LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Database Systems Group • Prof. Dr. Thomas Seidl

Exercise 8: Clustering, Outlier Detection

Knowledge Discovery in Databases I

SS 2016

Exercise 8-1: DBSCAN

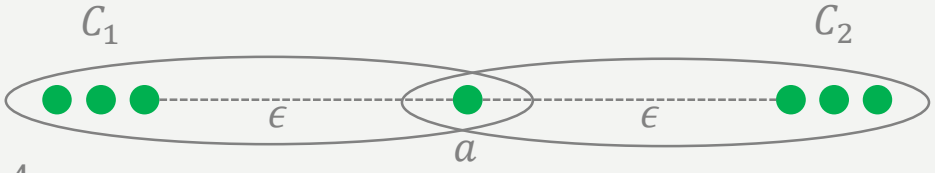
Let $\{C_1, C_2\}$ be a clustering result of DBSCAN. Prove or disprove:

a) *a* is a core object in both C_1 and $C_2 \Rightarrow C_1 = C_2$.

- Any object in a density-based cluster is density-reachable from any of its core objects.
- By the maximality condition, it must hold that $C_1 = C_2$.

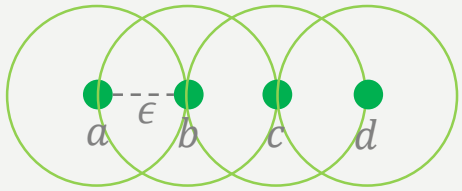
b) *a* is a border object in both both C_1 and $C_2 \Rightarrow |C_1 \cap C_2| > minPts$.

- Counterexample:



c) $a \in C_1$ is a core object $\Rightarrow$ $\bigcup_{x \in N_\epsilon(a)} N\epsilon(x) = C_1$.

- Counterexample:



$\{a, b, c\} \neq \{a, b, c, d\}$

$minPts = 2$

- Underlying reason: „density-reachable“ is the transitive closure of the „directly density-reachable“ relation. Here, we only consider two steps.

d) $a \in C_1$ is a border object $\Rightarrow$ there is at least one core object $b \in N_\epsilon(a)$.

- If there was no core object *b* in *a*'s ϵ-neighborhood, then *a* would not be density reachable from any other point in C_1.
- This would be a contradiction, since C_1 is a density-based cluster. In particular all points in C_1 are density-connected.

Exercise 8-1: DBSCAN

e) Let $C' = \{x \in C_1 \mid N_\epsilon(x) < minPts\}$ the set of all border objects in C_1 $\Rightarrow$ there exists a density-connected pair $(x, y) \in C' \times C'$.

- By definition of a density-based cluster, all pairs of objects in C_1 are density-connected. This includes all pairs of border objects.

f) C_1 has more border objects than core objects.

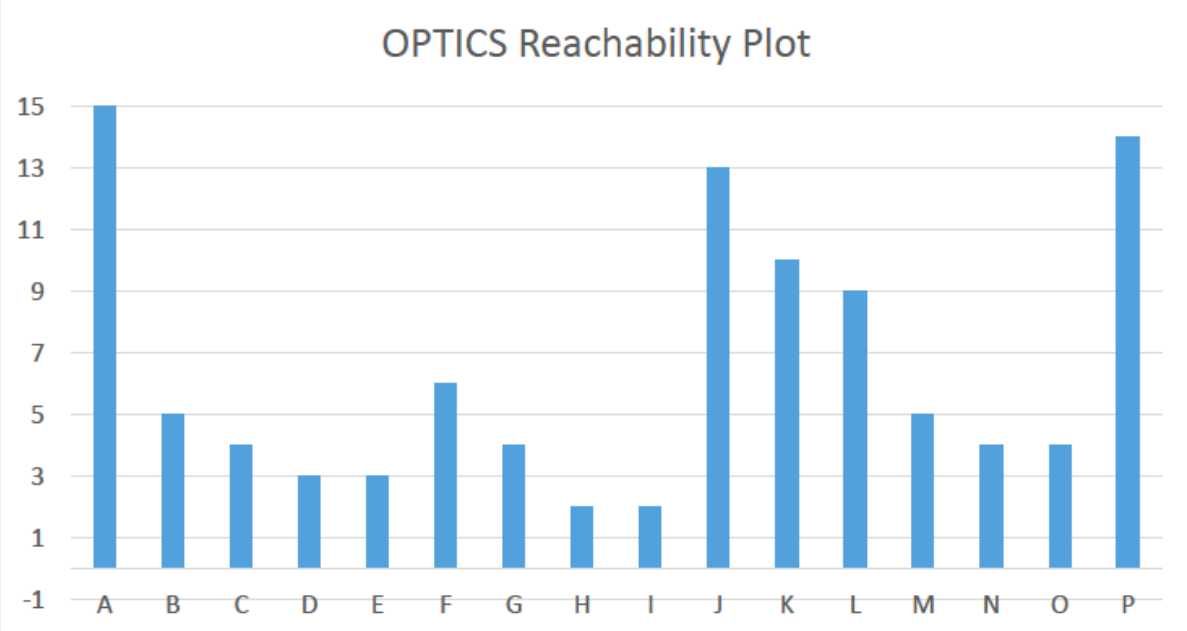
- This is not true in general: Just choose ϵ large enough, then all objects will become core objects. See also the counterexample for (c).

