Chapter 7
Anomaly Detection

Er. Shiva Ram Dam

Assistant Professor

Gandaki University

Content:

1. Introduction to Anomalies
2. Issues of Anomaly Detection
3. Approaches to Anomaly Detection
4. Statistical Approaches: Normal distribution and Box plot

7.1 Introduction to Anomalies

Anomalies

- Anomalies, also known as **outliers**, are data points that deviate significantly from the expected pattern or norm within a dataset.
- They can be considered as the **"odd ones out"** in a collection of data
- An **anomaly** in data warehousing and data mining **refers to data points, patterns, or behaviors that deviate significantly** from the expected or normal behavior within a dataset.
- Anomalies can arise due to errors, rare events, or significant deviations that may indicate important insights or problems.

- An **outlier** is a data point that lies significantly outside the range of the other observations in the dataset.
- Outliers can be much higher or lower than most other data points and may result from various causes, such as data entry errors, variability in the data, or novel occurrences.
- Identifying and handling outliers is an important part of data preprocessing, as outliers can distort statistical analyses and machine learning models if not properly dealt with.
- Example: In a dataset of **salaries** of employees in a company: : [35, 40, 50, 50, 60, 55, 7000],
- The value 7000 would be an outlier as it deviates significantly from the rest of the values. It could be a data entry error or a exceptional salary.

Applications

- **Fraud Detection:** detecting fraudulent usage of credit cards and loan applications
- **Fault Diagnosis:** monitoring engineering processes to detect faults in equipment or finished products in the production line
- **Time Series Analysis**: identifying variations in the parameter trends w.r.t time. For example stock performance, sales data
- **Medical conditions:** identifying medical conditions based on the study of molecular structures, MRI images, etc.,

Types of outliers:

- Univariate outliers
- Multivariate outliers
- Global Outliers
- Contextual Outliers

Types	Description	Example
a) Univariate Outliers	• These outliers appear in a single variable or feature of the dataset. • They are identified by observing the spread or distribution of data within one dimension.	◦ A dataset of student test scores: [50, 55, 52, 48, 51, 1000] ◦ The value 1000 is an outlier because it is significantly higher than all other test scores. It may be a data entry error or an exceptional case.
b) Multivariate Outliers	• These outliers arise from the combination of multiple variables. • They occur when the pattern or relationship between variables deviates from the norm	◦ In a dataset of height and weight: Height (in cm): [160, 170, 180, 150, 300] Weight (in kg): [55, 65, 70, 60, 50] ◦ The person with height 300 cm and weight 50 kg is an outlier in the multivariate context, as this combination of height and weight is unusual compared to others.

Types	Description	Example
c) Global Outliers (Point Outliers)	• These outliers are data points that are significantly different from the rest of the data, regardless of context or relationships with other features.	◦ Example: In a dataset of test scores, a score of 0 out of 100 might be an extreme outlier when the rest of the scores are above 50.
d) Contextual Outliers (Conditional Outliers)	• These outliers occur when a value is considered extreme in a specific context but may be normal in others. • This is often context-dependent and requires domain knowledge to recognize.	◦ In a weather dataset: Temperature in December: 18°C, 17°C, 16°C, 20°C, 5°C Temperature in June: 35°C, 37°C, 36°C, 40°C, 30°C ◦ A temperature of 40°C in June may be normal, but in December, it would be an outlier

7.2 Issues of Anomaly Detection

Issues of Anomaly Detection

- Anomaly detection is essential for identifying fraud, security threats, and system failures. However, it comes with several challenges:
  1. **Wrong Alerts (False Positives & False Negatives)** – Sometimes normal data is marked as a problem, or real problems are missed.
  2. **Changing Normal Behavior** – What is normal today may not be normal tomorrow.
  3. **Rare Anomalies** – Unusual events happen rarely, making them hard to detect.
  4. **Slow Detection** – Finding problems in large, fast-changing data can take too long.
  5. **Expensive Processing** – Detecting anomalies in big data needs a lot of computing power.
  6. **Different Solutions for Different Problems** – A method that works for one industry may not work for another.
  7. **Data Quality:** Our input datasets may have several problems – incomplete entries, inconsistent formats, duplicates, different benchmarks for measurement, human error.

