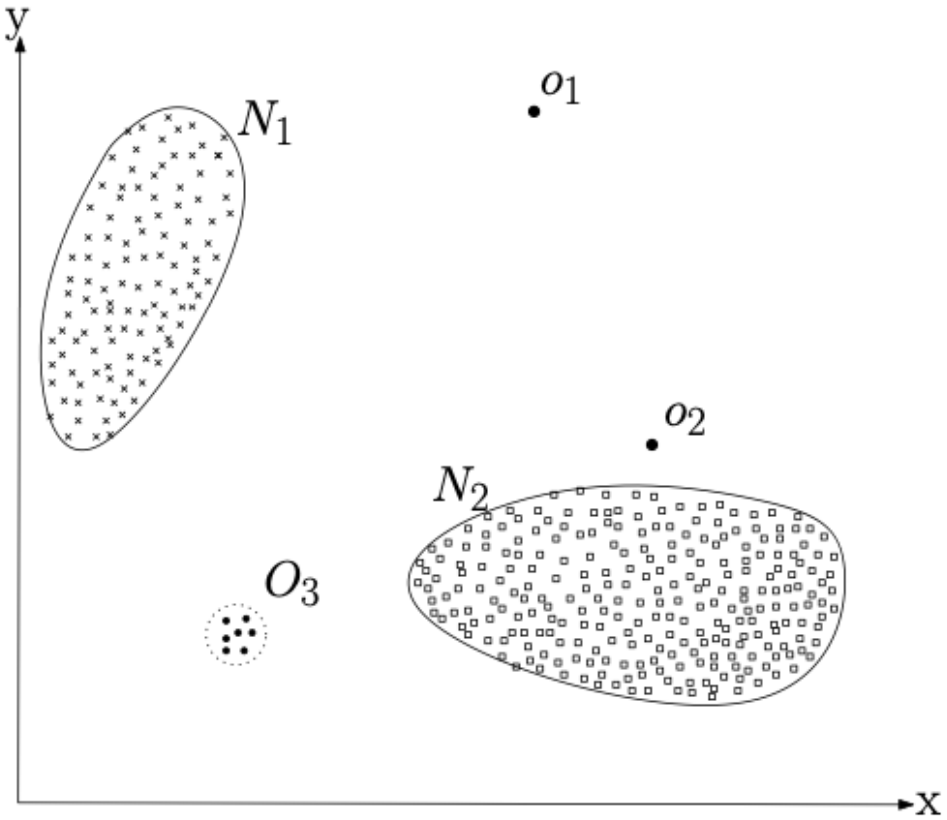
Anomaly Detection

Exercise 5

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Anomaly Detection
Definition of Outliers



Univariate Anomaly Detection

Interquartile Range

Definitions:

- Q1: $x \geq Q1$ holds for 75% of all x
- Q3: $x \geq Q3$ holds for 25% of all x
- 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

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

Median Absolute Deviation (MAD)

$$MAD := median_i\left(X_i - median_j\left(X_j\right)\right)$$

- 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

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

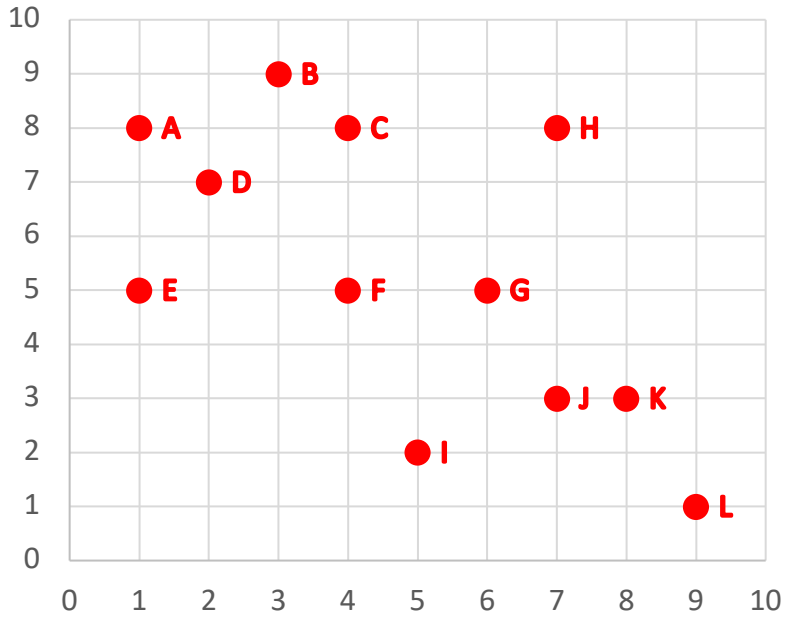
Multivariate Anomaly Detection
k-NN and Local Outlier Factor

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!



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.