Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Max Berrendorf, Julian Busch

Knowledge Discovery and Data Mining I

WS 2018/19

Exercise 8: Outlier Scores

Exercise 8-1 Monotonicity of Simple Outlier Scores

Proof or give an counterexample for the following claims:

(a) If o is an $D(\epsilon, \pi)$-outlier, it is also an $D(\epsilon', \pi)$-outlier for $\epsilon' \leq \epsilon$.

The statement is true. Let o be an $D(\epsilon, \pi)$-outlier. Then,

$$\pi\,|D| \overset{\clubsuit}{\geq} |\{q \in D \mid dist(o,q) < \epsilon\}| \overset{\heartsuit}{\geq} \left|\{q \in D \mid dist(o,q) < \epsilon'\}\right|$$

where $\clubsuit$ is the definition of $D(\epsilon, \pi)$-outlier, and $\heartsuit$ holds due to the transitivity of $<$ and $\leq$:

$$dist(o,q) < \epsilon' \wedge \epsilon' \leq \epsilon \implies dist(o,q) < \epsilon$$

(b) If o is an $D(\epsilon, \pi)$-outlier, it is also an $D(\epsilon, \pi')$-outlier for $\pi' \geq \pi$.

This statement is also true.

$$\pi'|D| \geq \pi|D| \geq |\{q \in D \mid dist(o,q) < \epsilon\}|$$

(c) If o is an kNN-outlier for threshold τ, it is also an k'NN-outlier for the same threshold with $k' > k$.

Let $nndist(o,k)$ denote the k-distance of o. As the k-distance is the kth smallest distance to an object in the database, we clearly have $nndist(o,k) \leq nndist(o,k+1)$ (the $(k+1)$-smallest distance cannot be larger than the k-smallest). Hence,

$$nndist(o,k') \geq nndist(o,k) > \tau,$$

i.e. o is also a k'NN outlier for threshold τ.

(d) If o is an kNN-outlier for threshold τ, it is also an kNN-outlier for threshold $\tau' < \tau$.

Let $nndist(o,k)$ denote the k-distance of o. Then,

$$nndist(o,k) > \tau > \tau'$$

i.e. o is also a kNN outlier for threshold τ'.

(e) The local density is monotonously decreasing in k, i.e. $ld_k(o) \geq ld_{k'}(o)$ for $k' > k$.

This statement is true. Let $nndist(o, k)$ denote the k-distance of o, i.e. the distance between o and its kth nearest neighbor. Then, we have

$$k' \geq k \implies nndist(o, k') \geq nndist(o, k)$$

i.e. the k-distance is monotonously increasing in k. With this notation, we can note the (*reciprocal*) local density $ld_k(o)$ by

$$(ld_k(o))^{-1} = \frac{1}{k}\sum_{i=1}^{k} nndist(o, i)$$

Moreover, we can apply the following sequence of equivalence transformations of the inequality of interest

$$\begin{aligned}
ld_k(o) &\geq ld_{k+1}(o) \\
\iff (ld_k(o))^{-1} &\leq (ld_{k+1}(o))^{-1} \\
\iff \frac{1}{k}\sum_{i=1}^{k} nndist(o, i) &\leq \frac{1}{k+1}\sum_{i=1}^{k+1} nndist(o, i) \\
\iff (k+1)\sum_{i=1}^{k} nndist(o, i) &\leq k\sum_{i=1}^{k+1} nndist(o, i) \\
\iff k\sum_{i=1}^{k} nndist(o, i) + \sum_{i=1}^{k} nndist(o, i) &\leq k\sum_{i=1}^{k+1} nndist(o, i) \\
\iff \sum_{i=1}^{k} nndist(o, i) &\leq k \cdot nndist(o, k+1)
\end{aligned}$$

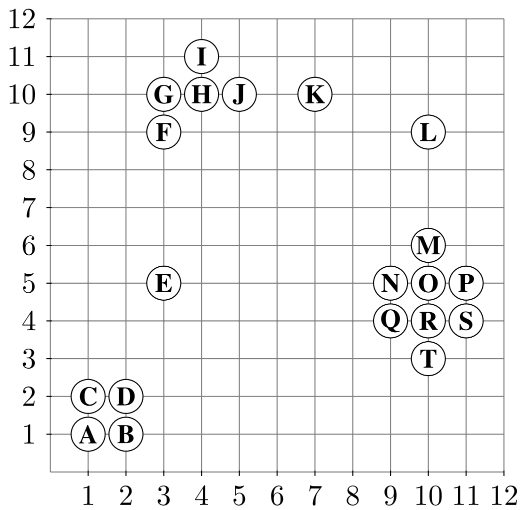
The last inequality holds due to

$$\sum_{i=1}^{k} nndist(o, i) \overset{\spadesuit}{\leq} \sum_{i=1}^{k} nndist(o, k+1) = k \cdot nndist(o, k+1)$$

where ♠ uses the monotonicity of the k-distance.

