Introduction

k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

- **Notations:**
  - Patterns: $x_1, x_2, \ldots, x_n \in \Re^m$
  - Cluster Centers: $c_1, c_2, \ldots, c_k \in \Re^m$
  - Euclidean distance: $||\mathrm{x}_i - \mathrm{c}_j||^2$

MacQueen, J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.

Partition

Based on minimum within cluster measurement.

Let there are A1,A2,A3,…. Ac Clusters/partition

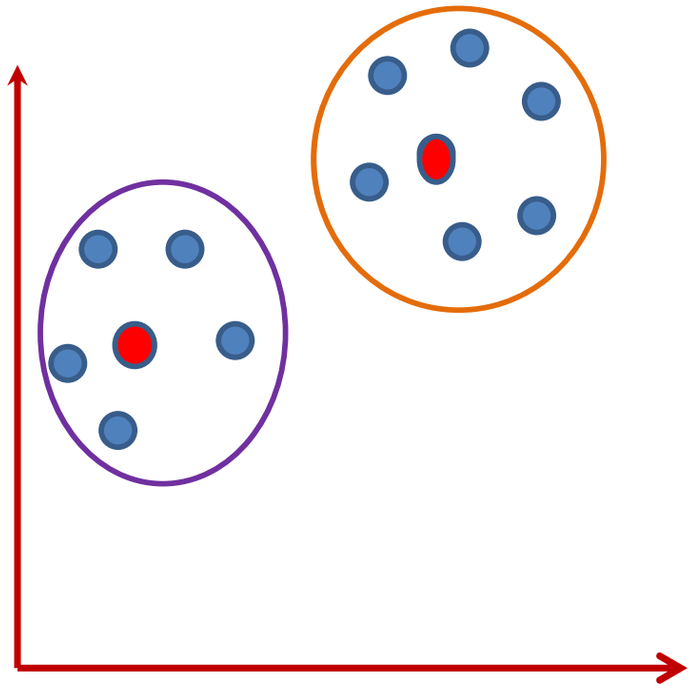
Then they should satisfies these constraints

- $A_i \cap A_j = \emptyset$ $\forall i \neq j$
- $\bigcup_1^c A_i = S$ s=(x1,x2,x3,…xn}
- $A_i \neq \emptyset$

Algorithm

There are 4 steps in the algorithm.

- Randomly select *k* centers from the given data.
- Assign each object to the cluster with the nearest center point. Compute the centers of the clusters of the current partition (i.e., *mean point*, of the cluster).
- Go back to Step 2, stop when no more new assignment.



New Centers will be calculated by

Center1=(x1+x2+x3+x4+x5)/5
Center2=(x6+x7+x8+x9+x10+x11)/6

Algorithm

We have given S={ x1,x2,x3,……., Xn) $\in R^m$

Step1: Choose c points y11,y12,y13,… y1c $\in R^m$

Step2: A_{2i} = { x$\in$S: d($\underline{x}, y_{1i}$) <= d($\underline{x}, y_{1j}$) $\forall$ j$\neq$i }

