COMP5121
Data Mining and Data Warehousing Applications

Week 10: Outlier Detection

Dr. Fengmei Jin

- Email: fengmei.jin@polyu.edu.hk
- Office: PQ747 (+852 3400 3327)
- Consultation Hours: 2.30-4.30pm every Thursday

Outline

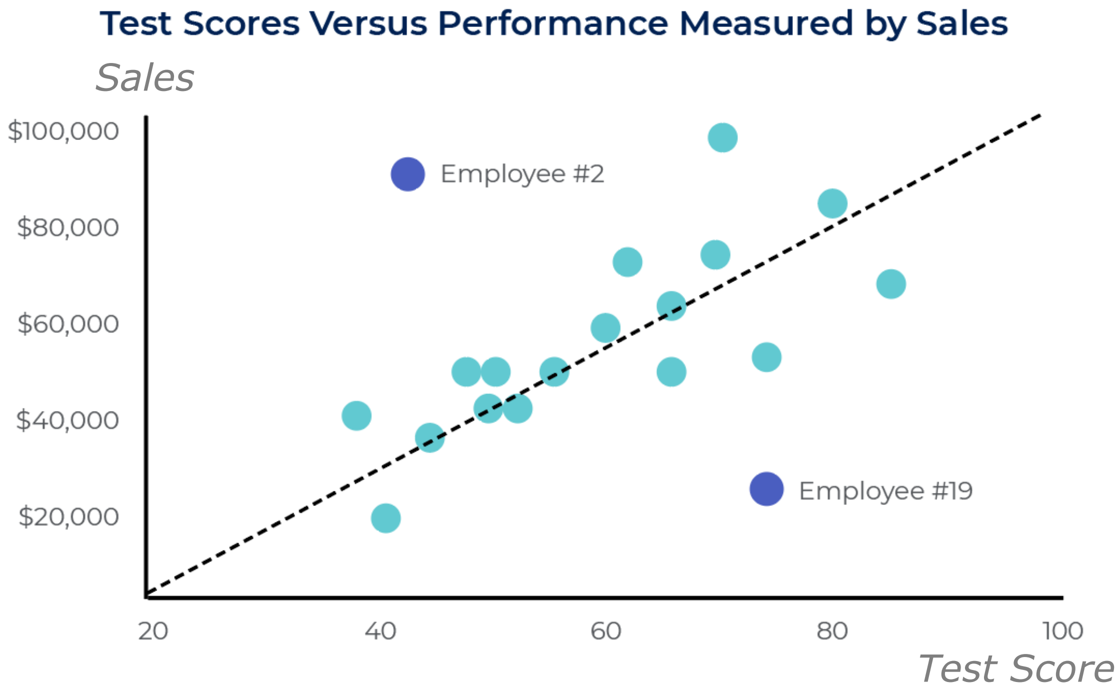
- Outliers and Outlier Analysis
- Outlier Detection Methods Categories
- Statistical Methods
- Proximity-Based Methods
- Clustering-Based Methods
- Mining Contextual and Collective Outliers
- Outlier Detection in High-Dimensional Space

OUTLIERS AND OUTLIER ANALYSIS

What are Outliers?

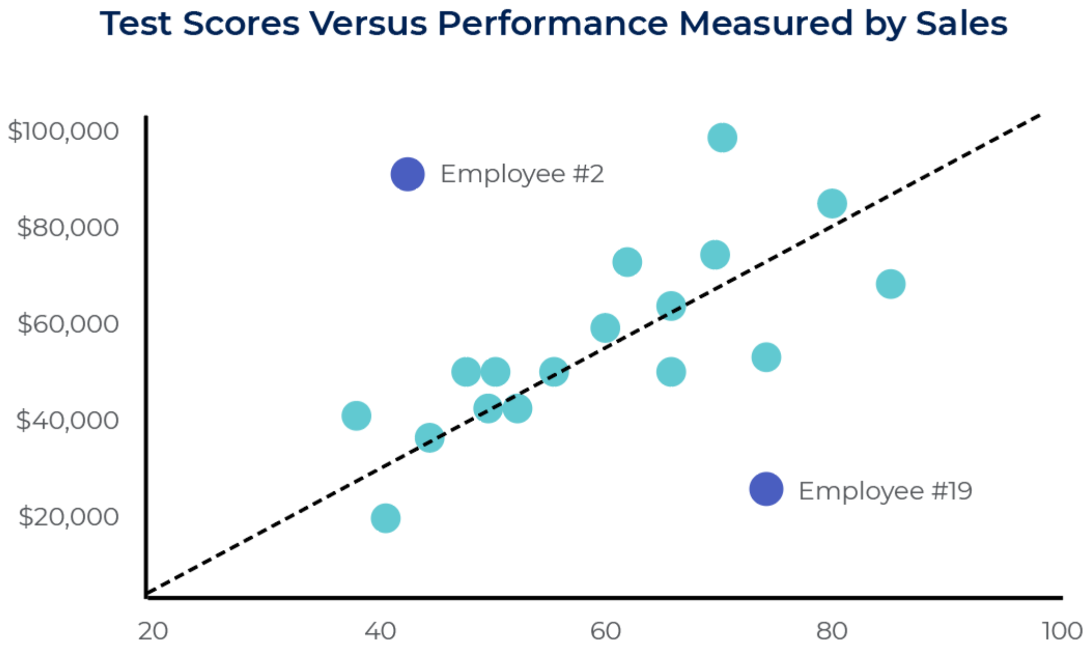
- **Outlier**: A data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism
  - Unusual transaction target/amount
  - Temperature
  - …