Exercise 8-2 Outlier Scores

Given the following 2 dimensional data set:



As distance function, use Manhattan distance $L_1(a,b) := |a_1 - b_1| + |a_2 - b_2|$. The following table summarises the pairwise distances.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	0	1	1	2	6	10	11	12	13	13	15	17	14	12	13	14	11	12	13	11
B	1	0	2	1	5	9	10	11	12	12	14	16	13	11	12	13	10	11	12	10
C	1	2	0	1	5	9	10	11	12	12	14	16	13	11	12	13	10	11	12	10
D	2	1	1	0	4	8	9	10	11	11	13	15	12	10	11	12	9	10	11	9
E	6	5	5	4	0	4	5	6	7	7	9	11	8	6	7	8	7	8	9	9
F	10	9	9	8	4	0	1	2	3	3	5	7	10	10	11	12	11	12	13	13
G	11	10	10	9	5	1	0	1	2	2	4	8	11	11	12	13	12	13	14	14
H	12	11	11	10	6	2	1	0	1	1	3	7	10	10	11	12	11	12	13	13
I	13	12	12	11	7	3	2	1	0	2	4	8	11	11	12	13	12	13	14	14
J	13	12	12	11	7	3	2	1	2	0	2	6	9	9	10	11	10	11	12	12
K	15	14	14	13	9	5	4	3	4	2	0	4	7	7	8	9	8	9	10	10
L	17	16	16	15	11	7	8	7	8	6	4	0	3	5	4	5	6	5	6	6
M	14	13	13	12	8	10	11	10	11	9	7	3	0	2	1	2	3	2	3	3
N	12	11	11	10	6	10	11	10	11	9	7	5	2	0	1	2	1	2	3	3
O	13	12	12	11	7	11	12	11	12	10	8	4	1	1	0	1	2	1	2	2
P	14	13	13	12	8	12	13	12	13	11	9	5	2	2	1	0	3	2	1	3
Q	11	10	10	9	7	11	12	11	12	10	8	6	3	1	2	3	0	1	2	2
R	12	11	11	10	8	12	13	12	13	11	9	5	2	2	1	2	1	0	1	1
S	13	12	12	11	9	13	14	13	14	12	10	6	3	3	2	1	2	1	0	2
T	11	10	10	9	9	13	14	13	14	12	10	6	3	3	2	3	2	1	2	0

(a) Calculate the $D(\epsilon, \pi)$-outliers using

(i) $\epsilon = 2$ with $n\pi = 1$ and $n\pi = 2$.

(ii) $\epsilon = 4$ with $n\pi = 1$, $n\pi = 3$ and $n\pi = 4$.

(iii) $\epsilon = 6$ with $n\pi = 4$, $n\pi = 5$ and $n\pi = 6$.

For the $D(\epsilon, \pi)$ outliers we have to check whether at most π percent of all points have a distance less than ϵ. Hence, we count per column how many times the distance is less than ϵ yielding

ϵ	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**
2	3	3	3	3	1	2	3	4	2	2	1	1	2	3	5	3	3	5	3	2
4	4	4	4	4	1	5	5	6	5	6	3	2	9	8	8	8	8	8	8	8
6	4	5	5	5	6	7	7	6	6	6	7	7	9	9	9	9	8	9	8	8

Finally, we check if the number divided by the number of objects $n = 20$ is at most the threshold π. We obtain the following outliers:

(i) For $(\epsilon, n\pi) = (2, 1)$: EKL. For $(\epsilon, n\pi) = (2, 2)$: EFIJKLMT.

(ii) For $(\epsilon, n\pi) = (4, 1)$: E. For $(\epsilon, n\pi) = (4, 3)$: EKL. For $(\epsilon, n\pi) = (4, 4)$: ABCDEKL.