7.3 Approaches to Anomaly Detection

Approaches to Anomaly detection

- There are several ways to detect anomalies, depending on the type of data and the problem being solved.
- Below are the main approaches:
  1. Statistical methods
  2. Machine Learning-Based methods
  3. Proximity-based methods
  4. Density-based methods
  5. Deep Learning-based methods
  6. Rule-based methods

Approach	Idea	Best For	Limitation	Examples
1.Statistical Methods	Assumes that normal data follows a known statistical distribution, and anomalies deviate from it.	Small datasets with well-defined distributions.	Doesn't work well when data patterns change over time.	• Z-Score (Standard Deviation) • Grubbs' Test • Gaussian Mixture Models (GMM)
2. Machine Learning Based Methods	Uses algorithms to learn patterns from normal data and identify anomalies.	Large, complex datasets.	Requires good training data and high computation power.	• Supervised Learning : Decision Trees, SVM, NN • Unsupervised Learning: K-means clustering, DBSCAN clustering,
3. Proximity-Based methods	Anomalies are far from normal data points.	Data with clear patterns.	Doesn't work well for high-dimensional data.	• K-NN, Local Outlier Factor (LOF)
4. Density-Based methods	Finds anomalies by identifying regions with low data density.	Data with irregular distributions.	Struggles with varying data densities.	• DBSCAN, GMM

1. Statistical Methods

- Parametric approaches use Gaussian distribution-based methods
- Non-parametric approaches employ histogram-based methods
- Time series anomaly detection utilizes ARIMA models for sequential data
- Examples include
  - Z-score method for identifying outliers in normally distributed data
  - Interquartile Range (IQR) for detecting outliers in skewed distributions

2. Machine Learning Algorithms

- Supervised techniques require labeled training data
  - Support Vector Machines (SVM) separate normal and anomalous data points
  - Random Forests combine multiple decision trees for classification
- Unsupervised methods operate without labeled data
  - Clustering-based approaches (K-means, DBSCAN) group similar data points
  - Dimensionality reduction techniques (PCA, autoencoders) identify anomalies in lower-dimensional spaces
- Semi-supervised algorithms use a combination of labeled and unlabeled data
- Ensemble methods combine multiple models
  - Isolation Forest isolates anomalies using random partitioning
  - Random Cut Forest builds an ensemble of trees for anomaly detection

3. Advanced Techniques

- Deep learning approaches leverage neural networks for complex pattern recognition
  - Long Short-Term Memory (LSTM) networks analyze sequential data for anomalies
  - Autoencoders learn compact representations to detect deviations
- Hybrid methods combine statistical and machine learning techniques
- Real-time anomaly detection systems process streaming data
  - Sliding window approaches analyze recent data points
  - Adaptive algorithms update models as new data arrives

7.4 Statistical Approaches: Normal distribution and Box plot

1. Statistical approach

- The statistical approach assumes that **normal data follows a predictable pattern or distribution** (e.g., normal distribution), and anomalies are data points that **significantly deviate** from this pattern.
- This method helps identify **outliers** by measuring how much a data point differs from the majority of data.

Z-score for Anomaly detection

- **Idea:** Measures how many standard deviations a data point is from the mean.
- **Formula:**

$$Z = \frac{(X - \mu)}{\sigma}$$

- Where:
  - X = Data point
  - μ = Mean of dataset
  - σ= Standard deviation
- **If |Z| > 3, the data point is considered an anomaly** (assuming a normal distribution).
- **Example:**
- A bank finds that average daily transactions are **$500**, with a standard deviation of **$50**.
- A transaction of **$700** gives a Z-score of:

$$Z = \frac{(700 - 500)}{50} = 4$$

- Since **Z = 4 > 3**, this is flagged as an anomaly.

Example with Z-score method:

- Consider a dataset: [50, 55, 52, 48, 51, 60, 62, 54, 53, 47, 65, 90]
- Step 1: Calculation of mean

$$\mu = \frac{\text{Sum of all values}}{\text{Number of values}}$$

$$\mu = \frac{50 + 55 + 52 + 48 + 51 + 60 + 62 + 54 + 53 + 47 + 65 + 90}{12}$$

$$\mu = \frac{737}{12} = 55.08$$

- Step 2: Computation of standard deviation and z-score of each points.

