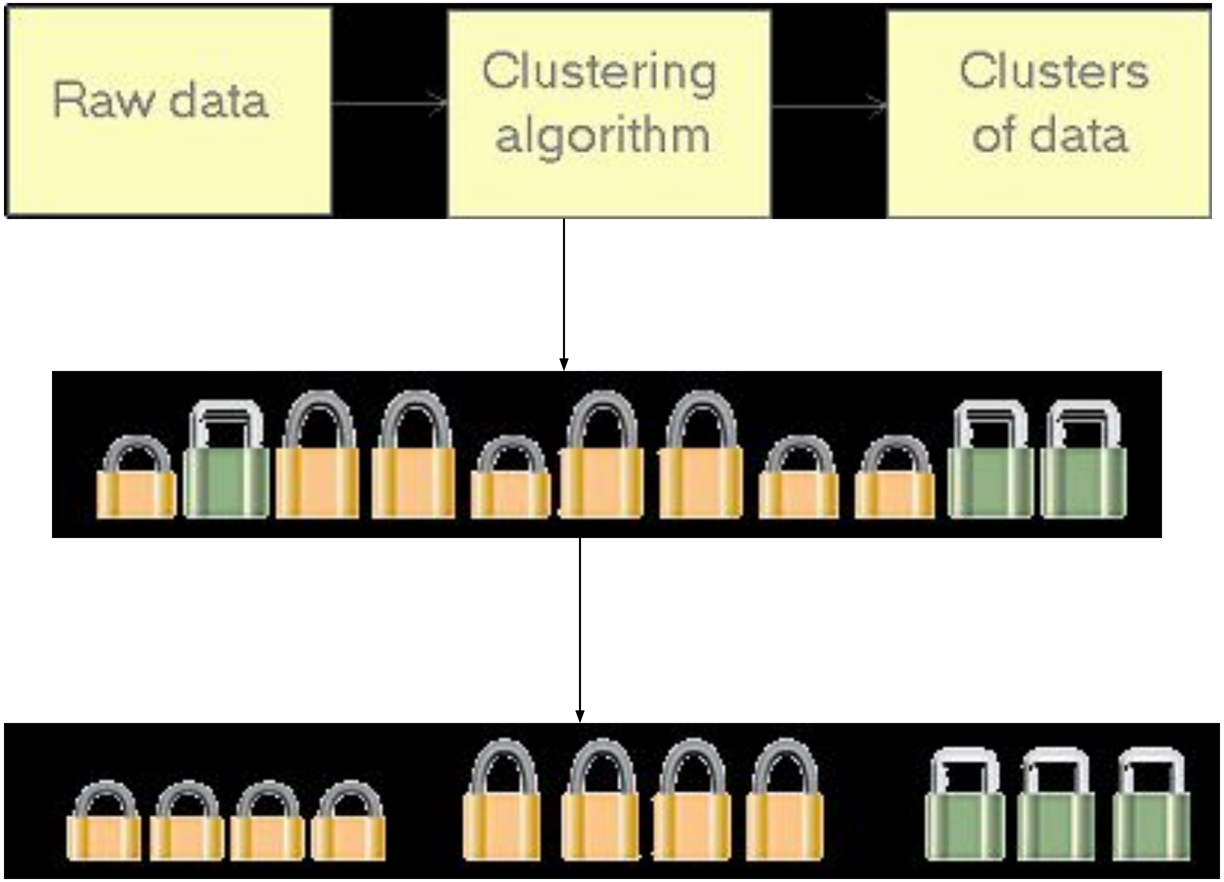
OUTLINE

1. Definition of Clustering
2. Applications of Clustering
3. Distance Measure
4. k-means CLUSTERING
5. KNN Clustering
6. Maximin distance Clustering
7. Conclusion

DEFINITION OF CLUSTERING

- Clustering is "the process of organizing objects into groups whose members are similar in some way".

DEFINITION OF CLUSTERING



Raw data

Clustering algorithm

Clusters of data

EXAMPLES OF CLUSTERING APPLICATIONS

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.
- Land use: Identification of areas of similar land use in an earth observation database.
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost.
- City-planning: Identifying groups of houses according to their house type, value, and geographical location.
- and many others,…

DISTANCE MEASURE

k-nerst

- Euclidean distance

$$d(g_1, g_2) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Manhattan distance

k-mean

$$d(g_1, g_2) = \sum_{i=1}^{n} \left|(x_i - y_i)\right|$$

- Minkowski distance

$$d(g_1, g_2) = \sqrt[m]{\sum_{i=1}^{n} (x_i - y_i)^m}$$

k-means CLUSTERING

- Input: n objects (or points) and a number k
- Algorithm

1. Randomly place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the stopping criteria is met

k-means CLUSTERING (Cont..)