We often refer the rest of the object as normal data and outliers as abnormal data.



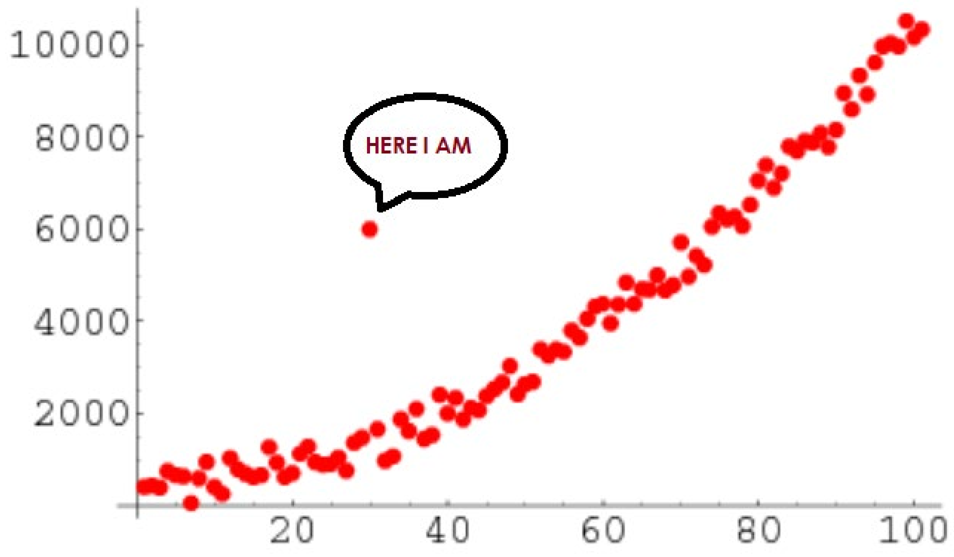
What are Outliers?

- Outliers are different from the noises
  - Noises are random errors or variance in a measurement process.
  - Noises can mislead data analysis and need to be removed.
- Outliers are interesting
  - Provide new knowledge
  - Potentially be influential
  - Need to be handled with care



Types of Outliers – Global

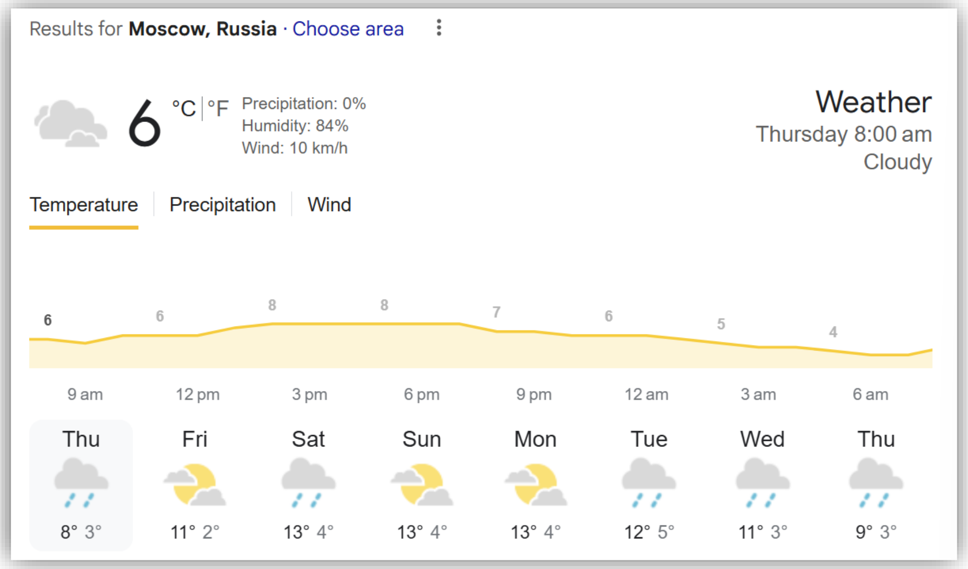
- Global outlier: A data object that deviates significantly from the entire dataset



