Chapter 8 : Applications and case studies

Sunil Regmi, Lecturer, DoAI, KU

Anomaly/Outlier Detection

- Is the process to localize **objects that are different from other objects** (anomalies).
- The set of **data points that are considerably different than the remainder of the data** are anomalies/outliers.
- **Anomaly detection is the process of detecting something unusual relative to something expected**.
- The goal of anomaly detection is **to identify cases that are unusual within data that is seemingly homogeneous**.

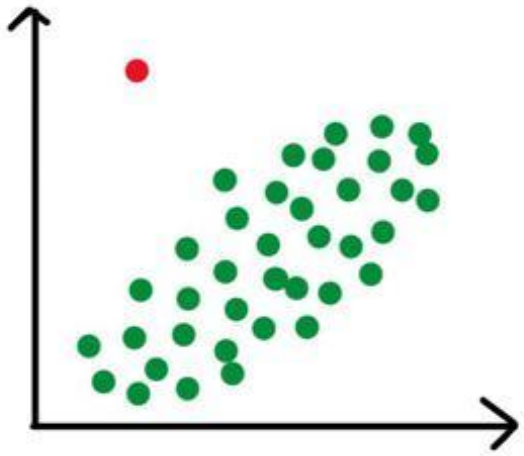
Anomaly Detection ….

- **Variants of Anomaly/Outlier Detection Problems**
  - Given a database D, find all the data points $x \in D$ with **anomaly scores greater than some threshold t**
  - Given a database D, find all the data points $x \in D$ having the **top-n largest anomaly scores f(x)**
  - Given a database D, containing mostly normal (but unlabeled) data points, and a **test point x, compute the anomaly score of x with respect to D**
- Why is Anomaly Detection important?
  - **to detect problems**
  - **to detect new phenomenon**
  - **to discover unusual behavior in data**

Examples of interesting application for Anomaly Detection

- **Fraud Detection** - looking for buying patterns different from typical behavior
- **Intrusion Detection** - monitoring systems and networks for unusual behavior
- **Ecosystem Disturbances** - try to predict events like hurricanes and floods
- **Public Health** - use medical statistic reports for diagnosis
- **Medicine** - use unusual symptoms or test result to indicate potential health problems

Types of Outliers



1. Global Outliers (Point Anomalies): These are individual data points that significantly deviate from the rest of the dataset.

- They stand out as exceptionally different from the majority of the data.

Example: In a dataset of exam scores, a score of 2 out of 100 would be a global outlier if most other scores are in the 70-90 range.

- Errors in data collection, measurement errors, or truly unusual events can result in global outliers.

- Impact: Global outliers can distort data analysis results and affect machine learning model performance.

...

4. Detection: Techniques include statistical methods (e.g., z-score, Mahalanobis distance), machine learning algorithms (e.g., isolation forest, one-class SVM), and data visualization techniques.

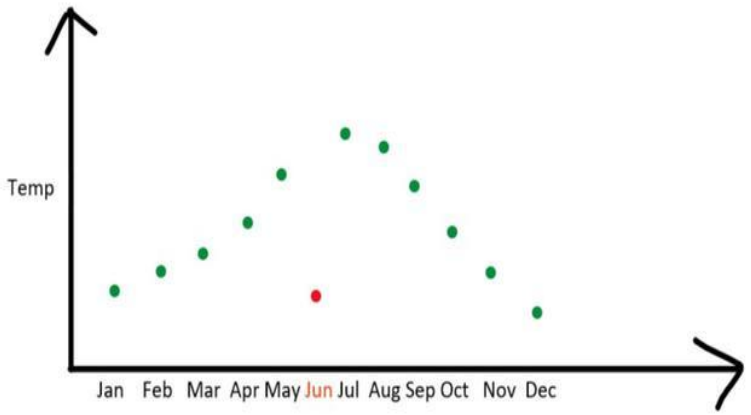
5. Handling: Options may include removing or correcting outliers, transforming data, or using robust methods.

6. Considerations: Carefully considering the impact of global outliers is crucial for accurate data analysis and machine learning model outcomes.

Types of Outliers

2. Contextual Outliers (Conditional Anomalies): These outliers are unusual within a specific context or condition, but might not be considered outliers in a broader context.

Example: A temperature of 80°F might be an outlier in the winter but not in the summer.



- Contextual information such as time, location, or other relevant factors are crucial in identifying contextual outliers.

- Contextual outliers may not be outliers when considered in the entire dataset, but they exhibit unusual behavior within a specific context or subgroup.

...

