CSCI3230 (ESTR3108)
Fundamentals of Artificial Intelligence

Tutorial 4. Clustering Algorithms

Yiyao Ma

Email: yyma23@link.cuhk.edu.hk
Office: Room 1024, 10/F, SHB

Dept. of Computer Science & Engineering
The Chinese University of Hong Kong

Outline

Part 1. Overview

Overview

In the tutorial, there are two exercise examples for today

- K-Means exercise
- DBSCAN exercise

Note: Hierarchical clustering is also important for the course. We already show a detailed exercise in our lecture 5 notes.

Part 2. K-Means exercise

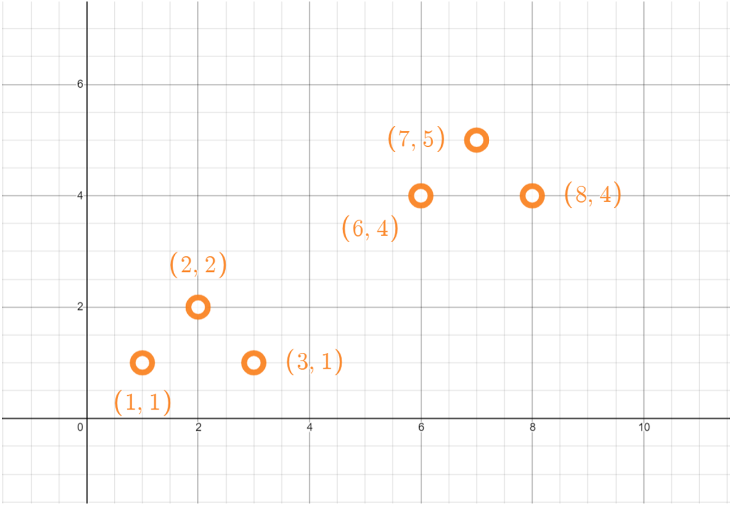
K-Means Review

Intuitively, the K-Means algorithm works as follows:

1. Choose K (random) data points (as seeds) to be the initial **centroids** as cluster centers.
2. Assign each data point to the closest centroid.
3. Re-compute the centroids using the current cluster memberships.
4. If a convergence criterion is not met, repeat steps 2 and 3.

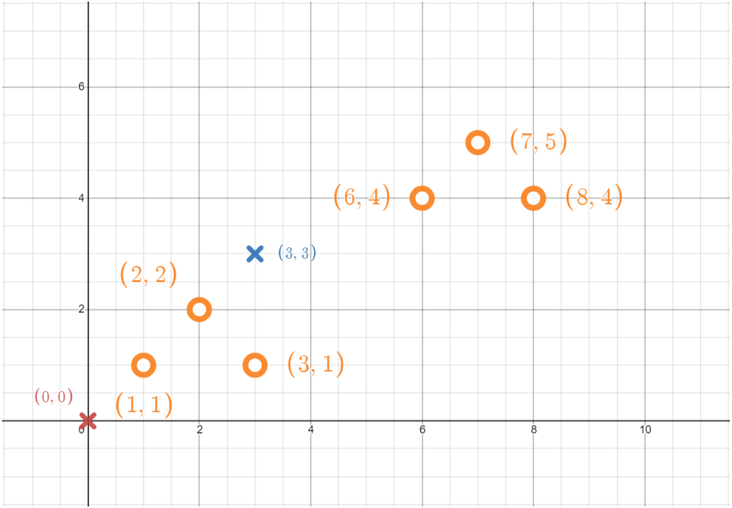
K-Means - an example

K-Means is an iterative clustering algorithm.



K-Means

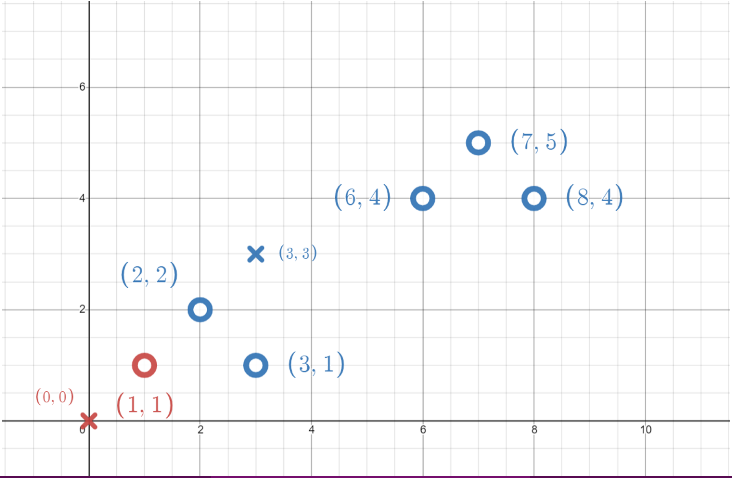
Initialize: Pick K (number of clusters) random points as cluster centroids. Assume K=2 and random centroids are (0,0) and (3,3)



K-Means

Iteration: Assign points to the closest cluster centroid (Euclidean Dist).

