Clustering: Unsupervised Learning

IF-3270 Pembelajaran Mesin

Teknik Informatika ITB

Modul 7: Clustering

Fariska Z. Ruskanda, S.T., M.T.
(fariska@informatika.org)

KK IF -Teknik Informatika - STEI ITB

03 Density-based Clustering

IF3270 - Pembelajaran Mesin
(Machine Learning)

Outline

DBSCAN

Density-based Clustering

Example

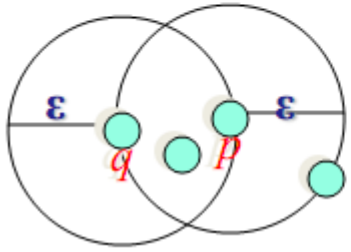
Density-based Clustering

- Clusters are dense regions in the data space, separated by regions of lower object density
- A cluster is defined as a maximal set of density-connected points

ε-Neighborhood – Objects within a radius of ε from an object.

$$N_\varepsilon(p) : \{q \mid d(p,q) \leq \varepsilon\}$$

"High density" - ε-Neighborhood of an object contains at least *MinPts* of objects.



ε-Neighborhood of p
ε-Neighborhood of q

Density of p is "high" (MinPts = 4)

Density of q is "low" (MinPts = 4)

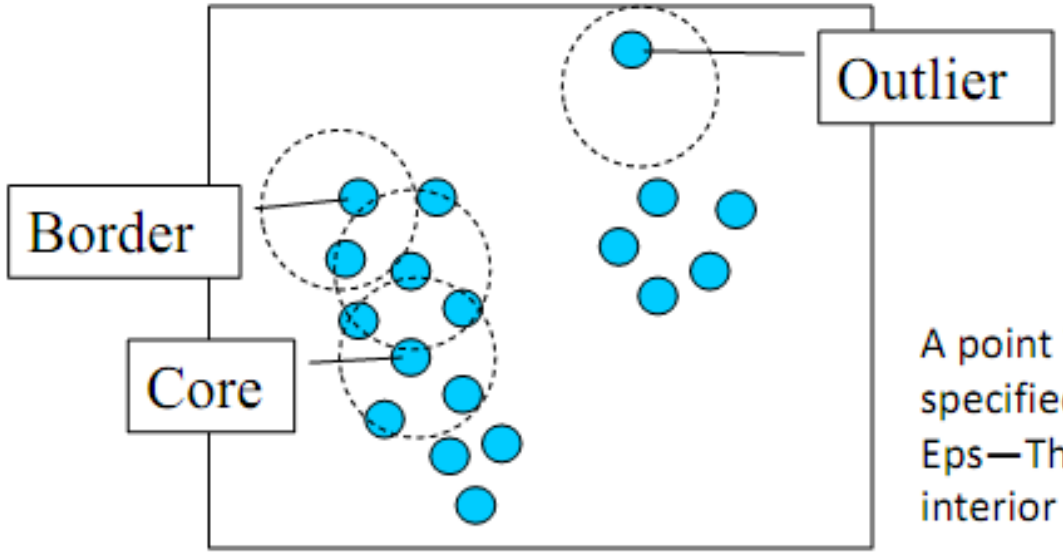
DBSCAN (Density Based Spatial Clustering of Applications with Noise)

- DBSCAN starts with an arbitrary point p and retrieves all points density-reachable from p wrt. Eps and MinPts.
  - If p is a core point, this procedure yields a cluster wrt. Eps and MinPts.
  - If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.