Types of Outlier – Contextual

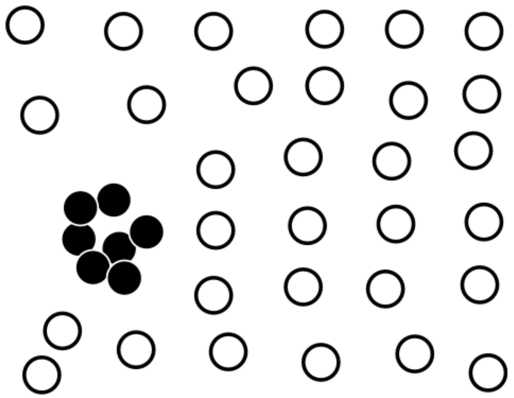
- Contextual outlier: A data object deviating significantly with respect to a specific context of the object

- Example: Is 25℃ in March an outlier?
  - In Hong Kong, it is normal.
  - In Moscow, …



Types of Outliers – Collective

- Collective outlier: A subset of data objects that collectively deviates significantly from the entire dataset
- **Key point**: A single data point may not be an outlier on its own, but their combined behavior makes them unusual.
  - Example: A sudden spike in network traffic from a group of devices might indicate a cyber attack.



Types of Outliers

- A dataset can have multiple types of outliers

- Different outliers may be used in different applications
  - **Global**: simplest but may not be accurate
  - **Contextual**: require domain knowledge
  - **Collective**: model the behavior of a group of data objects

Challenges of Outlier Detection (I)

- Modeling normal objects and outliers properly
  - The quality of detection depends on how well we model normal data and outliers.
  - It is almost impossible to enumerate all normal data in a dataset.
  - The boundary between "normal" and "abnormal" is not clear.

- Application-specific outlier detection
  - The choice of distance measure and relationship between objects are often application-dependent.
  - It is impossible to develop a universal outlier detection method.

Challenges of Outlier Detection (II)

- Handling noise in outlier detection
  - Outlier provides valuable insights while noise doesn't.
  - Noise may distort the normal objects and blur the distinction between normal objects and outliers, making detection hard.
- Understandability
  - Understand why these are outliers: justification of the detection
  - Specify the degree of an outlier: how unlikely it is for the object to be generated by a normal mechanism

CATEGORIZATION: OUTLIER DETECTION METHODS

Categorization: Different Criteria

- Based on the data labels
  - Supervised, Unsupervised, Semi-supervised (partial labels)

- Based on assumption regarding normal data *vs* outliers
  - **Statistical**: normal data are generated from a statistical model
  - **Proximity-based**: outliers are far away from their nearest neighbors compared to normal data
  - **Clustering-based**: normal data belong to large, dense clusters; outliers belong to small, sparse clusters, or no clusters

(I) Supervised Methods

- Modeling outlier detection as a classification problem
  - Samples examined by domain experts used for training & testing
  - To learn a classifier for outlier detection effectively:
    - Model normal objects and report those not matching the model as outliers, or
    - Model outliers and treat those not matching the model as normal

- Challenges
  - Imbalanced classes: Outliers are rare → Boost the outlier class by generating some artificial outliers for training
  - Recall > Precision: Catch as many outliers as possible, even if it means misclassifying some normal objects as outliers

(I) Unsupervised Methods

- **Intuition**: assume the normal objects are somewhat clustered into multiple groups, each having some distinct features
  - Outliers are expected to be far away from any normal groups
- **Weakness**
  - Normal objects may not share any strong patterns, but the collective outliers may share high similarity in a small area
  - Unsupervised methods may have a high false positive rate but still miss many real outliers.
  - Hard to distinguish noise from outliers
  - Clustering is expensive, but far fewer outliers than normal objects

(I) Semi-Supervised Methods

- **Situation**: in many applications, # labeled data is often limited
  - Labels could be on outliers only, normal objects only, or both.
- If labeled normal objects are available:
  - Use the labeled examples and the nearby unlabeled objects to train a model for normal objects
  - Those not fitting the normal model are flagged as outliers
- If labeled outliers are available:
  - A small number of labeled outliers may not represent all outliers
  - Combine with unsupervised methods to learn a model of normal objects and improve detection accuracy.

