K-means

Mengdi Huai

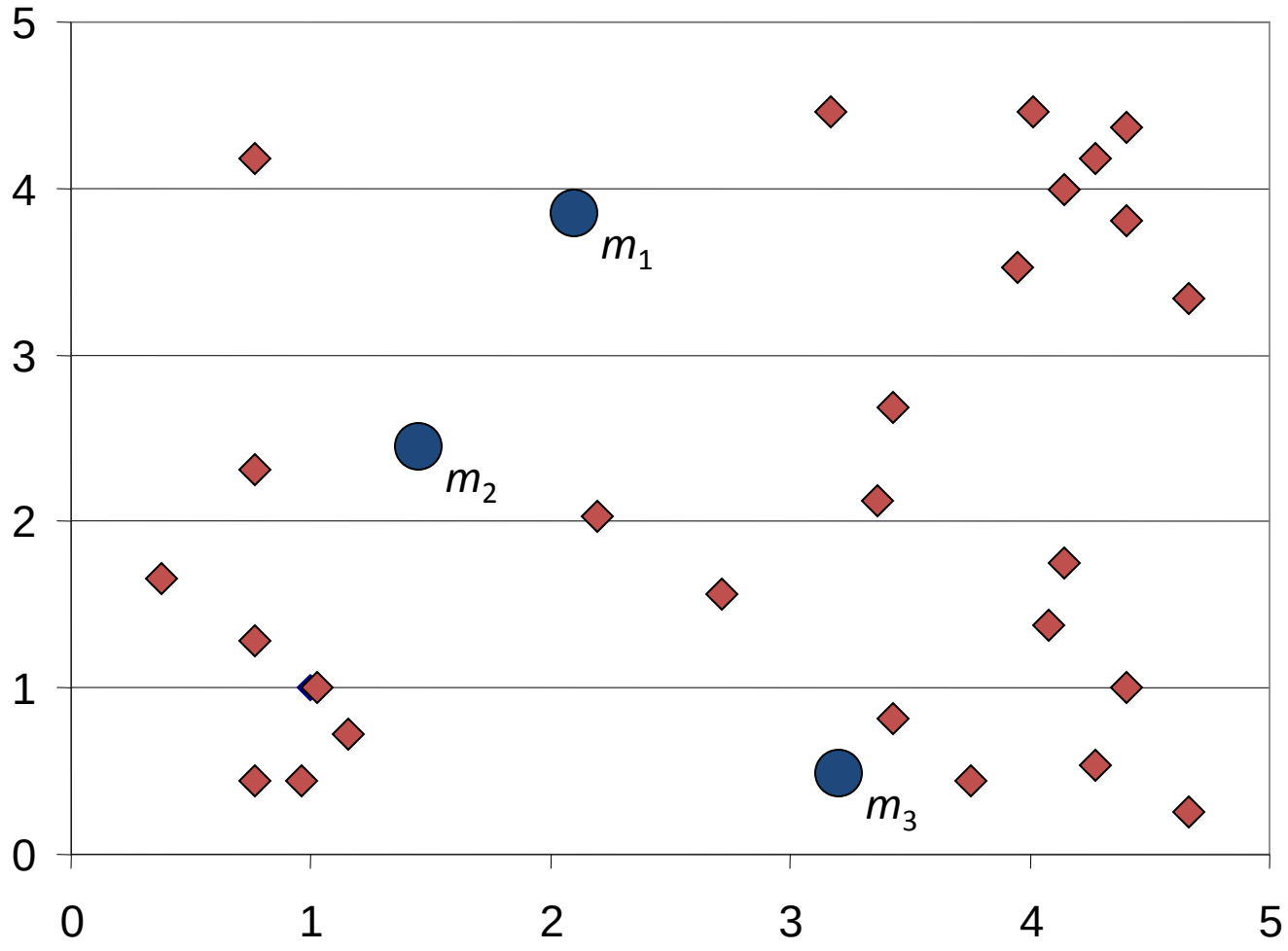
SUNY Buffalo

K-means

- **Partition $\{x_1,\ldots,x_n\}$ into K clusters**
  - K is predefined
- **Initialization**
  - Specify the initial cluster centers (centroids)
- **Iteration until no change**
  - For each data x_i
    - Calculate the distances between x_i and the K centroids
    - (Re)assign x_i to the cluster whose centroid is the closest to x_i
  - Update the cluster centroids based on current assignment

