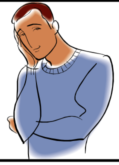
Clustering

UE 141 Spring 2013

Jing Gao
SUNY Buffalo

Clustering

- **Data**

user	items
	orange, banana, apple, water
	orange, apple, water
	rice, bread, milk, eggs
	bread, milk, eggs, water
	yogurt, milk, eggs

- **Goal**
  - Increase profit while maintaining advertising cost!

Clustering: From Data to Knowledge to Decision

	user	items
Group 1		orange, banana, apple, water
		orange, apple, water
		rice, bread, milk, eggs
Group 2		bread, milk, eggs, water
		yogurt, milk, eggs

Group 1: Bob and Alice bought lots of fruits

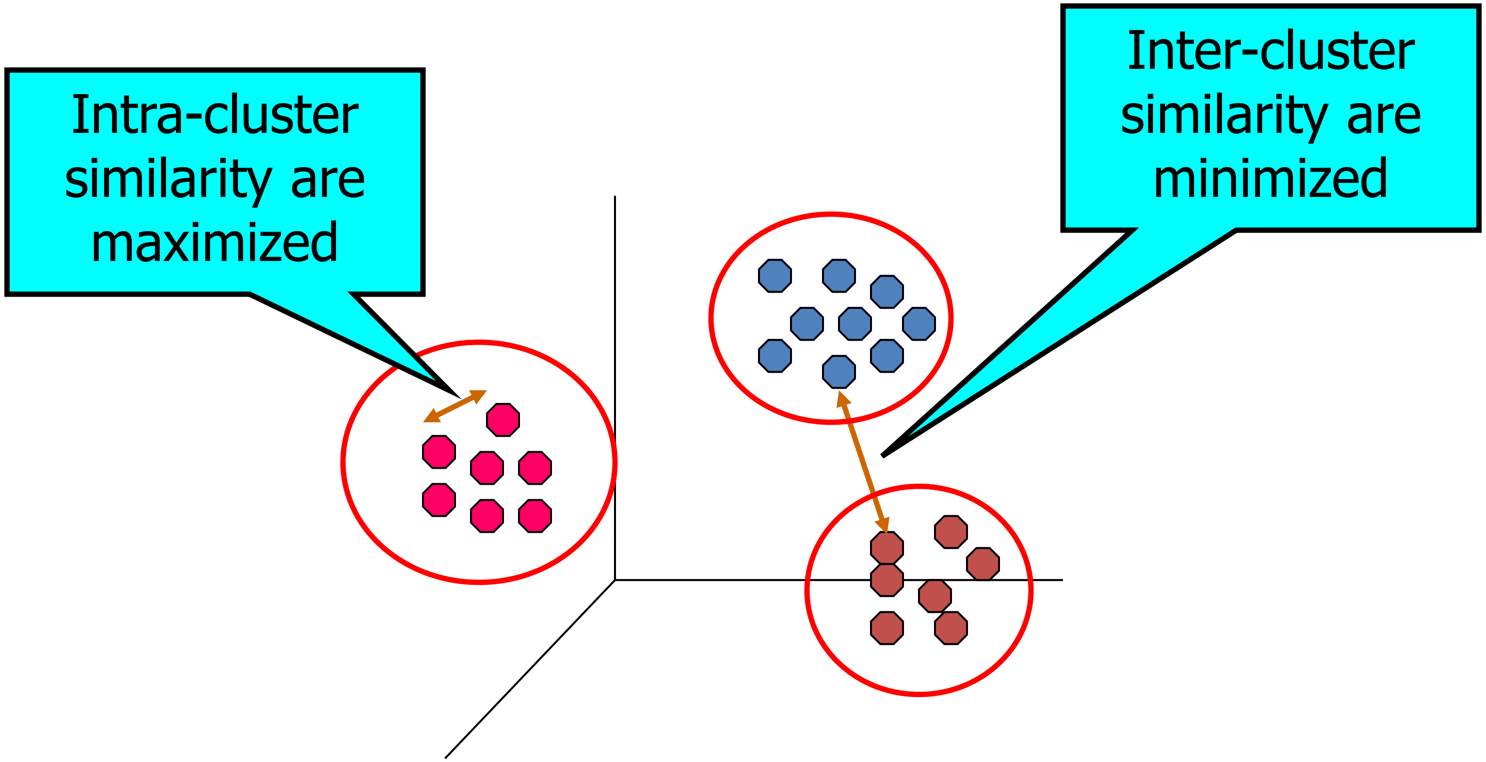
Group 2: Mary, Mike and Joe bought bread, eggs, milk

Target marketing!