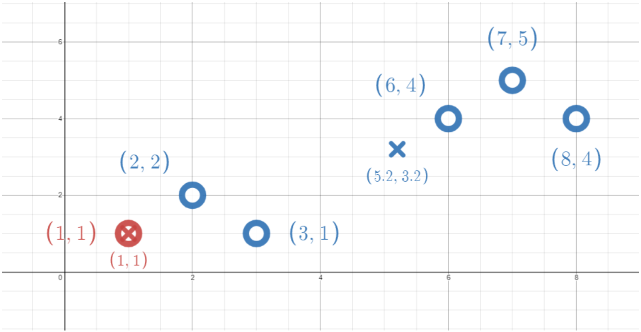
- Cluster 0: (1,1)
- Cluster 1: (2,2), (3,1), (6,4), (7,5), (8,4)



K-Means

Iteration: Update the cluster centroid to the mean of its assigned points.

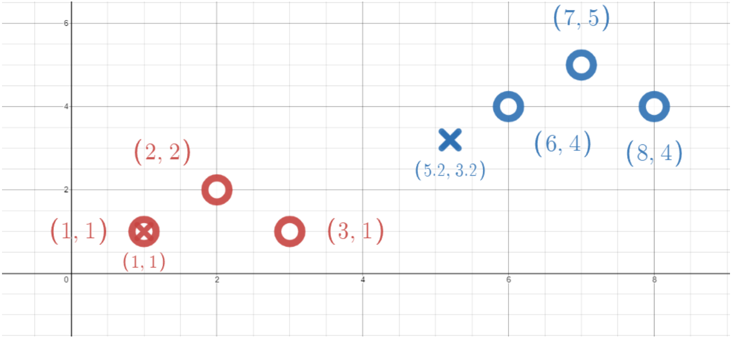
- Cluster 0: (1,1) $\rightarrow$ cluster 0 centroid is (1,1)
- Cluster 1: (2,2), (3,1), (6,4), (7,5), (8,4) $\rightarrow$ cluster 1 centroid is $((2+3+6+7+8)/5, (2+1+4+5+4)/5)$=(5.2,3.2)



K-Means

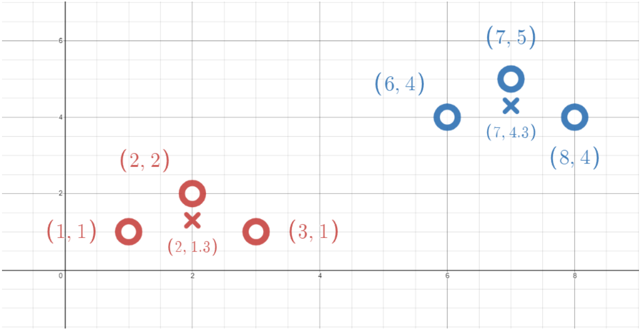
Iteration: Assign data points to the (new) closest cluster centroids.

- Cluster 0: (1,1), (2,2), (3,1)
- Cluster 1: (6,4), (7,5), (8,4)



Iteration: Update the cluster centroid to the mean of its assigned points.

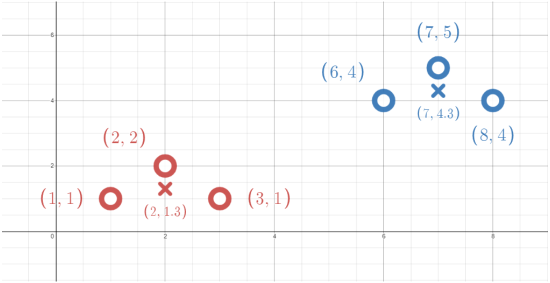
- Cluster 0: (1,1), (2,2), (3,1) $\rightarrow$ cluster 0 centroid is $((1+2+3)/3, (1+2+1)/3) = (2, 1.3)$
- Cluster 1: (6,4), (7,5), (8,4) $\rightarrow$ cluster 1 centroid is $((6+7+8)/3, (4+5+4)/3) = (7, 4.3)$



K-Means

Iteration: Assign data points to the (new) closest cluster centroid.