- Problem: Cluster the following eight points (with (x, y) representing locations) into three clusters A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9). Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

k-means CLUSTERING (Cont..)

Solution:

Iteration 1

		(2, 10)	(5, 8)	(1, 2)	
	Point	Dist Mean 1	Dist Mean 2	Dist Mean 3	Cluster
A1	(2, 10)				
A2	(2, 5)				
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

k-means CLUSTERING (Cont..)

- point mean1
- $x1, y1$ $x2, y2$
- (2, 10) (2, 10)
- $\rho(a, b) = |x2 - x1| + |y2 - y1|$
- $\rho(point, mean1) = |x2 - x1| + |y2 - y1|$
- $= |2 - 2| + |10 - 10|$
- $= 0 + 0$
- $= 0$

k-means CLUSTERING (Cont..)

- point mean2
- $x1, y1$ $x2, y2$
- (2, 10) (5, 8)
-
- $\rho(a, b) = |x2 - x1| + |y2 - y1|$
-
- $\rho(point, mean2) = |x2 - x1| + |y2 - y1|$
- $= |5 - 2| + |8 - 10|$
- $= 3 + 2$
- $= 5$

k-means CLUSTERING (Cont..)

- point mean3
- $x1, y1$ $x2, y2$
- (2, 10) (1, 2)
- $\rho(a, b) = |x2 - x1| + |y2 - y1|$
- $\rho(point, mean2) = |x2 - x1| + |y2 - y1|$
- $= |1 - 2| + |2 - 10|$
- $= 1 + 8$
- $= 9$

k-means CLUSTERING (Cont..)

So, we fill in these values in the table:

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	**(2, 10)**	0	5	9	1
A2	(2, 5)				
A3	(8, 4)				
A4	**(5, 8)**				
A5	(7, 5)				
A6	(6, 4)				
A7	**(1, 2)**				
A8	(4, 9)				

k-means CLUSTERING (Cont..)

Iteration 1

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	**(2, 10)**	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)	12	7	9	2
A4	**(5, 8)**	5	0	10	2
A5	(7, 5)	10	5	9	2
A6	(6, 4)	10	5	7	2
A7	**(1, 2)**	9	10	0	3
A8	(4, 9)	3	2	10	2

k-means CLUSTERING (Cont..)

Cluster 1	Cluster 2	Cluster 3
(2, 10)	(8, 4)	(2, 5)
	(5, 8)	(1, 2)
	(7, 5)	
	(6, 4)	
	(4, 9)	

k-means CLUSTERING (Cont..)

- Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

k-means CLUSTERING (Cont..)

- For Cluster 1, we only have one point A1(2, 10), which was the old mean, so the cluster center remains the same

- For Cluster 2, we have ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)