Increase profit!!

Definition of Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Intra-cluster similarity are maximized

Inter-cluster similarity are minimized

Two Important Aspects

- **Properties of input data**
  - Define the similarity or dissimilarity between points
- **Requirement of clustering**
  - Define the objective and methodology

Similarity/Dissimilarity for Simple Attributes

p and *q* are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Categorical	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Continuous	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

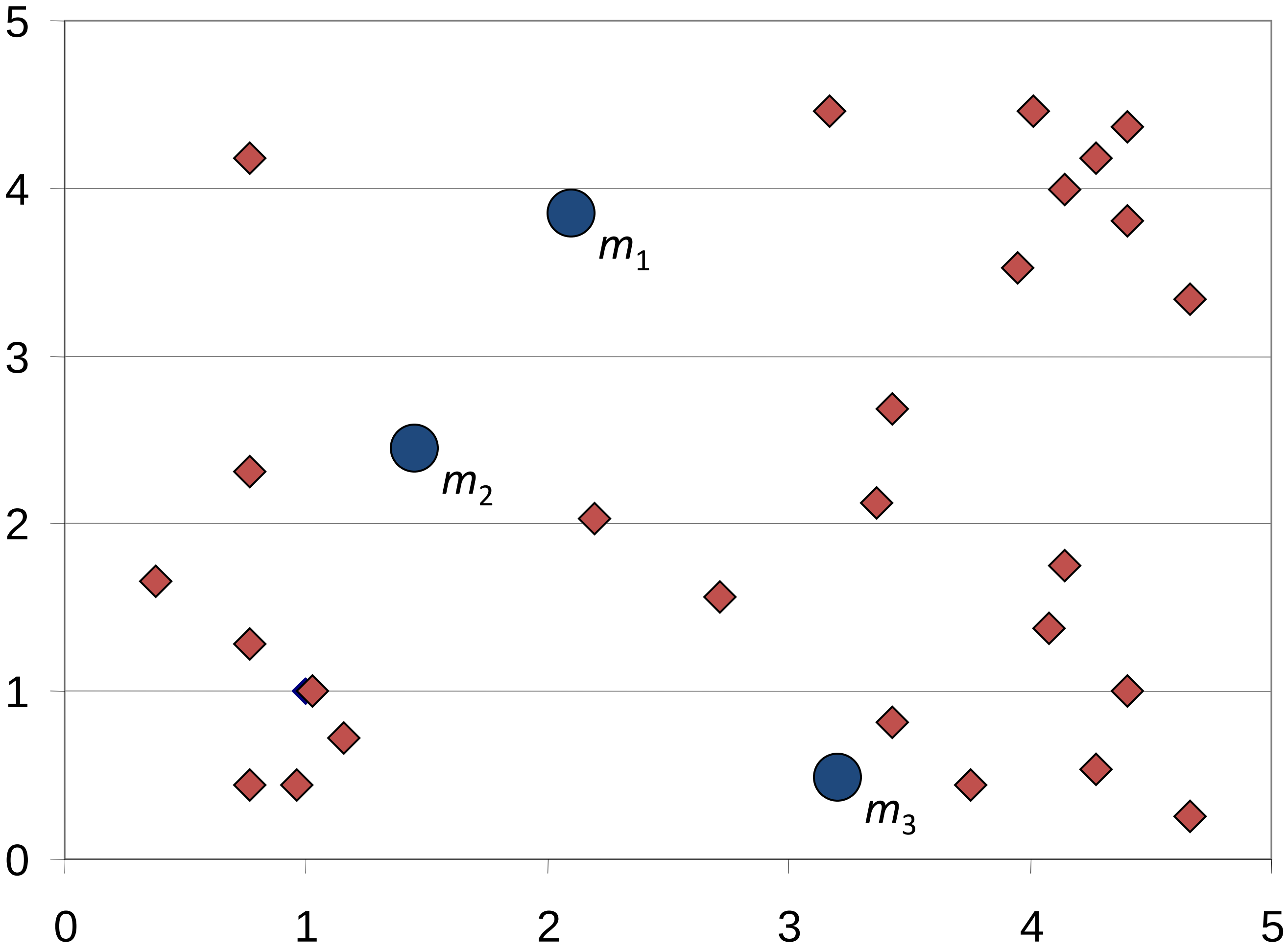
Dissimilarity and similarity between p and q

