DATA MINING 2
Exercises – Outliers Detection

Riccardo Guidotti

a.a. 2019/2020

UNIVERSITÀ DI PISA

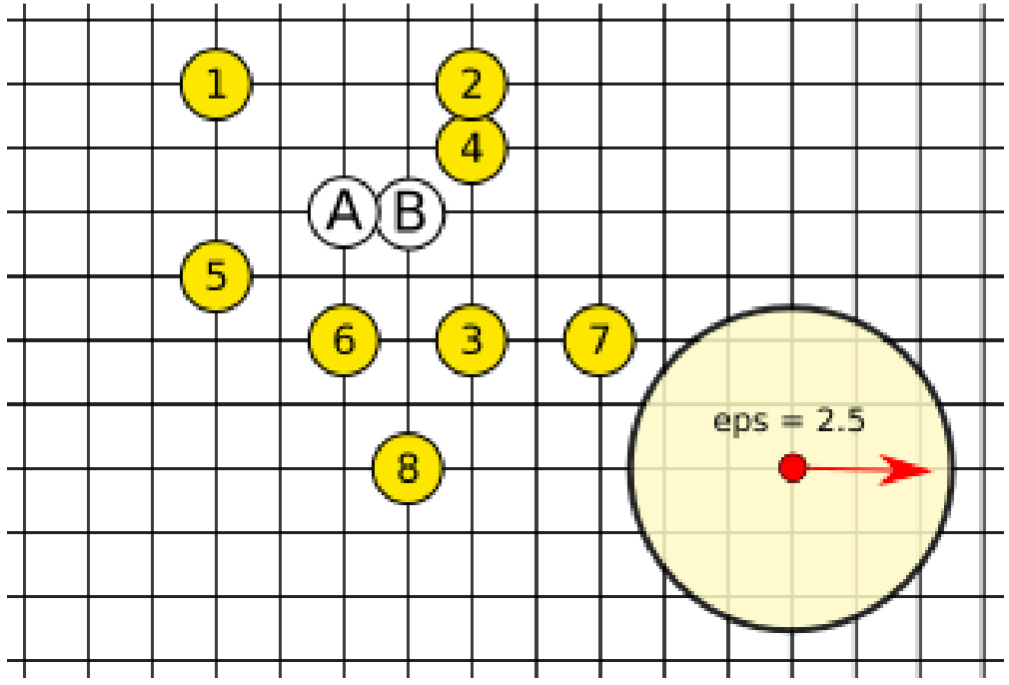
Outlier Detection – Exercise 1

Given the dataset of 10 points below, consider the outlier detection problem for points A and B, adopting the following three methods:

a) Distance-based: DB(ε,π) **(2 points)**
Are A and/or B outliers, if thresholds are forced to ε = 2.5 and π = 0.15 ? The point itself should not be counted.

b) Density-based: LOF **(2 points)**
Compute the LOF score for points A and B by taking k=2, i.e. comparing each point with its 2 NNs (not counting the point itself). In order to simplify the calculations, the reachability-distance used by LOF can be replaced by the simple Euclidean distance.

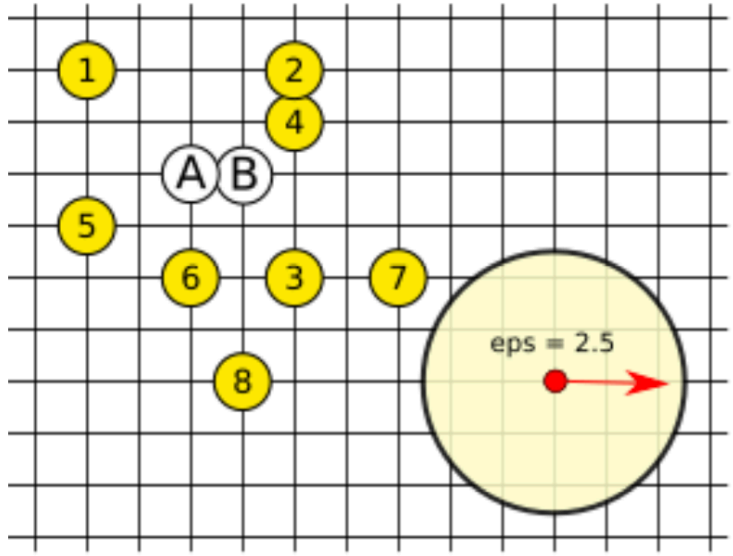
c) Depth-based **(2 points)**
Compute the depth score of all points.