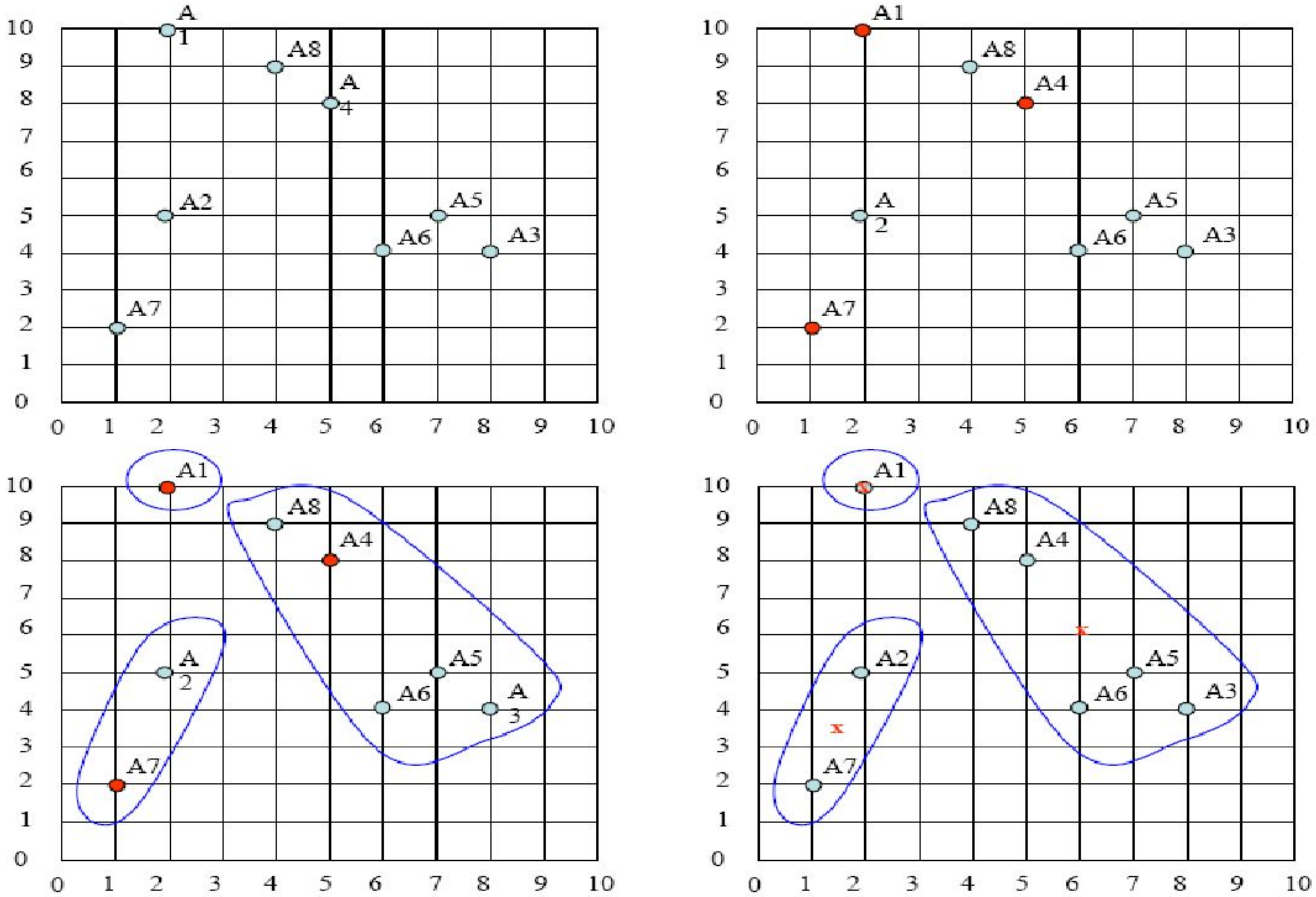
- For Cluster 3, we have ((2+1)/2, (5+2)/2) = (1.5, 3.5)

k-means CLUSTERING (Cont..)

new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

b) centers of the new clusters:
C1= (2, 10), C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

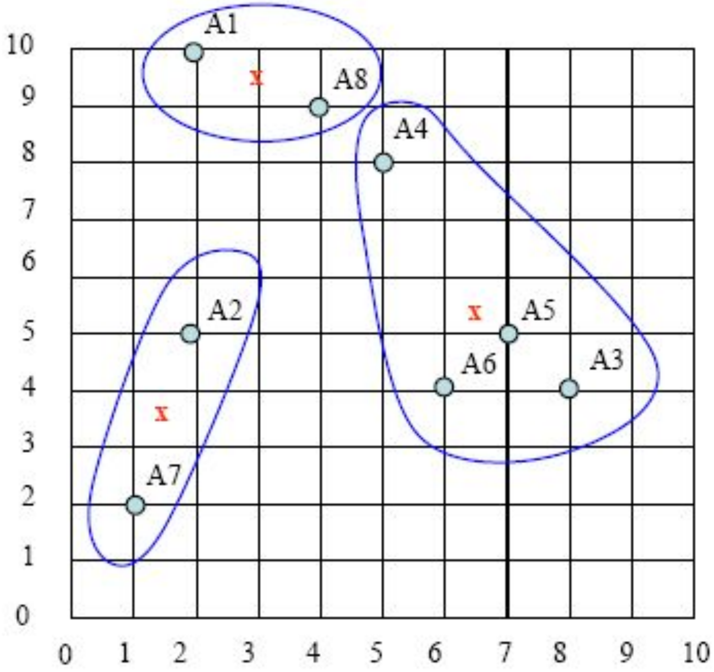
c)