Value (x)	Deviation ($x - \mu$)	Squared Deviation ($(x - \mu)^2$)	Z-Score ($z = \frac{x-\mu}{\sigma}$)
50	$50 - 55.08 = -5.08$	25.81	-0.445
55	$55 - 55.08 = -0.08$	0.0064	-0.007
52	$52 - 55.08 = -3.08$	9.49	-0.27
48	$48 - 55.08 = -7.08$	50.16	-0.62
51	$51 - 55.08 = -4.08$	16.63	-0.36
60	$60 - 55.08 = 4.92$	24.2	0.43
62	$62 - 55.08 = 6.92$	47.92	0.61
54	$54 - 55.08 = -1.08$	1.17	-0.094
53	$53 - 55.08 = -2.08$	4.33	-0.18
47	$47 - 55.08 = -8.08$	65.28	-0.71
65	$65 - 55.08 = 9.92$	98.4	0.87
90	$90 - 55.08 = 34.92$	1219.57	3.06

- Here: $\sum(x_i - \mu)^2 = 1562.08$

$$\sigma = \sqrt{\frac{1562.08}{12}} = \sqrt{130.17} = 11.41$$

- Step 3: Identify outliers:
  - Threshold for outliers: $|z| > 3$
  - The value 90 is an outlier because $z=3.06$

- Step 4: Handle Outliers
  - Replace 90 with the mean (55.08).
  - Updated dataset: [50,55,52,48,51,60,62,54,53,47,65,55.08]

IQR method for Anomaly detection

Consider the following dataset of exam scores: [50, 55, 52, 48, 51, 60, 62, 54, 53, 47, 65, 90] and detect the outlier, and also plot the box-plot.

Step 1: Sort the dataset in ascending order:

[**47**, 48, 50, 51, 52, 53, 54, 55, 60, 62, **65**, 90]

Step 2: Calculate Q1 (First Quartile) and Q3 (Third Quartile) **Quartiles:**

Q1: The median of the lower half of the data.

Lower half: [47, 48, 50, 51, 52, 53]

Median = (50 + 51) / 2 = 50.5

Q3: The median of the upper half of the data

Upper half: [54, 55, 60, 62, 65, 90]

Median = (60 + 62) / 2 = 61

Step 3: Calculate the IQR

IQR = Q3−Q1 = 61−50.5 = 10.5

Step 4: Determine the Outlier Boundaries Use 1.5 times the IQR to calculate the lower and upper bounds:

- Lower Bound: $Q1 - 1.5 \times \text{IQR} = 50.5 - (1.5 \times 10.5) = 50.5 - 15.75 = 34.75$
- Upper Bound: $Q3 + 1.5 \times \text{IQR} = 61 + (1.5 \times 10.5) = 61 + 15.75 = 76.75$

Step 5: **Identify Outliers** Any value below 34.75 or above 76.75 is considered an outlier:

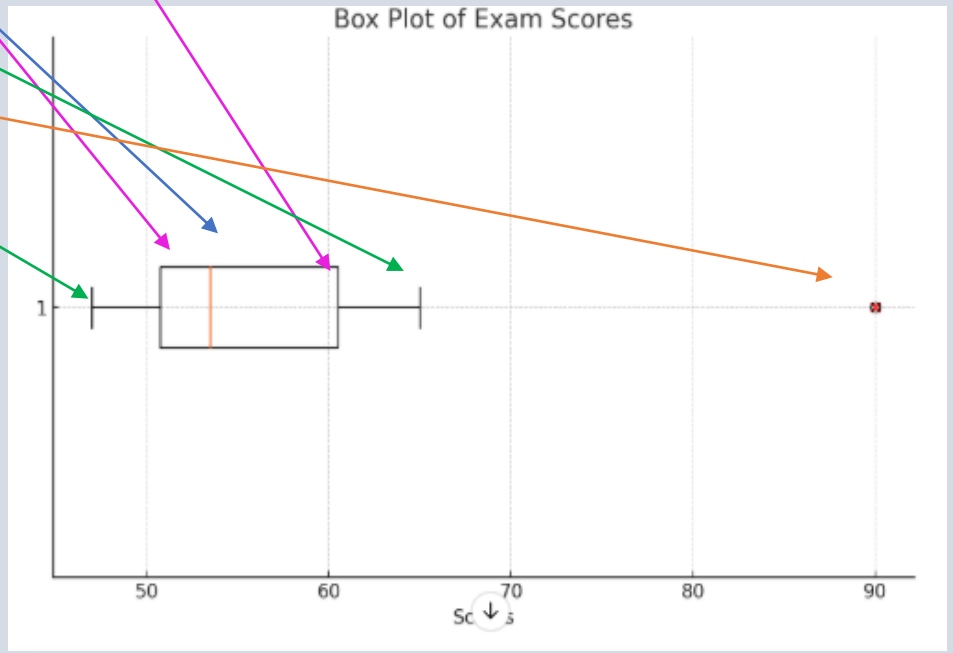
Outlier(s): 90

Step 6: Handle Outliers Depending on the context, you can:

Remove 90 from the dataset.