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 object 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

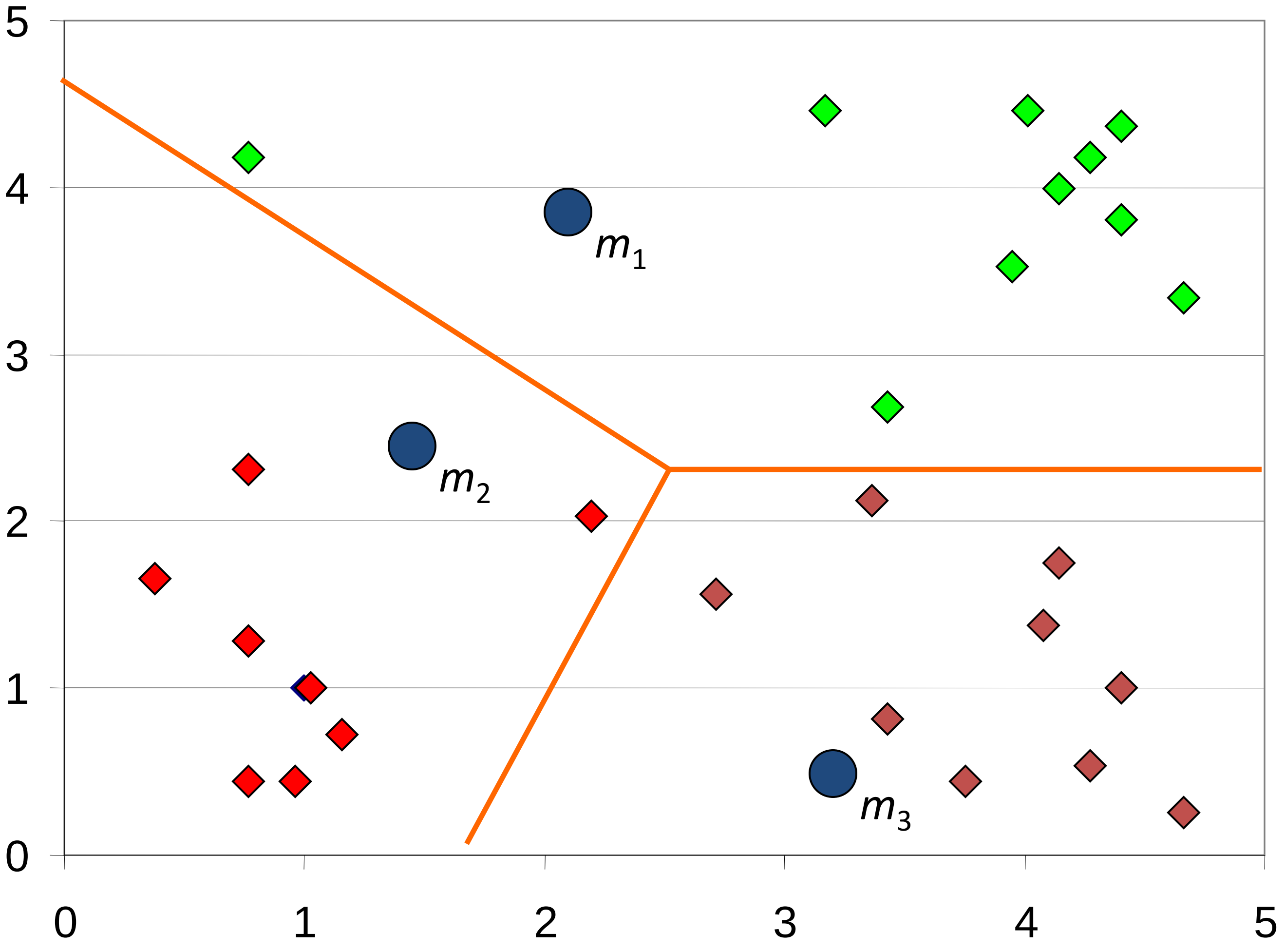
K-means: Initialization