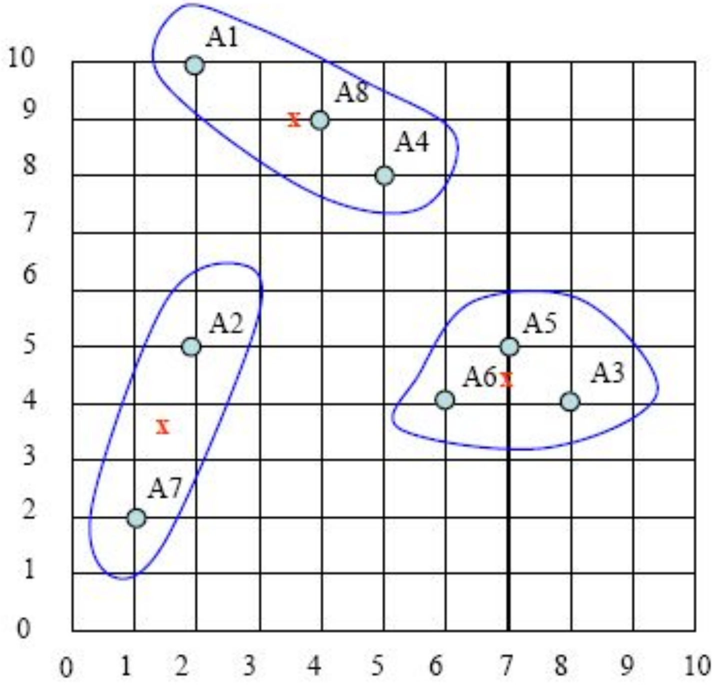


k-means CLUSTERING (Cont..)

d)

We would need two more epochs. After the 2nd epoch the results would be:
1: {A1, A8}, 2: {A3, A4, A5, A6}, 3: {A2, A7}
with centers C1=(3, 9.5), C2=(6.5, 5.25) and C3=(1.5, 3.5).
After the 3rd epoch, the results would be:
1: {A1, A4, A8}, 2: {A3, A5, A6}, 3: {A2, A7}
with centers C1=(3.66, 9), C2=(7, 4.33) and C3=(1.5, 3.5).





k-means CLUSTERING (Cont..)

- Stopping criteria:
  - No change in the members of all clusters
  - when the squared error is less than some small threshold value α
  - Centroids of newly formed clusters do not change.
  - Points remain in the same cluster.
  - the Maximum number of iterations is reached.

k-means CLUSTERING (Cont..)

- Pros:
  - Low complexity ,O(nkt), where t = #iterations
    - Relatively simple to implement.
    - Scales to large data sets.
    - Guarantees convergence.
    - Can warm-start the positions of centroids.
    - Easily adapts to new examples.
    - Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

k-means CLUSTERING (Cont..)

- Cons:
- Choosing k manually.

Use the “Loss vs. Clusters” plot to find the optimal (k),

- Being dependent on initial values.

For a low k, you can mitigate this dependence by running k-means several times with different initial values and picking the best result. As k increases, you need advanced versions of k-means to pick better values of the initial centroids (called k-means seeding)..

- Clustering data of varying sizes and density.

k-means has trouble clustering data where clusters are of varying sizes and density. To cluster such data, you need to generalize k-means.

k-means CLUSTERING (Cont..)

- Cons:
- Clustering outliers.

Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored. Consider removing or clipping outliers before clustering.

- Scaling with number of dimensions.

As the number of dimensions increases, a distance-based similarity measure converges to a constant value between any given examples. Reduce dimensionality either by using PCA on the feature data,