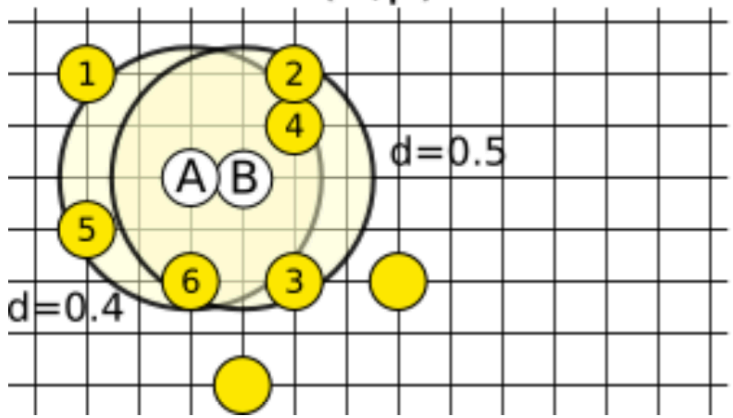
Outlier Detection – Exercise 1 – Solution

Distance-based

- No outliers because within their radius there are 0.4 and 0.5 points for A and B, respectively



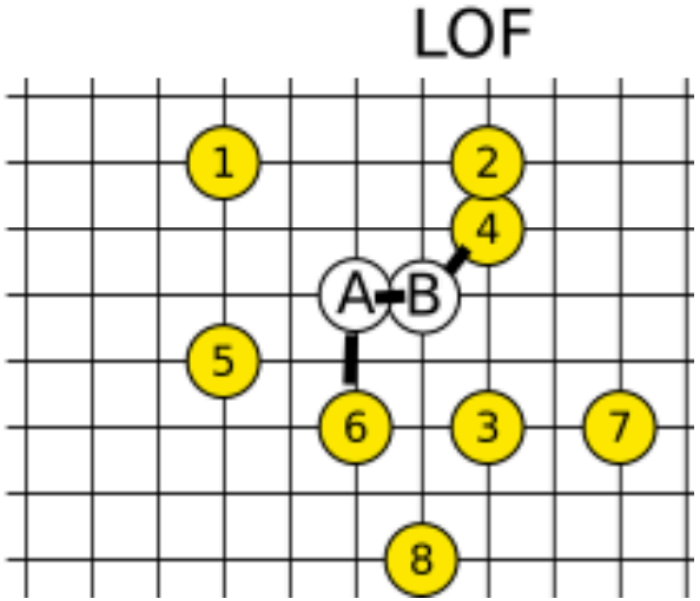
DB(e,p)



Outlier Detection – Exercise 1 – Solution

Density-based

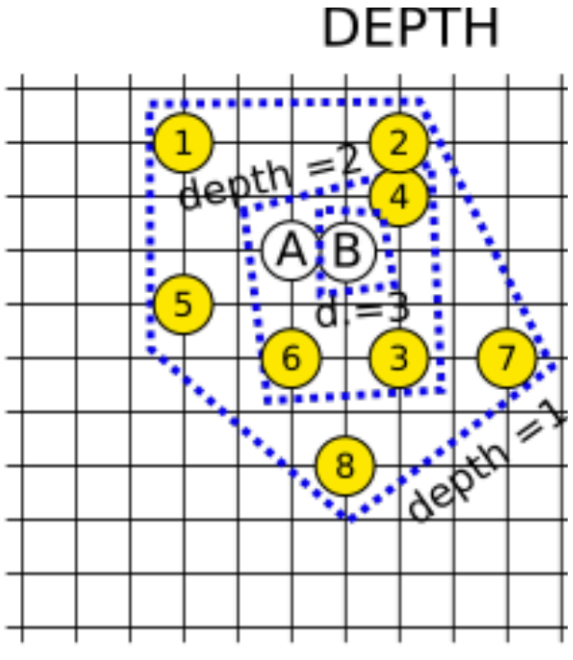
- LRD(A) = 1/ [(1 + 2)/2] = 0.666
- LRD(B) = 1/ [(1 + √2)/2] = 0.828
- LRD(6) = 1/ [(2 + 2)/2] = 0.500
- LOF(A) = ([LRD(B) + LRD(6)]/2) / LRD(A) = [(0.828 + 0.500) / 2] / 0.666 = 1.003
- LRD(4) = 1/ [(1 + √2)/2] = 0.828
- LOF(B) = ([LRD(A) + LRD(4)]/2) / LRD(B) = [(0.666 + 0.828) / 2] / 0.828 = 0.902
- Both are smaller or very close to 1, so they are most likely no outliers.



