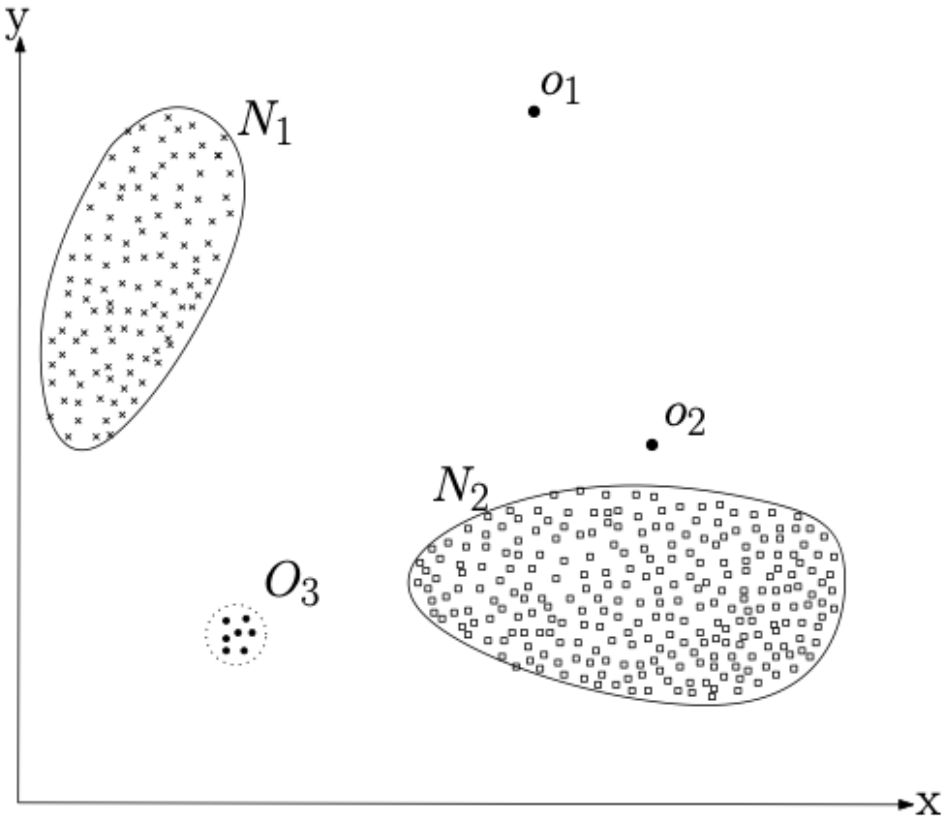
Anomaly Detection

Exercise 5

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Anomaly Detection
Definition of Outliers



Univariate Anomaly Detection

Interquartile Range

Definitions:

- Q1: x ≥ Q1 holds for 75% of all x = median of the lower half
- Q3: x ≥ Q3 holds for 25% of all x = median of the upper half
- IQR = Q3-Q1

Outlier detection:

- All values outside [median-1.5*IQR ; median+1.5*IQR]

TASK

Find outliers in **[3, 5, 6, 6, 8, 11, 21]** with IQR

Q1 = median

Q2

Q3 = median

Outlier

IQR: 11 – 5 = 6

Allowed interval: [6 – 1.5*6 ; 6 + 1.5*6] = [-3 ; 15]

Univariate Anomaly Detection
Interquartile Range

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

TASK

- Find outliers in **[-5, 3, 7, 11]** with IQR

- Find outliers in **[1, 4, 9]** with IQR

- Find outliers in **[-14, -12, 7, 10, 11, 12, 14, 16.5, 17, 38]** with IQR

Univariate Anomaly Detection

Interquartile Range

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

TASK

- Find outliers in **[-5, 3, 7, 11]** with IQR

Q1 = -1, Q2 = 5, Q3 = 9, IQR = 10, Interval = [-10 ; 20]

- Find outliers in **[1, 4, 9]** with IQR

Q1 = 1, Q2 = 4, Q3 = 9, IQR = 8, Interval = [-8 ; 16]

- Find outliers in **[-14, -12, 7, 10, 11, 12, 14, 16.5, 17, 33]** with IQR

Q1 = 7, Q2 = 11.5, Q3 = 16.5, IQR = 9.5, Interval = [-2.75 ; 25.75]

Univariate Anomaly Detection

Median Absolute Deviation (MAD)

$$MAD := median_i(X_i - median_j(X_j))$$

- all values that are k*MAD away from the median are considered to be outliers
- e.g., k=3

TASK

Find outliers in **[3, 5, 6, 6, 8, 11, 21]** with MAD

Outlier

median = 6

deviations = [3, 1, 0, 0, 2, 5, 15]