Step3: y_{2i}= Mean(A_{2i}), i=1,2,3,… c

Step4: if $||y_{1i} - y_{2i}||$<=ε

then

STOP with the output A_{2i}

else

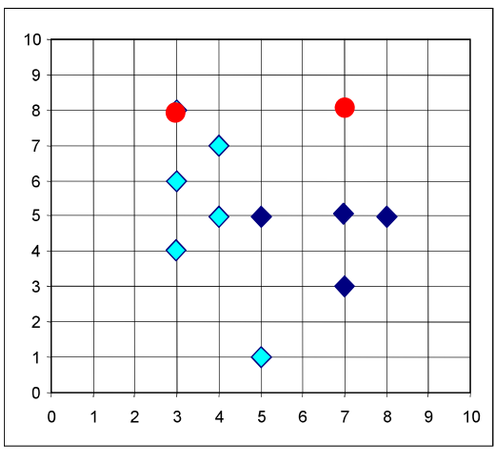
$A_{1i} = A_{2i}$

$A_{2i} = \emptyset$

y_{1i}=y_{2i}

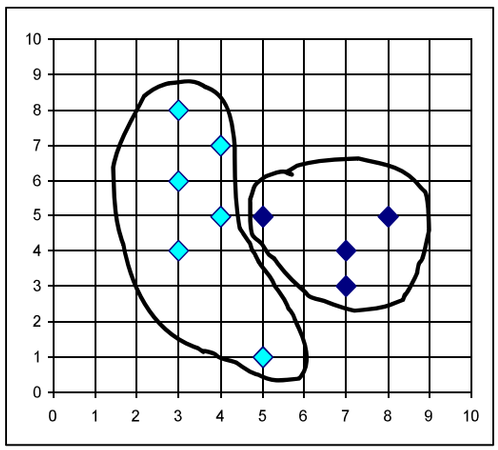
GO TO Step2

Example-1

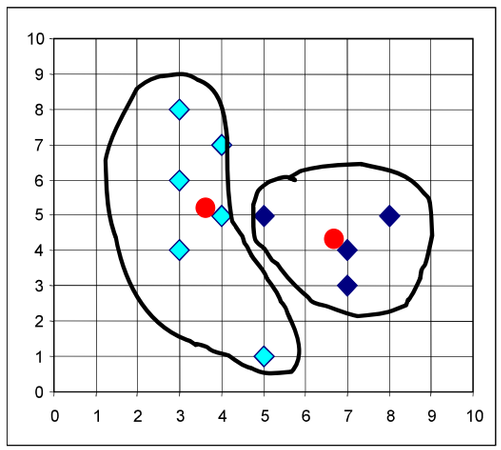


K=2

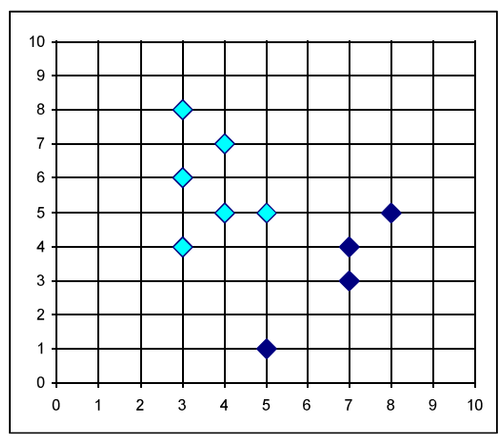
Arbitrarily choose K object as initial cluster center



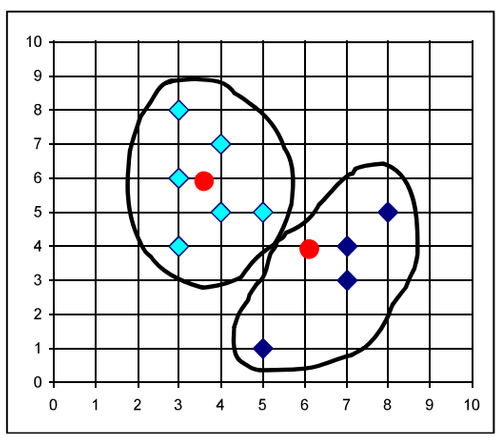
Assign each objects to most similar center