Replace 90 with the upper bound 76.75

- **Summary of the Box Plot:**
  - **Median (Q2)**: 52.5 — This represents the middle of the dataset.
  - **Interquartile Range (IQR)**: From 50.25 to 60.5, this range shows where the middle 50% of the data lies.
  - **Whiskers**: The whiskers extend from the quartiles to the most extreme values within the 1.5×IQR range, from **47** to **65**.
  - **Outliers**: The outlier is **90**, as it is far above the upper whisker.



7.5 Clustering Approaches for Anomaly detection

Example 1

- Perform DBSCAN on the given problem with ε = 2 and minpoint = 2 and identify the outliers in the dataset.

	x	y
A1	2	10
A2	2	5
A3	8	4
A4	5	8
A5	7	5
A6	6	4
A7	1	2
A8	4	9

Reference Video:
https://www.youtube.com/watch?v=3l1vpcRMGcc

Step 1: Calculation of Euclidian distance

Euclidean Distance	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	5	8.49	3.61	7.07	7.21	8.06	2.24
A2	5	0	6.08	4.24	5	4.12	3.16	4.47
A3	8.49	6.08	0	5	1.41	2	7.28	6.4
A4	3.61	4.24	5	0	3.61	4.12	7.21	1.41
A5	7.07	5	1.41	3.61	0	1.41	6.71	5
A6	7.21	4.12	2	4.12	1.41	0	5.39	5.39
A7	8.06	3.16	7.28	7.21	6.71	5.39	0	7.62
A8	2.24	4.47	6.4	1.41	5	5.39	7.62	0

Step 2: Count of points within ε = 2 and identify each points as core, border or noise point w.r.t. Minpts=2

Points	No of points	Remarks
A1	1 (A1)	Noise
A2	1 (A2)	Noise
A3	3 (A3, A5, A6)	**Core**
A4	2 (A4, A8)	**Core**
A5	3 (A3, A5, A6)	**Core**
A6	3 (A3, A5, A6)	**Core**
A7	1 (A7)	Noise
A8	2 (A4, A8)	**Core**

Here:
A1, A2 andA7 are Noise (i.e. outlier)
Cluster 1: A3, A5, A6
Cluster2: A4, A8

Example 2

- Perform DBSCAN on the given problem with ε = 3.5 and minpoint = 3 and identify the outliers in the dataset.
-

S1	5	7
S2	8	4
S3	3	3
S4	4	4
S5	3	7
S6	6	7
S7	6	1
S8	5	5

Reference: https://www.youtube.com/watch?v=jISFQ0I5Gj4

Step 1: Calculation of Euclidian distance

First, we compute the Euclidean distance between each pair of points. The Euclidean distance between two points (x1, y1) and (x2, y2) is given by:

distance= $\sqrt{(x2-x1)^2+(y2-y1)^2}$

	S1	S2	S3	S4	S5	S6	S7	S8
S1	0	4.24	4.47	3.16	2	1	6.08	2
S2	4.24	0	5.1	4	5.83	3.61	3.61	3.16
S3	4.47	5.1	0	1.41	4	5	3.61	2.83
S4	3.16	4	1.41	0	3.16	3.61	3.61	1.41
S5	2	5.83	4	3.16	0	3	6.71	2.83
S6	1	3.61	5	3.61	3	0	6	2.24
S7	6.08	3.61	3.61	3.61	6.71	6	0	4.12
S8	2	3.16	2.83	1.41	2.83	2.24	4.12	0

Step 2: Count of points within ε = 3.5 and identify each points as core, border or noise point w.r.t. Minpts=3

Points	No of points	Remarks
S1	5 (S1,S4,S5,S6,S8)	**Core**
S2	2 (S2,S8)	Border/Noise
S3	3 (S3,S4,S8)	**Core**
S4	5 (S4, S1,S5,S3,S8)	**Core**
S5	5 (S5,S1,S4,S6,S8)	**Core**
S6	4 (S6,S1,S5,S8)	**Core**
S7	1 (S7)	Border/Noise
S8	7 (S8,S1,S2,S3,S4,S5,S6)	**Core**

Points	No of points	Remarks
S1	5 (S1,S4,S5,S6,S8)	Core
S2	2 (S2,**S8**)	Border/Noise
S3	3 (S3,S4,S8)	Core
S4	5 (S4, S1,S5,S3,S8)	Core
S5	5 (S5,S1,S4,S6,S8)	Core
S6	4 (S6,S1,S5,S8)	Core
S7	1 (S7)	Border/Noise
S8	7 (S8,S1,**S2**,S3,S4,S5,S6)	Core

Step 3: Conversion of Noise to Border point

If density reachable condition is satisfied, convert noise to border point.