Outlier Detection – Exercise 1 – Solution

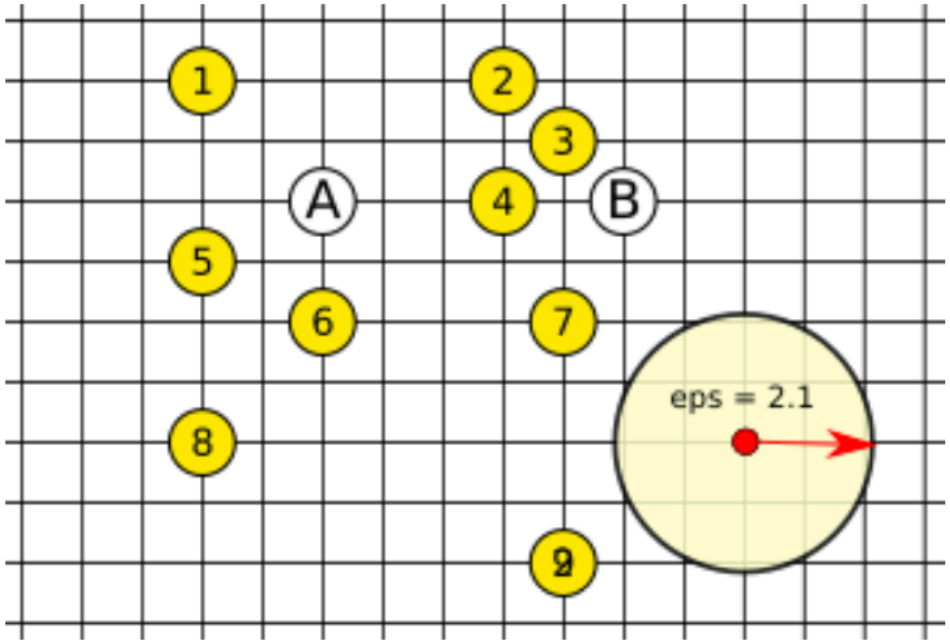
Depth-based

- A is an outlier for depth = 2
- For depth <= 1 neither A or B are outliers



Outlier Detection – Exercise 2

Given the dataset of 10 points below, consider the outlier detection problem for points A and B, adopting the following three methods:

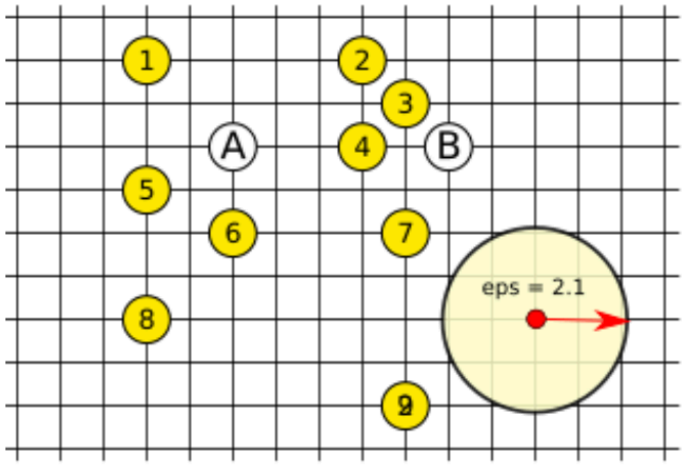


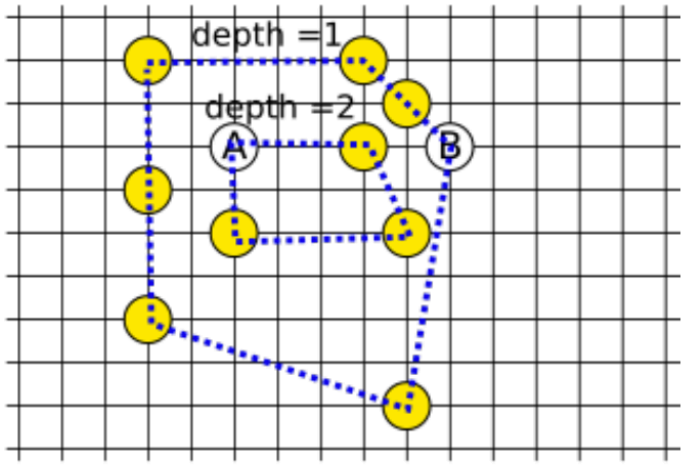
a) Distance-based: DB(ε,π) **(2 points)**
Are A and/or B outliers, if thresholds are forced to ε = 2.1 and π = 0.15 ? The point itself should not be counted.

b) Density-based: LOF **(2 points)**
Compute the LOF score for points A and B by taking k=2, i.e. comparing each point with its 2 NNs (not counting the point itself). In order to simplify the calculations, the reachability-distance used by LOF can be replaced by the simple Euclidean distance.