- Cluster 0: (1,1), (2,2), (3,1).
- Cluster 1: (6,4), (7,5), (8,4).
- No change anymore, therefore, the clustering is **finished**.



K-Means convergence (stopping) criteria

- no (or minimum) reassignment of data points to different clusters, or
- no (or minimum) change of centroids, or
- minimum decrease in the sum of squared error (SSE)
  - C_i is the i-th cluster
  - $\boldsymbol{\mu}_i$ is centroid of cluster C_i (the mean vector of all data points in C_i)
  - $\text{dist}(\boldsymbol{x}, \boldsymbol{\mu}_i)$ is the Eucledian distance between data point $\boldsymbol{x}$ and centroid $\boldsymbol{\mu}_i$:

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} \text{dist}(\boldsymbol{x}, \boldsymbol{\mu}_i)^2$$

K-Means convergence (stopping) criteria

- no (or minimum) reassignment of data points to different clusters, or
- no (or minimum) change of centroids, or
- minimum decrease in the sum of squared error (SSE)
  - C_i is the i-th cluster
  - $\boldsymbol{\mu}_i$ is centroid of cluster C_i (the mean vector of all data points in C_i)
  - dist$(\boldsymbol{x}, \boldsymbol{\mu}_i)$ is the Eucledian distance between data point $\boldsymbol{x}$ and centroid $\boldsymbol{\mu}_i$:

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} \text{dist}(\boldsymbol{x}, \boldsymbol{\mu}_i)^2$$

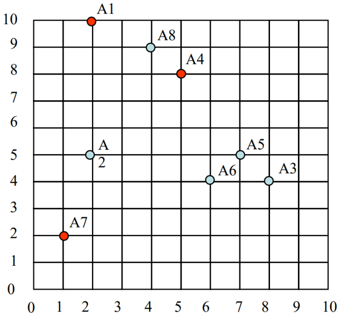
K-Means algorithm:

1. Select K points as the initial centroids.
2. **repeat**
   1) Form K clusters by assigning all points to the closest centroid.
   2) Recompute the centroid of each cluster
3. **until** fullfill the stopping condition

K-Means exercise

Use the k-means algorithm and Euclidean distance.

- $\boldsymbol{a_1}$ (2,10) ,$\boldsymbol{a_2}$ (2,5), $\boldsymbol{a_3}$ (8,4), $\boldsymbol{a_4}$ (5,8), $\boldsymbol{a_5}$ (7,5), $\boldsymbol{a_6}$ (6,4), $\boldsymbol{a_7}$ (1,2), $\boldsymbol{a_8}$ (4,9)
- Q1. Please show three clusters' centroids after the first iteration. Initial centroids are $\boldsymbol{\mu}_1$ (2,10), $\boldsymbol{\mu}_2$ (5,8), $\boldsymbol{\mu}_3$ (1,2)
- Q2. What are final three clusters?



K-Means exercise: Q1

- $\boldsymbol{a_1}$ (2,10) ,$\boldsymbol{a_2}$ (2,5), $\boldsymbol{a_3}$ (8,4), $\boldsymbol{a_4}$ (5,8), $\boldsymbol{a_5}$ (7,5), $\boldsymbol{a_6}$ (6,4), $\boldsymbol{a_7}$ (1,2), $\boldsymbol{a_8}$ (4,9)
- For point $\boldsymbol{a_1}$ (2,10)
  1. $\boldsymbol{\mu}_1$ (2,10). distance is $\sqrt{(2-2)^2+(10-10)^2}$=0
  2. $\boldsymbol{\mu}_2$ (5,8). distance is $\sqrt{(2-5)^2+(10-8)^2}=\sqrt{13}$
  3. $\boldsymbol{\mu}_3$ (1,2). distance is $\sqrt{(2-1)^2+(10-2)^2}=\sqrt{65}$

point	$\boldsymbol{\mu}_1$ (2,10)	$\boldsymbol{\mu}_2$ (5,8)	$\boldsymbol{\mu}_3$ (1,2)	Class index
$\boldsymbol{a_1}$ (2,10)	0	$\sqrt{13}$	$\sqrt{65}$	1
$\boldsymbol{a_2}$ (2,5)				
$\boldsymbol{a_3}$ (8,4)				
$\boldsymbol{a_4}$ (5,8)				
$\boldsymbol{a_5}$ (7,5)				
$\boldsymbol{a_6}$ (6,4)				
$\boldsymbol{a_7}$ (1,2)				
$\boldsymbol{a_8}$ (4,9)				