Initialization: Determine the three cluster centers



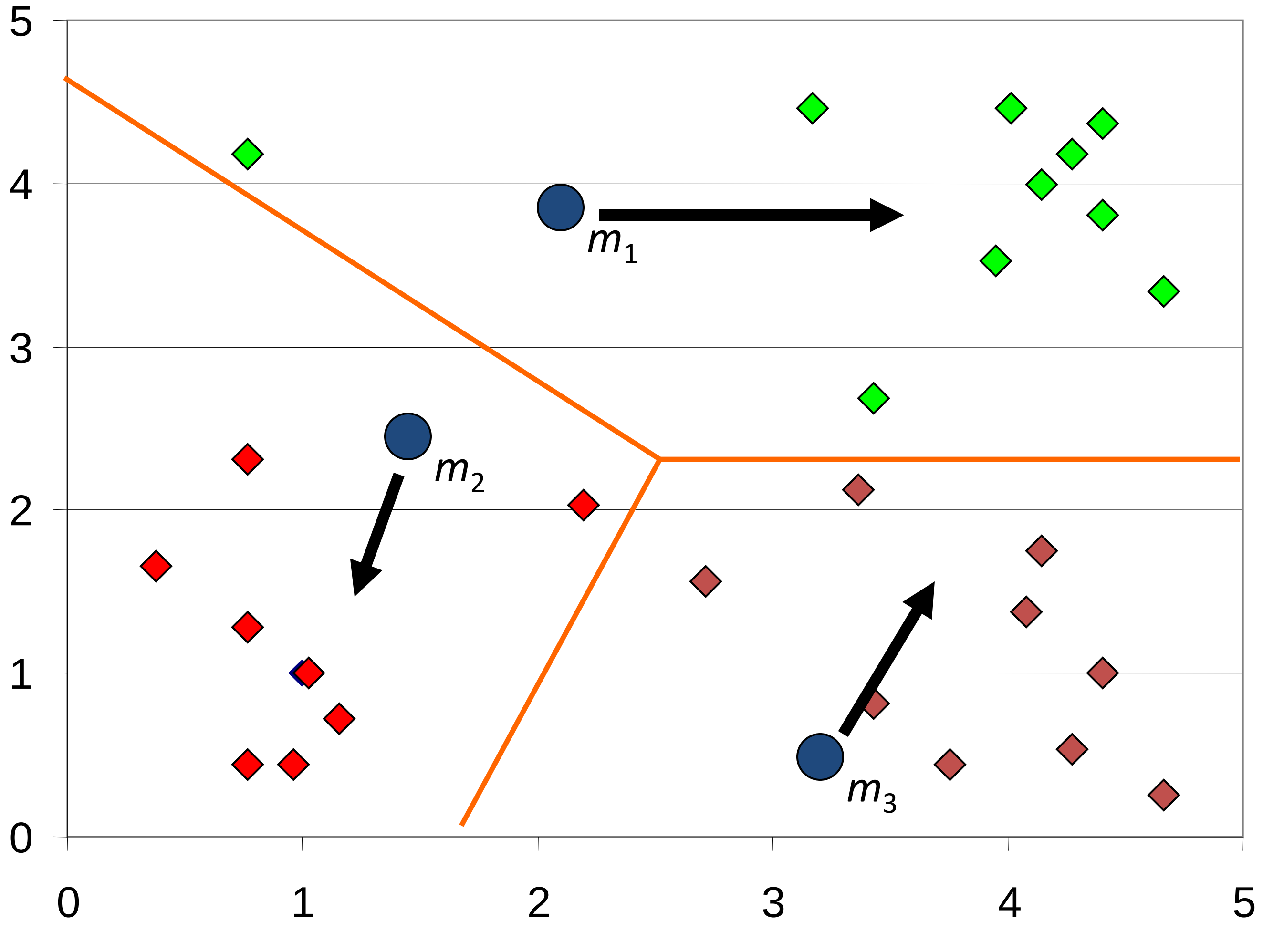
K-means Clustering: Cluster Assignment

Assign each object to the cluster which has the closet distance from the centroid to the object