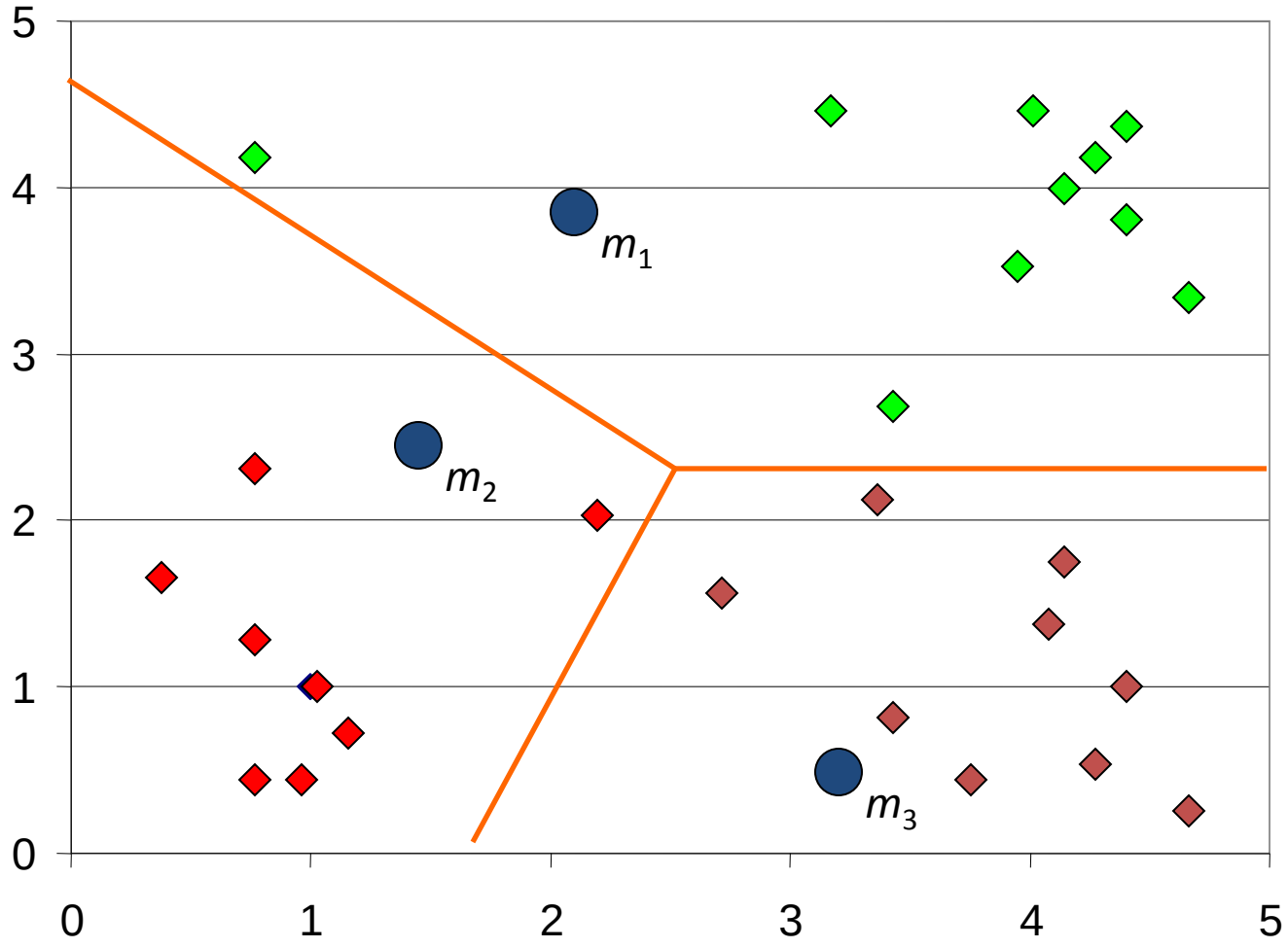
K-means: Initialization

Initialization: Determine the three cluster centers



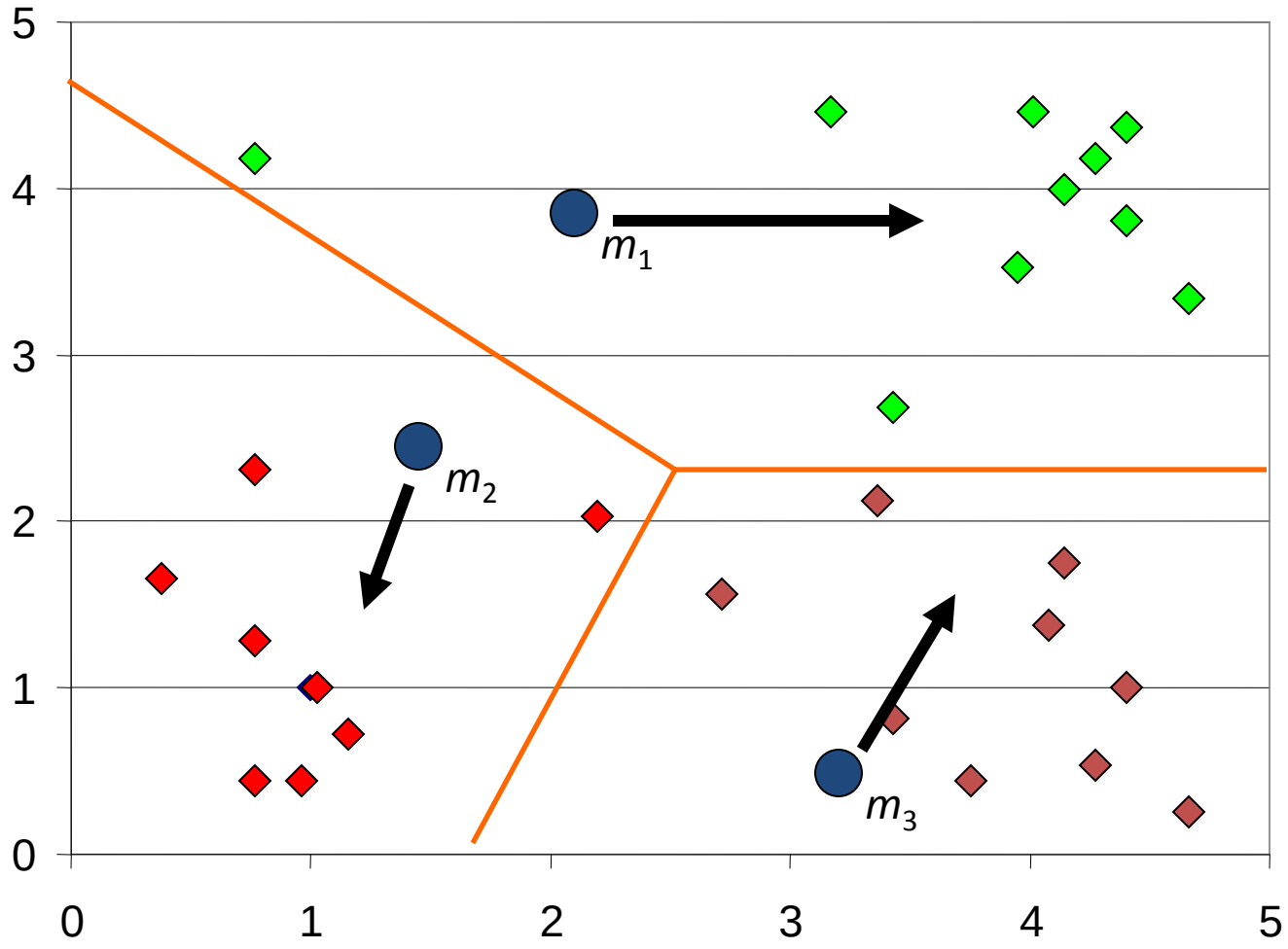
K-means Clustering: Cluster Assignment