K-Means exercise: Q1

- $\boldsymbol{a_1}$ (2,10) ,$\boldsymbol{a_2}$ (2,5), $\boldsymbol{a_3}$ (8,4), $\boldsymbol{a_4}$ (5,8), $\boldsymbol{a_5}$ (7,5), $\boldsymbol{a_6}$ (6,4), $\boldsymbol{a_7}$ (1,2), $\boldsymbol{a_8}$ (4,9)
- For point $\boldsymbol{a_2}$ (2,5)
  1. $\boldsymbol{\mu}_1$ (2,10). distance is $\sqrt{(2-2)^2+(5-10)^2}$=$\sqrt{25}$=5
  2. $\boldsymbol{\mu}_2$ (5,8). distance is $\sqrt{(2-5)^2+(5-8)^2}$=$\sqrt{18}$
  3. $\boldsymbol{\mu}_3$ (1,2). distance is $\sqrt{(2-1)^2+(5-2)^2}$=$\sqrt{10}$

point	$\boldsymbol{\mu}_1$ (2,10)	$\boldsymbol{\mu}_2$ (5,8)	$\boldsymbol{\mu}_3$ (1,2)	Class index
$\boldsymbol{a_1}$ (2,10)	0	$\sqrt{13}$	$\sqrt{65}$	1
$\boldsymbol{a_2}$ (2,5)	5	$\sqrt{18}$	$\sqrt{10}$	3
$\boldsymbol{a_3}$ (8,4)				
$\boldsymbol{a_4}$ (5,8)				
$\boldsymbol{a_5}$ (7,5)				
$\boldsymbol{a_6}$ (6,4)				
$\boldsymbol{a_7}$ (1,2)				
$\boldsymbol{a_8}$ (4,9)				

K-Means exercise: Q1

- $\boldsymbol{a_1}$ (2,10) ,$\boldsymbol{a_2}$ (2,5), $\boldsymbol{a_3}$ (8,4), $\boldsymbol{a_4}$ (5,8), $\boldsymbol{a_5}$ (7,5), $\boldsymbol{a_6}$ (6,4), $\boldsymbol{a_7}$ (1,2), $\boldsymbol{a_8}$ (4,9)
- For point $\boldsymbol{a_3}$ (8,4)
  1. $\boldsymbol{\mu}_1$ (2,10). distance is $\sqrt{(8-2)^2+(4-10)^2}=\sqrt{36}=6$
  2. $\boldsymbol{\mu}_2$ (5,8). distance is $\sqrt{(8-5)^2+(4-8)^2}=\sqrt{25}=5$
  3. $\boldsymbol{\mu}_3$ (1,2). distance is $\sqrt{(8-1)^2+(4-2)^2}=\sqrt{53}$

point	$\boldsymbol{\mu}_1$ (2,10)	$\boldsymbol{\mu}_2$ (5,8)	$\boldsymbol{\mu}_3$ (1,2)	Class index
$\boldsymbol{a_1}$ (2,10)	0	$\sqrt{13}$	$\sqrt{65}$	1
$\boldsymbol{a_2}$ (2,5)	5	$\sqrt{18}$	$\sqrt{10}$	3
$\boldsymbol{a_3}$ (8,4)	6	5	$\sqrt{53}$	2
$\boldsymbol{a_4}$ (5,8)				
$\boldsymbol{a_5}$ (7,5)				
$\boldsymbol{a_6}$ (6,4)				
$\boldsymbol{a_7}$ (1,2)				
$\boldsymbol{a_8}$ (4,9)				

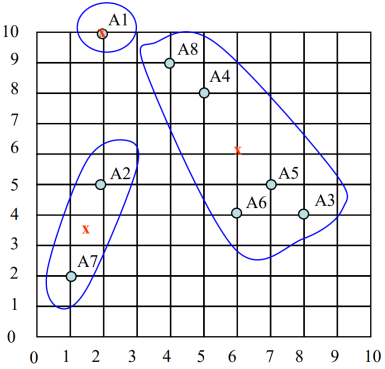
K-Means exercise: Q1

- $\boldsymbol{a_1}$ (2,10) ,$\boldsymbol{a_2}$ (2,5), $\boldsymbol{a_3}$ (8,4), $\boldsymbol{a_4}$ (5,8), $\boldsymbol{a_5}$ (7,5), $\boldsymbol{a_6}$ (6,4), $\boldsymbol{a_7}$ (1,2), $\boldsymbol{a_8}$ (4,9)
- Repeat the above steps for each point. We can finally get the following table.