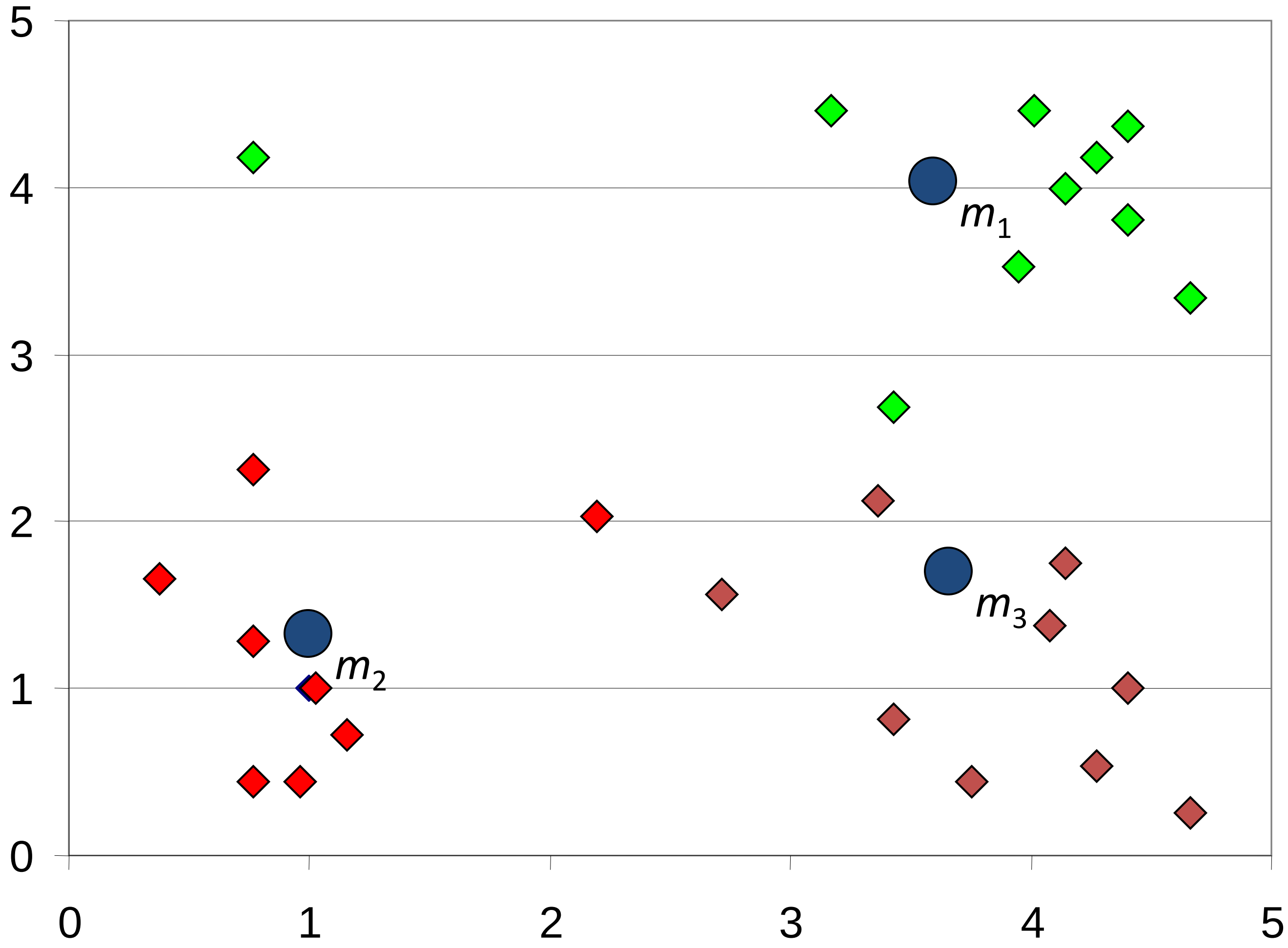
K-means Clustering: Update Cluster Centroid