Assign each data point to the cluster which has the closet distance from the centroid to the data point



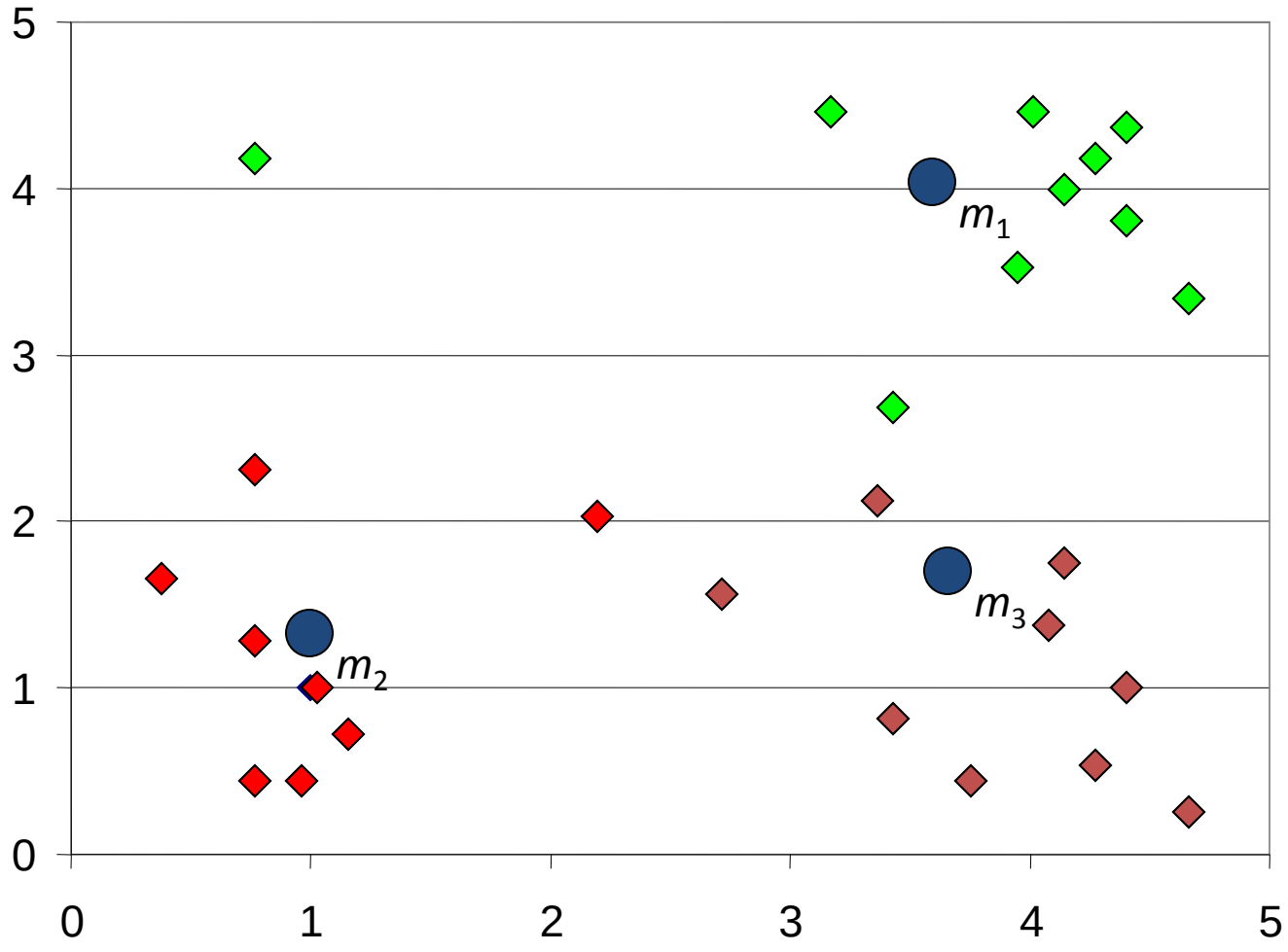
K-means Clustering: Update Cluster Centroid