point	$\boldsymbol{\mu}_1$ (2,10)	$\boldsymbol{\mu}_2$ (5,8)	$\boldsymbol{\mu}_3$ (1,2)	Class index
$\boldsymbol{a_1}$ (2,10)	0	$\sqrt{13}$	$\sqrt{65}$	1
$\boldsymbol{a_2}$ (2,5)	5	$\sqrt{18}$	$\sqrt{10}$	3
$\boldsymbol{a_3}$ (8,4)	6	5	$\sqrt{53}$	2
$\boldsymbol{a_4}$ (5,8)	$\sqrt{13}$	0	$\sqrt{50}$	2
$\boldsymbol{a_5}$ (7,5)	$\sqrt{50}$	$\sqrt{13}$	$\sqrt{45}$	2
$\boldsymbol{a_6}$ (6,4)	$\sqrt{52}$	$\sqrt{17}$	$\sqrt{29}$	2
$\boldsymbol{a_7}$ (1,2)	$\sqrt{65}$	$\sqrt{52}$	0	3
$\boldsymbol{a_8}$ (4,9)	$\sqrt{5}$	$\sqrt{2}$	$\sqrt{58}$	2

K-Means exercise: Q1

- a_1 (2,10) ,a_2 (2,5), a_3 (8,4), a_4 (5,8), a_5 (7,5), a_6 (6,4), a_7 (1,2), a_8 (4,9)
  1. Cluster 1: a_1. Hence μ_1=(2,10)
  2. Cluster 2: a_3, a_4, a_5, a_6, a_8. μ_2= (a_3+ a_4+ a_5+ a_6+a_8)/5= (((8+5+7+6+4)/5, (4+8+5+4+9)/5) =(6,6)
  3. Cluster 3: a_2, a_7. μ_3=(a_2+ a_7)/2= ((2+1)/2, (5+2)/2)=(1.5,3.5)



K-Means exercise: Q1

By the new $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$, we will have the following table for next iteration calculation

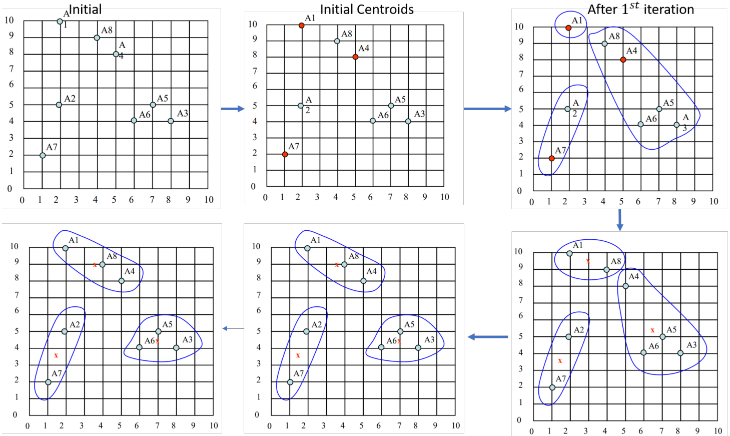
point	$\boldsymbol{\mu}_1$ (2,10)	$\boldsymbol{\mu}_2$ (6,6)	$\boldsymbol{\mu}_3$ (1.5,3.5)	Class index
$\boldsymbol{a_1}$ (2,10)				
$\boldsymbol{a_2}$ (2,5)				
$\boldsymbol{a_3}$ (8,4)				
$\boldsymbol{a_4}$ (5,8)				
$\boldsymbol{a_5}$ (7,5)				
$\boldsymbol{a_6}$ (6,4)				
$\boldsymbol{a_7}$ (1,2)				
$\boldsymbol{a_8}$ (4,9)				

