## Data Clustering

1

- Clustering is an unsupervised learning task in ML
  - Provides an intuition about the structure of the data
  - Problem: Given a set of data points and a similarity measure, partition the dataset into *k* disjoint subsets (clusters)
    - Data points in the same cluster are similar to each other

1

## K-means Clustering

2

- K-means Clustering
  - Simple iterative approach
  - The number of clusters, *K*, must be specified
  - Each cluster has a centroid (center point)
  - Each data point is assigned to the cluster with the closest centroid

1: Select $K$ points as the initial centroids.
2: **repeat**
3: Form $K$ clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

2

## K-means Clustering

3

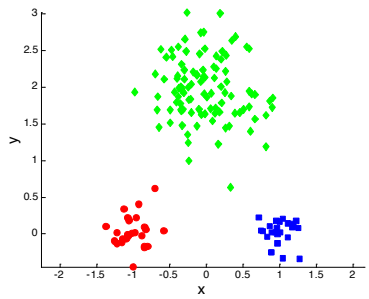
- K-means Clustering
  - Initially, the centroids are chosen randomly
  - Typically, the centroid is the mean of the data points in the cluster and Euclidean distance is used as "closeness measure"

3

## K-means Clustering

4

- K-means Clustering: An Example

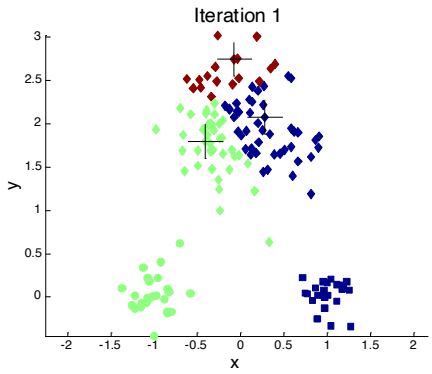



4

## K-means Clustering

5

- K-means Clustering: An Example
  - Initialization: case 1

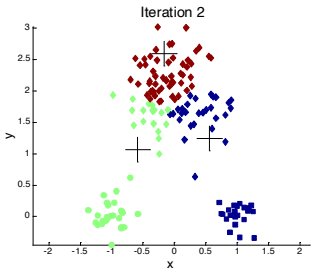


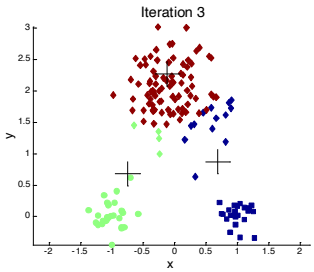

5

## K-means Clustering

6

- K-means Clustering: An Example
  - Iterations



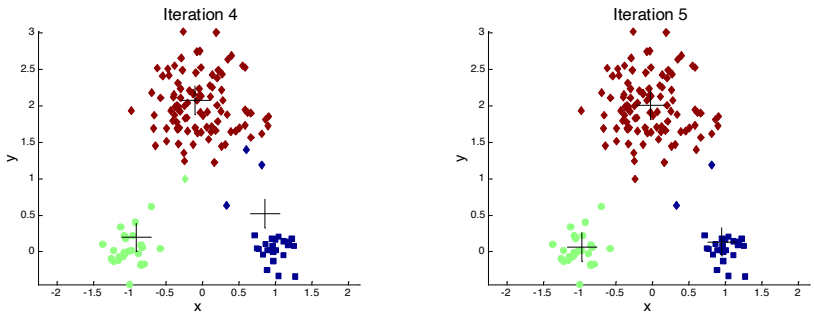