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



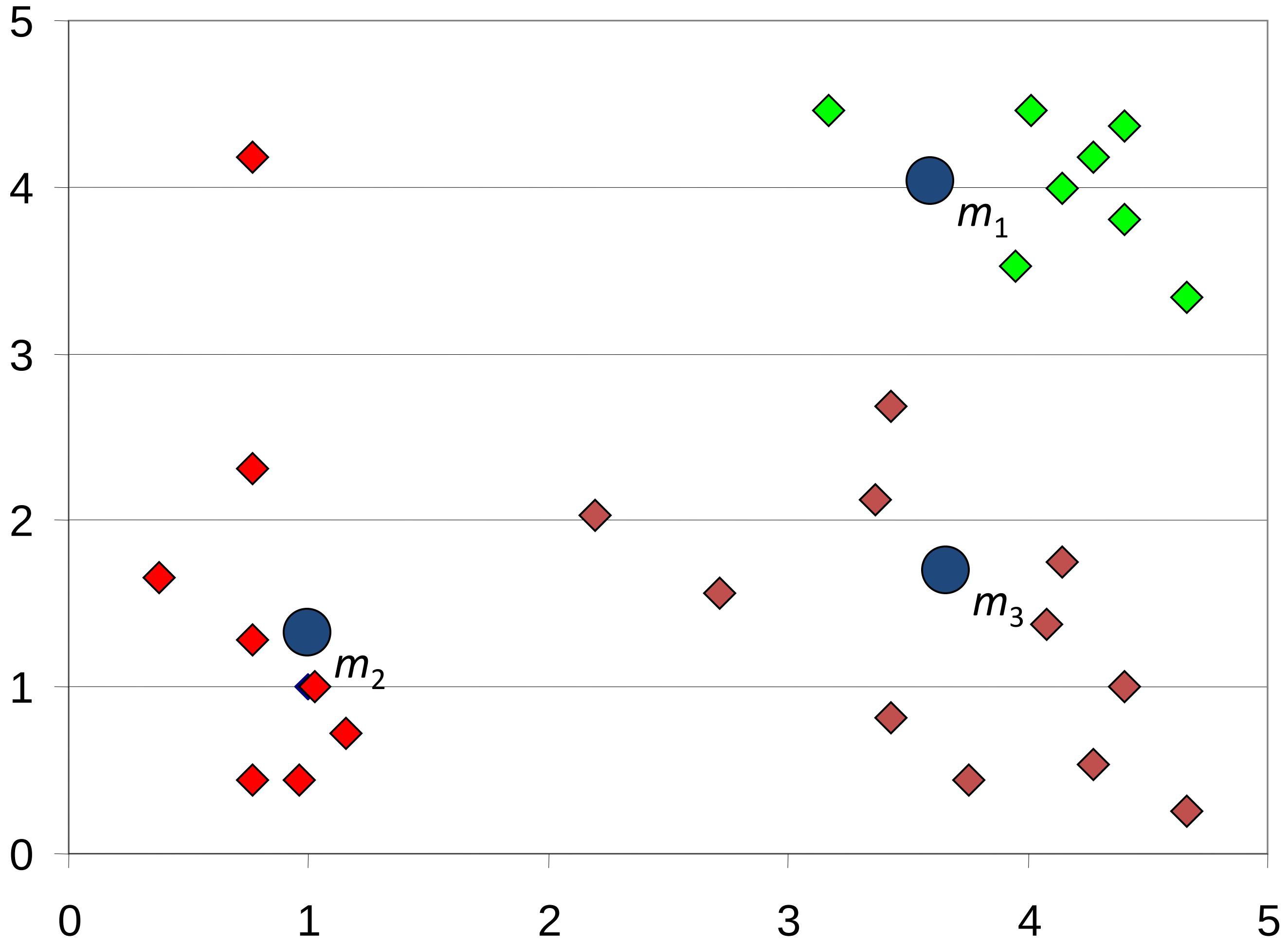
K-means Clustering: Update Cluster Centroid