(iii) For $(\epsilon, n\pi) = (6, 4)$: A. For $(\epsilon, n\pi) = (6, 5)$: ABCD. For $(\epsilon, n\pi) = (6, 6)$: ABCDEHIJ.

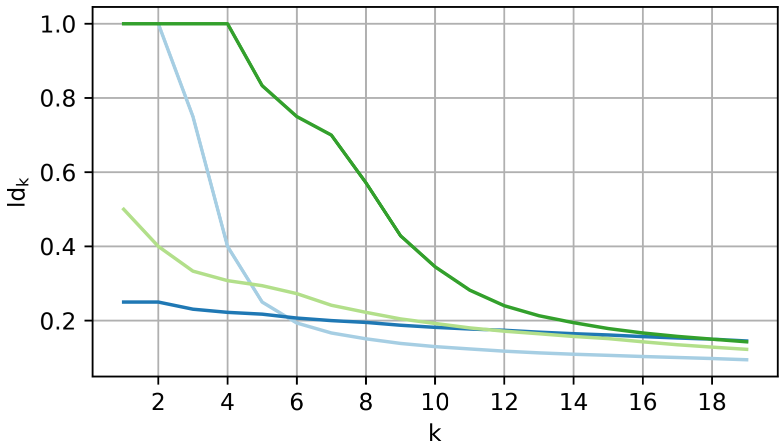
(b) Calculate the kNN based outliers for $(k, \tau) = (3, 3)$ and $(k, \tau) = (5, 8)$. The point itself is counted as the 0-nearest neighbour.

First, we compute the k-distances for each point.

k	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**
3	2	2	2	2	5	3	2	1	2	2	4	4	2	2	1	2	2	1	2	2
5	10	9	9	8	5	4	4	3	4	3	4	5	3	2	2	2	2	2	2	3

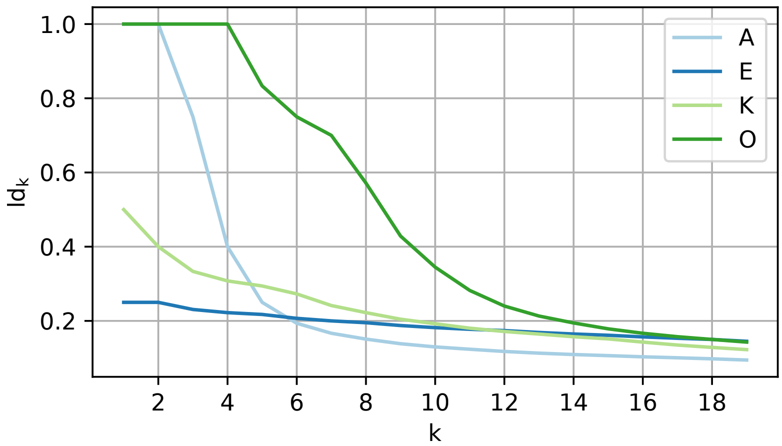
Finally, we obtain the outliers as those points whose k-distance exceeds the threshold τ, i.e. for $(k, \tau) = (3, 3)$ we have E, K, L, and for $(k, \tau) = (5, 8)$ we have A, B, C.

(c) Given the following curves of the local density ld_k for different values of k.



Can you identify which curve belongs to which point? Explain your mapping.

This is the ground truth mapping.



We can observe:

- The dark green line has a ld_k of one up to $k = 4$. Hence, the inverse average distance to the 4-nearest neighbours is 1, and equivalently, the average of distances of the 4-nearest neighbours is one. We can only find two points in the dataset fulfilling this requirement: O and R.
- The dark blue line has a ld_1 of 0.25, i.e. the 1-nearest neighbour has distance 4. This requirement is only fulfilled by E.
- For the light blue line we can observe that ld_k stays one until $k = 2$, i.e. there are two points with distance 1. This reduces the candidate set to ABCDG. As we observe a sharp drop afterwards, the point is likely to reside in ABCD. All of these points have a quite similar ld_k-line.

- The light green line is also in a region that has a low local density already for small k values. As it is still higher, as the light green line, we might suspect a point that has a slightly smaller 1-distance, such as K, or L. Using $ld_1 = 0.5$, we can conclude that the 1-distance is equal to 2, and hence only K possible.