K-Nearest Neighbor

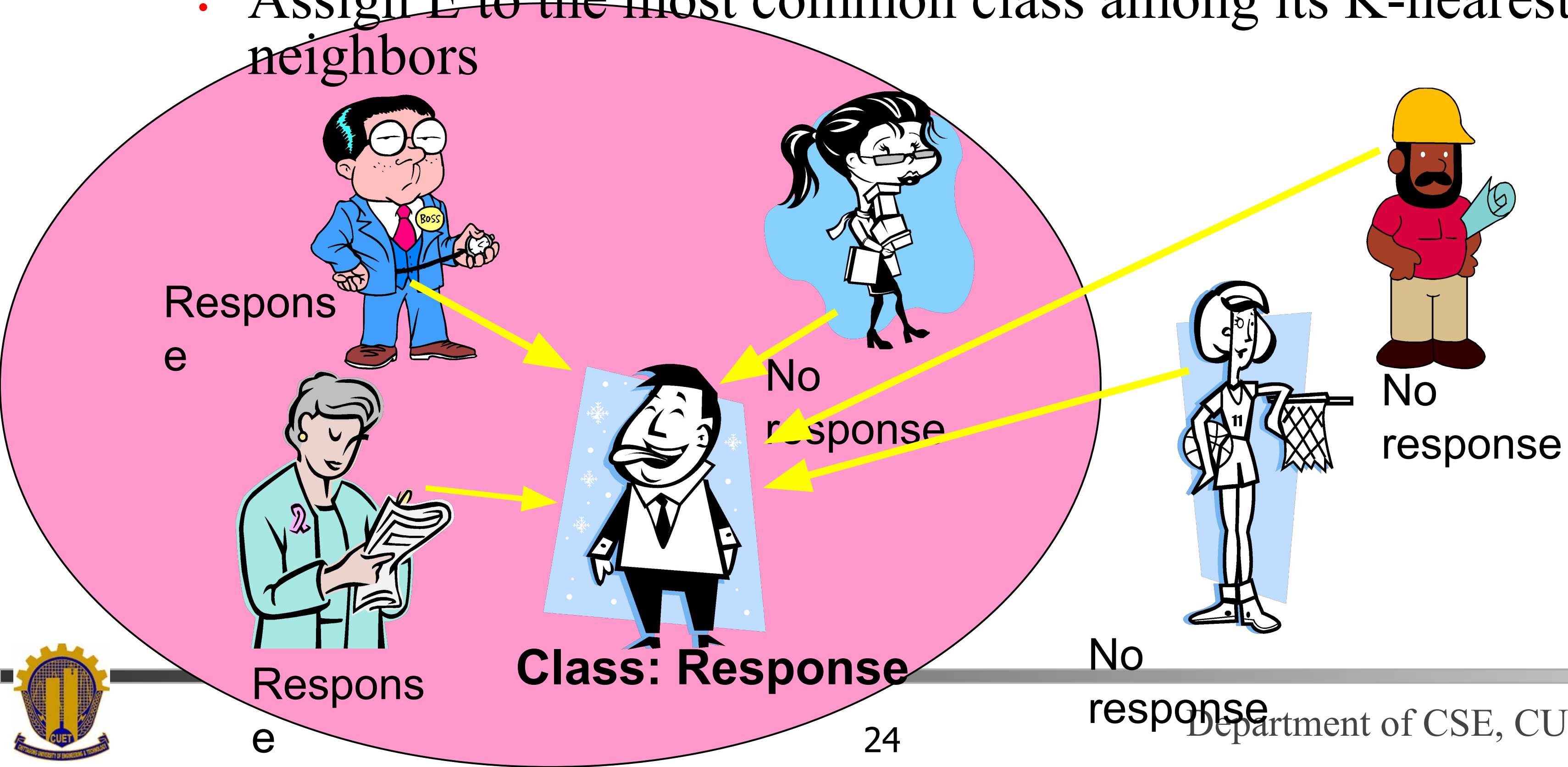
Learning by analogy:

Tell me who your friends are and I'll tell you who you are

A new example is assigned to the most common class among the (K) examples that are most similar to it.

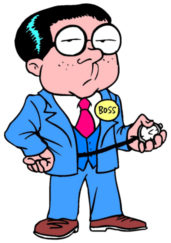
K-Nearest Neighbor Algorithm

- To determine the class of a new example E:
  - Calculate the distance between E and all examples in the training set
  - Select K-nearest examples to E in the training set
  - Assign E to the most common class among its K-nearest neighbors



Distance Between Neighbors

- Each example is represented with a set of numerical attributes



John:
Age=35
Income=95K
No. of credit cards=3

Rachel:
Age=41
Income=215K
No. of credit cards=2

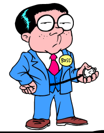
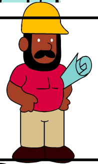
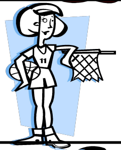
- "Closeness" is defined in terms of the *Euclidean* distance between two examples.
  - The Euclidean distance between X=(x_1, x_2, x_3,…x_n) and Y =(y_1,y_2, y_3,…y_n) is defined as:

$$D(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

  - Distance (John, Rachel)=sqrt [$(35\text{-}41)^2$+$(95K\text{-}215K)^2$ +$(3\text{-}2)^2$]

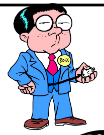
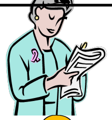
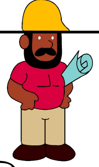
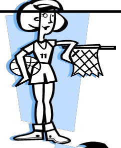
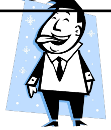
K-Nearest Neighbor

Example : 3-Nearest Neighbors

Customer	Age	Income	No. credit cards	Response
John 	35	35K	3	No
Rachel	22	50K	2	Yes
Hannah	63	200K	1	No
Tom 	59	170K	1	No
Nellie 	25	40K	4	Yes
David	37	50K	2	?