(II) Statistical Methods (model-based)

- Assume normal data follow some statistical/stochastic models.
  - Data that do not conform to the model are outliers.

- **Effectiveness**: highly depends on whether the assumption of statistical model holds in the real data

- Statistical modeling
  - **Parametric**: Assume a specific distribution (e.g., Gaussian).
  - **Non-parametric**: Do not assume a specific distribution, offering more flexibility.

(II) Proximity-Based Methods

- An object is an **outlier** if its nearest neighbors of the object are farther away compared to most other objects
  - Proximity: measured by comparing its distance to its neighbors.
  - If the object's proximity significantly deviates from the proximity of most other objects in the same set, it is flagged as an outlier.
- **Effectiveness**: highly relies on the proximity measure
  - Defining proximity measures can be difficult in some applications.
  - Struggles with groups of outliers that are close to each other.
  - Two types: distance-based vs. density-based (density of objects in the surrounding area)

(II) Clustering-Based Methods

- Normal data belong to large, dense clusters
- Outliers belong to small or sparse clusters, or no clusters

- Challenges
  - Clustering is expensive: Clustering methods often have high computational costs, especially for large datasets.
  - Scalability: Straightforward clustering may not scale well to large or high-dimensional datasets.

parametric *vs* non-parametric

STATISTICAL METHODS

Statistical Methods

- Assume that the normal objects in a data set are generated by a stochastic process or a generative model

- Categories
  - Parametric method assumes that the normal data objects are generated by a parametric distribution with parameter θ
    - Example: Gaussian distribution, Poisson distribution.
  - Non-parametric does not assume an a priori statistical model
    - Example: Kernel Density Estimation (KDE), histogram-based

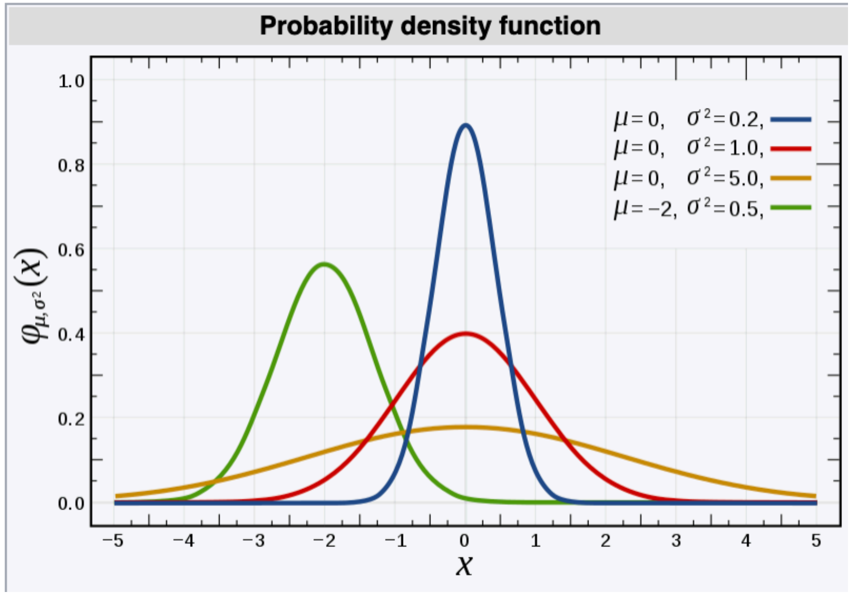
Parametric Method – Normal Distribution

- Widely used in statistics and natural/social sciences to model real-valued random variables with unknown distribution
  - Represented by the probability density function (PDF):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

  - Notation: $X \sim N(\mu, \sigma^2)$

Normal distributions are common in:
- Adult heights
- IQ scores
- Measurement errors
- …

