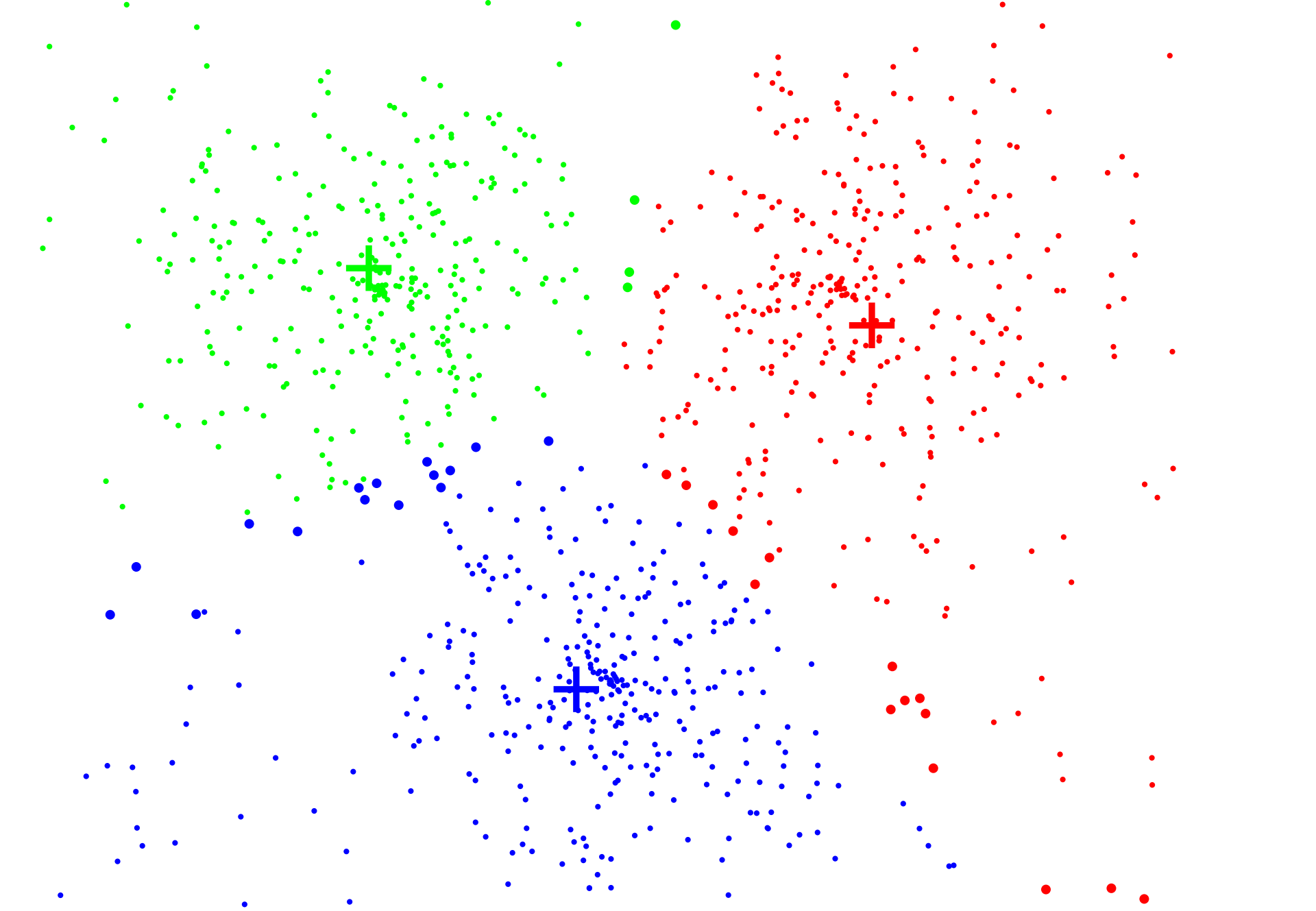
k-Clustering with k = 3, after third iteration



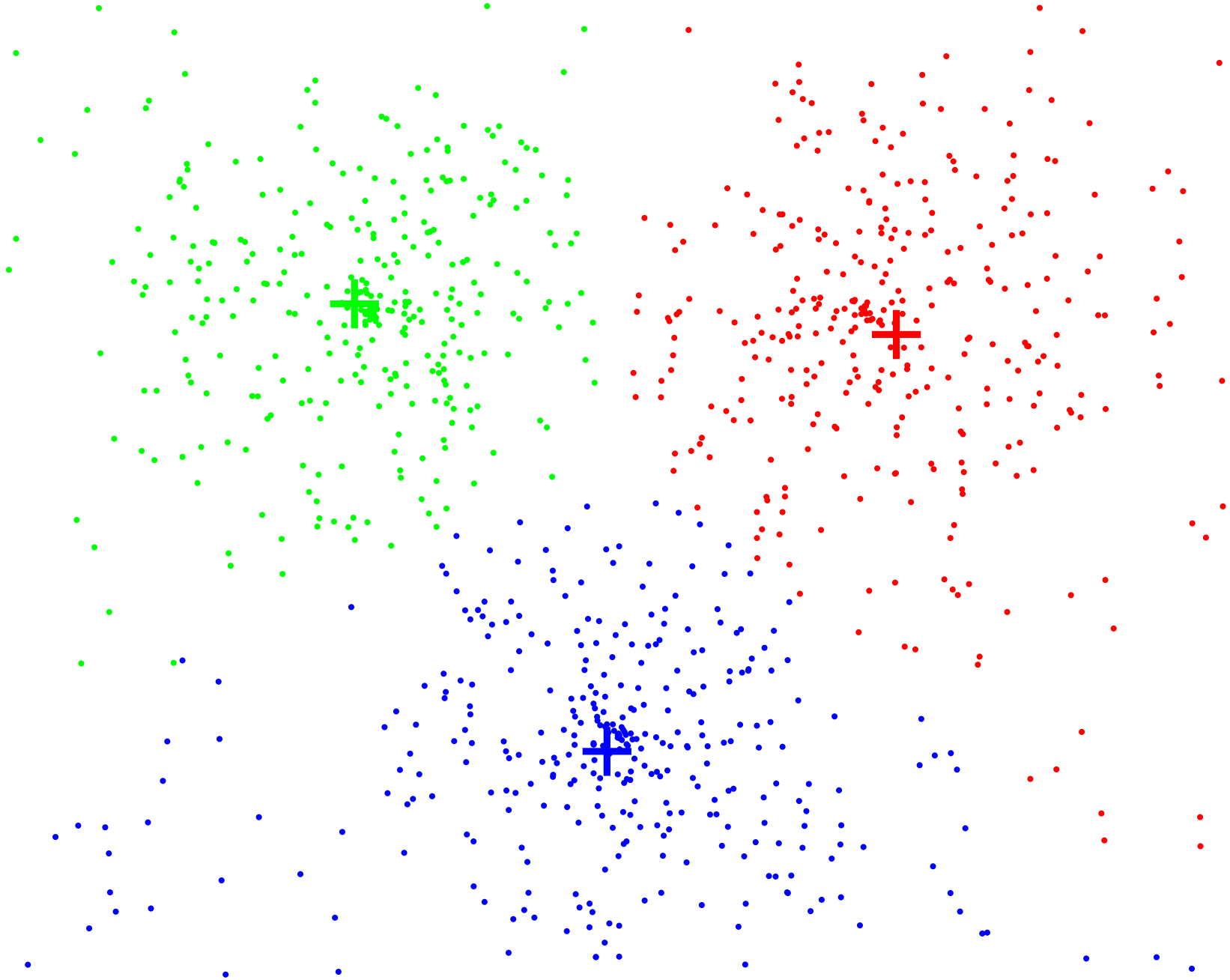
k-Clustering with k = 3, points changing cluster



