Unit 5
Outlier Detection

Preliminaries,
Outliers and types
Approaches
(Statistical, Proximity-based, Clustering-based, Classification based)

Objective

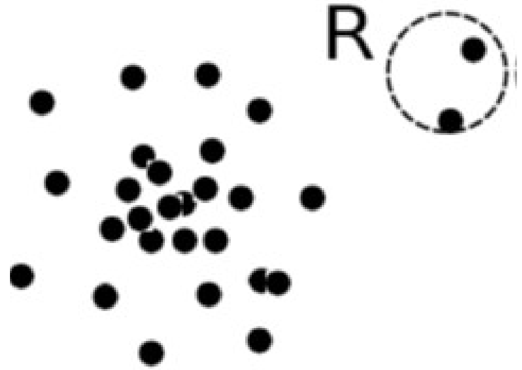
- Preliminaries
- Outliers and types
- Approaches
  - Statistical
  - Proximity-based
  - Clustering-based
  - Classification based

Anomaly/Outlier Detection

- Similar terms
  - Anomaly
  - Deviation
  - Extreme Value
  - Noise
  - Novelty → something new, off-track, unusual
  - Fluctuation
  - Discrepancy

Anomaly/Outlier

- What are anomalies/outliers?
  - The set of data points that are **considerably different** than the remainder of the data
  - An outlier is a data point that **significantly differs** from other data points in a dataset.
- Outliers are different from the noise data
  - Noise is random error or variance in a measured variable
  - Noise should be removed before outlier detection



Anomaly/Outlier

- How outliers are formed?
  - Outliers can occur due to various reasons, such as errors in data collection, measurement variability, or genuine rare occurrences in the data.
- Identifying and handling outliers is important in data analysis and statistics because they can distort statistical analyses and machine learning models.

Characteristics of Outliers

- Unusual Value
  - Outliers are data points that are significantly different from the majority of other data points in the dataset.
  - They may be unusually high or low compared to the rest of the data.
- Impact on Summary Statistics
  - Outliers can significantly affect summary statistics such as the mean and standard deviation.
  - For example, a single extremely high or low value can skew the mean and increase the variance.
- Visual Detection
  - Outliers can often be detected visually in plots such as histograms, box plots, or scatter plots.
  - They may appear as points that are far from the main cluster of data points.
- Quantitative Detection
  - Outliers can also be detected quantitatively using statistical methods such as the z-score, which measures how many standard deviations a data point is away from the mean.
- Impact on Models
  - Outliers can have a significant impact on statistical analyses and machine learning models.
  - They can lead to biased estimates, reduced model performance, or incorrect conclusions if not properly handled.

Sources of Outliers

- Data Entry Errors
  - Outliers may occur due to human errors during data entry, recording, or transcription.
  - For example, typos, misreadings, or mistakes in data collection can lead to erroneous data points that deviate significantly from the true values.
- Sampling Errors
  - Outliers may arise from sampling errors when the sample size is small or not representative of the population.
  - Sampling variability, sampling bias, or outliers in the population itself can lead to outliers in the sample data.
- Natural Variability
  - In some cases, outliers may occur naturally due to inherent variability or heterogeneity in the data-generating process.
  - Natural processes, such as biological variation, environmental factors, or random fluctuations, can produce outliers that deviate from the central tendency of the data

Sources of Outliers

- Measurement Variability
  - Variability in measurement instruments or techniques can lead to outliers in the data.
  - Measurement errors, calibration issues, or sensor malfunctions may result in data points that are inconsistent or inaccurate compared to the majority of observations.
- Measurement Units
  - Outliers may arise from inconsistencies in measurement units or scales used to record data.
  - Incompatible units, unit conversions, or mixing of different measurement scales can result in outliers that do not conform to the expected distribution of values.
- Data Collection Methods
  - Outliers may be introduced by limitations or biases in data collection methods.
  - Biased sampling, non-random selection of subjects, or incomplete data collection can lead to outliers that do not accurately represent the underlying population.