K-Means exercise: Q2

- 1^{st} iteration, Cluster 1=$\{\boldsymbol{a_1}\}$, Cluster 2= $\{\ \boldsymbol{a_3},\ \boldsymbol{a_4},\ \boldsymbol{a_5},\ \boldsymbol{a_6},\ \boldsymbol{a_8}\ \}$. Cluster 3= $\{\ \boldsymbol{a_2},\ \boldsymbol{a_7}\ \}$ $\boldsymbol{\mu}_1$=(2,10), $\boldsymbol{\mu}_2$=(6,6), $\boldsymbol{\mu}_3$=(1.5,3.5)
- 2^{nd} iteration, Cluster 1=$\{\ \boldsymbol{a_1},\ \boldsymbol{a_8}\ \}$, Cluster 2= $\{\ \boldsymbol{a_3},\ \boldsymbol{a_4},\ \boldsymbol{a_5},\ \boldsymbol{a_6}\ \}$. Cluster 3= $\{\ \boldsymbol{a_2},\ \boldsymbol{a_7}\ \}$ $\boldsymbol{\mu}_1$=(3,9.5), $\boldsymbol{\mu}_2$=(6.5,5.25), $\boldsymbol{\mu}_3$=(1.5,3.5)
- 3^{rd} iteration, Cluster 1=$\{\ \boldsymbol{a_1},\ \boldsymbol{a_4},\ \boldsymbol{a_8}\ \}$, Cluster 2= $\{\ \boldsymbol{a_3},\ \boldsymbol{a_5},\ \boldsymbol{a_6}\ \}$. Cluster 3= $\{\ \boldsymbol{a_2},\ \boldsymbol{a_7}\ \}$ $\boldsymbol{\mu}_1$=(3.66,9), $\boldsymbol{\mu}_2$=(7,4.33), $\boldsymbol{\mu}_3$=(1.5,3.5)
- 4^{th} iteration, Cluster 1=$\{\ \boldsymbol{a_1},\ \boldsymbol{a_4},\ \boldsymbol{a_8}\ \}$, Cluster 2= $\{\ \boldsymbol{a_3},\ \boldsymbol{a_5},\ \boldsymbol{a_6}\ \}$. Cluster 3= $\{\ \boldsymbol{a_2},\ \boldsymbol{a_7}\ \}$ $\boldsymbol{\mu}_1$=(3.66,9), $\boldsymbol{\mu}_2$=(7,4.33), $\boldsymbol{\mu}_3$=(1.5,3.5)
- Terminate because point assignment after 3^{rd} iteration is the same as the point assignment after 4^{th} iteration

K-Means exercise: Q2

- Terminate because point assignment after 3^{rd} iteration is the same as the point assignment after 4^{th} iteration



After 4^{th} iteration

After 3^{rd} iteration

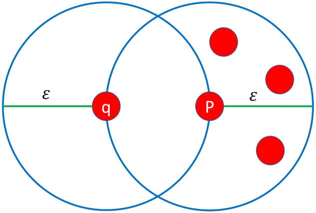
After 2^{nd} iteration

Part 3. DBSCAN exercise

DBSCAN - Review

DBSCAN: Density-based spatial clustering of applications with noise

- Two parameters:
  - ϵ: maximum radius of the neighborhood
    ϵ-Neighbor: data points within a radius of ϵ from a data point (including the point itself)
  - *MinPts*: minimum number of points required in an ϵ-Neighbor
- Density definition:
  - density $=$ number of points within a specified radius ϵ
  - "high density": data point's ϵ-Neighbor contains at least *MinPts* data



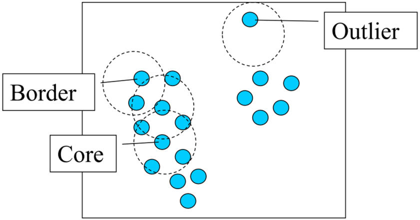
$MinPts = 4$

$\epsilon - Neighbthorhood\ of\ p$
$\epsilon - Neighbthorhood\ of\ q$
$Density\ pf\ p\ is$ "high" $(MinPts = 4)$
$Density\ pf\ q\ is$ "low" $(MinPts = 4)$

DBSCAN - Review

Given ϵ and *MinPts*, categorize the data points to three exclusive groups.

- **Core point:** has more than or equal to MinPts within ϵ. These are points that are at the interior of a cluster.
- **Border point:** has fewer than MinPts within ϵ, but is in the neighborhood of a core point.
- **Noise point (or outlier):** any point that is neither a core point nor a border point.



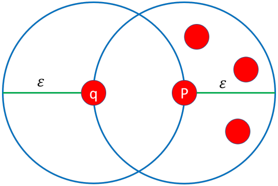
ε = 1unit, MinPts = 5

Density-reachable

Directly density-reachable

- An object q is directly density-reachable from object p if p is a core object and q is in p's ϵ-neighborhood.