6

## K-means Clustering

7

- K-means Clustering: An Example
  - Iterations

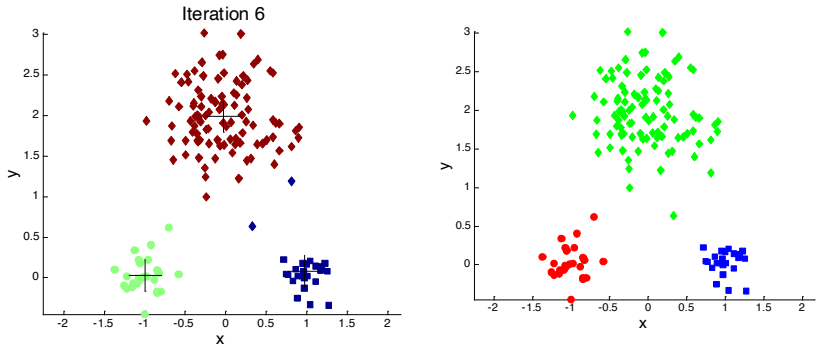



7

## K-means Clustering

8

- K-means Clustering: An Example
  - Convergence

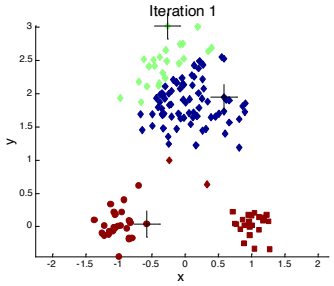



8

## K-means Clustering

9

- K-means Clustering: An Example
  - Initialization: case 2

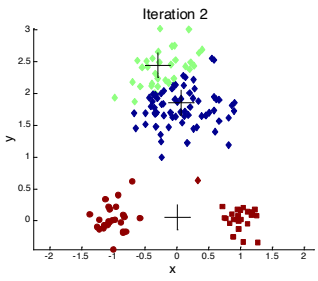


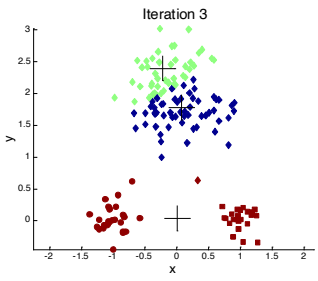

9

## K-means Clustering

10

- K-means Clustering: An Example
  - Iterations



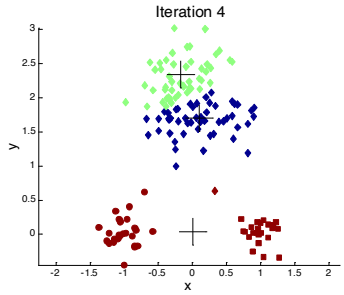



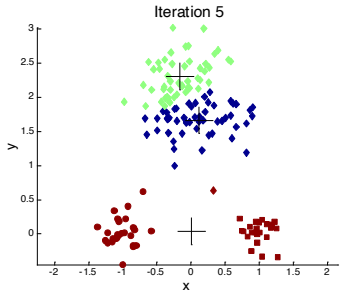

10

## K-means Clustering

11

- K-means Clustering: An Example
  - Iterations



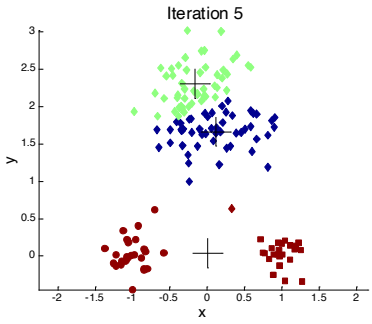



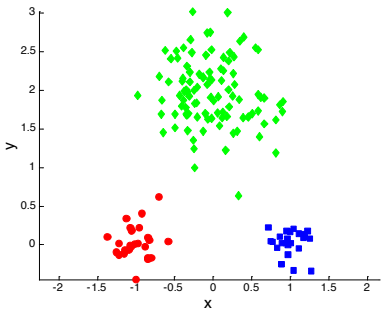

11

## K-means Clustering

12

- K-means Clustering: An Example
  - Convergence



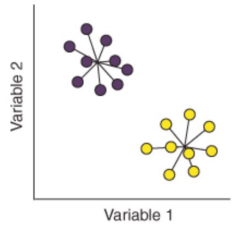




12

## K-means Clustering: Weaknesses and Solutions

13

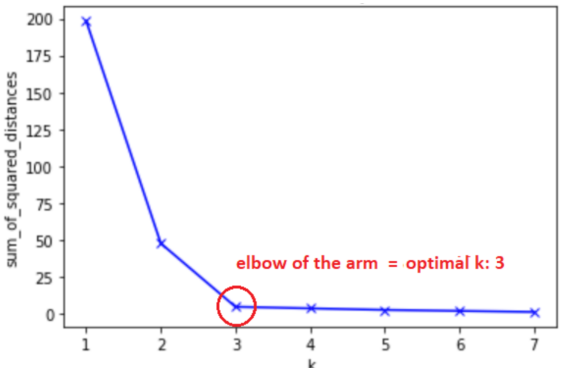
- The number of clusters needs to be known beforehand
  - Elbow method
- Sensitive to initial cluster centers
  - Compute K-means several times with different random initializations (cluster centers) and select the best result correspondina to the one with the lowest within-cluster variation.



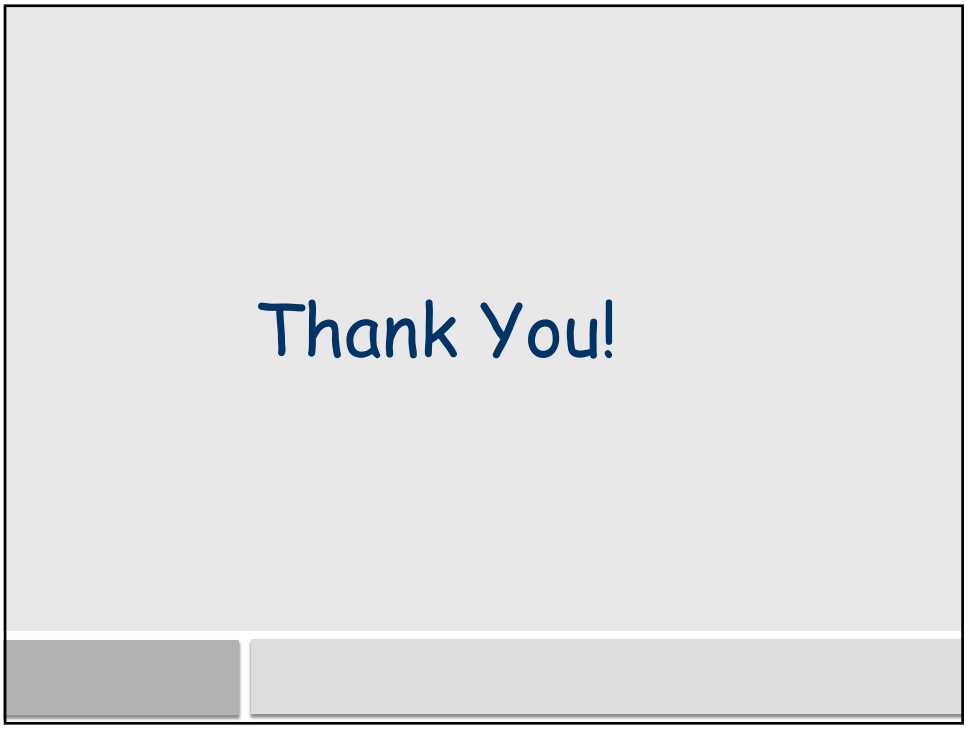

13

## K-means Clustering: Elbow Method

14




14




15