

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

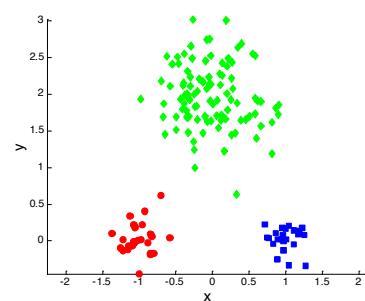
□ K-means Clustering

- 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

□ K-means Clustering: An Example



4

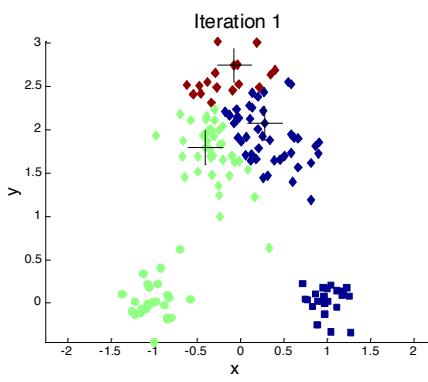
2

K-means Clustering

5

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

Iteration 1



x

y

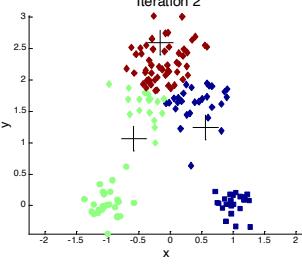
5

K-means Clustering

6

- K-means Clustering: An Example
 - Iterations

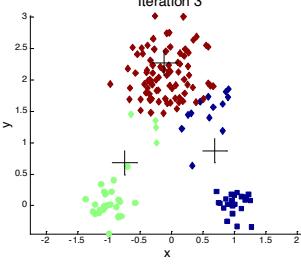
Iteration 2



x

y

Iteration 3



x

y

6

K-means Clustering

7

- K-means Clustering: An Example
 - Iterations

Iteration 4

A scatter plot with x and y axes ranging from -2 to 2. It shows three distinct clusters of points: red, green, and blue. Each cluster has a centroid marked by a cross (+). The red cluster is centered around (-0.2, 1.8), the green cluster around (-1.2, 0.2), and the blue cluster around (0.8, 0.2).

Iteration 5

A scatter plot with x and y axes ranging from -2 to 2. The clusters have shifted slightly compared to Iteration 4. The red cluster is centered around (-0.2, 2.0), the green cluster around (-1.2, 0.5), and the blue cluster around (0.8, 0.5).

7

K-means Clustering

8

- K-means Clustering: An Example
 - Convergence

Iteration 6

A scatter plot with x and y axes ranging from -2 to 2. The clusters have shifted significantly. The red cluster is centered around (-0.2, 2.2), the green cluster around (-1.2, 0.8), and the blue cluster around (0.8, 0.8).

A scatter plot with x and y axes ranging from -2 to 2. The final state of the clustering shows three well-defined, non-overlapping clusters: red, green, and blue. The red cluster is centered around (-0.2, 2.2), the green cluster around (-1.2, 0.8), and the blue cluster around (0.8, 0.8). The centroids are marked by crosses.

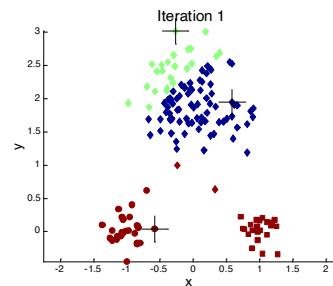
8

K-means Clustering

9

□ K-means Clustering: An Example

□ Initialization: case 2



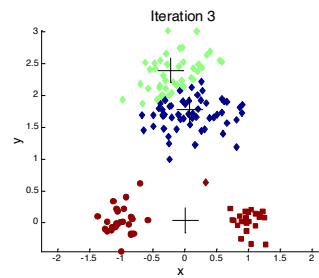
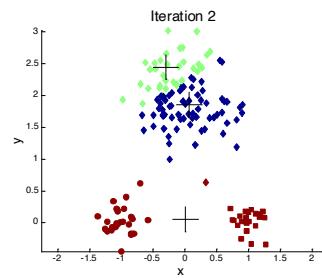
9