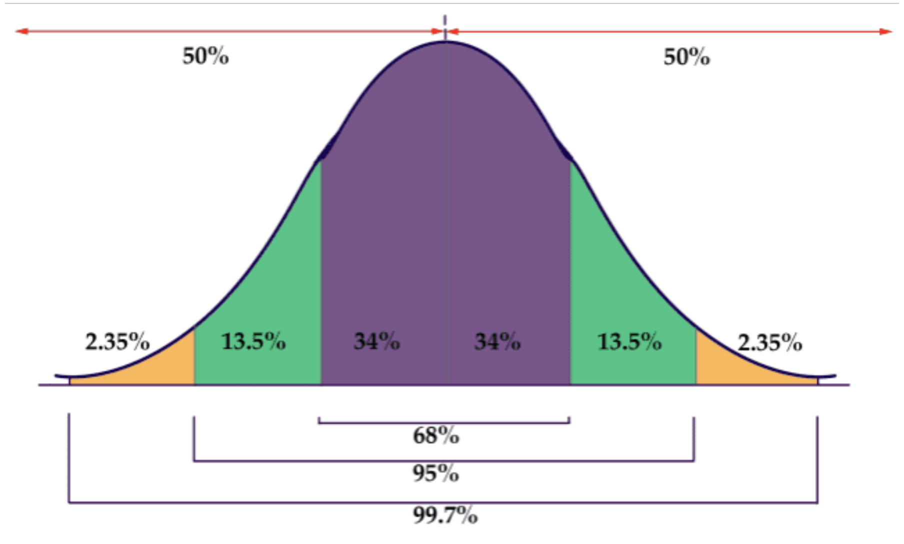


Parametric Method – Normal Distribution

68-95-99.7 Rule: describes the percentage of data falling within 1, 2, or 3 standard deviations of the mean:

- μ *is the mean (center)*
- σ^2 *is the variance (spread)*

- Given a random variable $X \sim N(\mu, \sigma^2)$
  - $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68.27\%$
  - $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95.45\%$
  - $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 99.73\%$



Parametric Method – Normal Distribution

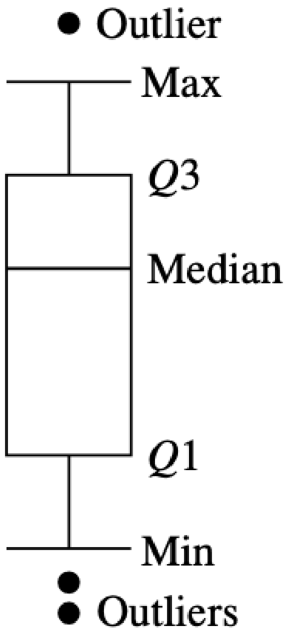
- Given a dataset $x_1, x_2, \ldots, x_n$:
  - Estimate mean: $\hat{\mu} = \bar{x} = \frac{1}{n}\sum x_i$
  - Estimate variance: $\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$

- Outlier detection using 65-95-99.7 rule
  - A data object x_i is considered an outlier if $|x_i - \bar{x}| > 3\sigma$
  - Only 0.3% of data lies beyond 3 standard deviations from mean.
  - This makes x_i highly unlikely belong to this normal distribution.

Parametric Method – IQR and Boxplot

- Given a dataset $x_1, x_2, \dots, x_n$
  - Calculate Q_1 (lower quartile), Q_2 (median), Q_3 (upper quartile)
  - Calculate the interquartile range $IQR = Q_3 - Q_1$
  - **Outliers**: Any data point outside $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$

- **Key idea:** Similar to 68-95-99.7 rule, the range captures most normal data
  - The parameter 1.5 is a typical threshold but could be adjusted accordingly



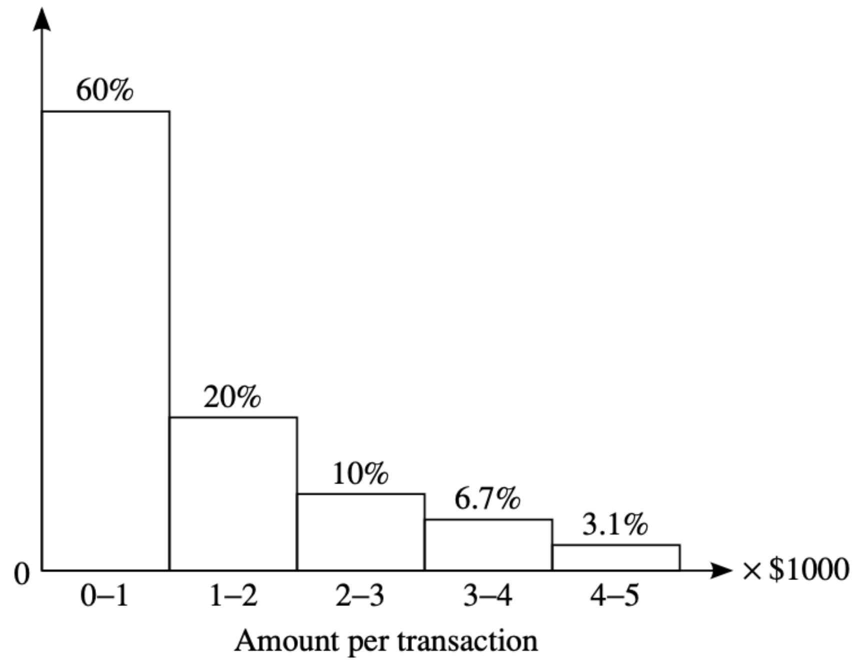
Parametric Method – χ^2 Statistic

- Multivariate data: data involving two or more attributes
  - Transform it to univariate statistic for easier outlier detection