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



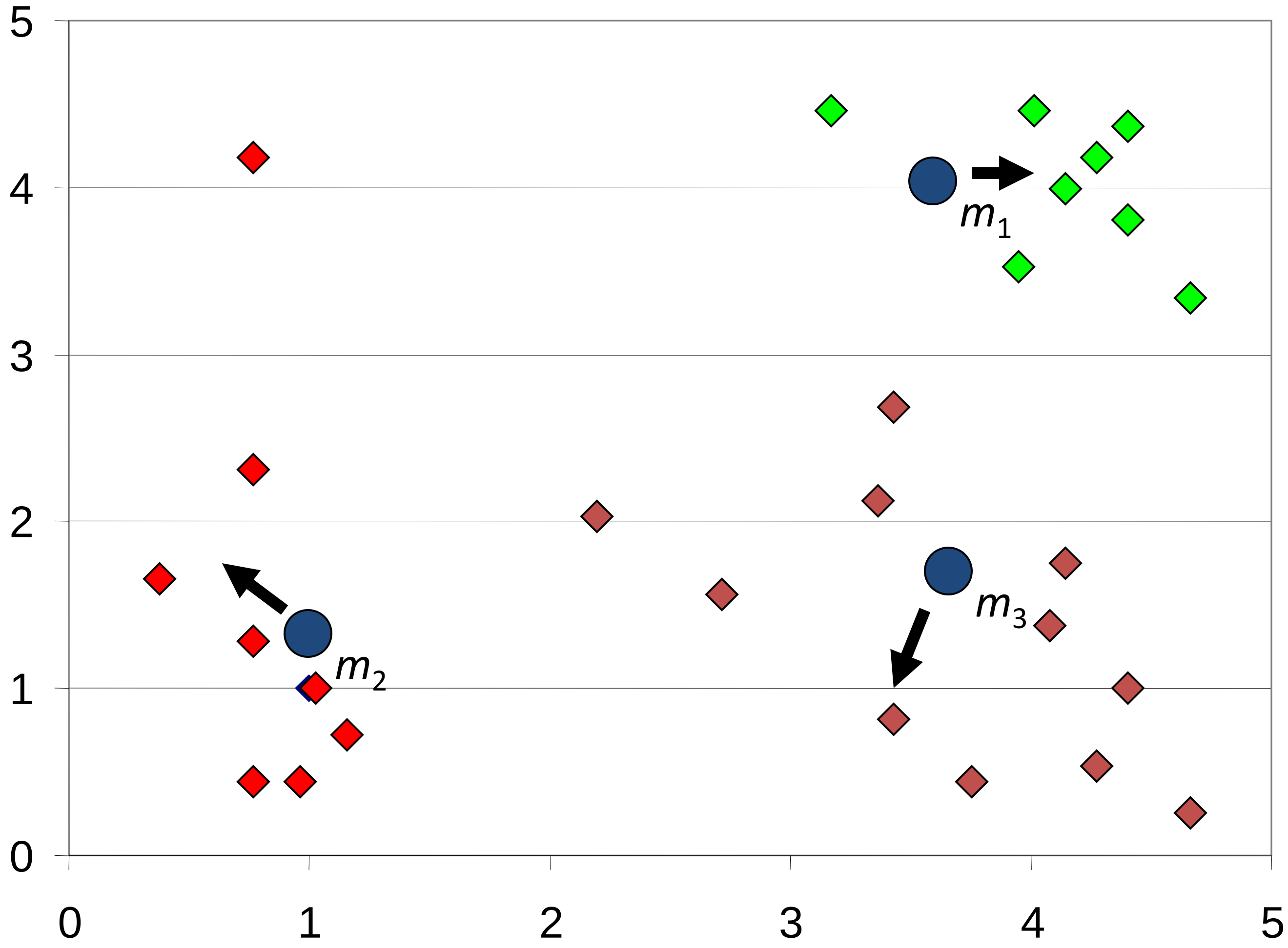
K-means Clustering: Cluster Assignment

Assign each object to the cluster which has the closet distance from the centroid to the object