g) C_1 has more core objects than border objects.

- This is also not true in general: For instance in the counterexample in (b), C_1 consists of 1 core object and 3 border objects

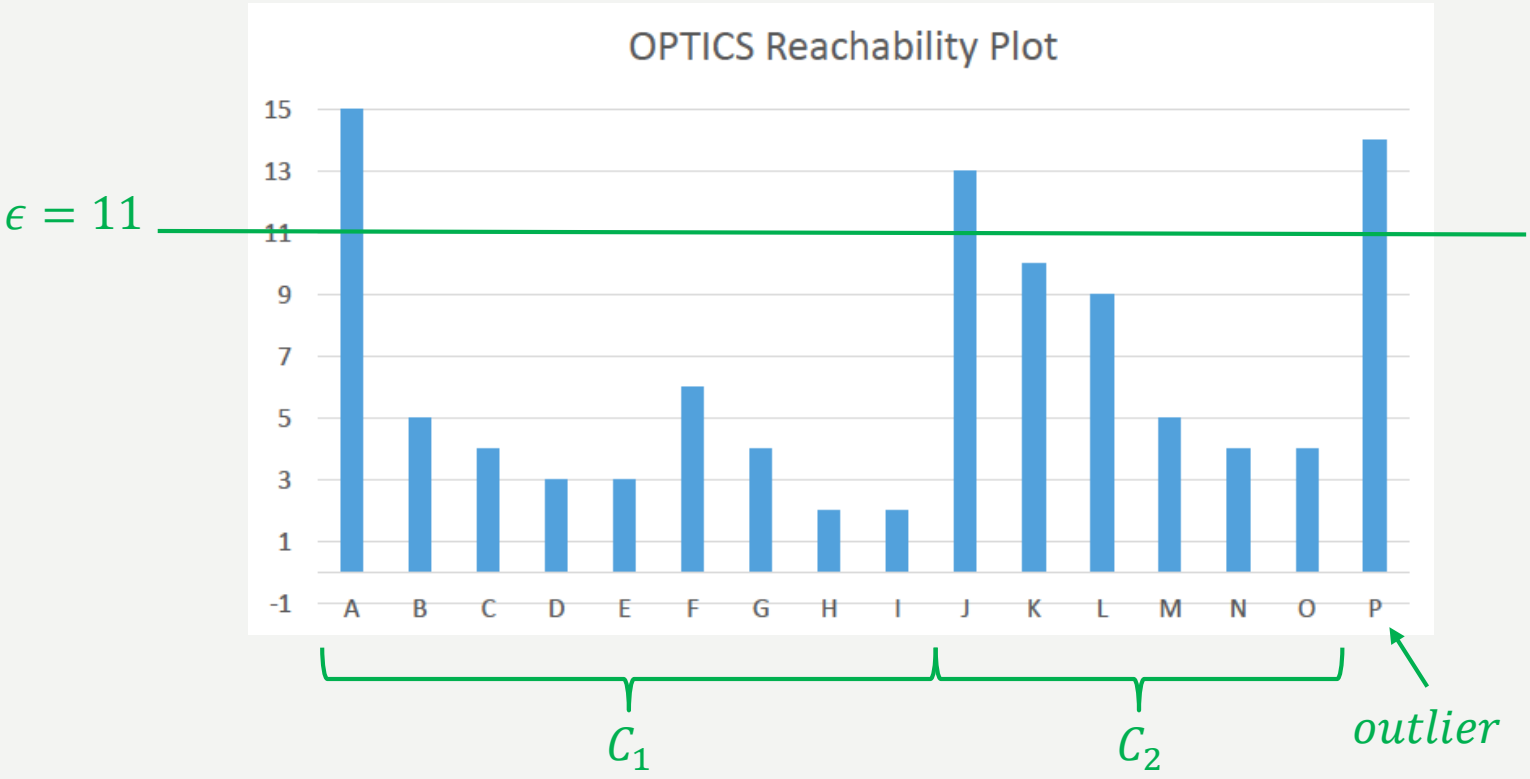
Exercise 8-2: OPTICS

