Clustering - DBSCAN Tutorial

Lin Guosheng
School of Computer Science and Engineering
Nanyang Technological University

Question 1

Q: Given the data points below and the parameters: *epsilon=3, minPoints=4*, write down which of the points are core points, border points and noise points, respectively.

A(2,2) B(4,3) C(3,5) D(1,5)
E(-5,3) F(5,1) G(-4,6) H(-3,4) I(-2,6)

Solution

DBSCAN parameters: *epsilon=3, minPoints=4*

To simplify calculation, here we use squared L2 distance,
so we need to compared the squared distance with squared epsilon: $\epsilon^2 = 3^2 = 9$
to identify the number of points in the epsilon-neighbourhood:

	A	B	C	D	E	F	G	H	I	#points within epsilon
A	0	5	10	10	50	10	52	29	32	2
B	5	0	5	13	81	5	73	50	45	4
C	10	5	0	4	68	20	50	37	26	3
D	10	13	4	0	40	32	26	17	10	2
E	50	81	68	40	0	104	10	5	18	2
F	10	5	20	32	104	0	106	73	74	2
G	52	73	50	26	10	106	0	5	4	3
H	29	50	37	17	5	73	5	0	5	4
I	32	45	26	10	18	74	4	5	0	3

The epsilon neighbourhood is defined using L2 distance.

If we want to use **squared** L2 distance to simplify the calculation, we need to use **squared** epsilon for distance comparison to identify the neighbouring points

$$\epsilon^2 = 3^2 = 9$$

L2 distance: $$d_{L2}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Squared L2 distance: $$\|x - y\| = \sum_{i=1}^{d} (x_i - y_i)^2$$

Solution

Base on the core points B, H,
we can find the directly reachable points as the border points.

We can look at the columns in the distance table to find the border points:
Data points within the eps-neighbourhood of B: {A, C, F}
Data points within the eps-neighbourhood of H: {E, G, I}

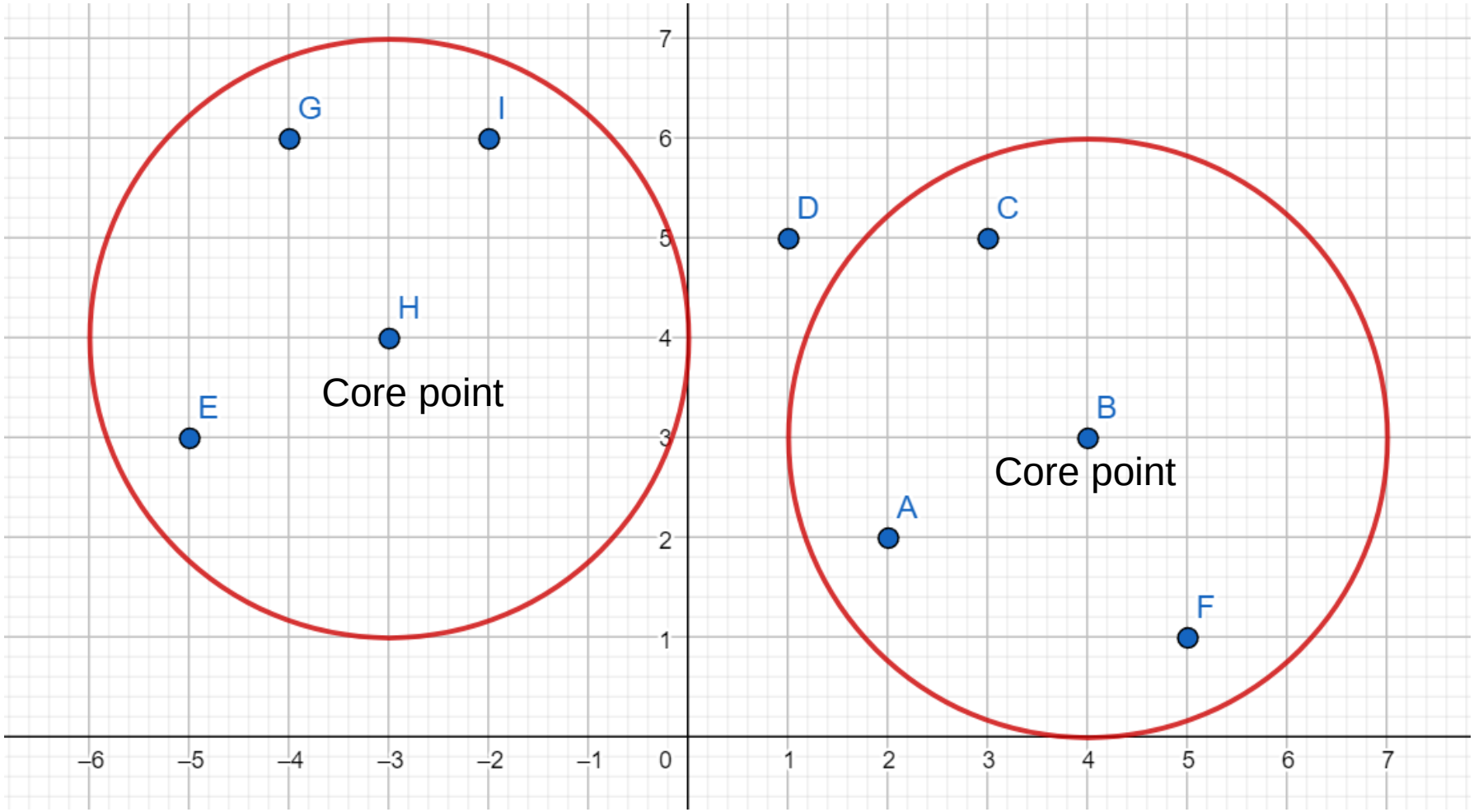
	A	B	C	D	E	F	G	H	I	#points within epsilon
B	5	0	5	13	81	5	73	50	45	4
H	29	50	37	17	5	73	5	0	5	4