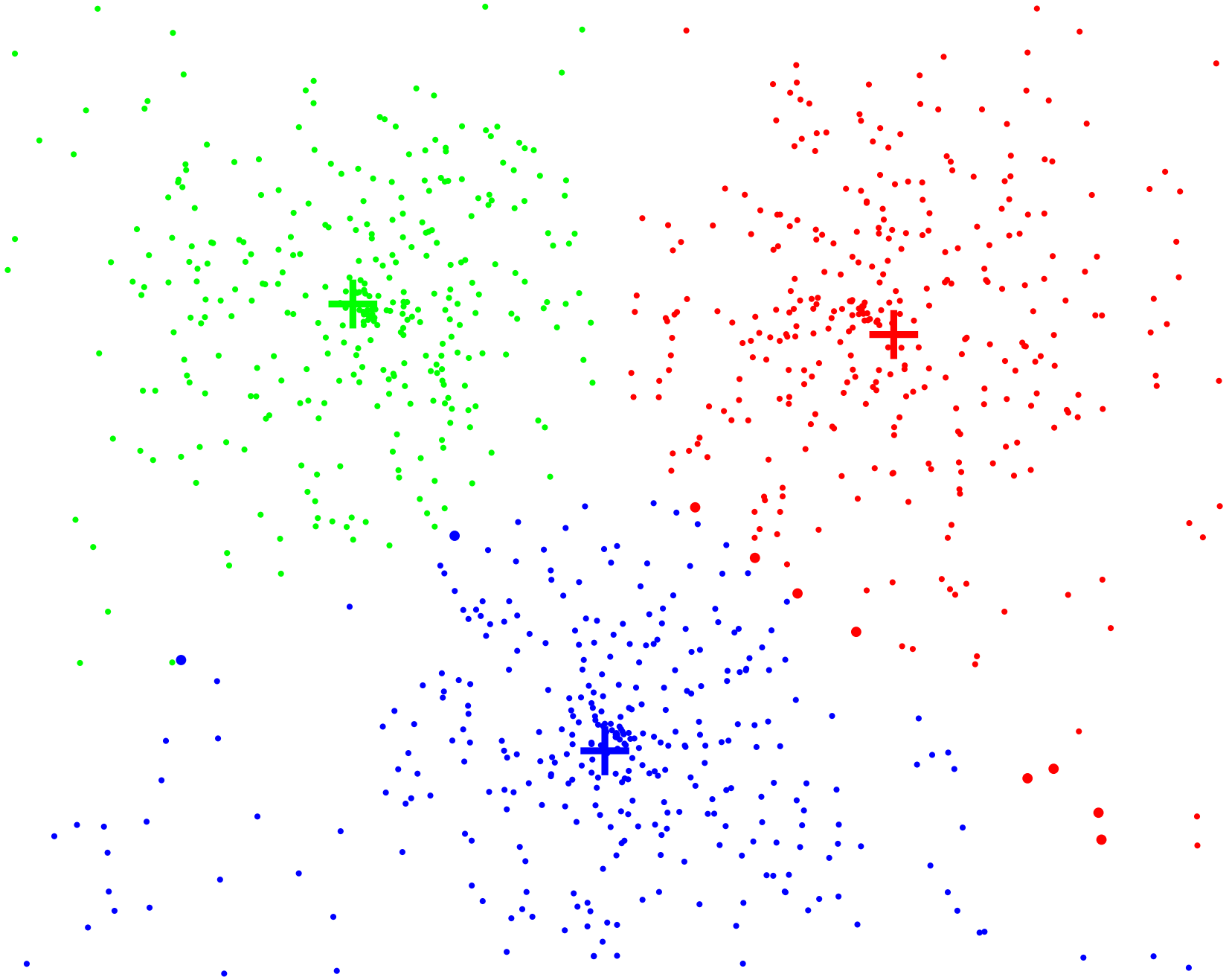
k-Clustering with k = 3, after fourth iteration