Sources of Outliers

- Genuine Extreme Values
  - Outliers may represent genuine extreme values or rare events in the data.
  - Unusual phenomena, extreme weather events, or rare occurrences may produce outliers that are legitimate observations but deviate from the typical patterns in the data.
- Errors in Data Processing
  - Outliers can result from errors or anomalies introduced during data processing or manipulation.
  - Mistakes in data cleaning, transformation, or aggregation processes may lead to outliers that distort the analysis or modeling results.
- Data Corruption or Tampering
  - Outliers may be introduced deliberately or accidentally through data corruption or tampering.
  - Malicious attacks, data breaches, or data manipulation can alter the integrity of the data and introduce outliers that compromise its quality and reliability.

Types of Outliers

- Three kinds
  - Global
  - Contextual
  - Collective outliers

Global Outlier (Point Anomaly)

- Object is O_g if it significantly deviates from the rest of the data set
- Ex. Intrusion detection in computer networks
- Issue
  - Find an appropriate measurement of deviation

Contextual Outlier (Conditional Outlier)

- Object is O_c if it deviates significantly based on a selected context
  - Ex. 25 Degree C in Kathmandu:
    - Outlier?
    - depending on summer or winter
- Attributes of data objects should be divided into two groups
  - Contextual attributes: defines the context, e.g., time & location
  - Behavioral attributes: characteristics of the object, used in outlier evaluation, e.g., temperature
- Can be viewed as a generalization of local outliers
  - whose density significantly deviates from its local area
- Issue
  - How to define or formulate meaningful context?

Collective Outliers

- A subset of data objects collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers
- Applications: E.g., intrusion detection:
    - When a number of computers keep sending denial-of-service packages to each other
- Detection of collective outliers
    - Consider not only behavior of individual objects, but also that of groups of objects
    - Need to have the background knowledge on the relationship among data objects, such as a distance or similarity measure on objects

Outlier Detection Problem

- Variants of Anomaly/Outlier Detection Problems
  - Given a database D, find all the data points x ϵ D with anomaly scores greater than some threshold t
  - Given a database D, find all the data points x ϵ 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

Applications

- The outlier detection techniques are applicable in various fields
  - Credit card fraud detection
  - Telecommunication fraud detection
  - Network intrusion detection
  - Fault detection
  - ……

Challenges

- How many outliers are there in the data?
- Method is unsupervised
  - Validation can be quite challenging (just like for clustering)
- Finding needle in a haystack

Anomaly Detection Schemes

- 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 differs significantly from the normal profile

Types of Anomaly Detection Schemes

- Based on user-labeled data
  - Supervised Methods, Semi-Supervised
- Based on assumptions about normal data and outliers
  - Unsupervised Method
  - Graphical and Statistical Based
  - Distance Based
  - Model Based

Supervised Method

- Modeling outlier detection as a classification problem
  - Samples examined by domain experts used for training & testing
- Methods for Learning a classifier for outlier detection effectively:
  - Model normal objects & report those not matching the model as outliers, or
  - Model outliers and treat those not matching the model as normal
- Challenges
  - Imbalanced classes, i.e., outliers are rare
    - Boost the outlier class and make up some artificial outliers
  - Catch as many outliers as possible, i.e., recall is more important than accuracy (i.e., not mislabeling normal objects as outliers)

Unsupervised Method

- Assume the normal objects are somewhat ‘clustered’ into multiple groups, each having some distinct features
- An outlier is expected to be far away from any groups of normal objects
- Weakness: Cannot detect collective outlier effectively
  - Normal objects may not share any strong patterns, but the collective outliers may share high similarity in a small area

Unsupervised Method

- Ex. In some intrusion or virus detection, normal activities are diverse
  - Unsupervised methods may have a high false positive rate but still miss many real outliers.
  - Supervised methods can be more effective, e.g., identify attacking some key resources
- Many clustering methods can be adapted for unsupervised methods
  - Find clusters, then outliers: not belonging to any cluster
    - Problem 1: Hard to distinguish noise from outliers
    - Problem 2: Costly since first clustering: but far less outliers than normal objects
      - Newer methods: tackle outliers directly