In the following example:



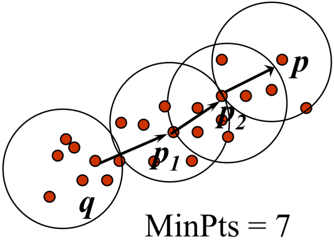
$MinPts = 4$

- q is directly density-reachable from p
- p is not directly density-reachable from q
- Note that density-reachability is asymmetric

Density-connected

Indirectly density-reachable (a.k.a. density-connected)

- A point p is directly density-reachable from p_2
- p_2 is directly density-reachable from p_1
- p_1 is directly density-reachable from q
- $p \leftarrow p_2 \leftarrow p_1 \leftarrow q$ form a chain
- p is (indirectly) density-reachable from q
  (a.k.a. p is density-connected from q)
- q is not density-reachable from p (why?)

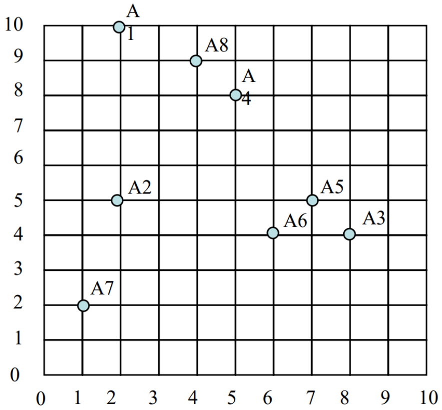


DBSCAN algorithm

- Label data points into core, border and noise
- Eliminate noise points
- For every core point p that has not been assigned to a cluster
  - Create a new cluster with the point p and all the points that are density-connected to p
- Assign border points to the cluster of the closest core point.

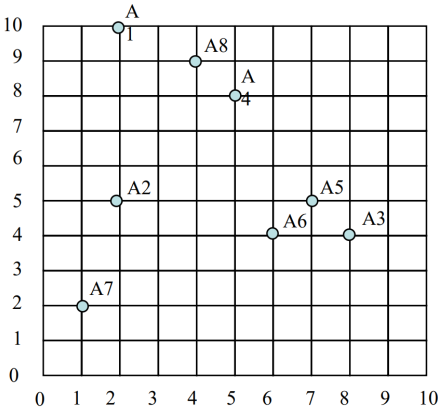
DBSCAN exercise

- If ϵ is $\sqrt{5}$ and minpoint is 3, what are the clusters that DBSCAN would discover with the following 8 points?
- $\boldsymbol{a_1}$ (2,10) ,$\boldsymbol{a_2}$ (2,5), $\boldsymbol{a_3}$ (8,4), $\boldsymbol{a_4}$ (5,8), $\boldsymbol{a_5}$ (7,5), $\boldsymbol{a_6}$ (6,4), $\boldsymbol{a_7}$ (1,2), $\boldsymbol{a_8}$ (4,9)



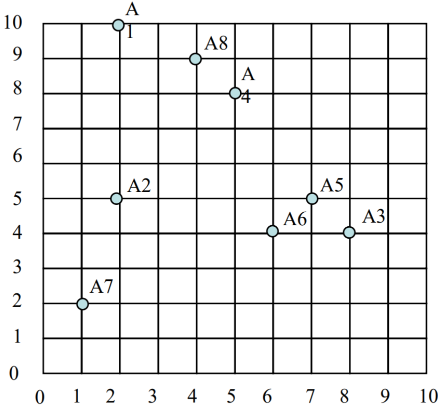
DBSCAN exercise

- $\boldsymbol{a_1}$ (2,10) ,$\boldsymbol{a_2}$ (2,5), $\boldsymbol{a_3}$ (8,4), $\boldsymbol{a_4}$ (5,8), $\boldsymbol{a_5}$ (7,5), $\boldsymbol{a_6}$ (6,4), $\boldsymbol{a_7}$ (1,2), $\boldsymbol{a_8}$ (4,9)
- neighbors of $\boldsymbol{a_1}$ (2,10) $\boldsymbol{a_1}$, $\boldsymbol{a_8}$.
- neighbors of $\boldsymbol{a_2}$ (2,5) $\boldsymbol{a_2}$