- Detection: Techniques for detecting contextual outliers include contextual clustering, contextual anomaly detection, and context-aware machine learning approaches.

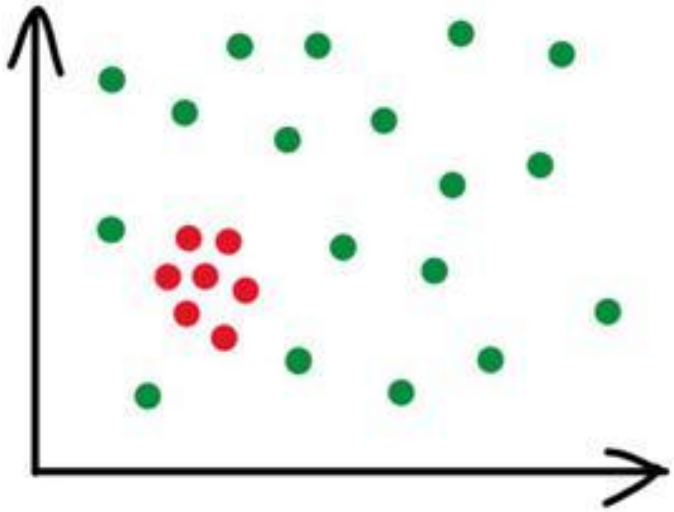
- Impact: Contextual outliers can represent unusual or anomalous behavior within a specific context, which may require further investigation or attention.

- Handling: Handling contextual outliers may involve considering the contextual information, contextual normalization or transformation of data, or using context-specific models or algorithms.

- Considerations: Proper understanding of the context and domain-specific knowledge is crucial for accurate detection and interpretation of contextual outliers, as they may vary based on the specific context or subgroup being considered.

...

3. Collective Outliers: Collective outliers are groups of data points that collectively deviate significantly from the overall distribution of a dataset.



Example: Several consecutive days of unusually low temperatures, even if each individual day's temperature is within the normal range for that time of year.

- Collective outliers may not be outliers when considered individually, but as a group, they exhibit unusual behavior.

...

- **Detection**: clustering algorithms, density-based methods, and subspace-based approaches.

- **Impact:** Collective outliers can represent interesting patterns or anomalies in data that may require special attention or further investigation.

- **Handling:** Handling collective outliers depends on the specific use case and may involve further analysis of the group behavior, identification of contributing factors, or considering contextual information.

- **Considerations:** Detecting and interpreting collective outliers can be more complex than individual outliers, as the focus is on group behavior rather than individual data points.
  - Proper understanding of the data context and domain knowledge is crucial for effective handling of collective outliers.

Challenges in Outlier Detection

1. Modeling Normal Behavior

“You can’t find what’s abnormal until you clearly define what is normal.”

- Building a robust model of “normal” data is challenging due to:
  - Variability in normal behavior
  - Context dependence
- Some models assign binary labels: normal or outlier.
- Others use **scoring functions** (e.g., anomaly scores, LOF).
- **Challenge**: Setting the threshold or boundary between "normal" and "abnormal" is subjective and difficult without domain knowledge.

...

2. Application-Specific Requirements

- Outlier definitions vary across applications:
  - **Clinical Data**: Even small deviations can be critical.
  - **Marketing Data**: Larger variations are tolerated.

Algorithms must be tailored to:

- Data format (numerical, text, categorical)
- Business logic and domain sensitivity

No one-size-fits-all outlier detection technique.

...

3. Noise vs. Outliers

Noise is often mistaken for outliers, but not all noise is useful.

- **Noise** refers to random errors or inconsistencies in data (e.g., missing values, typos).

- **Outliers** may carry meaningful deviations (e.g., fraud, system failure).

- **Challenge**: Noisy data can obscure real outliers or falsely flag normal points as anomalous.
- Requires robust preprocessing (data cleaning, imputation, deduplication).

...

4. Understandability (Interpretability)

- Clients or users may ask:
"Why was this labeled as an outlier?"
- Methods like **proximity-based** outlier detection are easier to explain:
  - Outliers are far from the cluster or nearest neighbors.
  - Complex methods (like deep learning) lack transparency.

- **Challenge**: Need for explainable outlier detection, especially in regulated domains (e.g., healthcare, finance).

...