- Given a data object $o = (o_1, o_2, \ldots, o_d)$
  - Calculate $\chi^2 = \sum \frac{(o_i - E_i)^2}{E_i}$
    - o is the observed value and E is the expected value
  - The larger χ^2 is, the more likely o is an outlier

Non-parametric Method – Histogram

- Construct a histogram from the dataset using bins
- **Example**: A transaction over $5,000 can be an outlier since only 0.2% of transactions is over $5,000



- **Challenge**: hard to choose bin size
  - Too small ⇒ Normal data in rare bins
  - Too big ⇒ Outliers in frequent bins

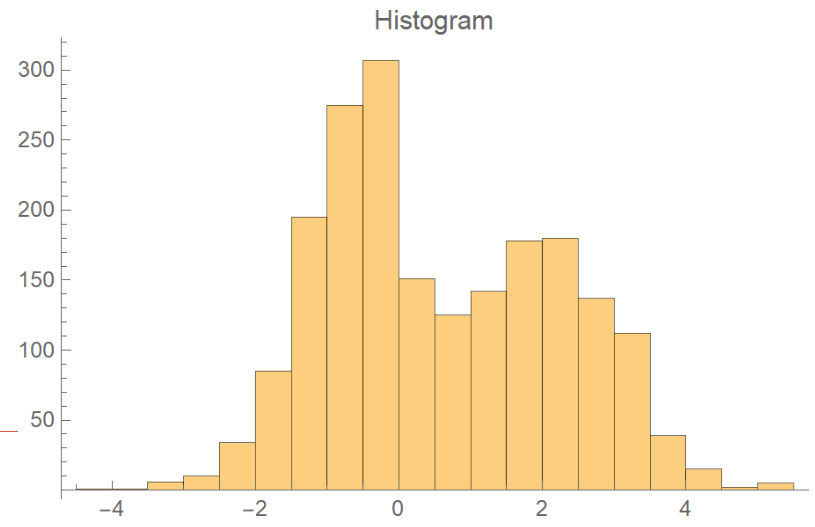
Non-parametric Method – KDE

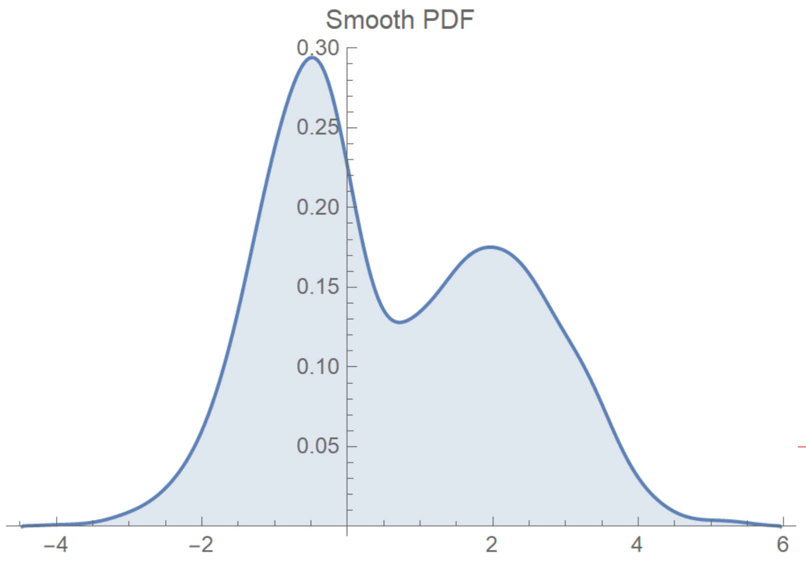
- **Kernel Density Estimation**: A method used to estimate the probability density distribution of the data
  - Every data object contributes to the probability density of others.
  - The contribution of a data object to another object decreases as their distance increases.

- **KDE-based outlier detection:**
  - Given a dataset $x_1, x_2, \dots, x_n$, calculate the estimated probability density function, that is, $\hat{f}_h(x) = \frac{1}{nh} \sum K(\frac{x - x_i}{h})$
    - Bandwidth: h, the larger the bandwidth, the smoother the estimated pdf
  - The lower $\hat{f}_h(x)$ is, the more unlikely x is generated from the distribution

Non-parametric Method – KDE

- Compared to histogram:
  - **Smoothness**: KDE provides a continuous density estimation, while histograms are discrete.
  - **Flexibility**: KDE does not rely on fixed bin sizes, reducing the sensitivity to bin width.
  - **Edge effect**: KDE minimizes abrupt changes at the boundaries.