Given the following OPTICS reachability plot as a result, answer the following questions:

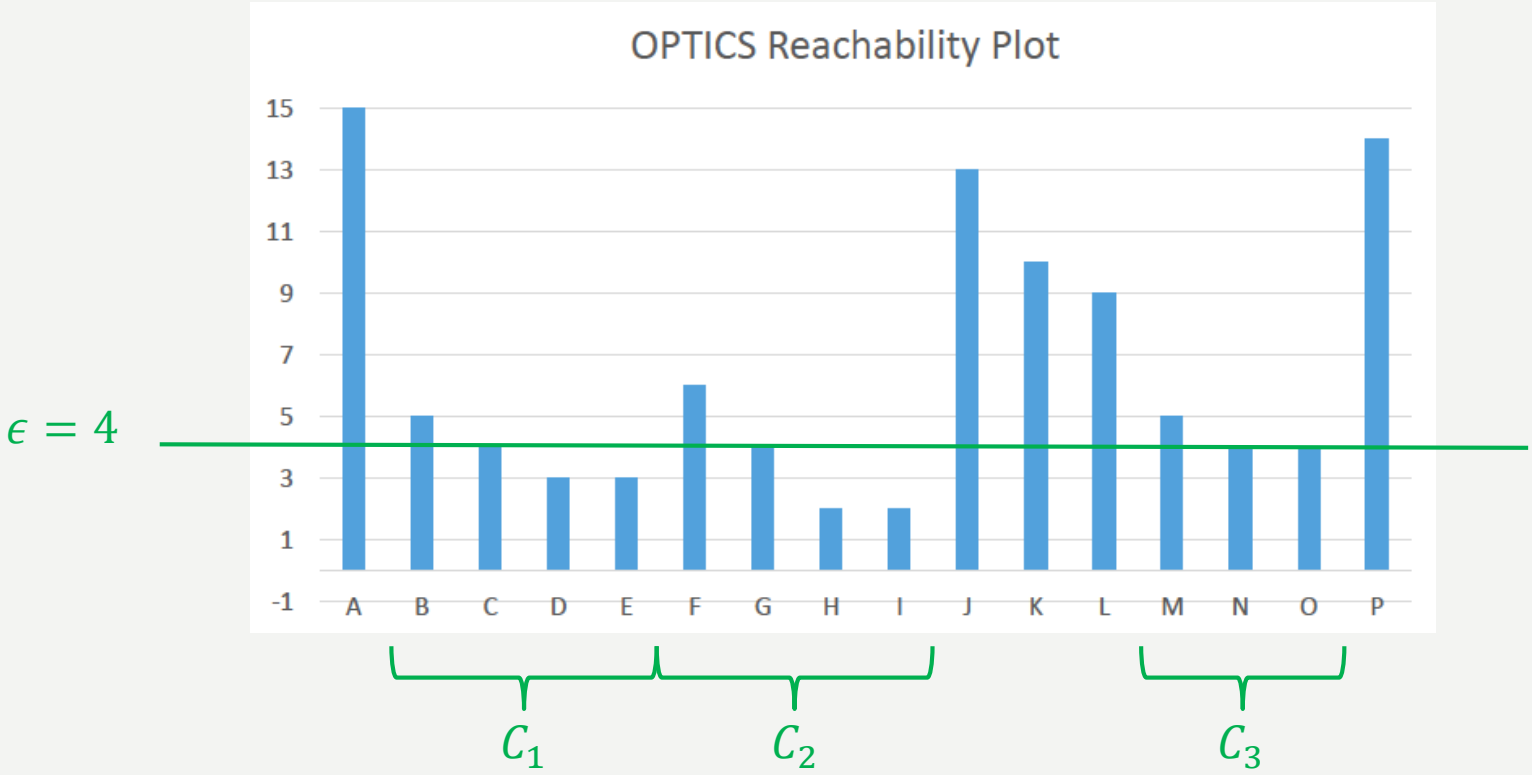


a) Which clustering can be obtained with $\epsilon = 11$?

b) Which objects are outliers?



c) What about $\epsilon = 4$?



