

Clustering Data with the K-means Algorithm

Objectives-

- Introduce the K-means algorithm
- Illustrate K-means with an example
- Identify considerations

Clustering: Organizing data in groups 2

given $\underline{a}_i \in \mathbb{R}^N, i=1, 2, \dots, M$, find centroids $\underline{\mu}_j, j=1, \dots, k$
and clusters $S_j = \{i \mid \underline{a}_i \text{ belongs to cluster } j\}$

Example:



$$\underline{A} = [\underline{a}_1 \ \underline{a}_2 \ \underline{a}_3 \ \underline{a}_4 \ \underline{a}_5] = \begin{bmatrix} 1 & 3 & 4 & 2 & 5 \\ 2 & 4 & 5 & 1 & 3 \end{bmatrix}$$

$$S_1 = \{2, 3, 5\}, \underline{\mu}_1 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}; S_2 = \{1, 4\}, \underline{\mu}_2 = \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}$$

Unsupervised learning: data w/o labels, # clusters unknown

Matrix factorization: $\underline{A} \approx \underline{I} \underline{W}^T, \underline{I} = [\underline{\mu}_1 \ \underline{\mu}_2 \ \dots \ \underline{\mu}_k]$

$$\underline{I} = [\underline{\mu}_1 \ \underline{\mu}_2], \underline{W}^T = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$[\underline{W}^T]_{lm} = \begin{cases} 1, & m \in S_l \\ 0, & m \notin S_l \end{cases}$$

The K-Means Algorithm (K clusters) 3

clusters: $S_j = \{i \mid \underline{a}_i \in \text{cluster } j\}$, $|S_j| = \# \underline{a}_i \text{ in } S_j$

centroids: $\underline{\mu}_j = \frac{1}{|S_j|} \sum_{i \in S_j} \underline{a}_i$ coherence: $c_j = \sum_{i \in S_j} \|\underline{a}_i - \underline{\mu}_j\|_2^2$

Overall coherence $C = \sum_{j=1}^K c_j = \sum_{j=1}^K \sum_{i \in S_j} \|\underline{a}_i - \underline{\mu}_j\|_2^2 = \|\underline{A} - \underline{I} \underline{W}^T\|_F^2$

1) Initialize: choose $\underline{\mu}_j^0$, $j=1, 2, \dots, K$ randomly from \underline{a}_i ; set $l=0$