DBSCAN: Object Type

- The neighborhood within a radius ε of a given object is called the **ε-neighborhood** of the object.
- If the ε-neighborhood of an object contains at least a minimum number, *MinPts*, of objects, then the object is called a **core object**.
- Given a set of objects, D, we say that an object $\boldsymbol{p}$ is **directly density-reachable** from object $\boldsymbol{q}$ if $\boldsymbol{p}$ is within the ε-neighborhood of $\boldsymbol{q}$, and $\boldsymbol{q}$ is a core object.
- An object $\boldsymbol{p}$ is **density-reachable** from object $\boldsymbol{q}$ with respect to ε and *MinPts* in a set of objects, D, if there is a chain of objects $\boldsymbol{p_1}, \ldots, \boldsymbol{p_n}$, where $\boldsymbol{p_1} = \boldsymbol{q}$ and $\boldsymbol{p_n} = \boldsymbol{p}$ such that $\boldsymbol{p_{i+1}}$ is directly density-reachable from $\boldsymbol{p_i}$ with respect to ε and *MinPts*, for $1 \leq i \leq n$, $\boldsymbol{p_i} \in D$.
- An object $\boldsymbol{p}$ is **density-connected** to object $\boldsymbol{q}$ with respect to ε and *MinPts* in a set of objects, D, if there is an object $\mathbf{o} \in D$ such that both $\boldsymbol{p}$ and $\boldsymbol{q}$ are density-reachable from $\boldsymbol{o}$ with respect to ε and *MinPts*.

Core, Border, Outlier



Given ε and *MinPts*, categorize the objects into three exclusive groups.

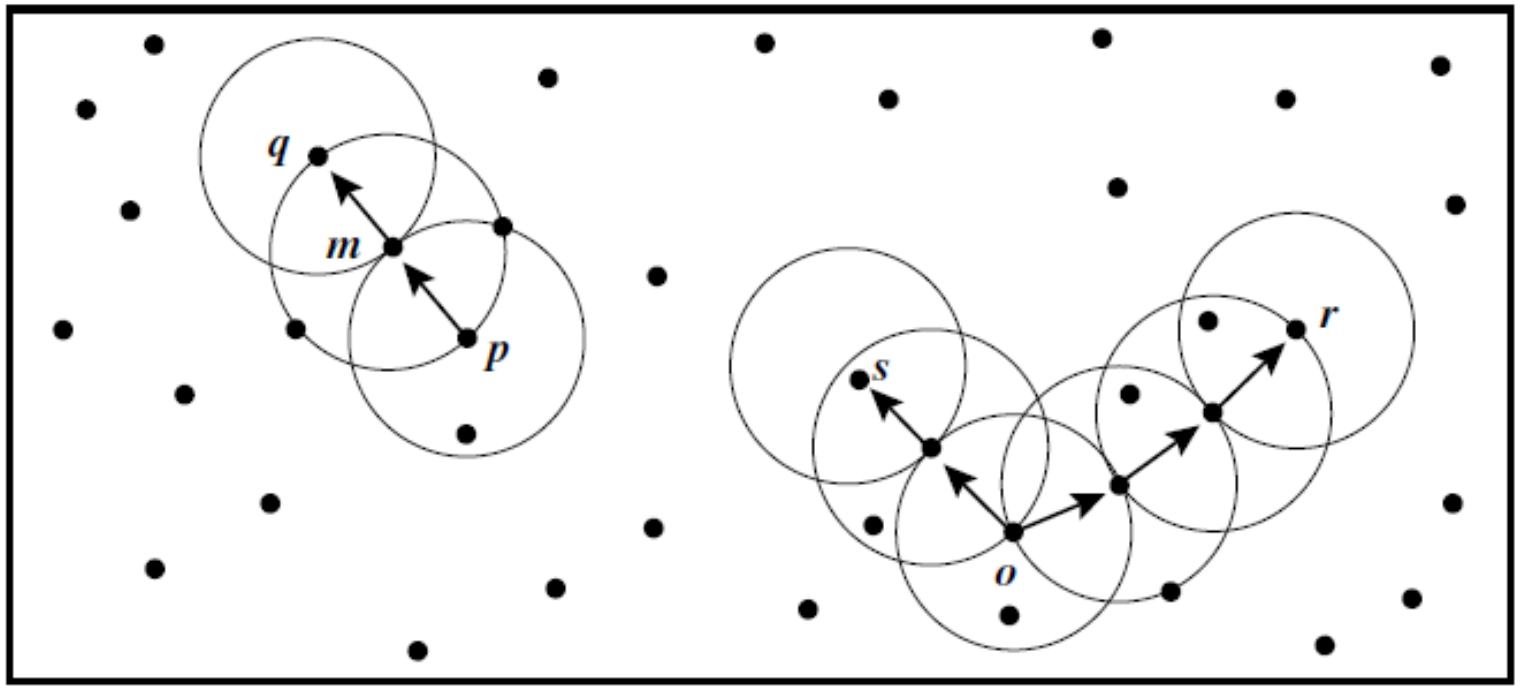
A point is a core point if it has more than a specified number of points (MinPts) within Eps—These are points that are at the interior of a cluster.