Compute cluster centroid as the center of the points in the cluster



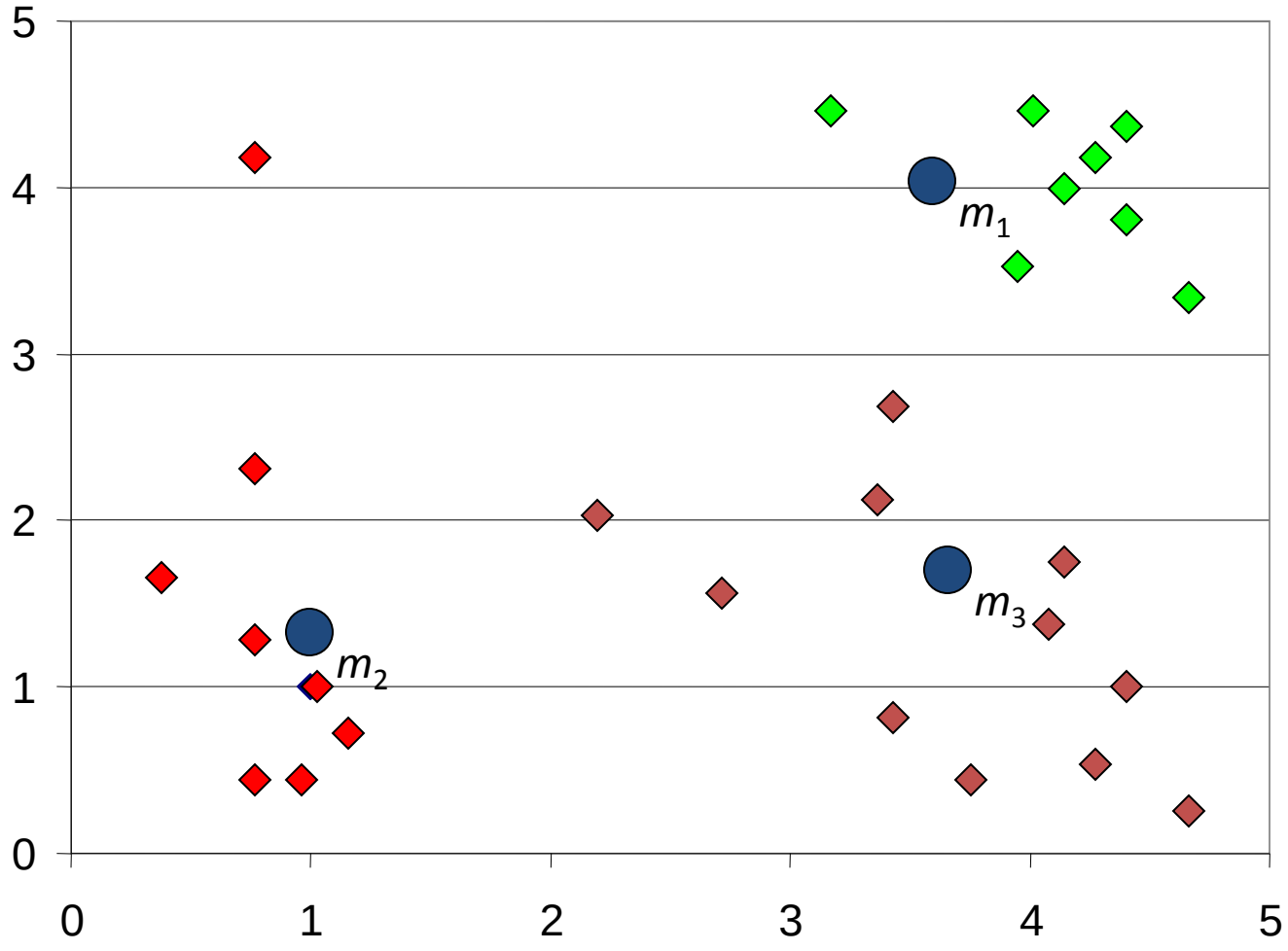
K-means Clustering: Update Cluster Centroid