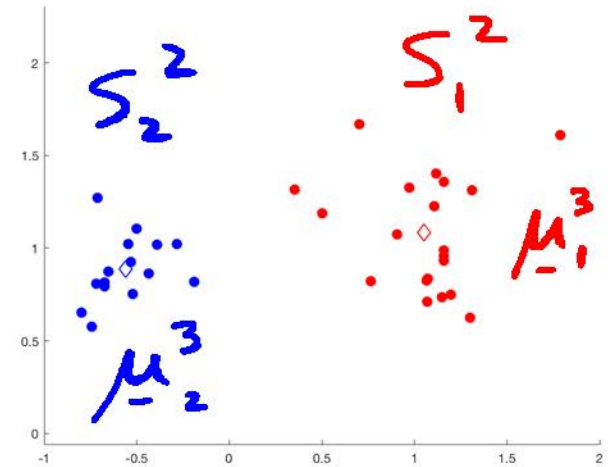
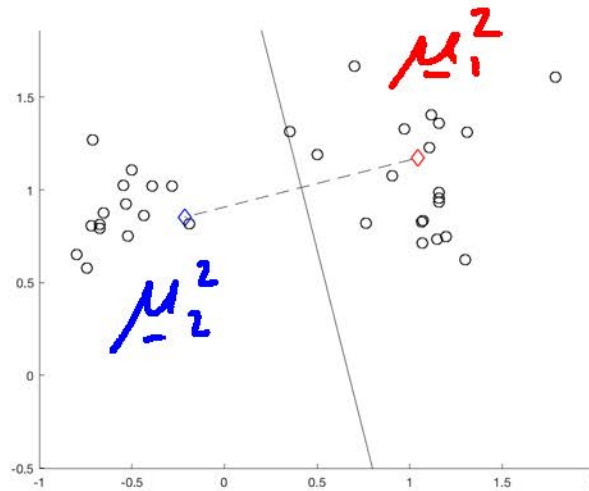
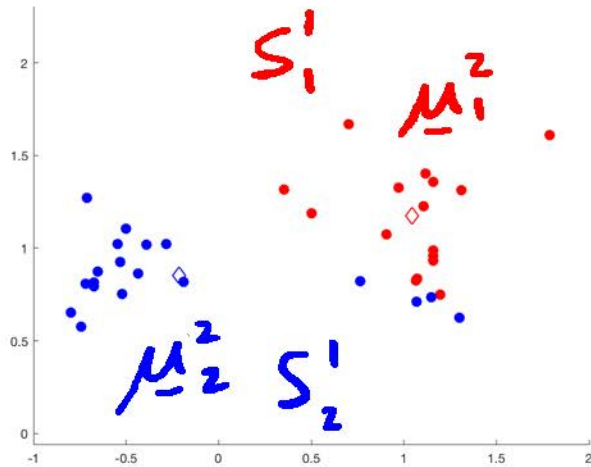
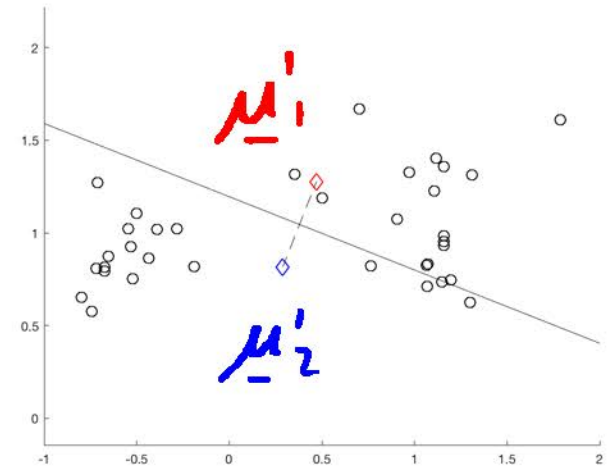
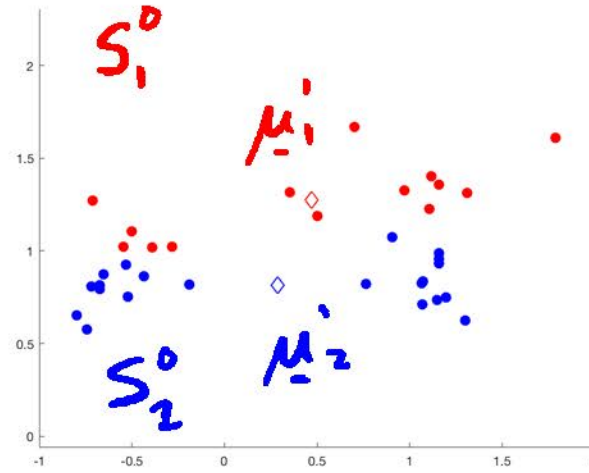
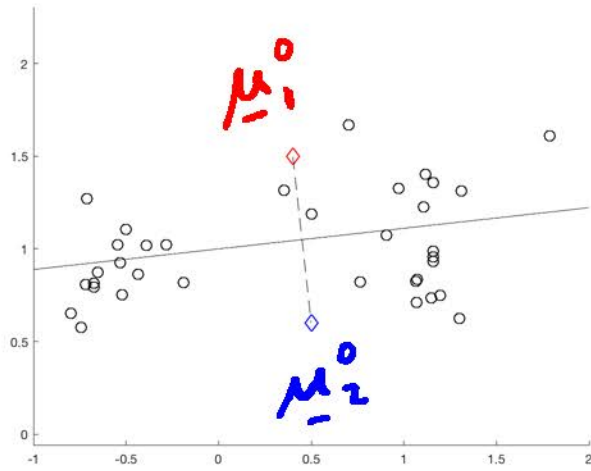
2) Assignment: put $\underline{a}_i \in S_j^l$ if \underline{a}_i is closest to $\underline{\mu}_j^l$ Iteration

3) Update Centroids: $\underline{\mu}_j^{l+1} = \frac{1}{|S_j^l|} \sum_{i \in S_j^l} \underline{a}_i$

4) If converged \rightarrow stop
else $\rightarrow l=l+1$, go to 2)

K-Means Algorithm Example

4



K-Means Algorithm Options

5

- Initialization: many variations
- Termination: change in clusters or overall coherence or fix iterations
- Use different norms to assign clusters

Challenges

- Convergence to local minima
repeat for multiple initializations
- Unknown K
try multiple values

Copyright 2019
Barry Van Veen