K-means Clustering

10

□ K-means Clustering: An Example

□ Iterations



10

K-means Clustering

11

- K-means Clustering: An Example
 - Iterations

Iteration 4

Iteration 5

11

K-means Clustering

12

- K-means Clustering: An Example
 - Convergence

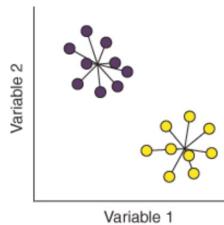
Iteration 5

12

K-means Clustering: Weaknesses and Solutions

13

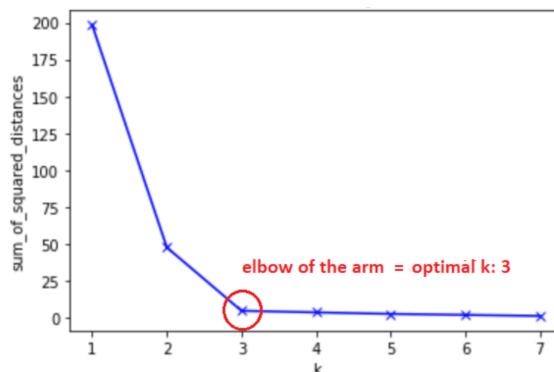
- The number of clusters needs to be known beforehand
 - Elbow method
- Sensitive to initial cluster centers
 - Compute K-means several times **with different random initializations** (cluster centers) and select the best result corresponding to the one with the **lowest within-cluster variation**.



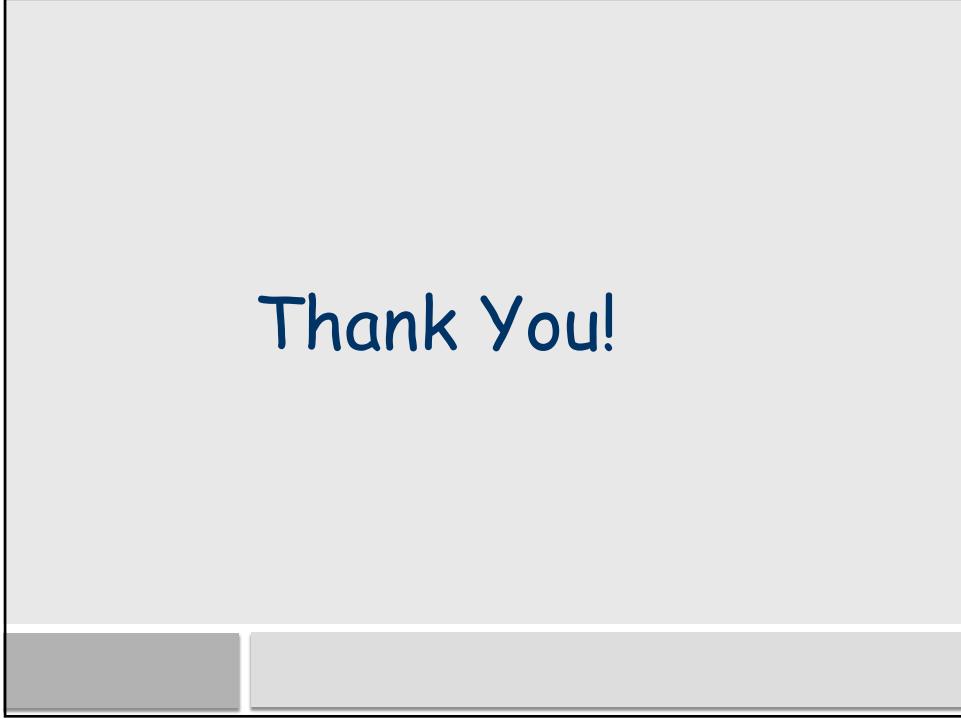
13

K-means Clustering: Elbow Method

14



14



Thank You!