Compute cluster centroid as the center of the points in the cluster



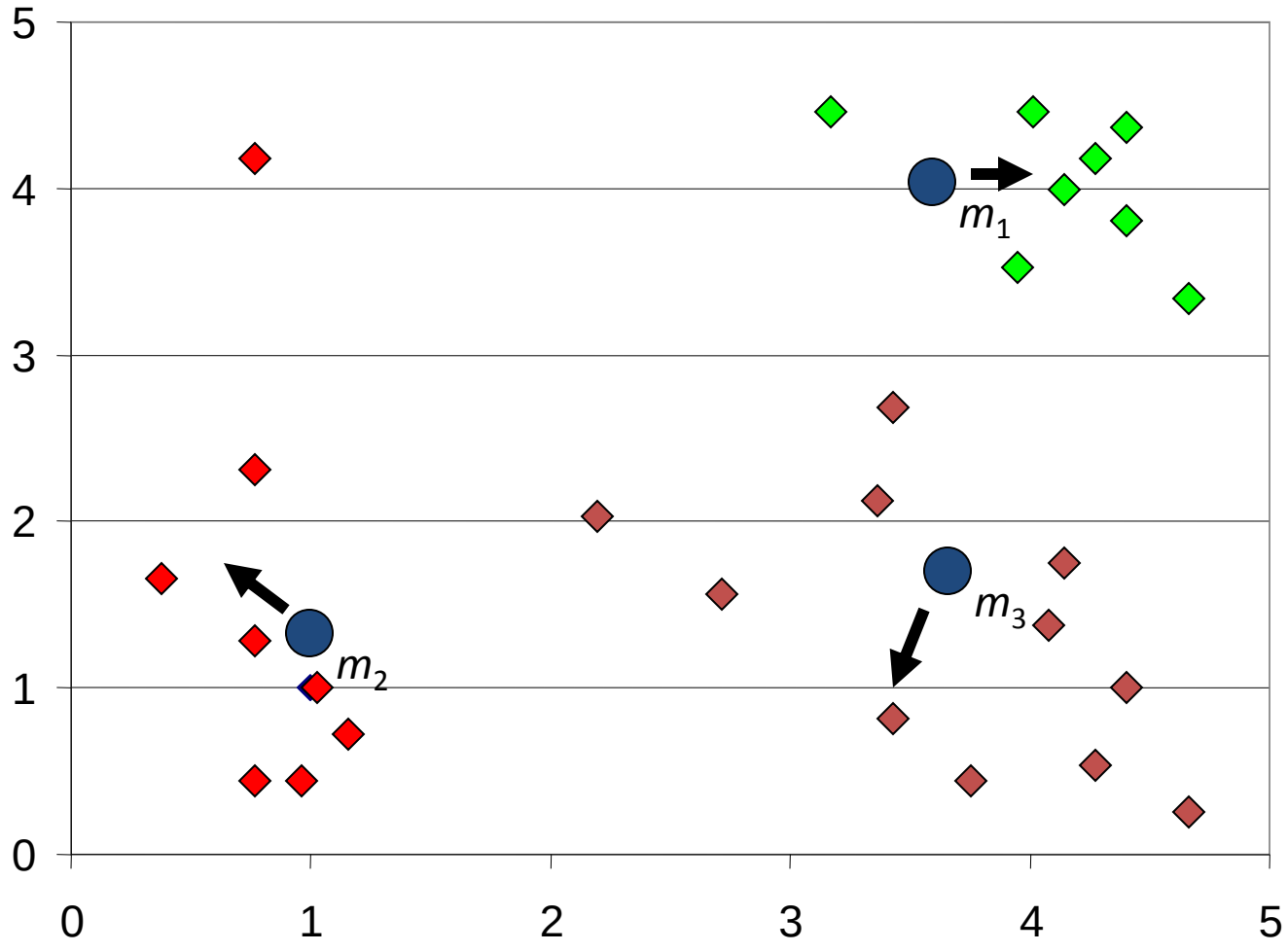
K-means Clustering: Cluster Assignment

Assign each data point to the cluster which has the closet distance from the centroid to the data point



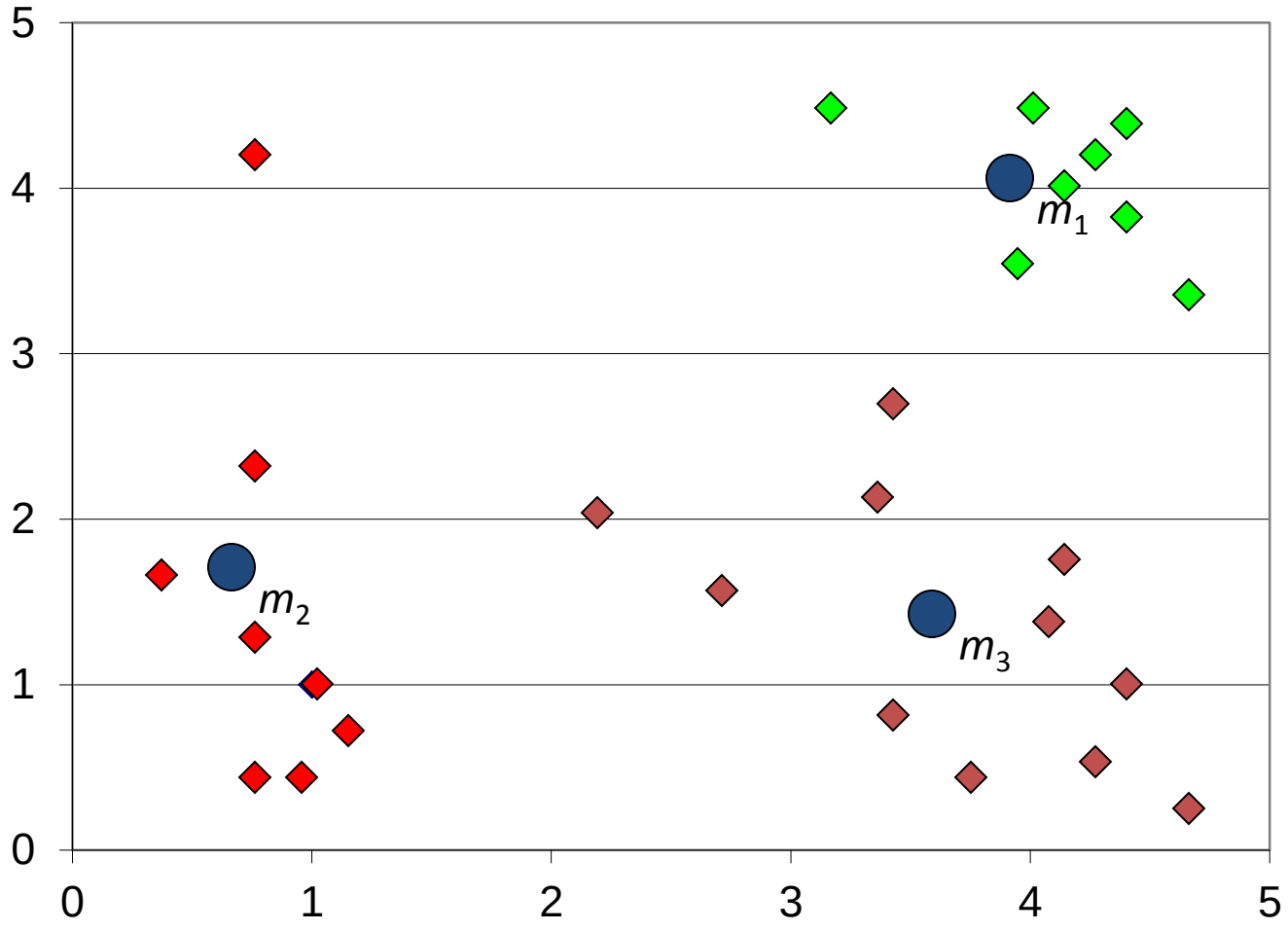
K-means Clustering: Update Cluster Centroid