K-Nearest Neighbor

Example

Customer	Age	Income (K)	No. cards	Response	Distance from David
John 	35	35	3	No	sqrt $[(35-37)^2+(35-50)^2+(3-2)^2]=$**15.16**
Rachel	22	50	2	Yes	sqrt $[(22-37)^2+(50-50)^2+(2-2)^2]=$**15**
Hannah 	63	200	1	No	sqrt $[(63-37)^2+(200-50)^2+(1-2)^2]=$**152.23**
Tom 	59	170	1	No	sqrt $[(59-37)^2+(170-50)^2+(1-2)^2]=$**122**
Nellie 	25	40	4	Yes	sqrt $[(25-37)^2+(40-50)^2+(4-2)^2]=$**15.74**
David 	37	50	2	**Yes**	

K-Nearest Neighbor Algorithm (Cont..)

- Use the Nearest Neighbor clustering algorithm and Euclidean distance to cluster the examples from the previous exercise: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).
- Suppose that the threshold t is 4.

K-Nearest Neighbor Algorithm (Cont..)

Solution:

A1 is placed in a cluster by itself, so we have K1={A1}.

We then look at A2 if it should be added to K1 or be placed in a new cluster.

d(A1,A2)= 25= 5 > t ⇒ K2={A2}

A3: we compare the distances from A3 to A1 and A2.

A3 is closer to A2 and d(A3,A2)= 36> t ⇒ K3={A3}

A4: We compare the distances from A4 to A1, A2 and A3.

A1 is the closest object and d(A4,A1)= 13 < t ⇒ K1={A1, A4}

K-Nearest Neighbor Algorithm (Cont..)

- A5: We compare the distances from A5 to A1, A2, A3 and A4.
- A3 is the closest object and d(A5,A3)= 2 < t ⇒K3={A3, A5}
-
- A6: We compare the distances from A6 to A1, A2, A3, A4 and A5.
- A3 is the closest object and d(A6,A3)= 2 < t ⇒K3={A3, A5, A6}
-
- A7: We compare the distances from A7 to A1, A2, A3, A4, A5, and A6.
- A2 is the closest object and d(A7,A2)= 10< t 4 K2={A2, A7)

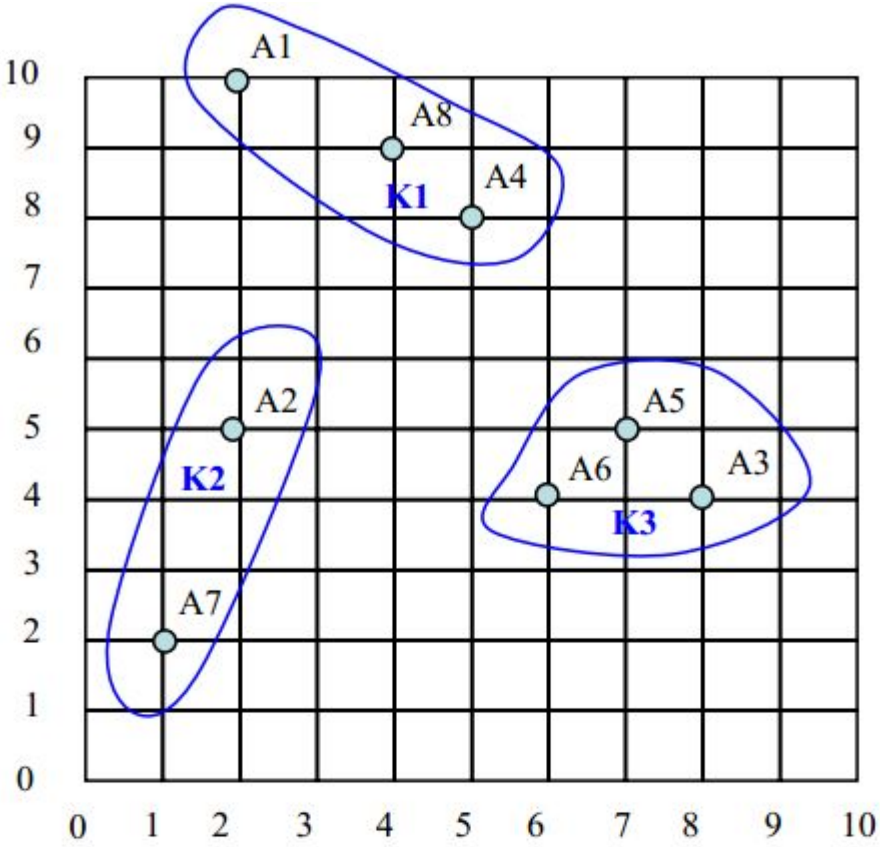
K1 = {A1} K2 ⇒ A2

K-Nearest Neighbor Algorithm (Cont..)