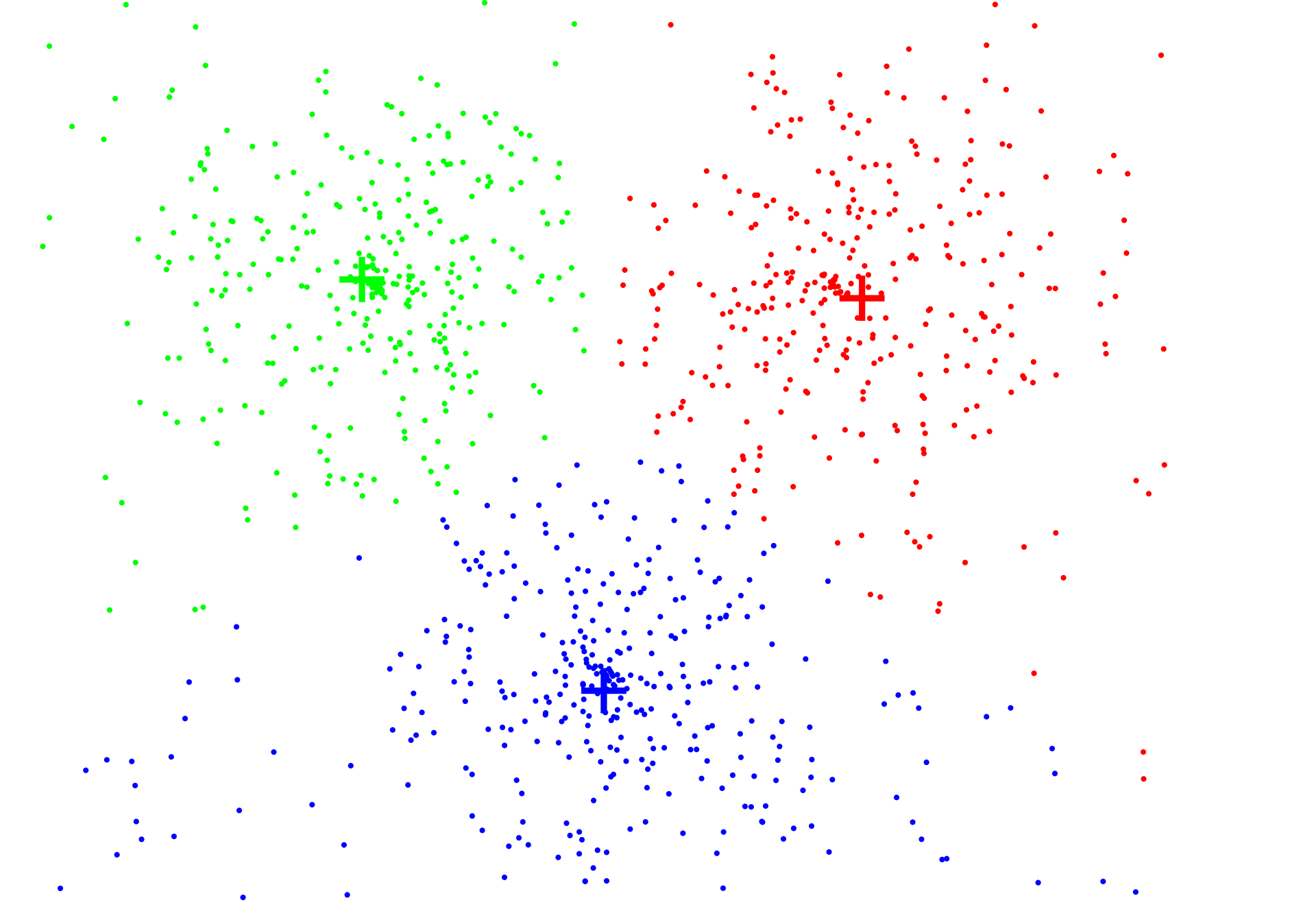
k-Clustering with k = 3, points changing cluster



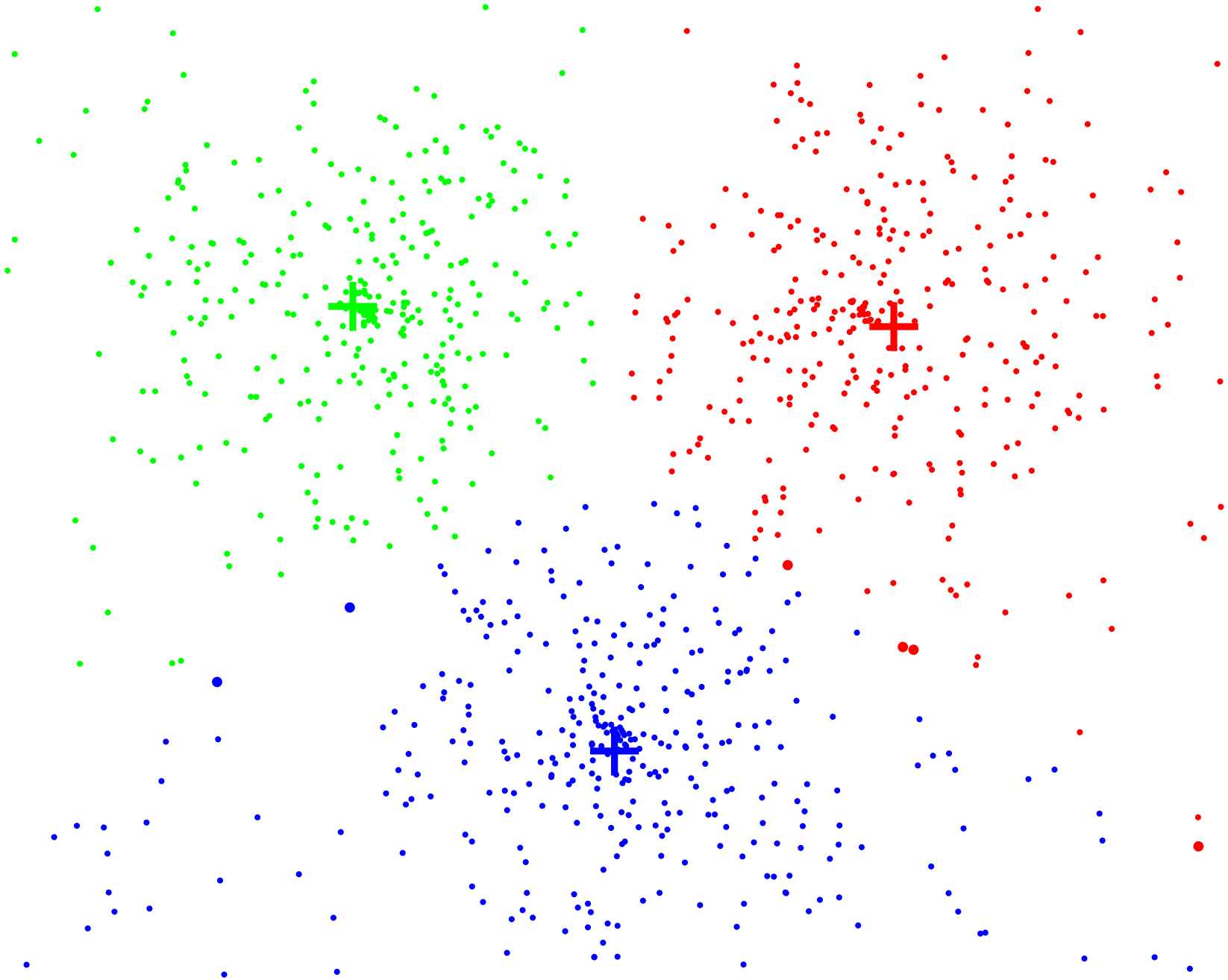
k-Clustering with k = 3, after fth iteration