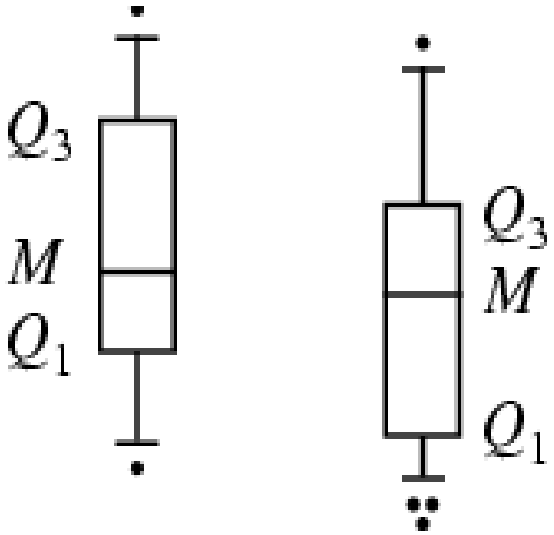
- **Heterogeneous Data**: Datasets often mix text, numeric, and categorical types, making it hard to design a universal detection method.
- **Scalability**: Large datasets demand efficient, scalable algorithms; traditional methods may become computationally infeasible.
- **High Dimensionality**: In high-dimensional spaces, distinguishing outliers from normal patterns is difficult due to the "curse of dimensionality."
- **Noise & Variability**: Noise can obscure true outliers or create false positives; robust methods are needed to handle data imperfections.
- **Parameter Sensitivity**: Many techniques rely on sensitive parameters (e.g., DBSCAN's ε), which can lead to inconsistent or subjective results.
- **Multi-modal Distributions**: In datasets with multiple clusters or modes, traditional methods may confuse minority clusters with outliers.

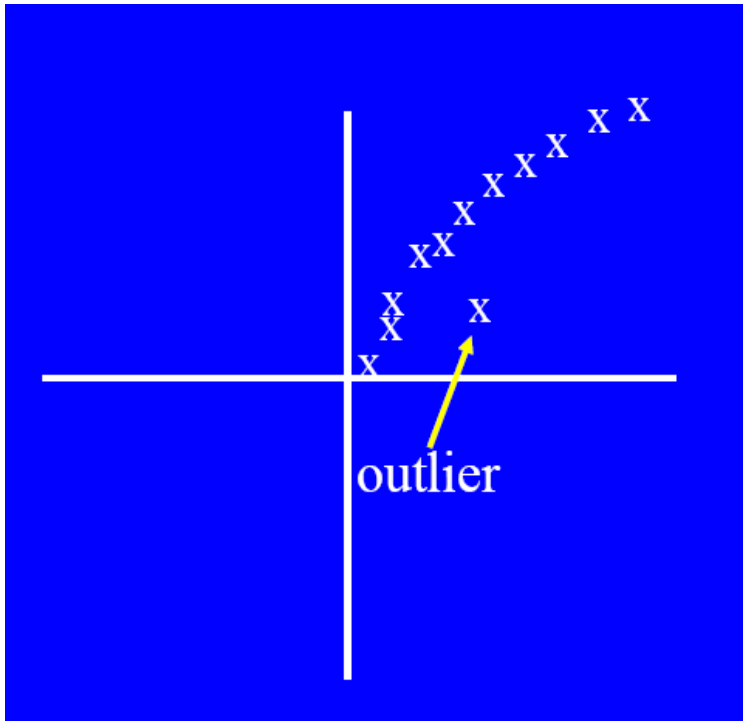
Anomaly Detection Schemes

- Working assumption:
  - There are considerably **more "normal" observations than "abnormal" observations** (outliers/anomalies) in the data
- General Steps
  - **Build a profile of the "normal" behavior**
    - Profile can be **patterns or summary statistics for the overall population**
  - **Use the "normal" profile to detect anomalies**
    - Anomalies are observations whose **characteristics differ significantly from the normal profile**
- Types of anomaly detection schemes
  - Graphical
  - Statistical-based
  - Distance-based
  - Model-based

Graphical Approaches

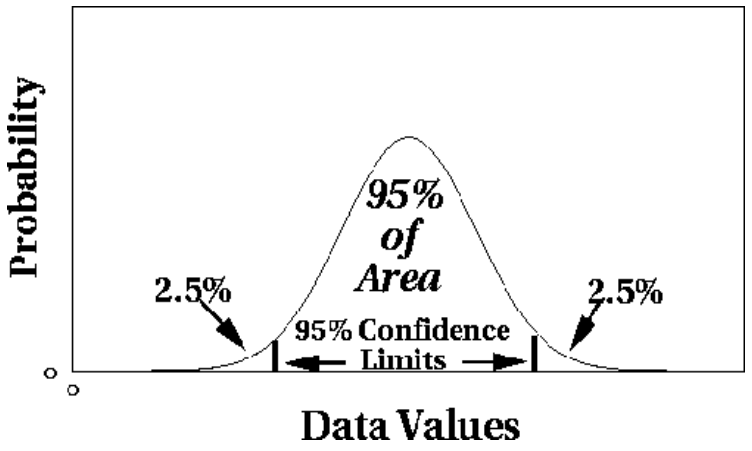
- Boxplot (1-D), Scatter plot (2-D)
- Limitations
  - Time consuming
  - Subjective