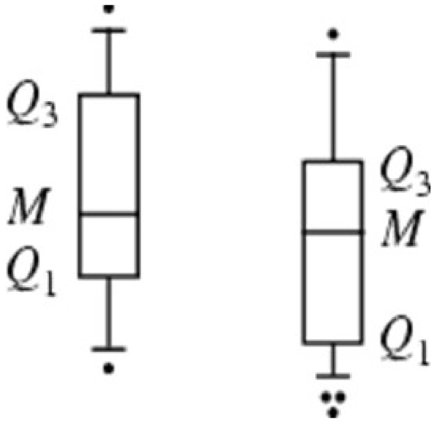
Semi-Supervised Methods

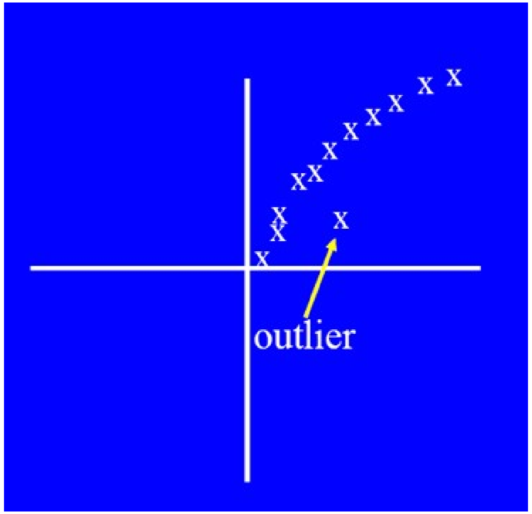
- Situation:
  - In many applications, the number of labeled data is often small: Labels could be on outliers only, normal objects only, or both
  - Semi-supervised outlier detection: Regarded as applications of semi-supervised learning
- If some labeled normal objects are available
  - Use the labeled examples and the proximate unlabeled objects to train a model for normal objects
  - Those not fitting the model of normal objects are detected as outliers
- If only some labeled outliers are available, a small number of labeled outliers many not cover the possible outliers well
  - To improve the quality of outlier detection, one can get help from models for normal objects learned from unsupervised methods

Graphical Approach

- Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D)

- Limitations
  - Time consuming
  - Subjective





Interquartile Range (IQR)

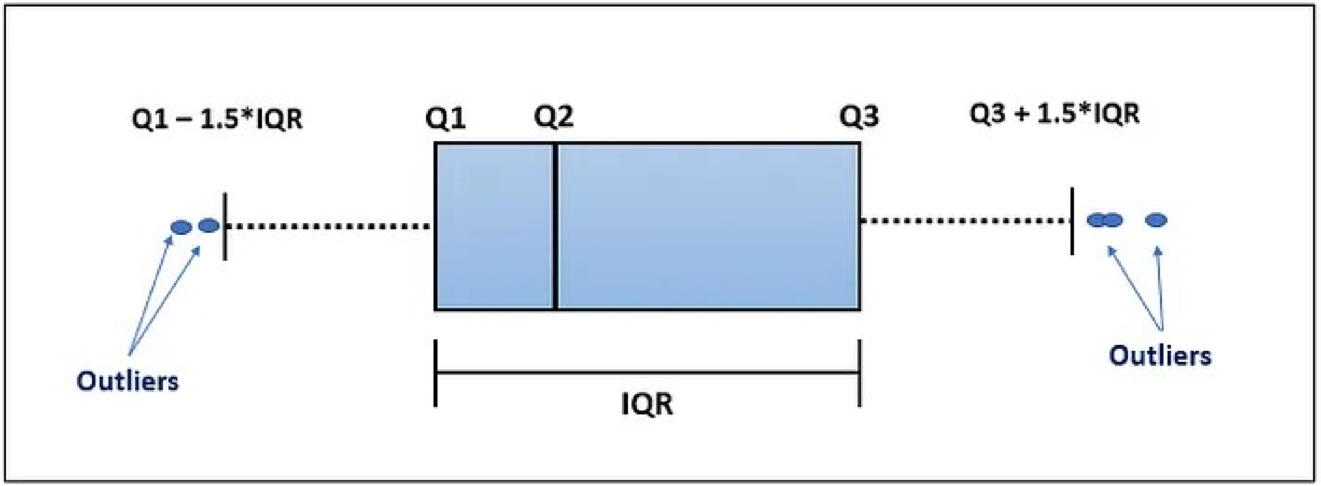
- The interquartile range (IQR) is a measure of statistical dispersion by dividing a data set into quartiles and is also called as Midspread or H-spread.
  - It shows how the data is spread about the median.
  - The data sorted in ascending order and then divided into quartiles.
- IQR is calculated as the difference between the 75th and 25th percentiles.
- This method is also called **Extreme Value analysis**