Summary: Statistical Methods

- **Pros**
  - Statistically justifiable: providing interpretable and reliable results
  - Once the distribution is learned, detection process is fast
- **Cons**
  - Learning process is slow especially for complex distributions
  - Not suitable for high-dimensional data

PROXIMITY-BASED METHODS

Proximity-based Outlier Detection

- Proximity: the degree of nearness/closeness between objects

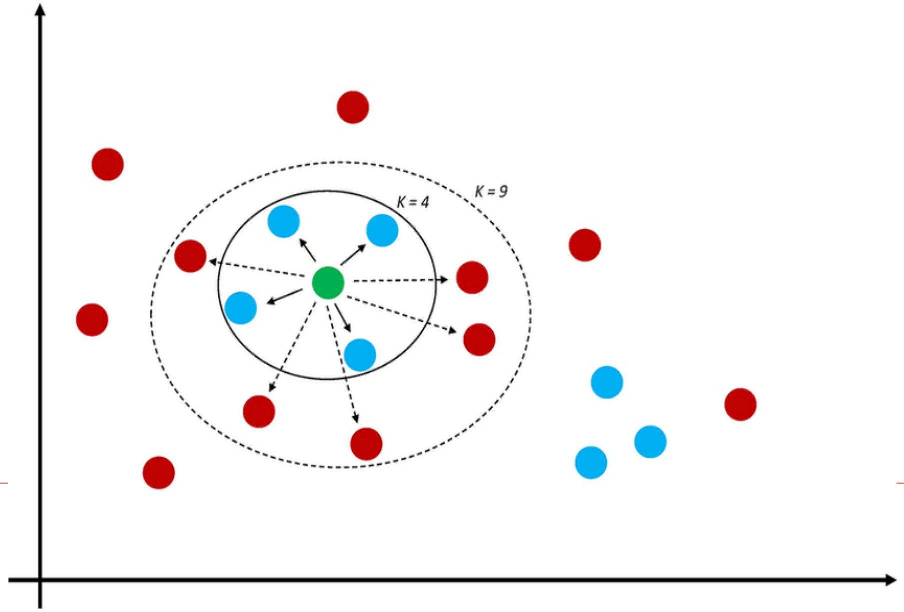
- In data science, it refers to similarity or dissimilarity (distance)
  - Euclidean distance, cosine similarity, jaccard similarity, etc.

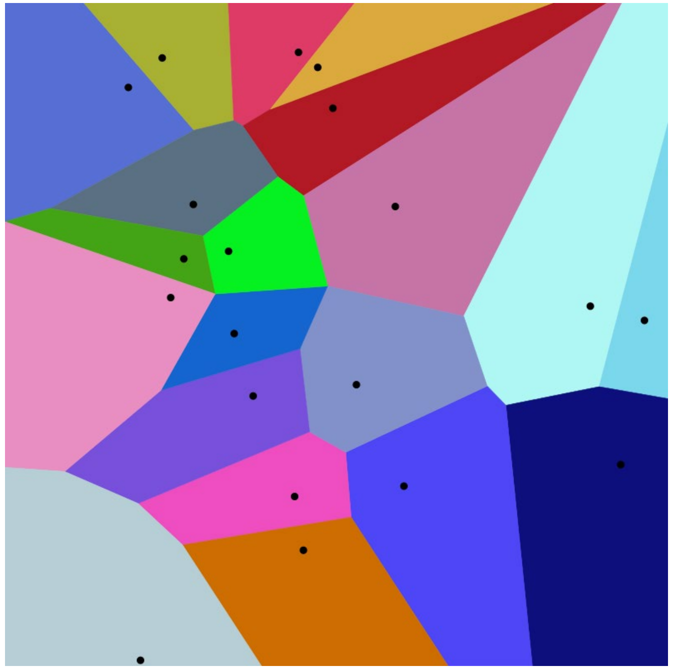
- **Assumption**: the proximity of an outlier object to its nearest neighbors significantly deviates from the proximity of the object to most other objects in the dataset

Nearest Neighbors (NN)

- The nearest neighbor to a data object o is the data object closest to o

- We can extend this concept to k-nearest neighbors.





Voronoi Diagram

Distance-Based Outlier

- Given a data object o and a distance threshold $r \geq 0$, its r-neighborhood is defined as $N_r = \{o' | o' \neq o \wedge dist(o', o) \leq r\}$

- A data object o is a $DB(r, \pi)$-outlier if $\frac{|N_r|}{|D|} < \pi$
  - Fraction threshold: $0 < \pi \leq 1$
  - It suggests o is an outlier if its r-neighborhood contains too few data points compared to the total dataset.