- A8: We compare the distances from A8 to A1, A2, A3, A4, A5, A6 and A7.
- A4 is the closest object and d(A8,A4)= 2 < t ⇒ K1={A1, A4, A8)

K-Nearest Neighbor Algorithm (Cont..)

Thus: K1={A1, A4, A8), K2={A2, A7), K3={A3, A5, A6)



K-Nearest Neighbor Algorithm (Cont..)

Strengths:

- Simple to implement and use
- Comprehensible – easy to explain prediction
- Robust to noisy data by averaging k-nearest neighbors.

Weaknesses:

- Need a lot of space to store all examples.
- Takes more time to classify a new example than with a model (need to calculate and compare distance from new example to all other examples).

Maximin distance algorithm

- 1) begin by identifying cluster regions that are farthest apart
- 2) define an initial threshold distance based on the separation of these cluster centers, and
- 3) continue selecting cluster centers and readjusting T until all possible cluster centers are
- established.
- Note : with the maximin algorithm, **NO** information is required from the user.

Maximin distance algorithm (Cont..)

1. Select a pixel, **x**, from the image at random.
2. Let **x** be the first cluster center, **x** ==> $\mathbf{z}_1$
3. Sort through the remaining pixels to find the pixel, **x**, which is farthest from $\mathbf{z}_1$.
4. Let the most distant pixel be the second cluster center, **x** ==> $\mathbf{z}_2$.
5. Find the distance, $T = |z_2 - z_1|$, between the two cluster centers.

   T will be an initial scaling distance used to determine the existence of the next cluster center.
6. Compute $D_{min}(\mathbf{x}_j) = \min[D_i(\mathbf{x}_j)]$ for i=1,2, for all remaining pixels in the image, i.e., find the distance to the closest cluster center for every pixel in the image.
7. Find $D_{max}(\mathbf{x}_m) = \max[D_{min}(\mathbf{x}_j)]$ for all j.

   Sort through all the distances determined in step 6 and select the maximum distance (select the maximum of the minimum distances). This procedure will find the pixel that is farthest (in measurement space) from either of the two cluster centers.

Maximin distance algorithm (Cont..)

8. If D_{max} > T/2, then let $\mathbf{x}_m$ ==> z_{n+1}, otherwise,

   if D_{max} < T/2, then terminate the procedure.

   In words, if the maximum distance is greater than half the distance between the two closest cluster centers, then $\mathbf{x}_m$ becomes a new cluster center, otherwise terminate the procedure.

9. Increment the number of cluster centers by 1: $N_c = N_c + 1$.

10. Reset the scaling distance: $T = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|z_i - z_j\right| \Big/ \sum_{k=1}^{n-1}\frac{k(k+1)}{2}$

11. Return to step 6.

Maximin distance algorithm (Cont..)

- Use the Maximin distance algorithm cluster the examples from the previous exercise: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

Maximin distance algorithm (Cont..)

- The initial cluster center is A1**(2, 10)**
- Distance from A1

		(2, 10)
	Point	**Dist**
A1	**(2, 10)**	0
A2	(2, 5)	5
A3	(8, 4)	12 **[MAX]**
A4	**(5, 8)**	5
A5	(7, 5)	10
A6	(6, 4)	10
A7	**(1, 2)**	9
A8	(4, 9)	3

Maximin distance algorithm (Cont..)

- Cluster 2 center is A3 AND T=12

		(2, 10)	(8, 4)	
	Point	**Dist 1**	**Dist 2**	**MIN Dist**
A1	**(2, 10)**	0	12	0
A2	(2, 5)	5	7	5
A3	**(8, 4)**	12	0	0
A4	(5, 8)	5	7	5
A5	(7, 5)	10	2	2
A6	(6, 4)	10	2	2
A7	**(1, 2)**	9	9	9 **[MAX]**
A8	(4, 9)	3	9	3

Maximin distance algorithm (Cont..)