Statistical Approaches

- Assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on
  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)



Statistical-based – Likelihood Approach

- Assume the data set D contains samples from a mixture of two probability distributions:
  - M (majority distribution)
  - A (anomalous distribution)
- General Approach:
  - Initially, assume all the data points belong to M
  - Let $L_t(D)$ be the log likelihood of D at time t
  - For each point x_t that belongs to M, move it to A
    - Let $L_{t+1}(D)$ be the new log likelihood.
    - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
    - If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Distance-based Approaches

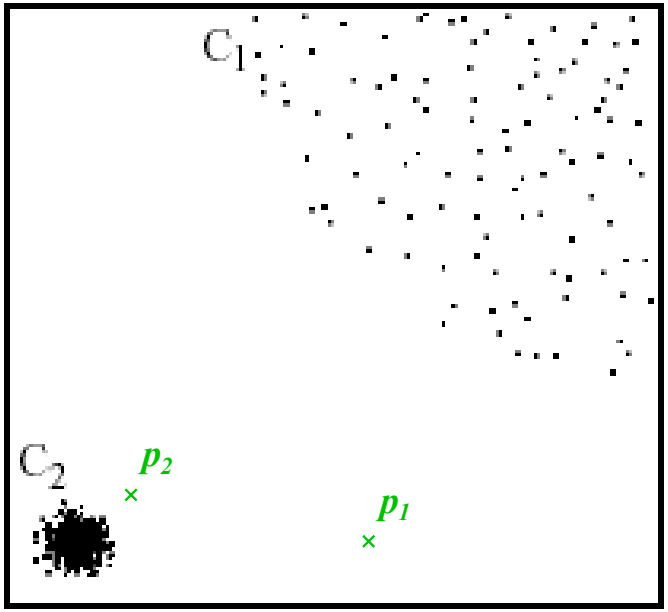
- Data is represented as a vector of features
- Three major approaches
  - Nearest-neighbor based
  - Density based
  - Clustering based

Nearest-Neighbor Based Approach

- Approach:
  - Compute the distance between every pair of data points
  - There are various ways to define outliers:
    - Data points for which there are **fewer than *p* neighboring points within a distance *D***
    - The **top n data points whose distance to the kth nearest neighbor is greatest**

Density-based: LOF approach

- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors
- Outliers are points with largest LOF value

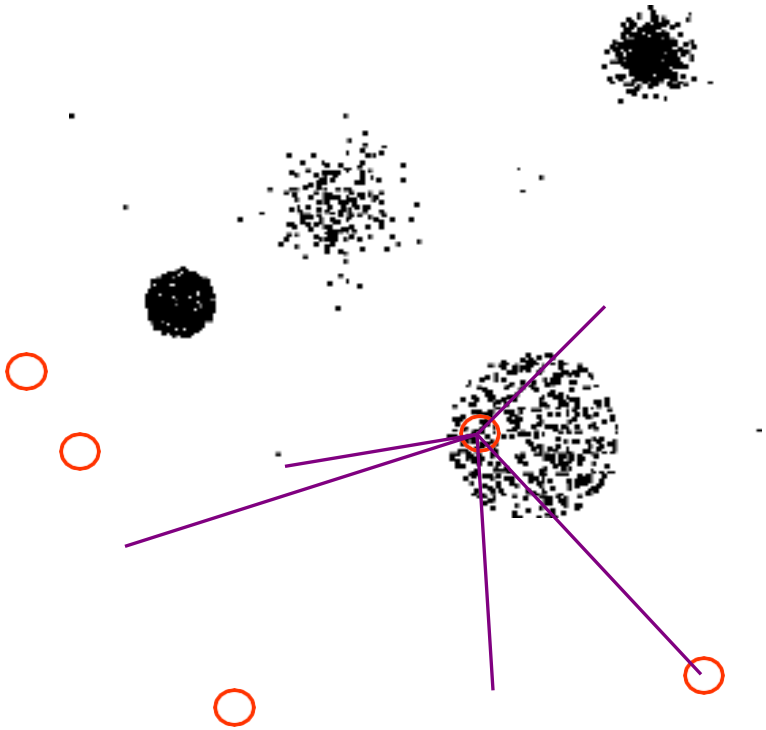


LOF(P) = AVG(Density of nearest Neighbors) ÷ Density of sample P

In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Clustering-Based

- Basic idea:
  - Cluster the data into groups of different density
  - Choose points in small cluster as candidate outliers
  - Compute the distance between candidate points and non-candidate clusters.
    - **If candidate points are far from all other non-candidate points, they are outliers**



Model-based

- An anomaly detection **model predicts whether a data point is typical for a given distribution or not**.