MAD = 2

interval = [6-3*2 ; 6+3*2] = [0 ; 12]

Univariate Anomaly Detection
Median Absolute Deviation (MAD)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

TASK **k = 3**

- Find outliers in **[-5, 3, 7, 11]** with MAD

- Find outliers in **[1, 4, 9]** with MAD

- Find outliers in **[-14, -12, 7, 10, 11, 12, 14, 16.5, 17, 38]** with MAD

Univariate Anomaly Detection
Median Absolute Deviation (MAD)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

TASK **k = 3**

- Find outliers in **[-5, 3, 7, 11]** with MAD

median = 5, deviations = [10, 2, 2, 6]

MAD = 4, interval = [-7 ; 17]

- Find outliers in **[1, 4, 9]** with MAD

median = 4, deviations = [3, 0, 5]

MAD = 3, interval = [-5 ; 13]

- Find outliers in **[-14, -12, 7, 10, 11, 12, 14, 16.5, 17, 38]** with MAD

median = 11.5, deviations = [25.5, 23.5, 4.5, 1.5, 0.5, 0.5, 2.5, 5, 5.5, 26.5]

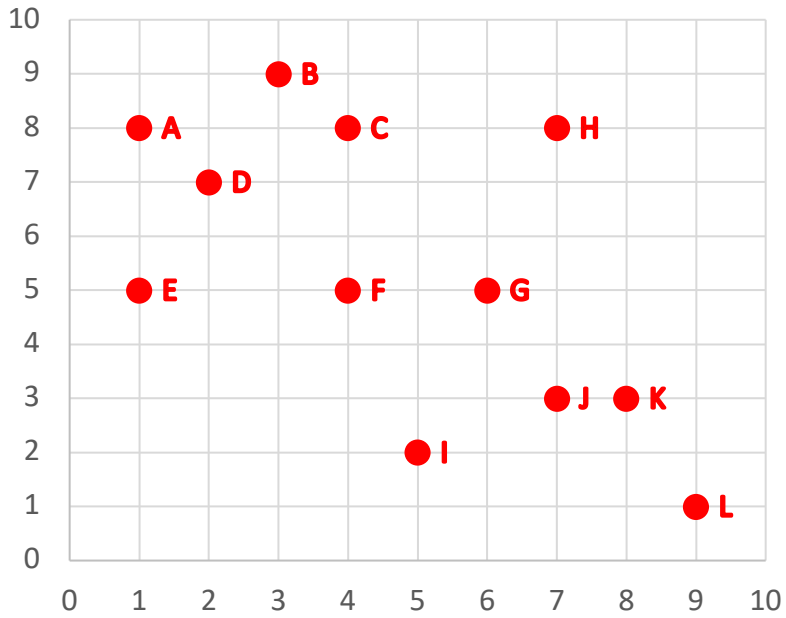
MAD = 4.75, interval = [-2.75 ; 25.75]

Multivariate Anomaly Detection
k-NN and Local Outlier Factor

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

TASK

1) Look up workings of k-NN and LOF*
2) Identify the top two outliers using k-NN approach with k=3. Use either the maximum or average distance
3) Compute the LOF outlier score for the two outliers identified in step 2 (with k=3). Which one is greater?



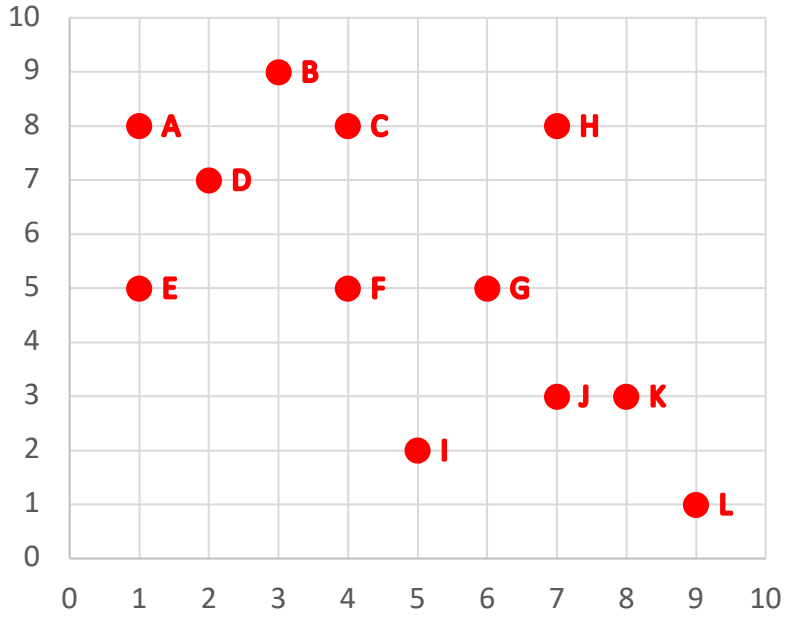
Hint: For convenience, use Manhattan distance as distance metric!