$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point.

A noise point is any point that is not a core point nor a border point.

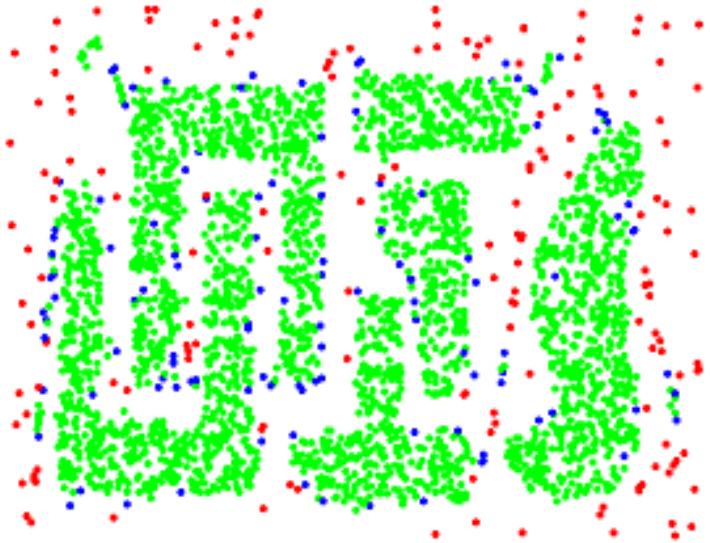
DBSCAN find clusters



DBSCAN searches for clusters by checking the e-neighborhood of each point in the database. If the e-neighborhood of a point ***p contains*** more than *MinPts, a new cluster with* ***p as a core object is created. DBSCAN then*** iteratively collects directly density-reachable objects from these core objects, which may involve the merge of a few density-reachable clusters. The process terminates when no new point can be added to any cluster.

DBSCAN Example

Original Points



Point types: core, border and outliers

ε = 10, MinPts = 4

DBScan – Example (1)

- If Epsilon is 2 and minpoint is 2, what are the clusters that DBScan would discover with the following 8 examples:

  A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

  Matriks jarak (kuadrat):

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	25	72	13	50	52	65	5
A2	25	0	37	18	25	17	10	20
A3	72	37	0	25	2	4	53	41
A4	13	18	25	0	13	17	52	2
A5	50	25	2	13	0	2	45	25
A6	52	17	4	17	2	0	29	29
A7	65	10	53	52	45	29	0	58
A8	5	20	41	2	25	29	58	0