Update the cluster means



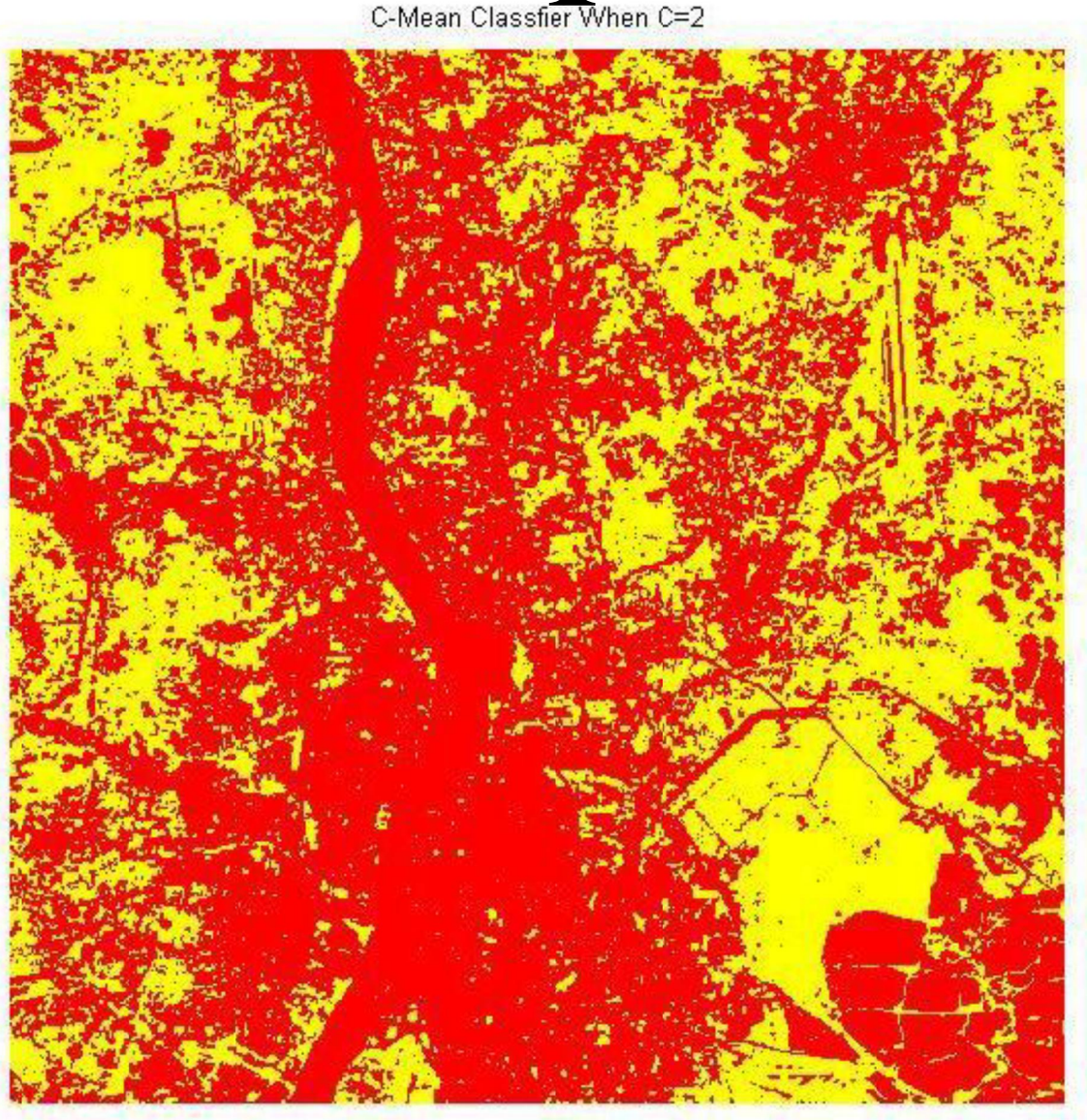
reassign



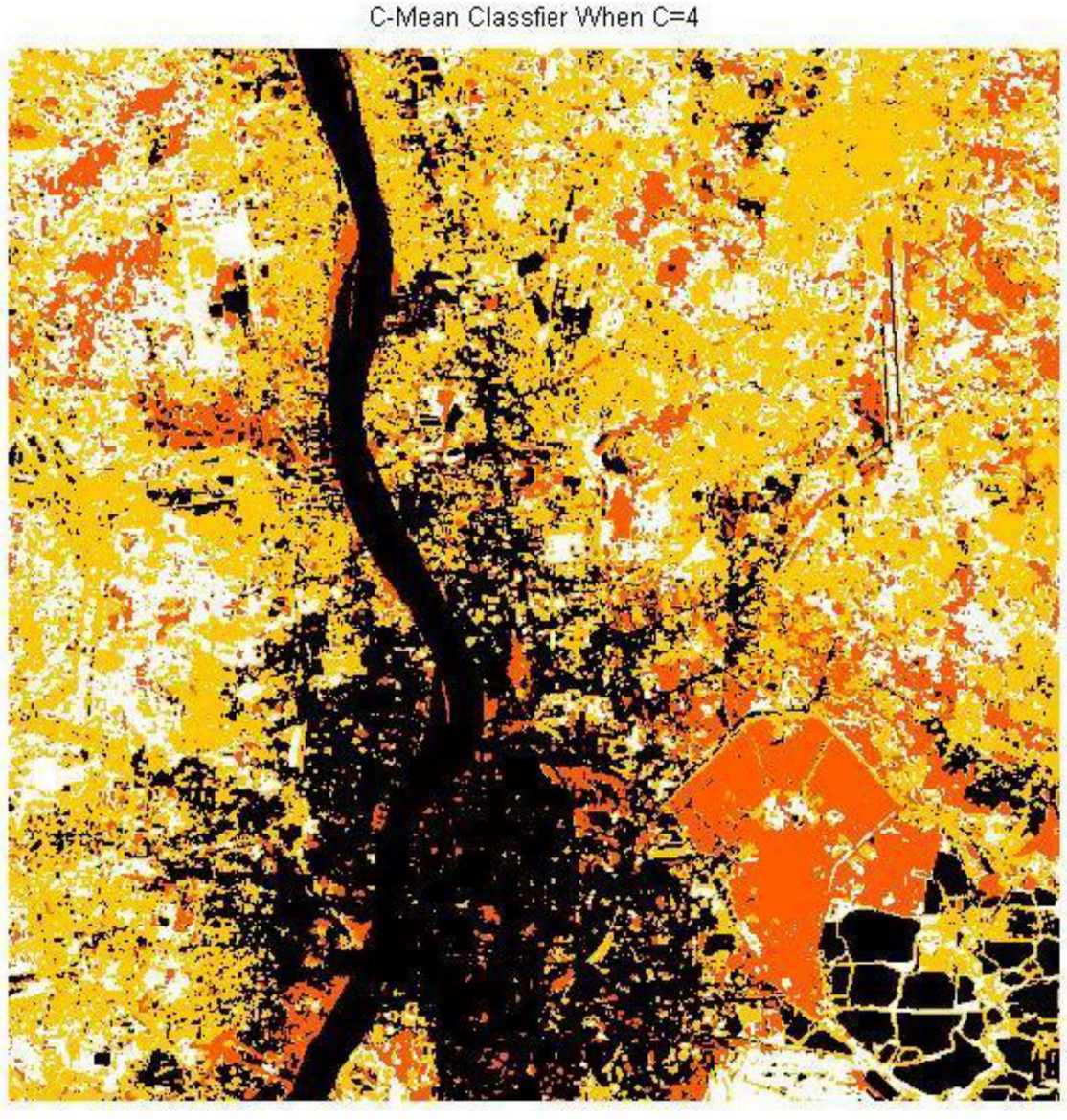
Update the cluster means

reassign

Example-2

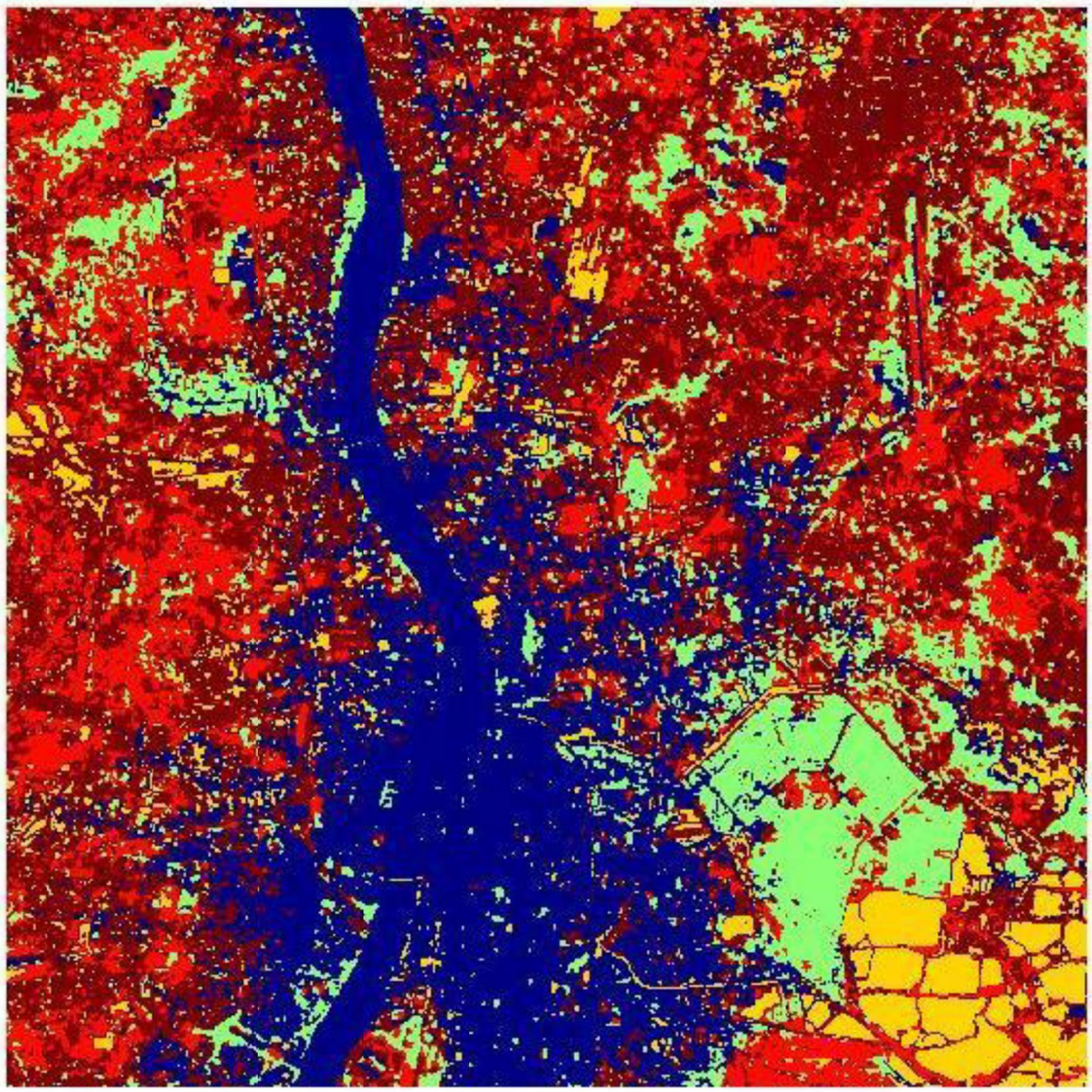


Example-2



Example-2

C-Mean Classfier When C=5



Remarks

- ✓ Algorithm usually converges
- ✓ Two different sets of initial seed point may sometimes give rise to two different final clustering.
- ✓ The algorithm tries to implement the Minimum Within Cluster Distance Criteria.
- ✓ We can start with the initial partition of the dataset (instead of the seed point)

Thanks