The remaining points are noise points: {D}

- Answer:
  - Core points: B, H
  - Border points: A, C, F, E, G, I
  - Noise points: D

Visualization of data points

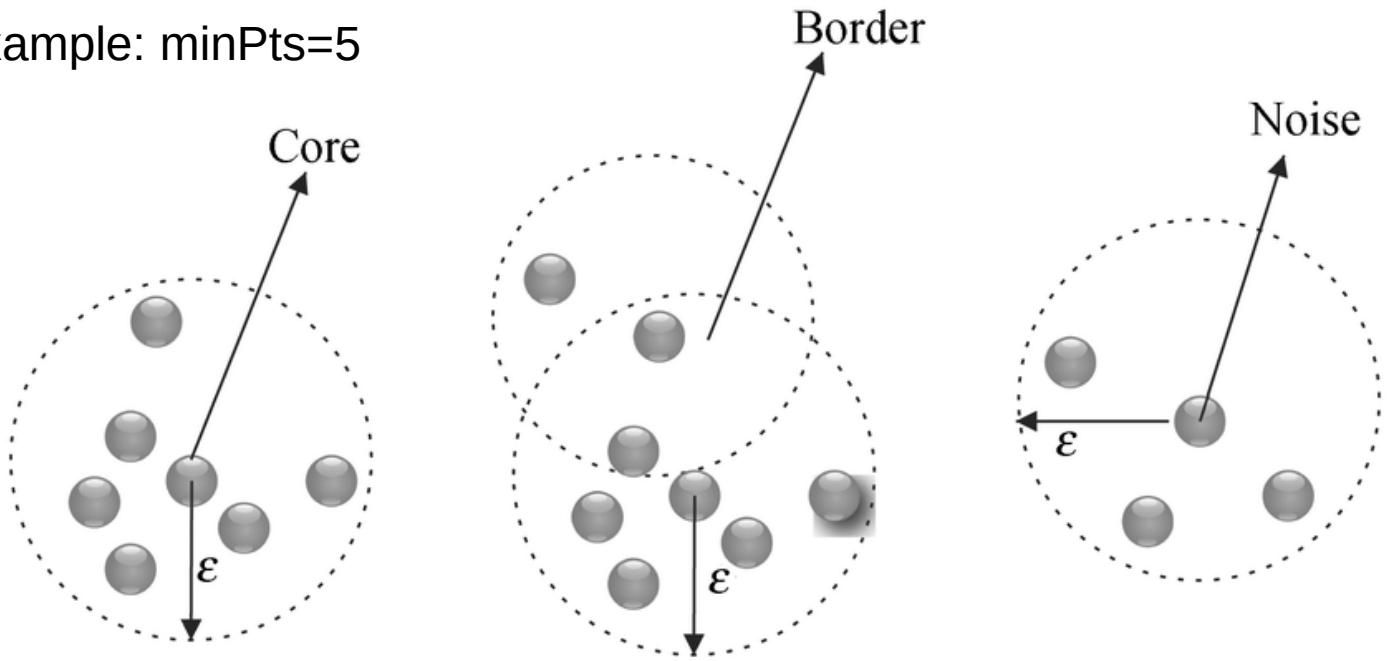
DBSCAN parameters: *epsilon=3, minPoints=4*



Review

- **Core point:** A point is a core point if there are at least minPts number of points (including the point itself) in its surrounding area with radius eps (within the epsilon-neighbourhood).
- **Border point**: A point is a border point if it is not a core point and it is directly reachable from a core point.
- **Outlier** (**noise point**): A point is an outlier if it is not a core point and not reachable from any core points.

Example: minPts=5



https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html