Compute cluster centroid as the center of the points in the cluster

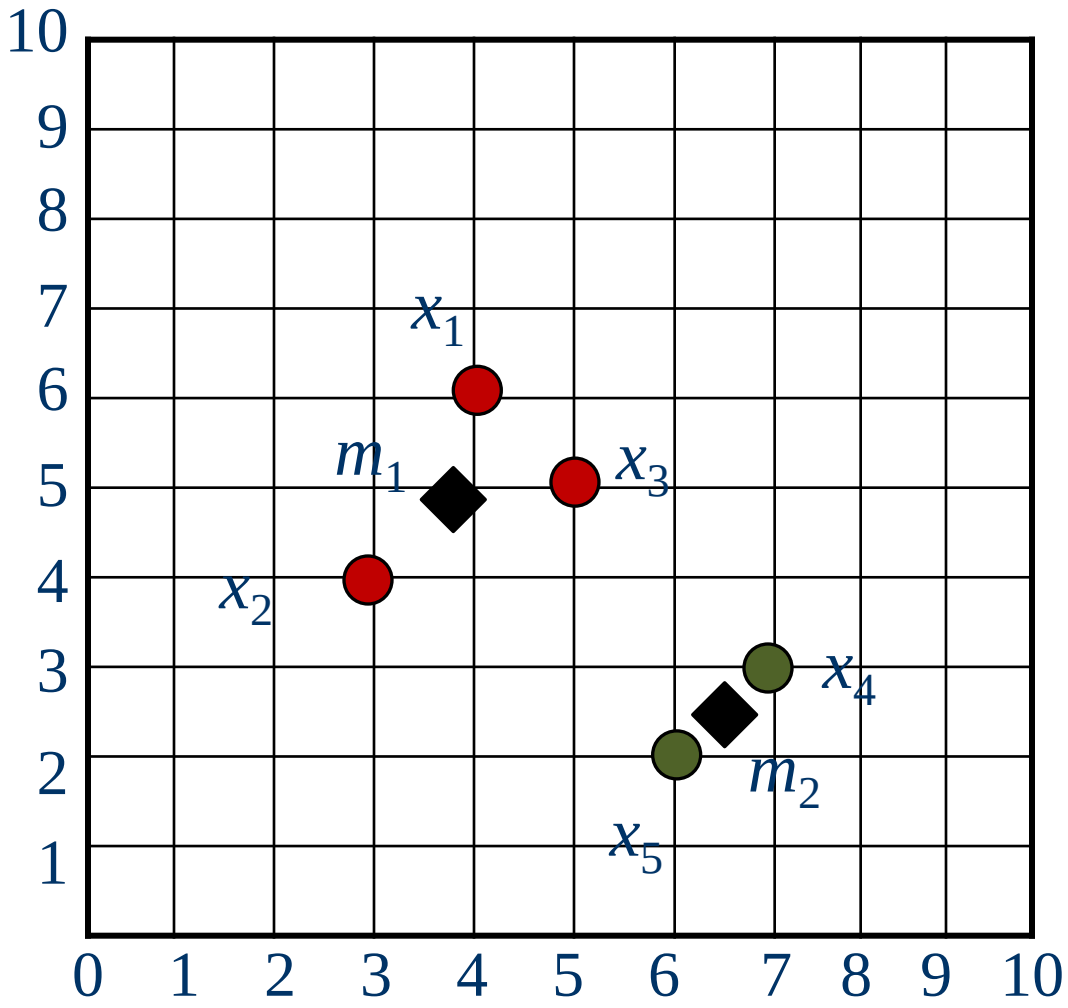


K-means Clustering: Update Cluster Centroid

Compute cluster centroid as the center of the points in the cluster



Example—Cluster Centroid Computation



Given the cluster assignment with two initial centers m1 and m2, compute the centers of the two clusters

Comments on the K-Means Method

- **Strength**
  - Efficient
  - Easy to implement
- **Issues**
  - Need to specify *K*, the number of clusters