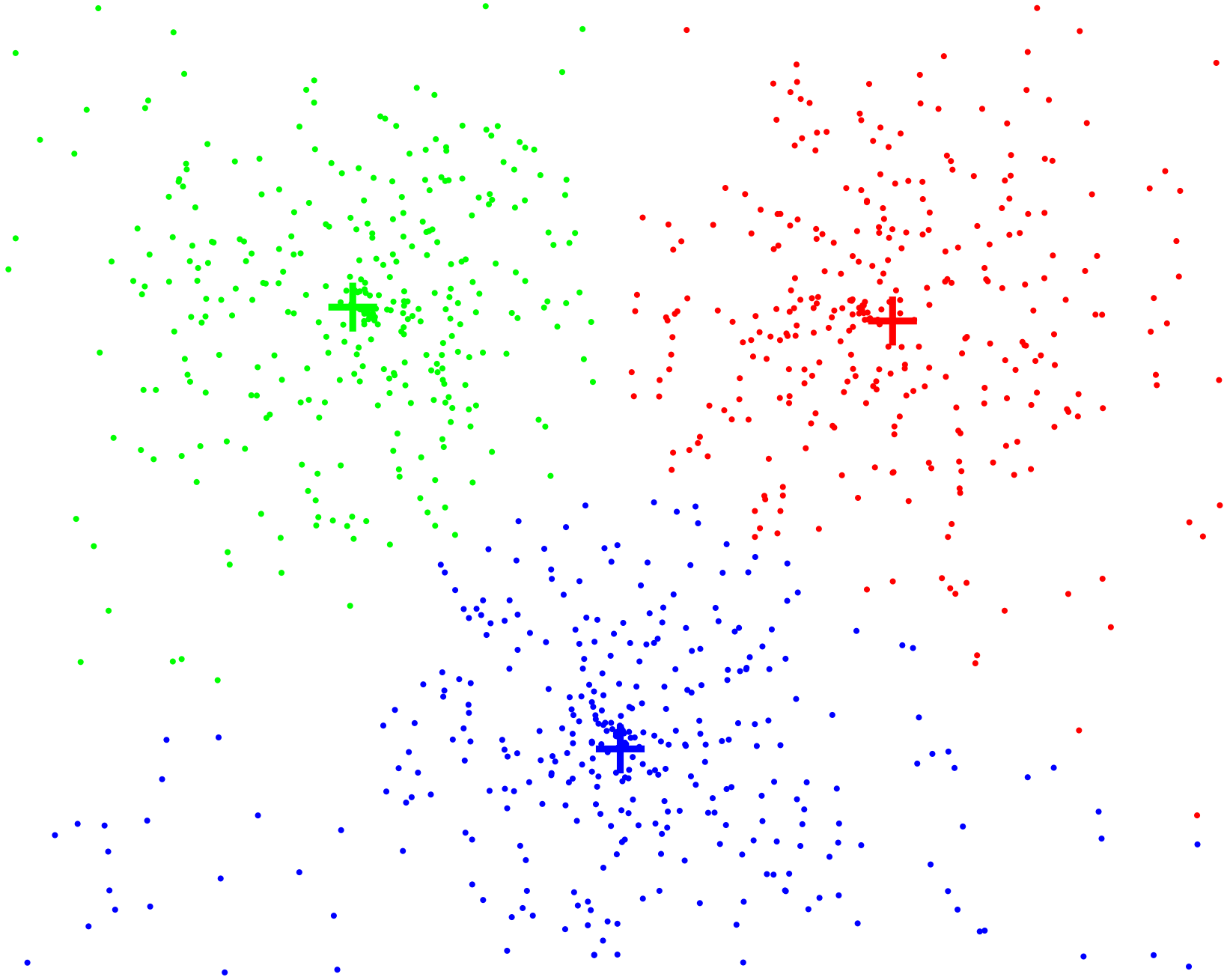
k-Clustering with k = 3, points changing cluster



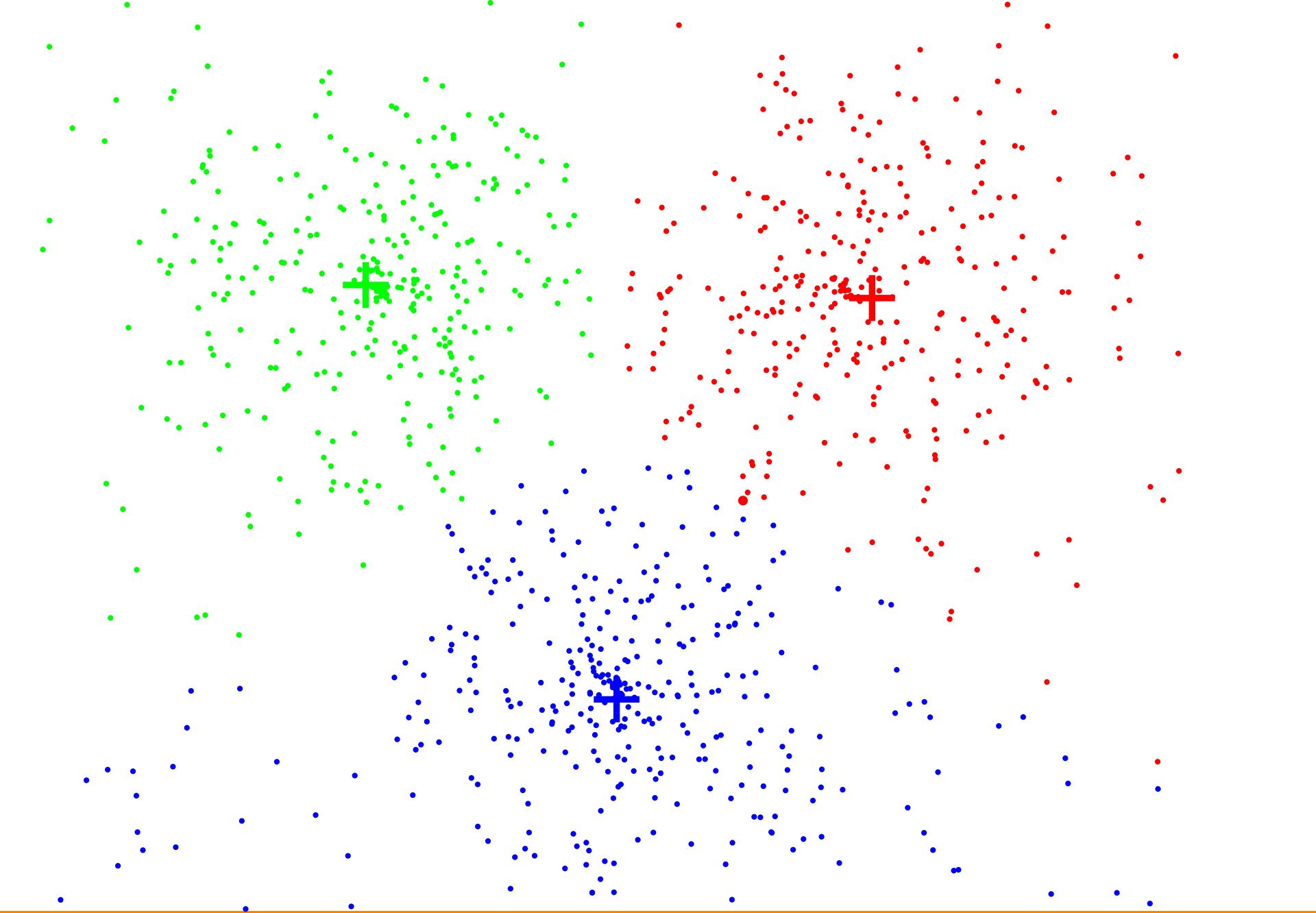
k-Clustering with k = 3, after sixth iteration