$outliers: A, J, K, L, P$

Exercise 8-2: OPTICS

d) For which ϵ is P not an outlier?

- For $\epsilon \geq r - dist(P) = 14$

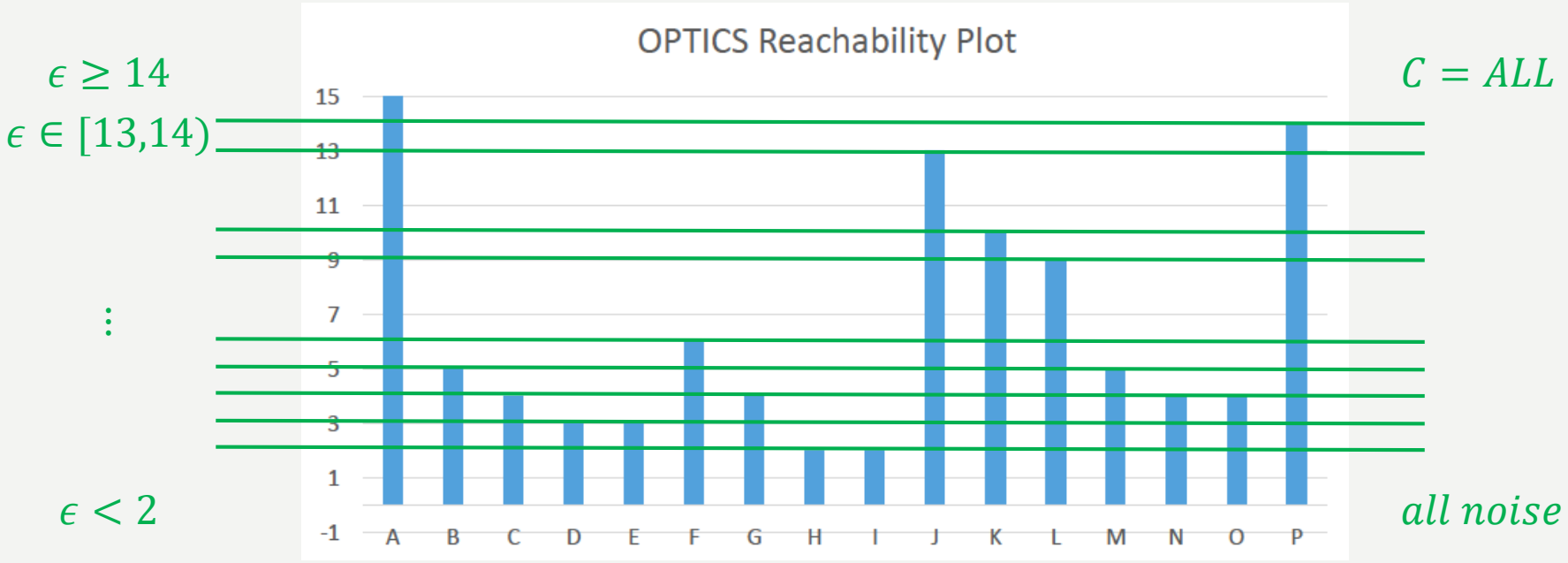
e) Which ϵ has to be used to form the cluster $\{G, H, I\}$?

- $\epsilon < r - dist(G) = 4$ (otherwise, *F* would be part of the cluster)
- $\epsilon \geq r - dist(H) = r - dist(I) = 2$ (otherwise, *H and I* would not be part of the cluster)

f) Is it possible to choose ϵ such that there is a one-element cluster?

- A one-element cluster is a noise object.
- We have seen noise objects for instance in (a) and (c).