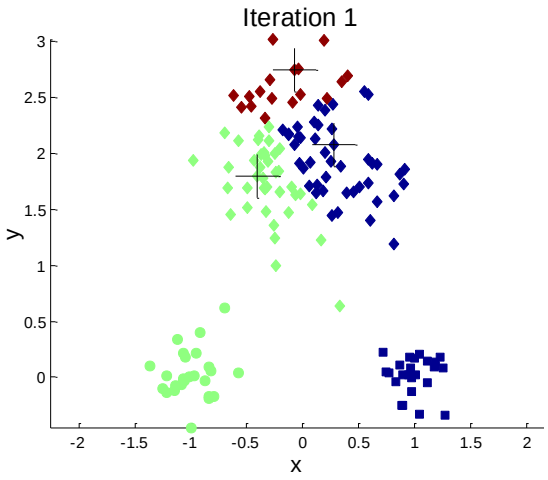
Problems with Selecting Initial Points

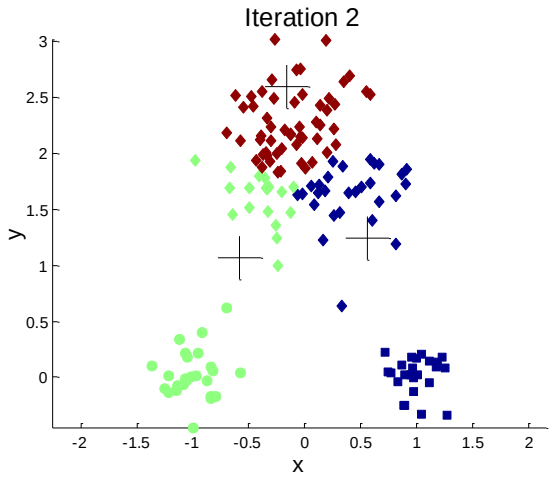
- If there are *K* 'real' clusters then the chance of selecting one centroid from each cluster is small
  - Chance is relatively small when *K* is large
  - If clusters are the same size, *n*, then

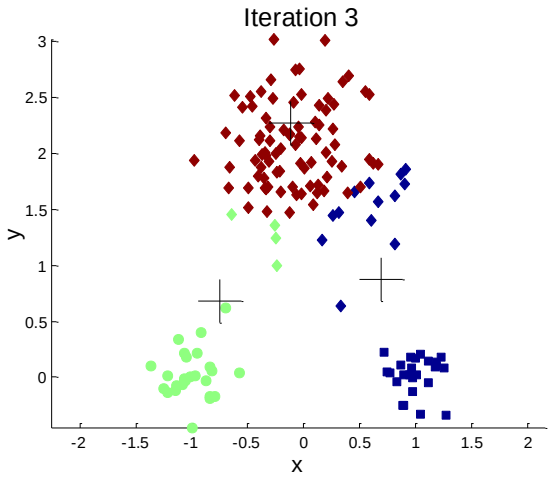
$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

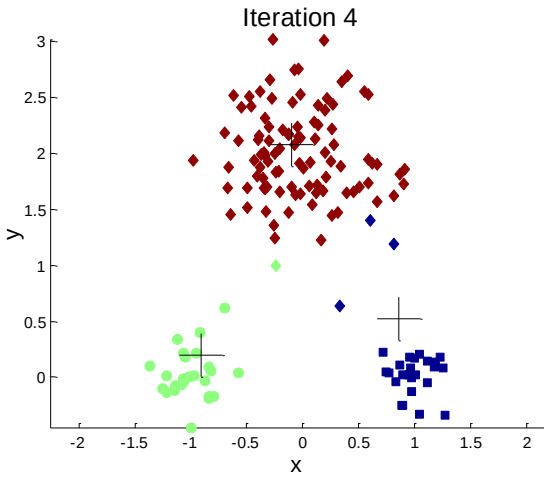
  - For example, if K = 10, then probability = $10!/10^{10}$ = 0.00036
  - Sometimes the initial centroids will readjust themselves in 'right' way, and sometimes they don't

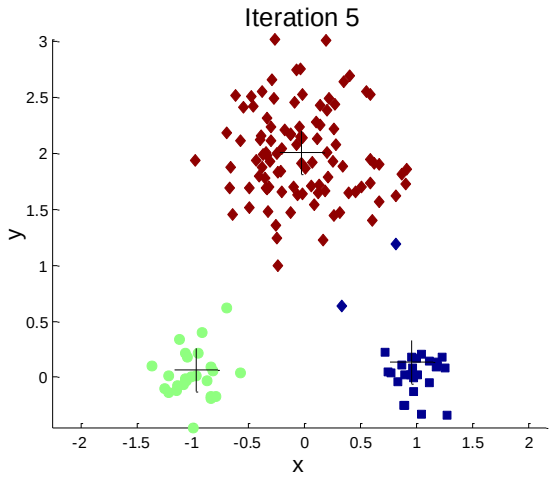
Importance of Choosing Initial Centroids