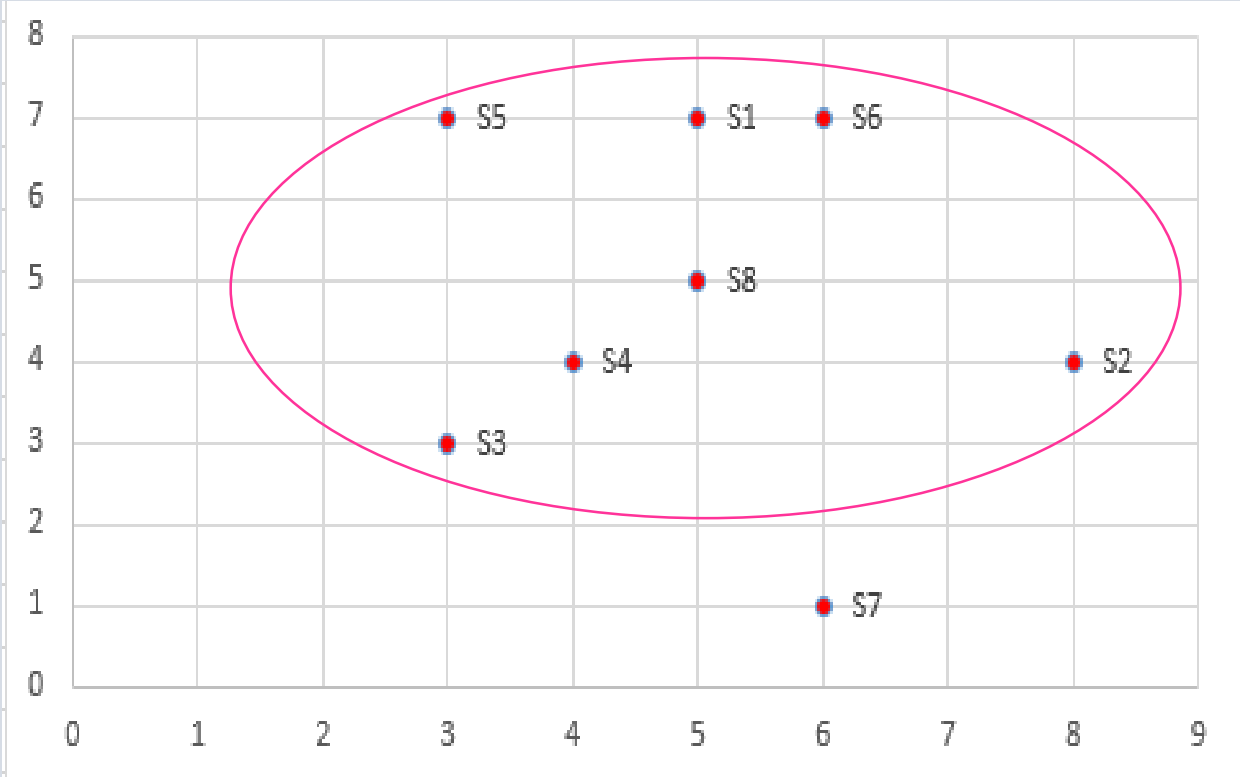
Here: S2 is converted to Border point since it has a core point S8 which its neighbor

Points	No of points	Remarks	Conversion
S1	5 (S1,S4,S5,S6,S8)	**Core**	
S2	2 (S2, **S8**)	Border/Noise	**Border**
S3	3 (S3,S4,S8)	**Core**	
S4	5 (S1, S3, S4,S5,S8)	**Core**	
S5	5 (S1, S5, S4,S6,S8)	**Core**	
S6	4 (S1,S6, S5,S8)	**Core**	
S7	1 (S7)	Border/Noise	**Noise**
S8	7 (S1,**S2**,S3,S4,S5,S6, S8)	**Core**	

Hence: Clusters are:
Cluster 1: {S1,S3,S4,S5,S6,S8, S2}
Noise: {S7}

Clusters are:

- **Cluster 1: {S1,S2,S3,S4,S5,S6,S8}**
- **Outlier: {S7}**



End of Chapter