Interquartile Range (IQR)

IQR = Q3 – Q1

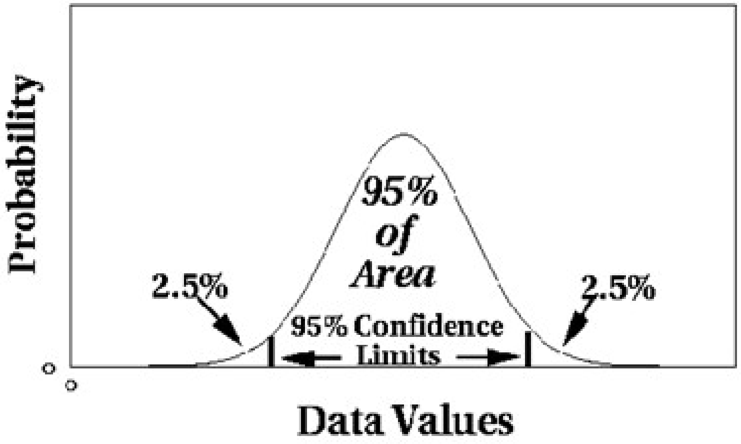
Upper Limit = Q3 + 1.5 IQR

Lower Limit = Q1 - 1.5 IQR



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)



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

- Data distribution, D = (1 -λ) M + λA
- M is a probability distribution estimated from data
  - Can be based on any modeling method
  - A is initially assumed to be uniform distribution
- Likelihood at time t:

$$L_t(D) = \prod_{i=1}^{N} P_D(x_i) = \left((1-\lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1-\lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

Limitation of Statistical Approaches

- Most of the tests are for a single attribute
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution

Distance-based Approaches

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

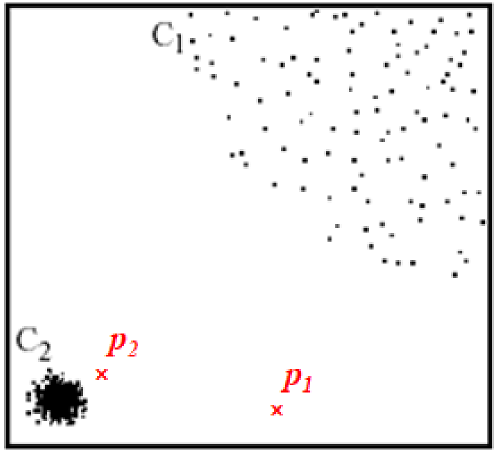
Nearest-Neighbor Based Approach

- 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
    - The top n data points whose average distance to the k nearest neighbors is greatest

Density Based: LOF Approach

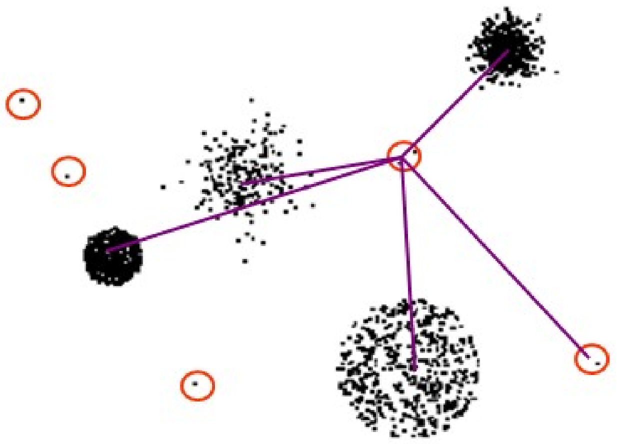
- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample p as the average of the ratios of the density of sample p and the density of its nearest neighbors
- Outliers are points with largest LOF value

In the NN approach, p2 is not considered as outlier,
while LOF approach find
both p1 and p2 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



Reference

- https://medium.com/subex-ai-labs/an-introduction-to-outlier-detection-methods-using-python-f2013824a7b7
-

...