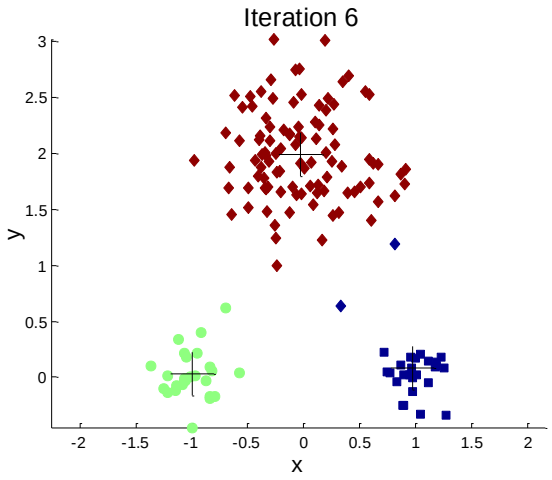




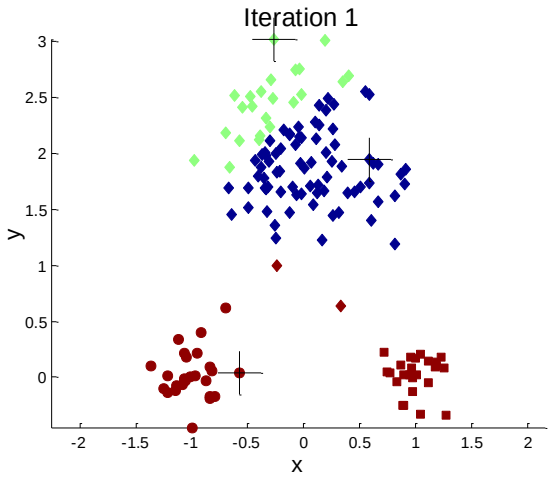


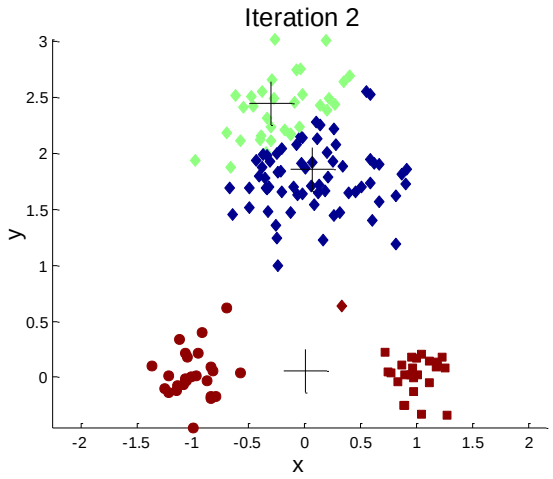


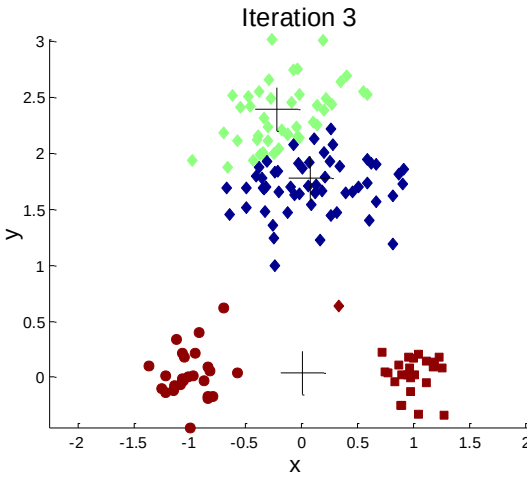


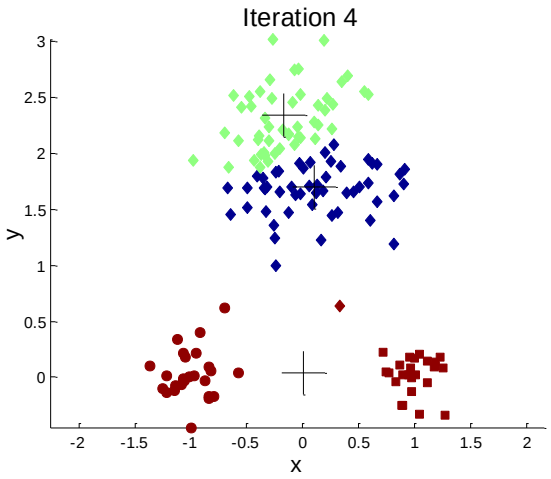


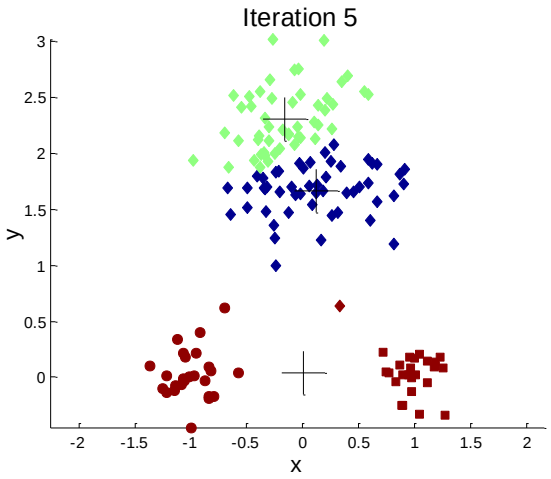
Importance of Choosing Initial Centroids











Solutions to Initial Centroids Problem

- Multiple runs
  - Average the results or choose the one that has the smallest sum of the squared errors
- Sample and use hierarchical clustering to determine initial centroids
- Select more than *K* initial centroids and then select among these initial centroids
  - Select most widely separated
- Postprocessing—Use K-means' results as other algorithms' initialization
- Bisecting K-means
  - Not as susceptible to initialization issues

Pre-processing and Post-processing

- **Pre-processing**
  - Normalize the data
  - Eliminate outliers
- **Post-processing**
  - Eliminate small clusters that may represent outliers