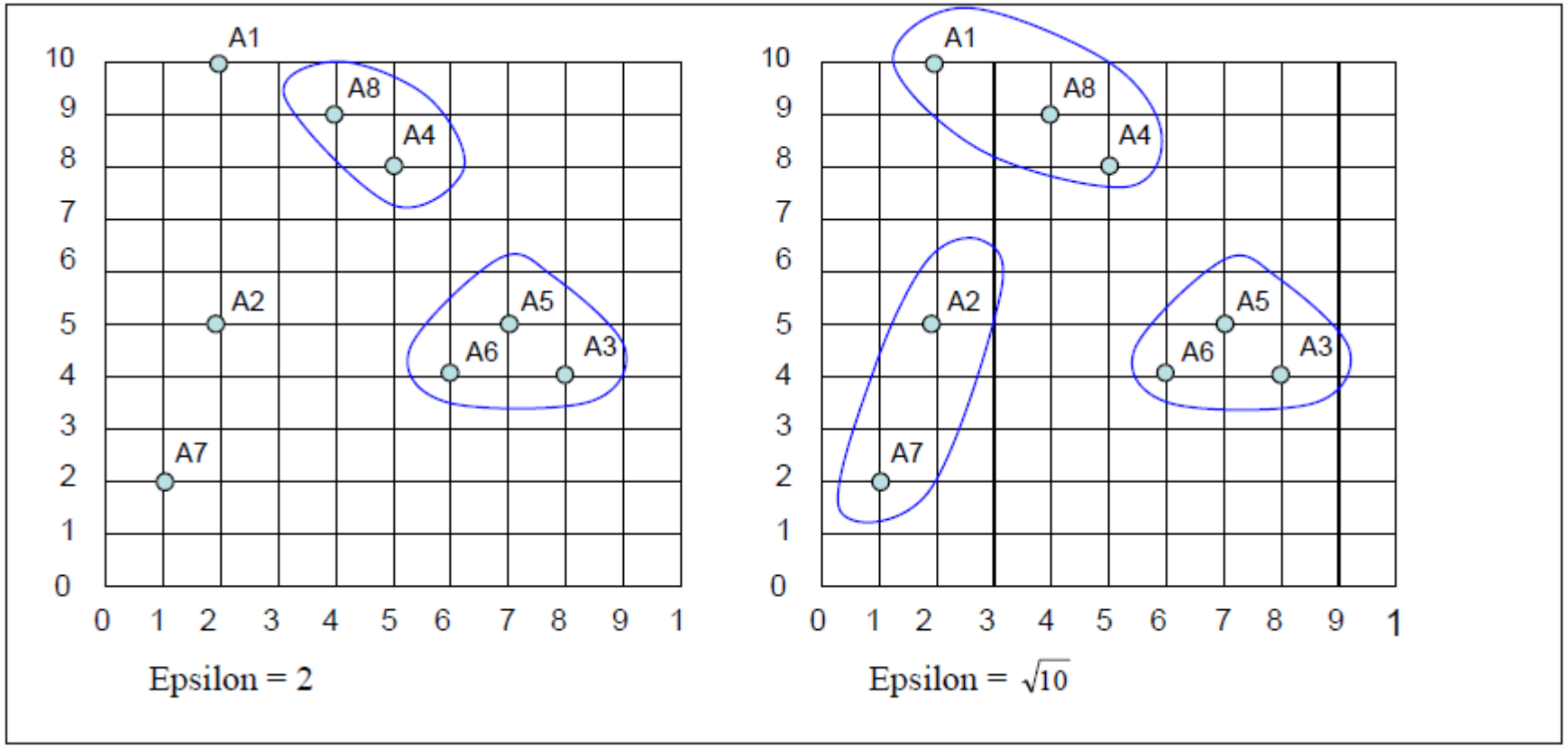
DBScan – Example (2)

- Solutions:
- Epsilon neighborhood of each point
  - N2(A1)={};
  - N2(A2)={};
  - N2(A3)={A5, A6};
  - N2(A4)={A8};
  - N2(A5)={A3, A6};
  - N2(A6)={A3, A5};
  - N2(A7)={};
  - N2(A8)={A4}
  - So A1, A2, and A7 are outliers, while we have two clusters C1={A4, A8} and C2={A3, A5, A6}

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	25	72	13	50	52	65	5
A2	25	0	37	18	25	17	10	20
A3	72	37	0	25	2	4	53	41
A4	13	18	25	0	13	17	52	2
A5	50	25	2	13	0	2	45	25
A6	52	17	4	17	2	0	29	29
A7	65	10	53	52	45	29	0	58
A8	5	20	41	2	25	29	58	0

- If Epsilon is $10^{1/2}$ then the neighborhood of some points will increase:
  - A1 would join the cluster C1 and A2 would joint with A7 to form cluster C3={A2, A7}.

DBScan – Example (3)



04 Hierarchical Clustering

IF3270 Pembelajaran Mesin