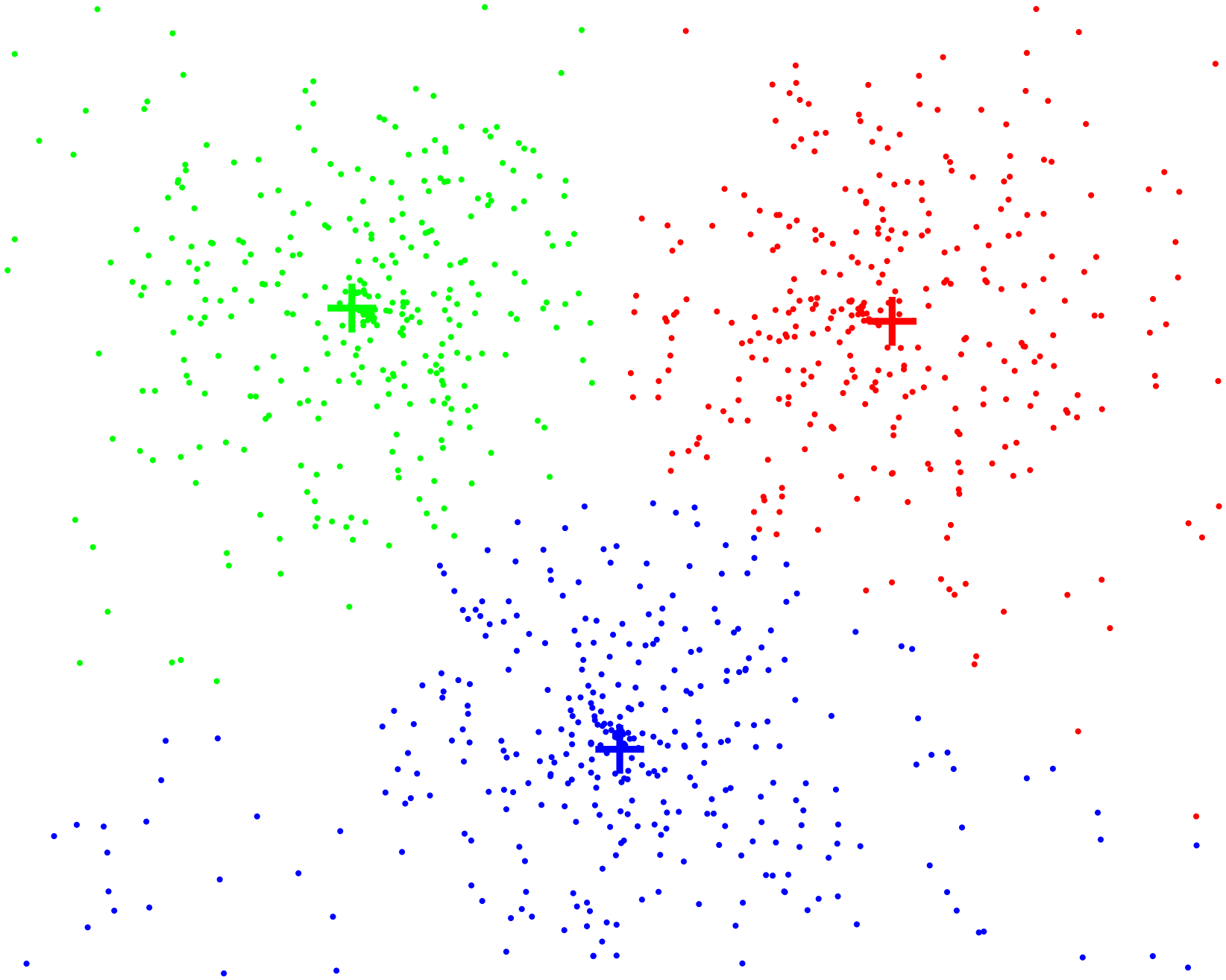
k-Clustering with k = 3, points changing cluster



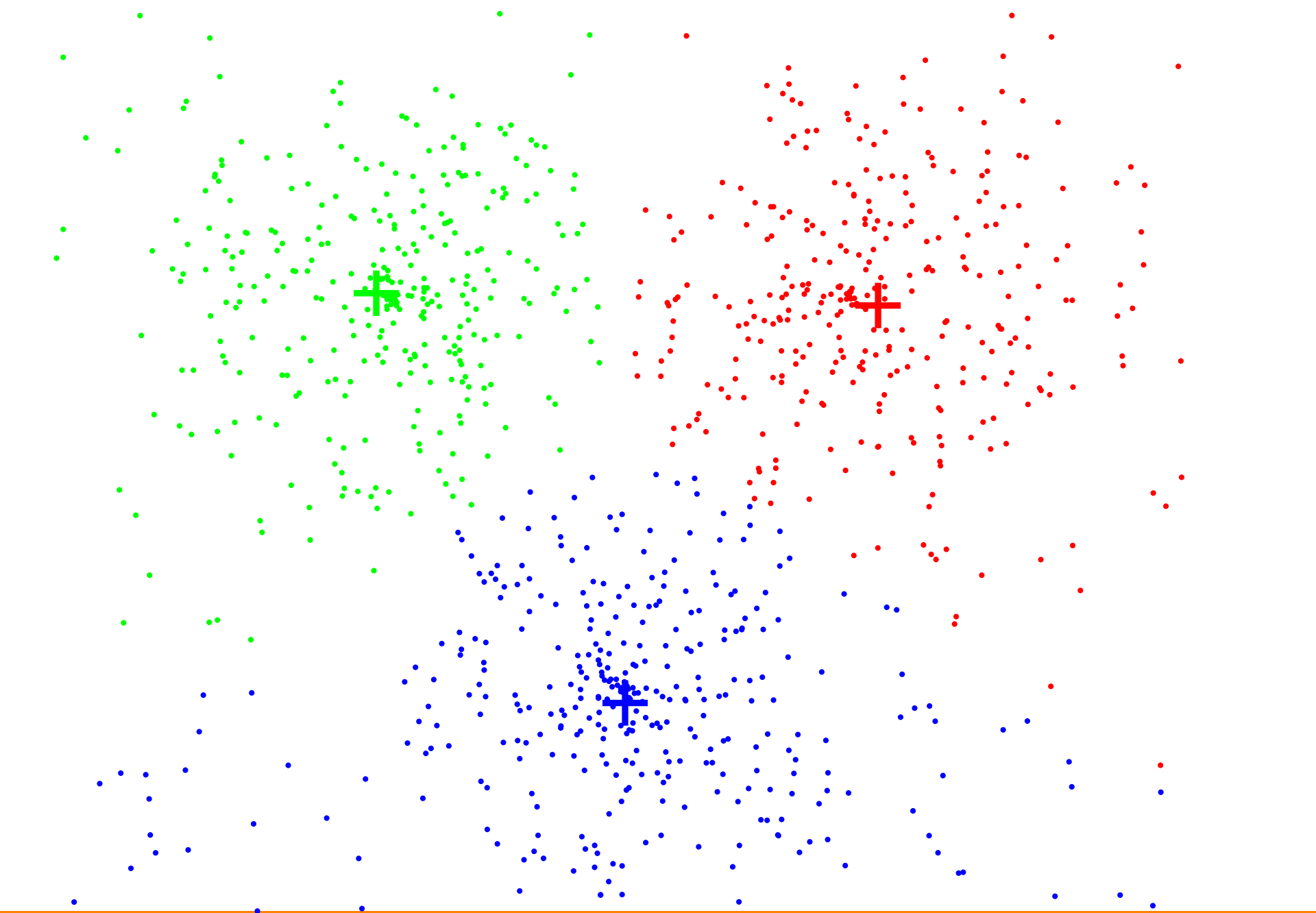
k-Clustering with k = 3, after seventh iteration