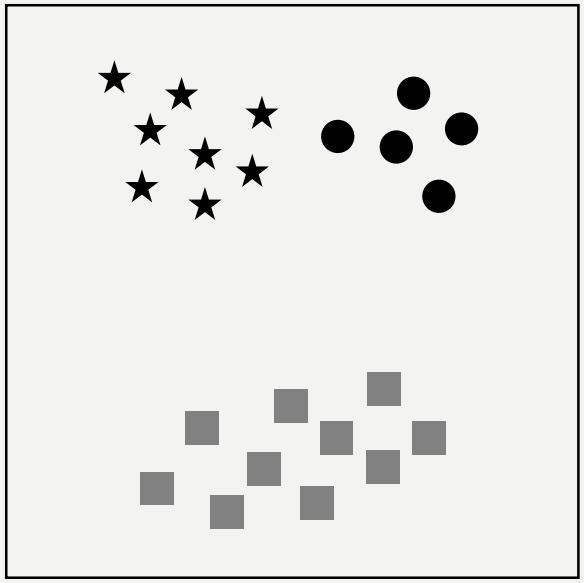
g) How many different cluster partitions can be found in this dataset?



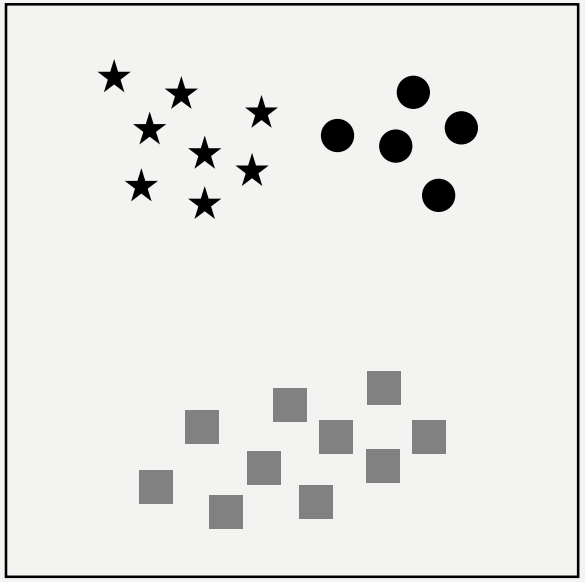
- In total, 10 different clusterings can be obtained.

Exercise 8-3: Evaluation of Clustering Results

Consider the following dataset, where the classes are represented by different colors and the clusters are indicated by the object shapes:



a) Construct the confusion matrix and compute the Rand Index, Precision, Recall and F1-Measure.



	Same cluster	Different clusters
Same class	$TP = \binom{10}{2} + \binom{8}{2} + \binom{5}{2} = 83$	$FN = 8 \cdot 5 = 40$
Different classes	$FP = 0$	$TN = 13 \cdot 10 = 130$

Exercise 8-3: Evaluation of Clustering Results

	Same cluster	Different clusters
Same class	$TP = \binom{10}{2} + \binom{8}{2} + \binom{5}{2} = 83$	$FN = 8 \cdot 5 = 40$
Different classes	$FP = 0$	$TN = 13 \cdot 10 = 130$

- Note: $TP + FN + FP + TN = \binom{23}{2} = 253$ — total number of pairs
- $RI = \frac{TP+TN}{TP+TN+FP+FN} = \frac{83+130}{253} = \frac{213}{253} \approx 84.19\%$ „accuracy“
- $Precision = \frac{TP}{TP+FP} = \frac{83}{83} = 100\%$ „fraction of pairs of the same cluster that are also in the same class“
- $Recall = \frac{TP}{TP+FN} = \frac{83}{83+40} = \frac{83}{123} \approx 67.48\%$ „fraction of pairs of the same class that also appear in the same cluster“
- $F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+recall} = \frac{166/123}{206/123} = \frac{166}{206} \approx 80.58\%$ harmonic mean

Exercise 8-3: Evaluation of Clustering Results

b) How would you evaluate the quality of the clustering result as an 'expert'? Does your assessment correspond with the external quality measures? In general, can you identify some potential problems or drawbacks of external cluster evaluation measures?

- Both indicated clusterings make sense.
- The clustering result can be viewed as a hierarchical clustering: The algorithm has detected two sub-concepts of the black class.
- Even though, the discovery is punished by the RI and Recall measures.
- In general:
  - External measures punish the discovery of novel structure in the data, in contradiction to the goal of unsupervised learning.
  - In some sense, the clustering is reduced to a retrieval task.
  - Also, ground truth labels are rarely available for data to be clustered: That's why we want to do clustering in the first place!

Exercise 8-4: Outlier Scores

There already exists a very nice solution to exercise 8-4 from the previous year. You can find the slides under the following link:

http://www.dbs.ifi.lmu.de/Lehre/KDD/SS15/uebung/Tutorial06.pdf