You can use “Simplified LOF” where the density of a point is 1/k-distance.

Multivariate Anomaly Detection
k-NN and Local Outlier Factor

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

k-NN	NN1	NN2	NN3	MAX	AVG
A	2	3	3	3	2.7
B	2	3	3	3	2.7
C	2	3	3	3	2.7
D	2	3	3	3	2.7
E	3	3	3	3	3.0
F	2	3	4	4	3.0
G	2	3	4	4	3.0
H	3	4	5	**5**	**4.0**
I	3	4	4	4	3.7
J	1	3	3	3	2.3
K	1	3	4	4	2.7
L	3	4	5	**5**	**4.0**



Multivariate Anomaly Detection
k-NN and Local Outlier Factor

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

LOF

Point H:

density = 1/5

avg. density of neighbors = avg. density(C,G,J) = avg(1/3, 1/4, 1/3) = 11/36

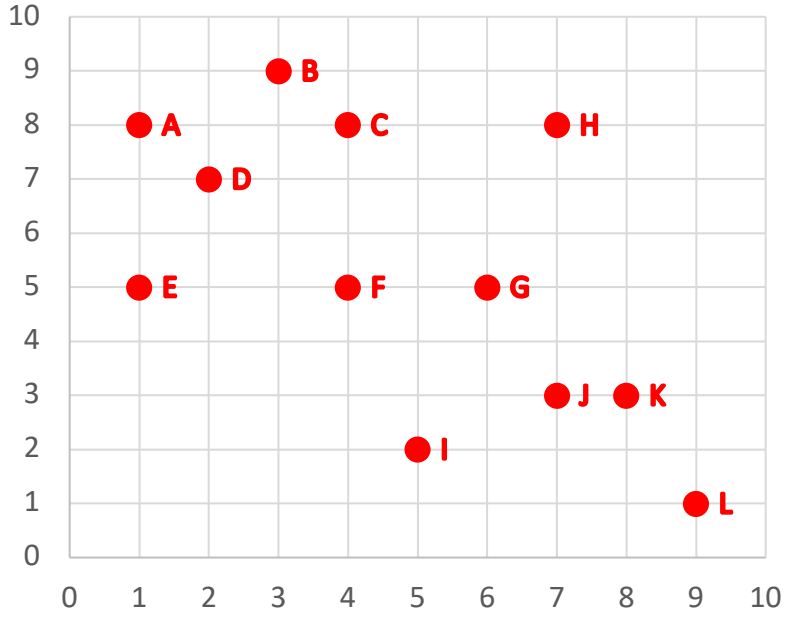
LOF = (11/36) / (1/5) = 1.53

Point L:

density = 1/5

avg. density of neighbors = avg. density(I,J,K) = avg(1/4, 1/3, 1/4) = 5/18

LOF = (5/18) / (1/5) = 1.39



=> H has a higher LOF than L.

Isolation Forests

Task

Using Isolation Forests, you want to find outliers in the dataset on the right.

A	(0, 10, 10, 10)
B	(5, 6, 6, 7)
C	(5, 4, 2, 6)
D	(10, 5, 0, 0)
E	(8, 1, 6, 9)
F	(5, 0, 3, 7)

TASK

Compute the outlier score (i.e., the probability of the data point ending in a leaf of height 1) for every point in the dataset.

Isolation Forests
Solution

A (0, 10, 10, 10)

~~B (5, 6, 6, 7)~~

~~C (5, 4, 2, 6)~~

D (10, 5, 0, 0)

~~E (8, 1, 6, 9)~~

F (5, 0, 3, 7)

B, **C**, and **E** do not have any min/max values => they can not be isolated in a leaf with one node.

A) P(Att1, V<5) + P(Att2, V>6) +
P(Att3, V>6) + P(Att4, V>9)
= 0.25 * 0.5 + 0.25 * 0.4 + 0.25 * 0.4 + 0.25 * 0.1
= 0.35

D) P(Att1, V>8) + P(Att3, V<2) + P(Att4, V<6)
= 0.25 * 0.2 + 0.25 * 0.2 + 0.25 * 0.6
= **0.25**

F) P(Att2, V<1) = 0.25 * 0.1 = **0.025**