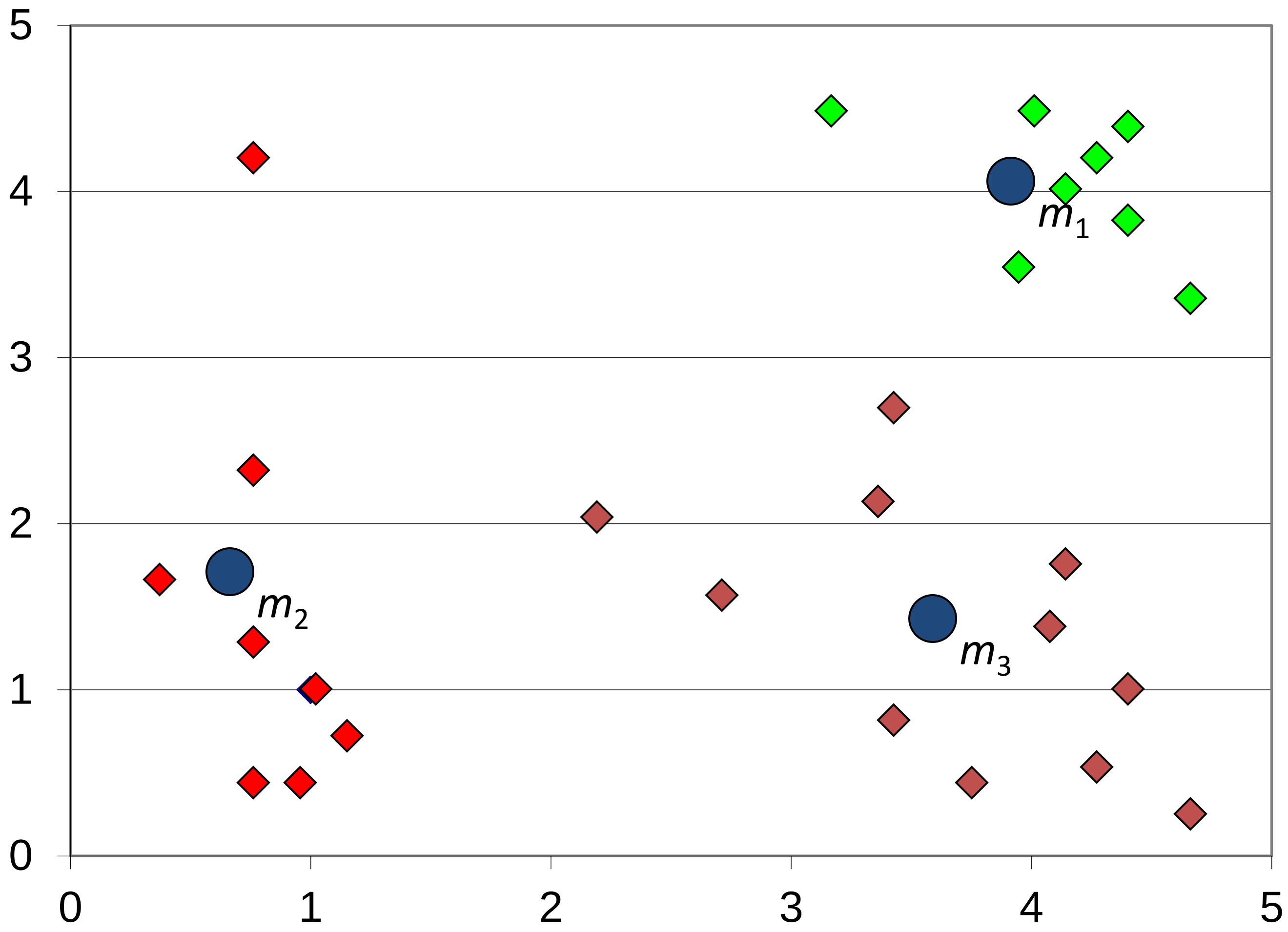
K-means Clustering: Update Cluster Centroid

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



Question

- **Evolution of clusters**
  - Feature values of objects evolve, so the clusters evolve accordingly
  - E.g., my affiliation changed from U Illinois to UB in 2012, so I belong to two different clusters at two different time
  - An interesting data mining question is to find the evolution of clusters
  - Can you discuss possible ways of cluster evolution?