CSCI 347, Data Mining

Clustering

K-means Clustering

1. Input k to indicate how many clusters are wanted
2. K points are randomly chosen within the space. These serve as the cluster centers
3. Loop while cluster centers are changing
   a. All instances are assigned to their closest cluster center
   b. Calculate the mean point of each cluster

Classical k-means clustering

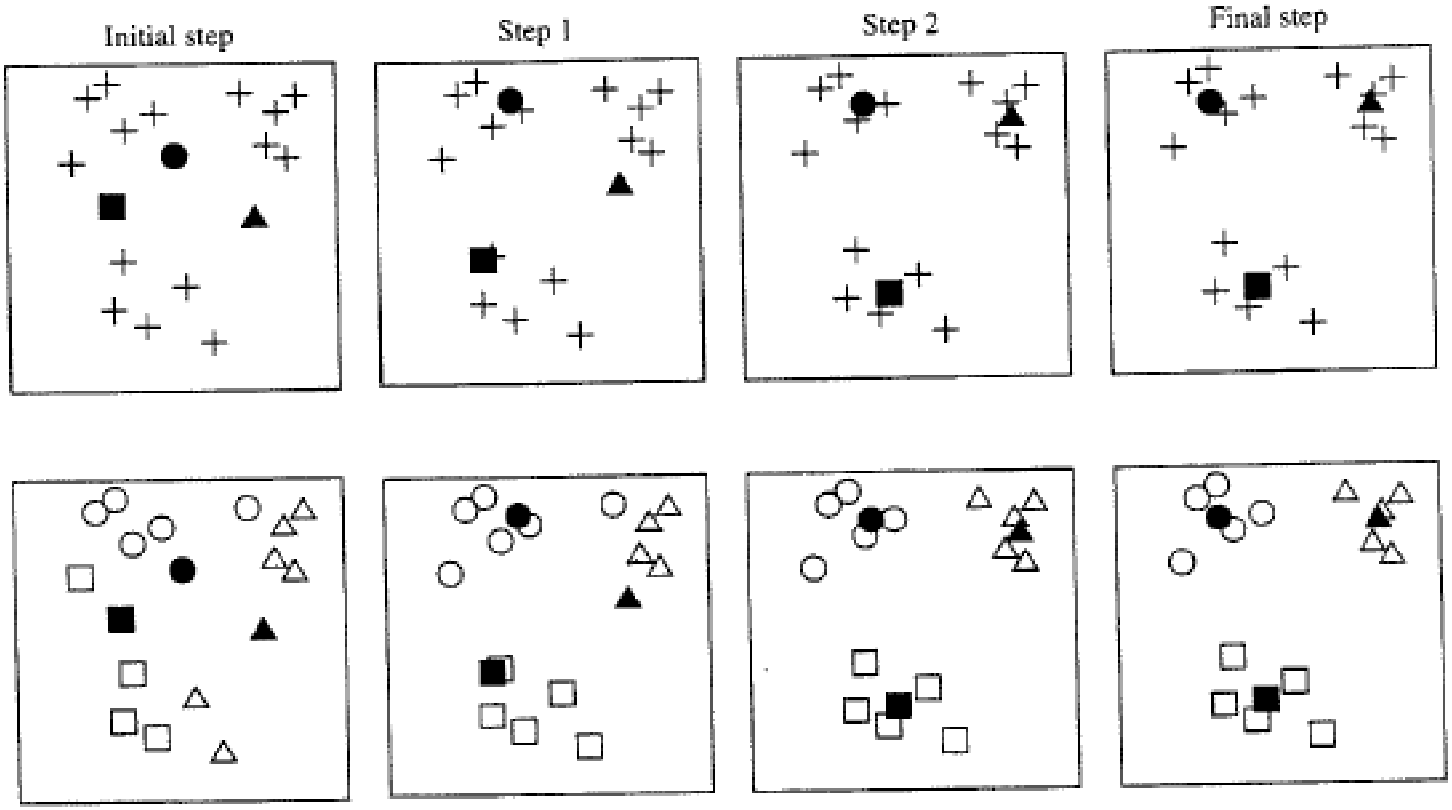
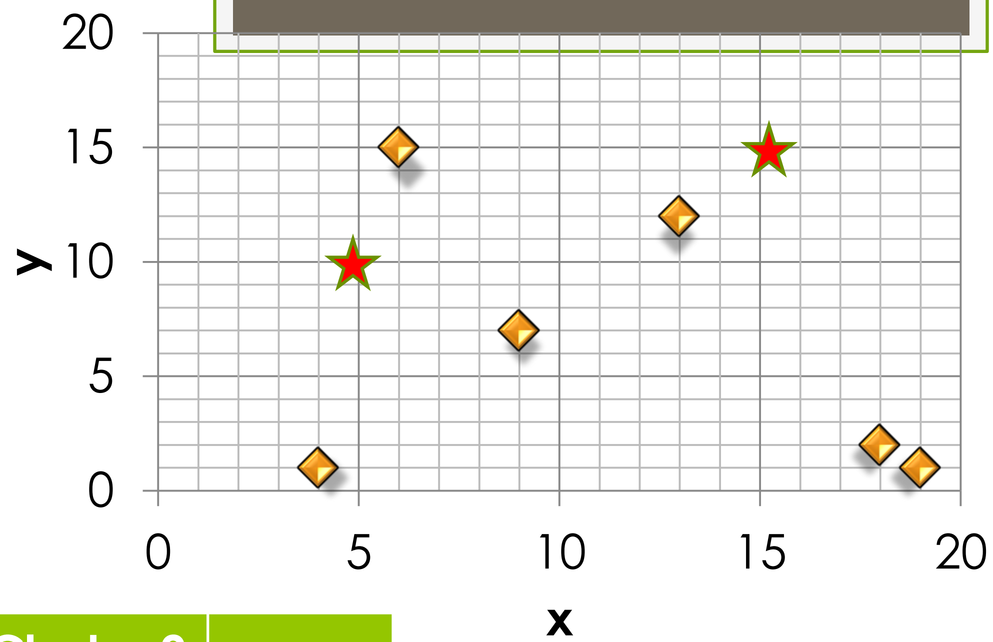


FIGURE 4.17

Iterative distance-based clustering.