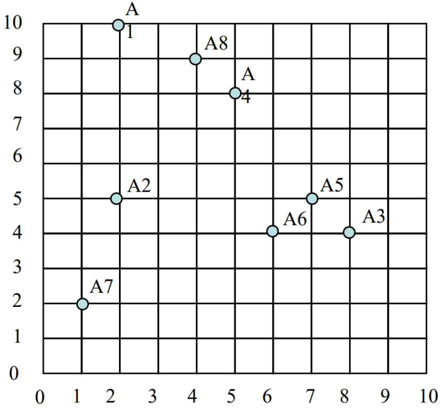
DBSCAN exercise

- $\boldsymbol{a_1}$ (2,10) ,$\boldsymbol{a_2}$ (2,5), $\boldsymbol{a_3}$ (8,4), $\boldsymbol{a_4}$ (5,8), $\boldsymbol{a_5}$ (7,5), $\boldsymbol{a_6}$ (6,4), $\boldsymbol{a_7}$ (1,2), $\boldsymbol{a_8}$ (4,9)
- neighbors of $\boldsymbol{a_3}$ (8,4) $\boldsymbol{a_3}$, $\boldsymbol{a_5}$, $\boldsymbol{a_6}$
- neighbors of $\boldsymbol{a_4}$ (5,8) $\boldsymbol{a_4}$, $\boldsymbol{a_8}$



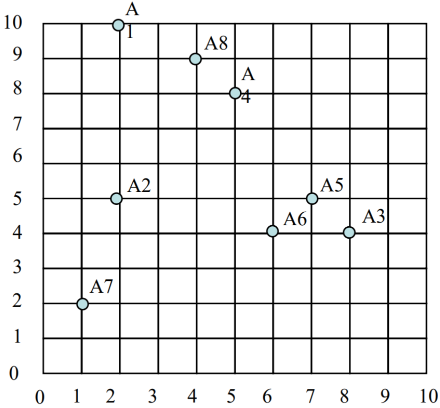
DBSCAN exercise

- $\boldsymbol{a_1}$ (2,10) ,$\boldsymbol{a_2}$ (2,5), $\boldsymbol{a_3}$ (8,4), $\boldsymbol{a_4}$ (5,8), $\boldsymbol{a_5}$ (7,5), $\boldsymbol{a_6}$ (6,4), $\boldsymbol{a_7}$ (1,2), $\boldsymbol{a_8}$ (4,9)
- neighbors of $\boldsymbol{a_5}$ (7,5) $\boldsymbol{a_5}$, $\boldsymbol{a_3}$, $\boldsymbol{a_6}$
- neighbors of $\boldsymbol{a_6}$ (6,4) $\boldsymbol{a_6}$, $\boldsymbol{a_3}$, $\boldsymbol{a_5}$



DBSCAN exercise

- a_1 (2,10) ,a_2 (2,5), a_3 (8,4), a_4 (5,8), a_5 (7,5), a_6 (6,4), a_7 (1,2), a_8 (4,9)
- neighbors of a_7 (1,2) a_7
- neighbors of a_8 (4,9) a_8,a_1, a_4



DBSCAN exercise

- $\boldsymbol{a_1}$ (2,10) ,$\boldsymbol{a_2}$ (2,5), $\boldsymbol{a_3}$ (8,4), $\boldsymbol{a_4}$ (5,8), $\boldsymbol{a_5}$ (7,5), $\boldsymbol{a_6}$ (6,4), $\boldsymbol{a_7}$ (1,2), $\boldsymbol{a_8}$ (4,9)
- Core point: $\boldsymbol{a_3}$,$\boldsymbol{a_5}$,$\boldsymbol{a_6}$,$\boldsymbol{a_8}$
- Border point: $\boldsymbol{a_1}$, $\boldsymbol{a_4}$
- Noise point: $\boldsymbol{a_2}$,$\boldsymbol{a_7}$

DBSCAN exercise

- Core point: $\boldsymbol{a_3}$,$\boldsymbol{a_5}$,$\boldsymbol{a_6}$,$\boldsymbol{a_8}$
- Border point: $\boldsymbol{a_1}$, $\boldsymbol{a_4}$
- Noise point: $\boldsymbol{a_2}$,$\boldsymbol{a_7}$
- Hence.
  1. Cluster 1. $\boldsymbol{a_1}$, $\boldsymbol{a_4}$, $\boldsymbol{a_8}$
  2. Cluster 2. $\boldsymbol{a_3}$, $\boldsymbol{a_5}$, $\boldsymbol{a_6}$