- An **atypical data point** can be either an **outlier or** an example of a previously **unseen class**.

- Normally, a classification **model must be trained** on data that includes both **examples and counter-examples** for each class so that the model can learn to distinguish between them.

- For example, a **model that predicts side effects of a medication** should be **trained on data that includes a wide range of responses to the medication.**

Issues on Anomaly Detection

Number of Attributes: Since an object may have many attributes, it may have anomalous values for some attributes; an object may be anomalous even if none of its attribute values are individually anomalous.

Global Vs Local Perspective: An object may seem unusual with respect to all objects, but not with respect to its local neighbors.

Degree of Anomaly: Some objects are more extreme anomalies than others;

One at Time Vs Many at Once: Is it better to remove anomalous objects one at a time or identify a collection of objects together?

Base Rate Fallacy example

- A group of police officers have breathalyzers displaying false drunkenness in 5% of the cases in which the driver is sober. However, the breathalyzers never fail to detect a truly drunk person. One in a thousand drivers is driving drunk. Suppose the police officers then stop a driver at random, and force the driver to take a breathalyzer test. It indicates that the driver is drunk. We assume you don't know anything else about him or her. How high is the probability he or she really is drunk?

Base Rate Fallacy

- The base-rate fallacy is **people's tendency to ignore base rates (general information) in favor of specific information** when such is available rather than integrate the two.
- This **tendency has important implications for understanding judgment phenomena** in many clinical, legal, and social-psychological settings.
- Base rate fallacy, also called **base rate neglect** or base rate bias, is a formal fallacy. If **presented with related base rate information and specific information**, the **mind tends to ignore the former and focus on the latter**.

Base Rate Fallacy (Axelsson, 1999)

The base-rate fallacy is best described through example.[2] Suppose that your doctor performs a test that is 99% accurate, i.e. when the test was administered to a test population all of whom had the disease, 99% of the tests indicated disease, and likewise, when the test population was known to be 100% free of the disease, 99% of the test results were negative. Upon visiting your doctor to learn the results he tells you he has good news and bad news. The bad news is that indeed you tested positive for the disease. The good news however, is that out of the entire population the rate of incidence is only 1/10000, i.e. only 1 in 10000 people have this ailment. What, given this information, is the probability of you having the disease? The reader is encouraged to make a quick "guesstimate" of the answer at this point.

Base Rate Fallacy

- Bayes theorem:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

- More generally:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{\sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)}$$

Base Rate Fallacy

$$P(S|P) = \frac{P(S) \cdot P(P|S)}{P(S) \cdot P(P|S) + P(\neg S) \cdot P(P|\neg S)}$$

$$P(S|P) = \frac{1/10000 \cdot 0.99}{1/10000 \cdot 0.99 + (1 - 1/10000) \cdot 0.01} = \\ = 0.00980\ldots \approx 1\%$$

- Even though the test is 99% certain, your chance of having the disease is 1/100, because the population of healthy people is much larger than sick people
- Browse below for additional examples: https://en.wikipedia.org/wiki/Base_rate_fallacy

▪▪▪

Scenario: Disease Detection Example
You are testing for a **rare disease** using a medical test.
📌 **Given:**
Disease prevalence (base rate): 1 in 1,000 people (0.1%)
Test accuracy:
True Positive Rate (Sensitivity) = 99%
False Positive Rate = 1%
You test **10,000 people**.

Actual disease cases (base rate):0.1% of 10,000 = 10 people actually have the disease
- **True Positives**:99% of 10 = 9.9 people correctly test positive
- **False Positives**:1% of 9,990 people without the disease = 99.9 people falsely test positive

Total Positive Test Results:True Positives+False Positives=9.9+99.9=109.8
- What's the probability that a person who tested positive actually has the disease?
- P(Disease|Positive Test)=9.9 /109.8≈0.09 or 9%
- P(Disease|Positive Test)= 109.89.9 ≈0.09 or 9%
- Inference: Even though the test is 99% accurate, a positive result only gives a 9% chance of actually having the disease.