EXAMPLE



Data		Cluster 1		Cluster 2	
X	Y	X=5	Y=10	X=15	Y=15
19	1				
13	12				
9	7				
6	15				
18	2				
4	1				

$$\sqrt{(a_1^{(1)}-a_1^{(2)})^2+(a_2^{(1)}-a_2^{(2)})^2+\dots(a_k^{(1)}-a_k^{(2)})^2}$$

EXAMPLE

Data		Cluster 1		Cluster 2	
X	Y	X=5	Y=10	X=15	Y=15
19	1	16.64		14.56	
13	12	8.25		3.61	
9	7	5.00		10.00	
6	15	5.10		9.00	
18	2	15.26		13.34	
4	1	9.06		17.80	

$$d(1) = \sqrt{(19-5)^2 + (1-10)^2} = 16.64$$

$$d(1) = \sqrt{(19-15)^2 + (1-15)^2} = 14.56$$

EXAMPLE

Data		Cluster 1		Cluster 2	
X	Y	X=5	Y=10	X=15	Y=15
19	1	16.64		14.56	
13	12	8.25		3.61	
9	7	5.00		10.00	
6	15	5.10		9.00	
18	2	15.26		13.34	
4	1	9.06		17.80	

Now we assign each instance to the cluster which it's closest to (highlighted
In the table.)

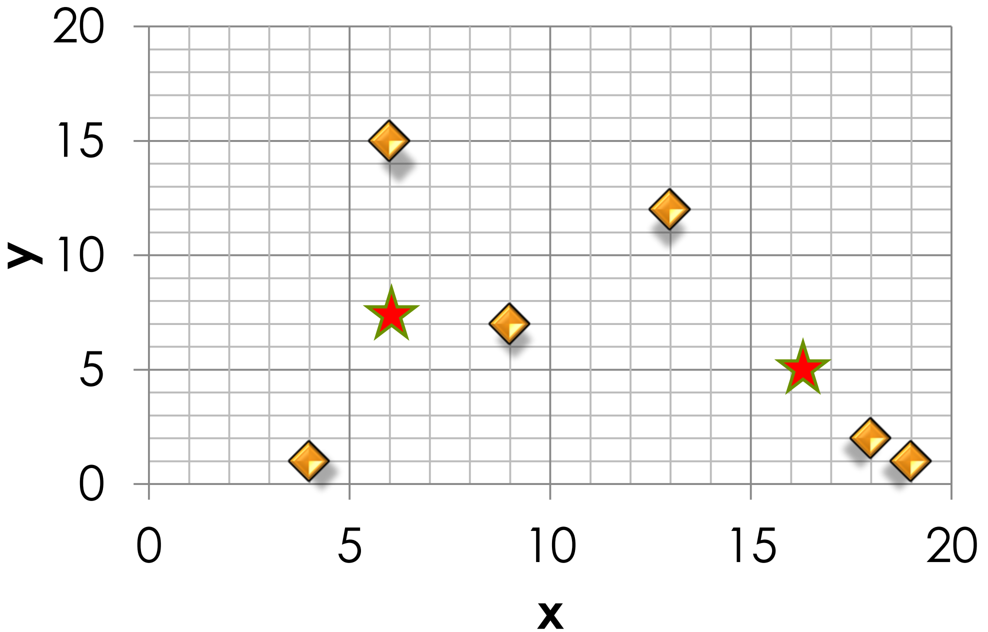
EXAMPLE

Data		Cluster 1		Cluster 2	
X	Y	X=5	Y=10	X=15	Y=15
19	1	16.64		14.56	
13	12	8.25		3.61	
9	7	5.00		10.00	
6	15	5.10		9.00	
18	2	15.26		13.34	
4	1	9.06		17.80	

Then we adjust the cluster centers to be the average of all of the instances
assigned to them. (This is called the centroid.)

Cluster Center 1, X = (9+6+4)/3 = 6.33
Y = (7+15+1)/3 = 7.67
Cluster Center 2, X = (19+13+18)/3 = 16.67
Y = (1+12+2)/3 = 5

EXAMPLE



We place the new cluster centers and do the entire process again. We repeat
this until no changes happen on an iteration.

Classical k-means clustering

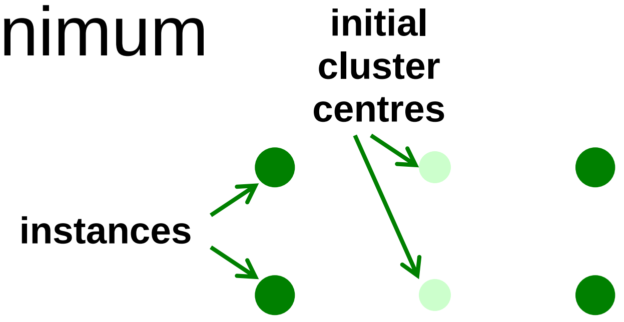
Algorithm minimizes squared distance to cluster centers

Result can vary significantly

- based on initial choice of seeds

Can get trapped in local minimum

- Example:



To increase chance of finding global optimum: restart with different random seeds

Can be applied recursively with $k = 2$