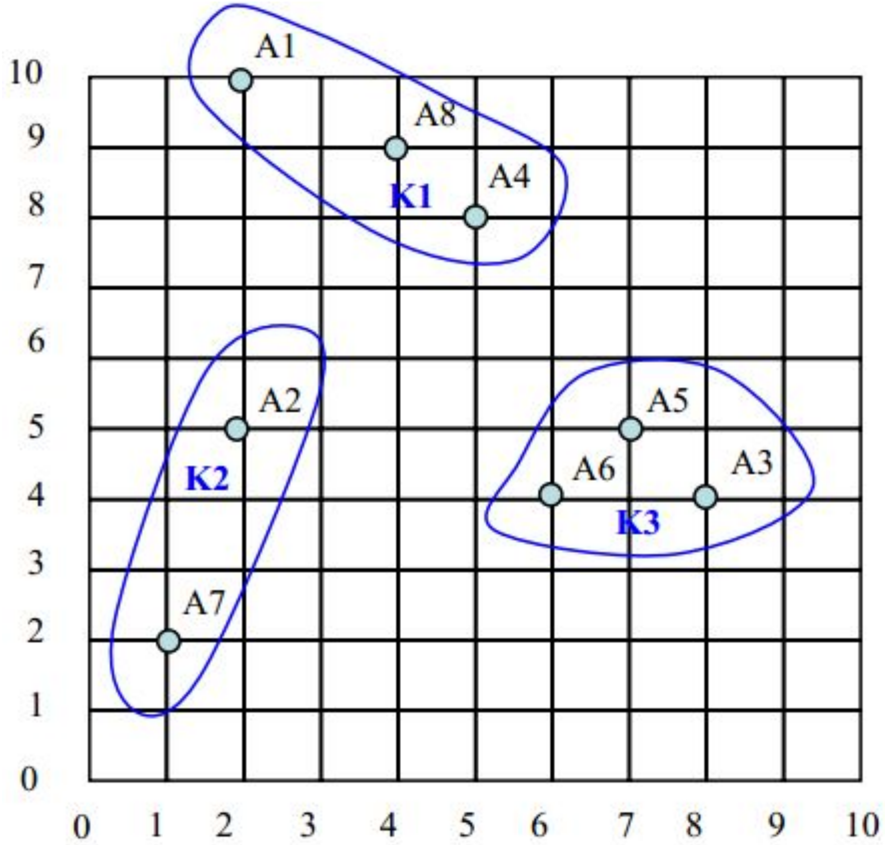
- 9 >T/2(6)
- So Cluster 3 center is A7(1,2)
- New value of T = 60/4=15

Maximin distance algorithm (Cont..)

- 5< T/2 so the procedure terminated

		(2, 10)	(8, 4)	(1, 2)	
	Point	**Dist 1**	**Dist 2**	**Dist 3**	**Min**
A1	**(2, 10)**	0	12	9	0
A2	(2, 5)	5	7	4	4
A3	**(8, 4)**	12	0	9	0
A4	(5, 8)	5	7	10	5 **[MAX]**
A5	(7, 5)	10	2	9	2
A6	(6, 4)	10	2	7	2
A7	**(1, 2)**	9	9	0	0
A8	(4, 9)	3	9	10	3

Maximin distance algorithm (Cont..)



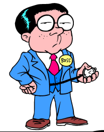
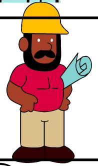
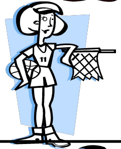
Assignment-1

Problem: Cluster the following ten points (with (x, y) representing locations) into three clusters using k-means

A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8)
A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9)
A9(9, 5), A10(1,4). Initial cluster centers are: A2(2, 5), A4(5, 8) and A7(1, 2).

Assignment-2

Determine the class for David using KNN

Customer	Age	Income	No. credit cards	Response
John 	35	35K	3	No
Rachel	22	50K	2	Yes
Hannah	63	200K	1	No
Tom 	59	170K	1	No
Nellie 	25	40K	4	Yes
David	43	83K	1	?

Assignment-3

- Use the Nearest Neighbor clustering algorithm and Euclidean distance to cluster the examples from the previous exercise: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9) A9=(9, 5), A10=(1,4).
- Suppose that the threshold t is 4.

Assignment-4

- Use the Maximin distance algorithm cluster the examples from the previous exercise: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9) A9=(9, 5), A10=(1,4).