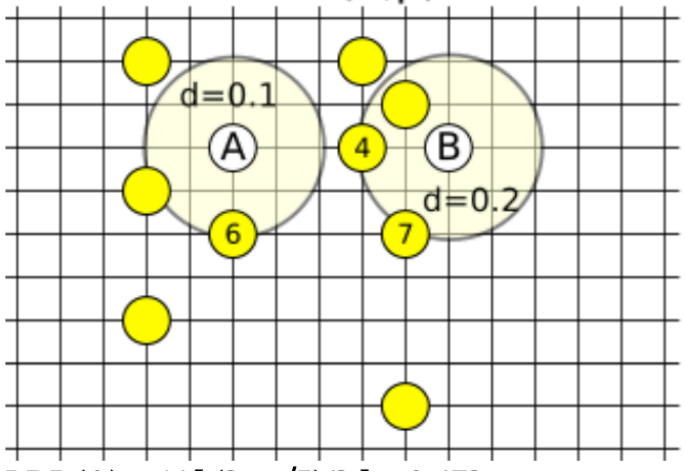
c) Depth-based **(2 points)**
Compute the depth score of all points. Are A and/or B outliers of depth 1?

Outlier Detection – Exercise 2 – Solution

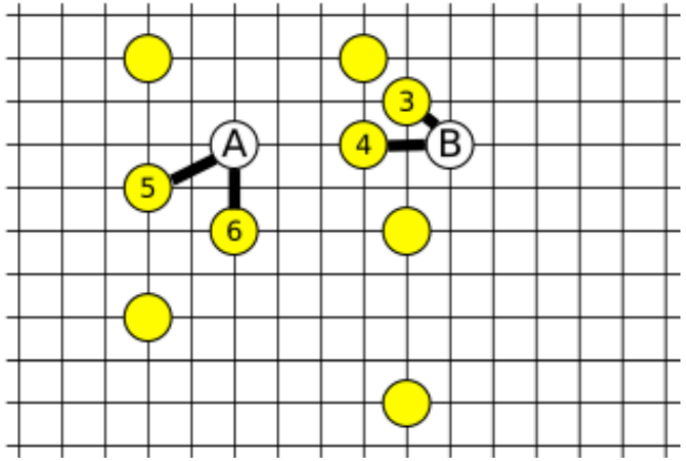




DB(e,p)



LOF



LRD(A) = 1/ [(2 + √5)/2] = 0.472

LRD(5) = 1/ [(√5 + √5)/2] = 0.447

LRD(6) = 1/ [(2 + √5)/2] = 0.472

LOF(A) = ([LRD(5) + LRD(6)]/2) / LRD(A)
= [(0.472 + 0.447) / 2] / 0.472 = 0.973

LRD(B) = 1/ [(2 + √2)/2] = 0.586

LRD(3) = 1/ [(√2 + √2 + √2)/3] = 0.707

LRD(4) = 1/ [(2 + 2 + √2)/3] = 0.554

LOF(B) = ([LRD(3) + LRD(4)]/2) / LRD(B)
= [(0.707 + 0.554) / 2] / 0.586 = 0.929

Outlier Detection – Exercise 3

Given the dataset of 10 points below (A, B, 1, 2, …, 8), consider the outlier detection problem for points A and B, adopting the following three methods:

a) Distance-based: DB(ε,π) **(2 points)**
Are A and/or B outliers, if thresholds are forced to ε = 2.5 and π = 0.3? Show the density of the two points. (Notice: in computing the density of a point P, P itself should not be counted as neighbour).

b) Density-based: LOF **(3 points)**
Compute the LOF score for points A and B by taking k=2, i.e. comparing each point with its 2-NNs (not counting the point itself). In order to simplify the calculations, the reachability-distance used by LOF can be replaced by the simple Euclidean distance.

c) Depth-based **(1 points)**
Compute the depth score of all points.