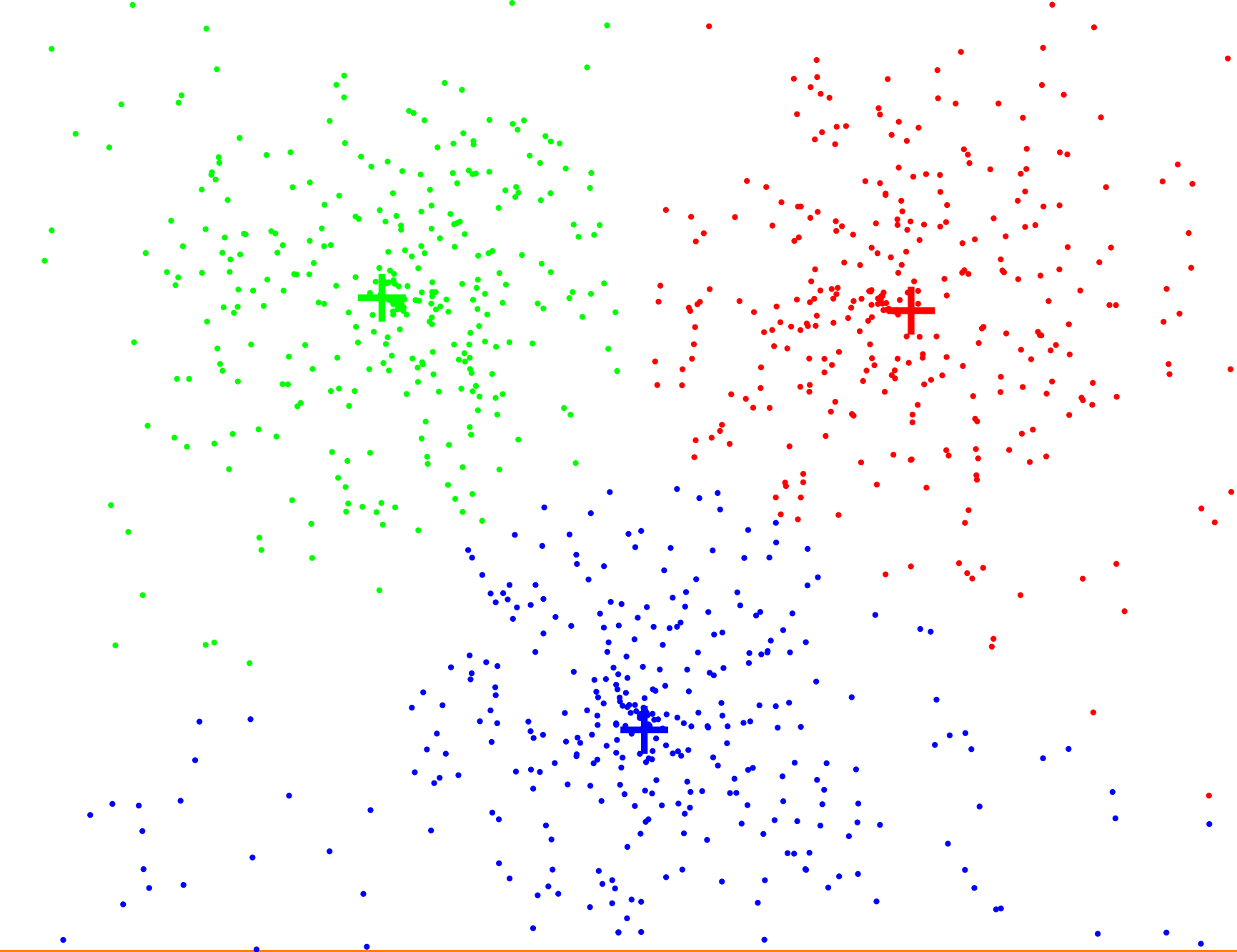
k-Clustering with k = 3, points changing cluster



k-Clustering with k = 3, after eighth iteration



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



k-Clustering with k = 3, nished.

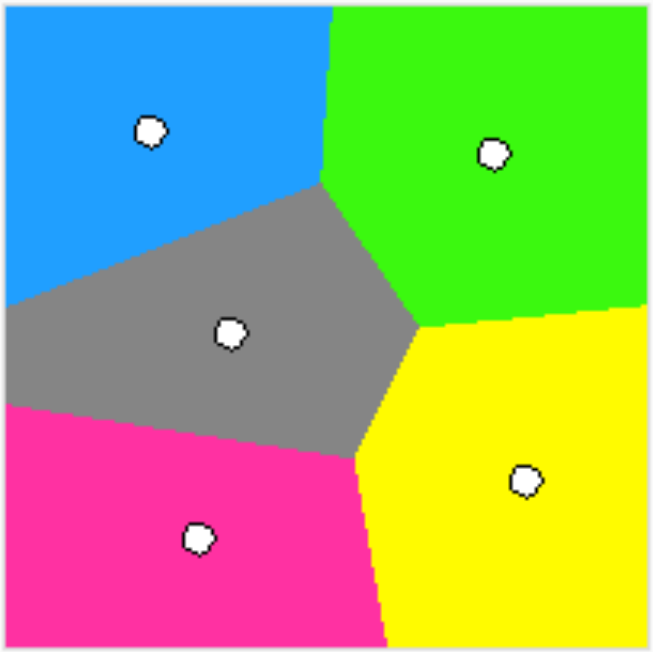
k-means clustering

Advantages:

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

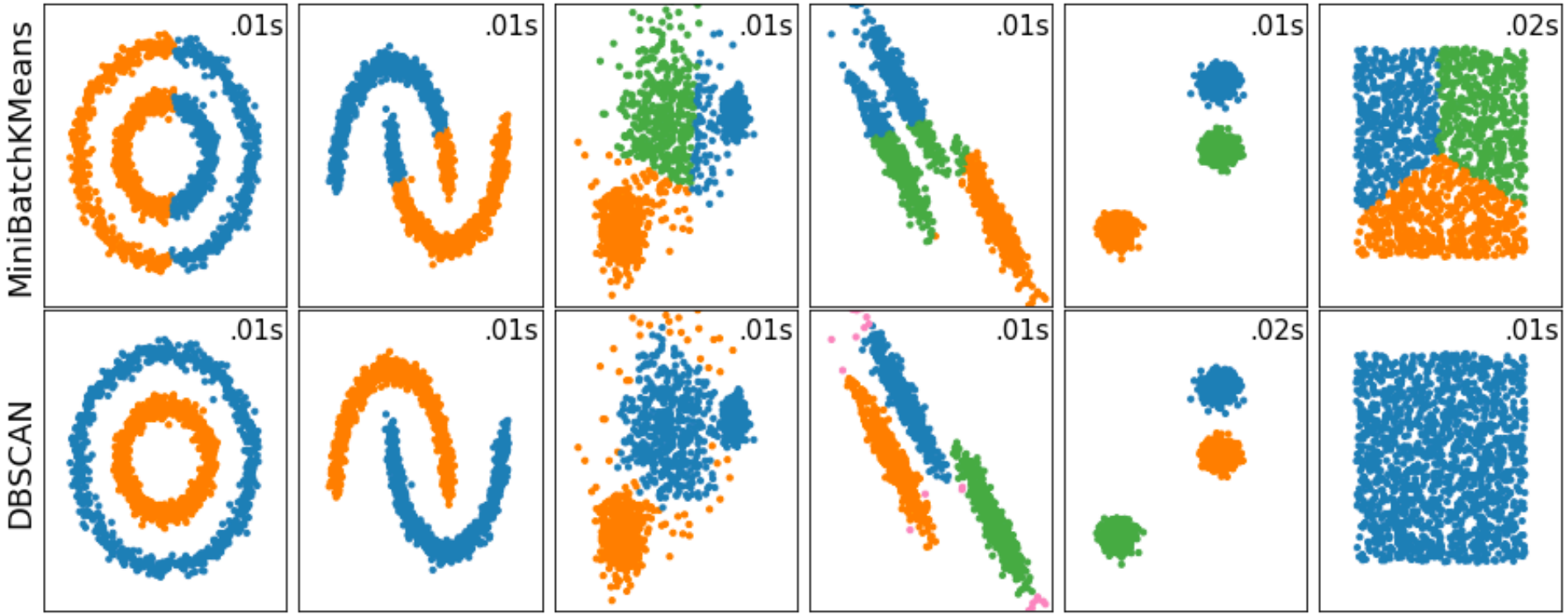
Disadvantages:

- 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



http://informationandvisualization.de/blog/kmeans-and-voronoi-tesselation-built-processing

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

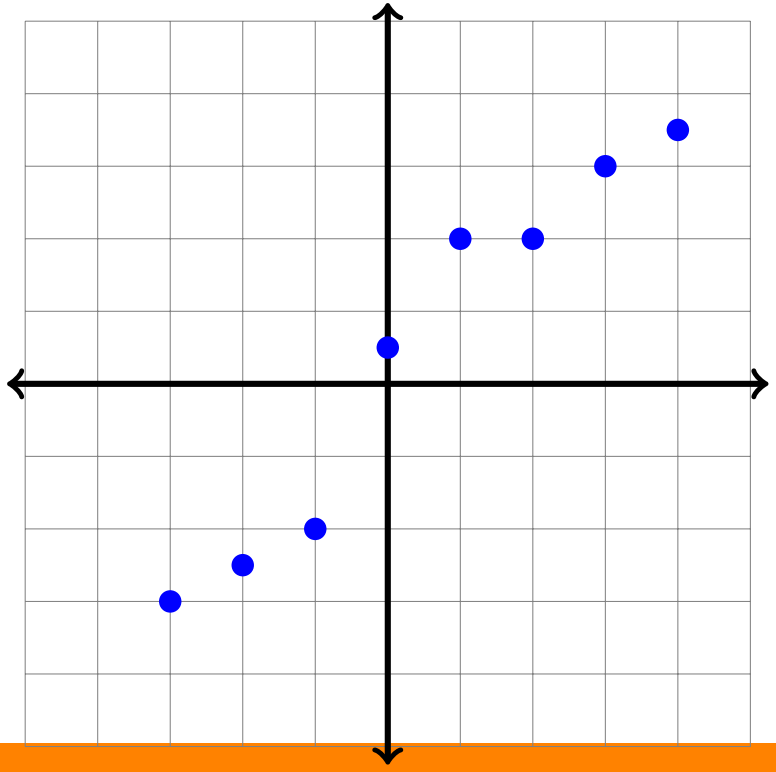
http://www.sthda.com/english/articles/29-cluster-validation-essentials/96-determining-the-optimal-number-of-clusters-3-must-know-methods/#elbow-method

Linear Regression

52

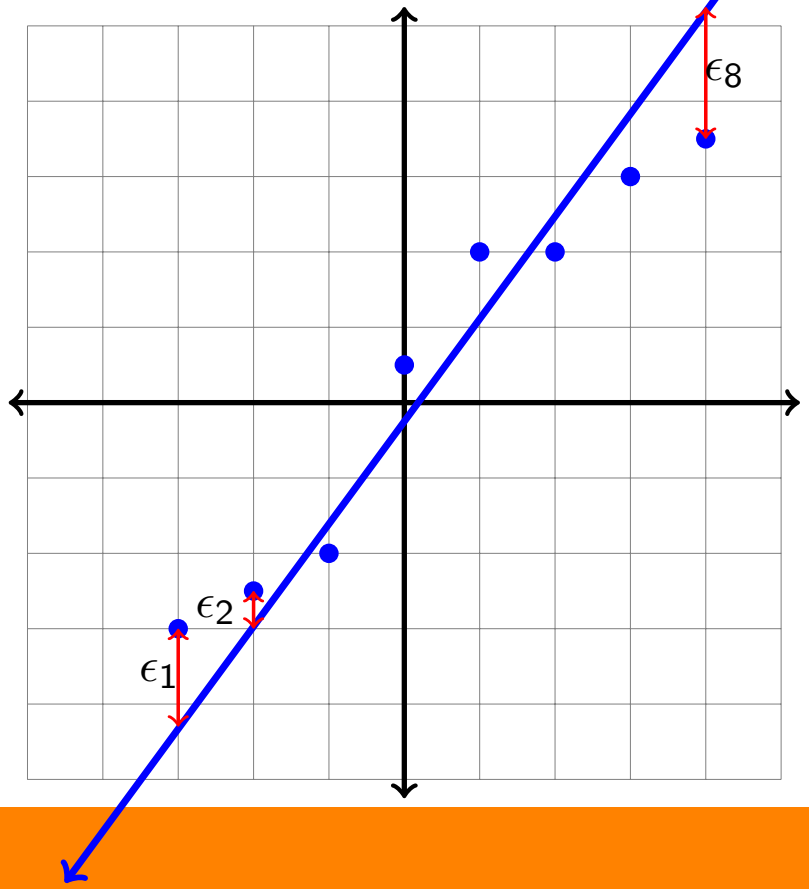
Linear regression

- Given a set of data points in the plane, (x1; y1); ….; (xn; yn), 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, $(x_1, y_1), \ldots, (x_n, y_n)$,
let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ 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:

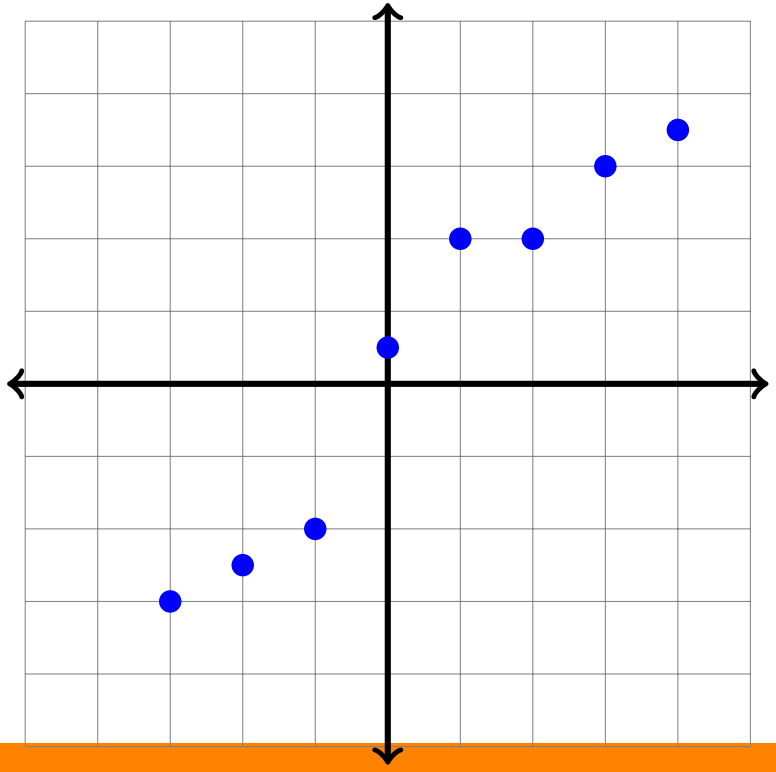
$$b = \bar{y} - m\bar{x}.$$

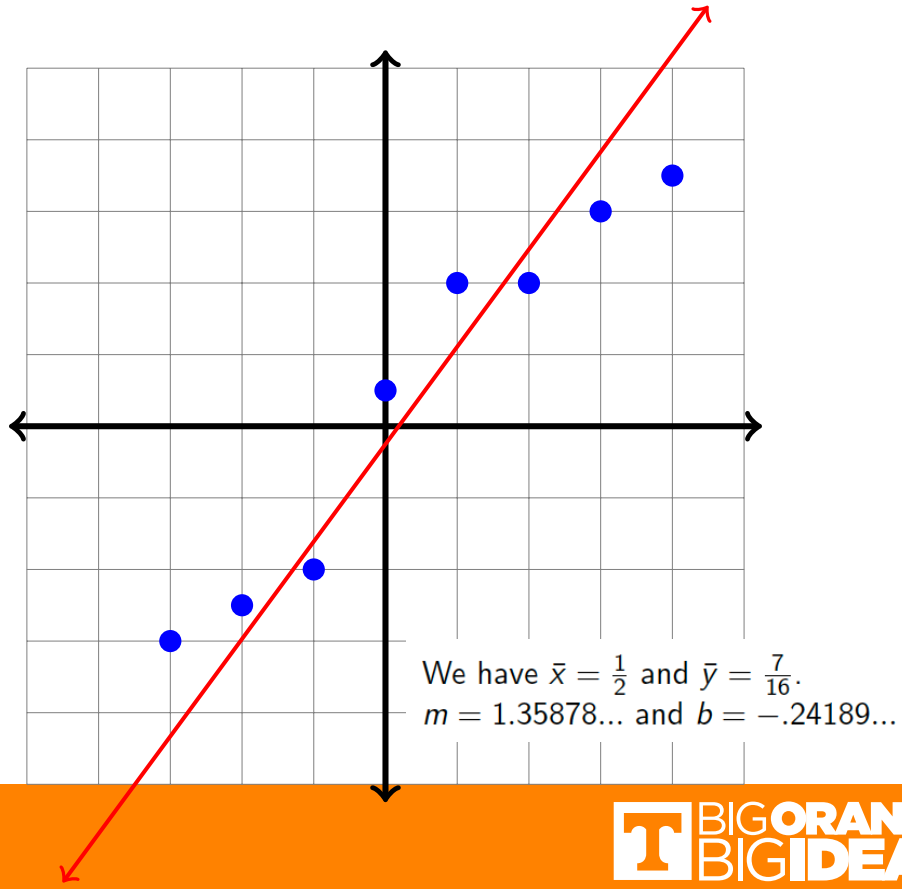
The **slope** of the *best fit line* is:

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

Linear regression

- Given a set of data points in the plane, (x1; y1); ….; (xn; yn), find an equation of the line y = mx + b that best describes the points





Example of Datasets

57

NHANES Dataset

58

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 Questions

- 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 f they are consuming high amounts of fiber on day one, are they also consuming high amounts on day 2?

Medicaid-Vital Statistics Data

62

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 Steps

65

Project (I)

Step 0: search for datasets

- Discussion in class of datasets identified

Problem 5: Find an interesting dataset

- Now that you've heard a bit about the projects we will be doing later in this class, find a dataset that you **could** use for the project. It should be large enough to allow for interesting analysis and non-trivial results. You don't have to download the data; make sure you do NOT add it to your GitHub repository.
- In the box below, describe the dataset you have selected in one or two paragraphs. Include its source (with url); how the data were collected; significance of the data; the number of rows, objects, or data points; what information is contained in each; data types (int, str, char, float, etc.) and numerical ranges where appropriate; and any details about the file that would be necessary for loading the data into a program.
- NOTE: You will not have to use this dataset for your project! This is an exercise in finding and describing data for research.

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

BIG ORANGE. BIG IDEAS.®

BIG ORANGE BIG IDEAS