Distance-Based Method: A Nested Loop Algorithm

- For each data object o_i, let $count \leftarrow 0$
  1. Calculate $dist(o_j, o_i)$ for $j \neq i$
  2. If $dist(o_j, o_i) \leq r$, then $count \leftarrow count + 1$
  3. If $count \geq \pi|D|$, exit
  4. Repeat from Step 1
- If not exit before, then o_i is a $DB(r, \pi)$-outlier

Distance-Based Method: A k-NN Algorithm

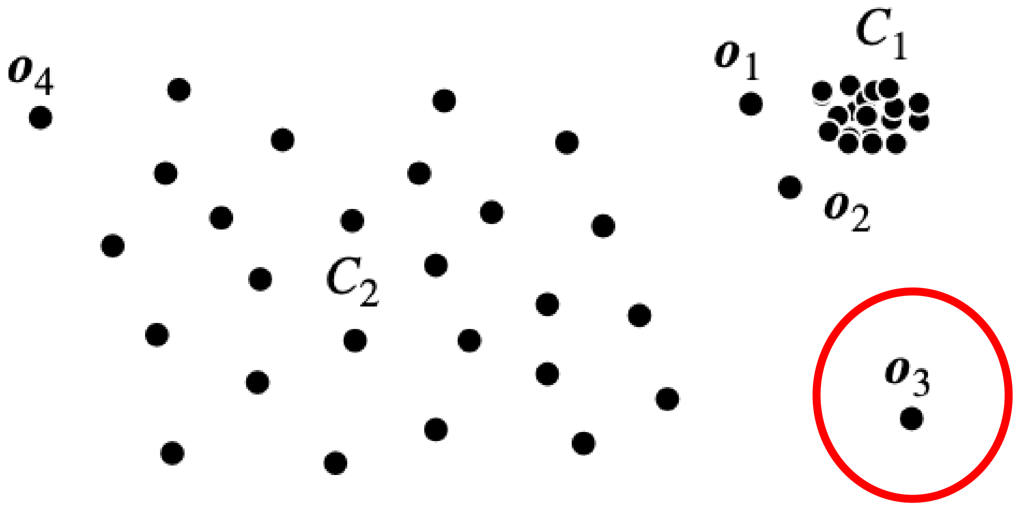
- Determine $DB(r, \pi)$-outlier with k-nearest neighbors
  - A data object o is an outlier if the distance to its k-th nearest neighbor exceeds the distance threshold r, i.e., $dist(o_k, o) > r$
  - The number of neighbors is determined by: $k = \lceil \pi |D| \rceil$

- **Advantages**: Simple and interpretable. Works well for datasets where proximity is meaningful.
- **Challenges**: Computationally expensive for large datasets and may struggle with high-dimensional data.

Density-Based Method

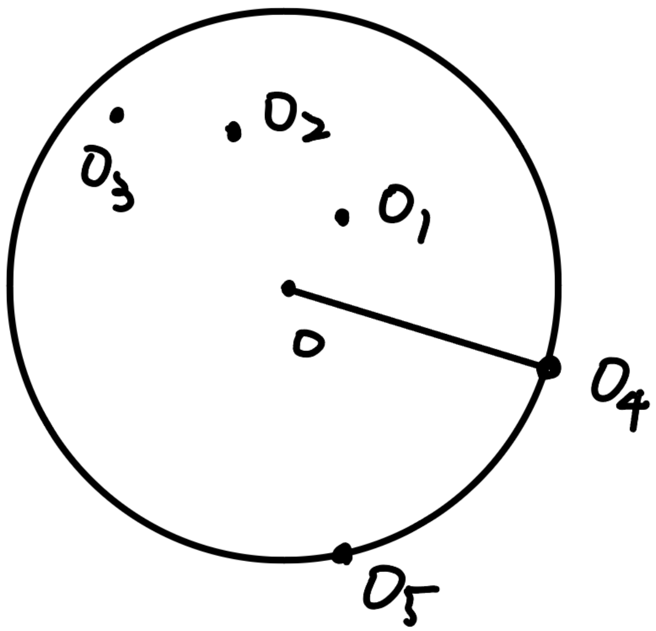
- Distance-based methods discover global outliers
  - The $DB(r, \pi)$-outlier is far from $(1 - \pi) \times 100\%$ of data objects
  - Controlled by two global parameters r and π

- **Density-based methods** assume the density around an outlier object is significantly different from the density around its neighbors.



Density-Based Method

- We define $dist_k(o)$ to be the distance between o and its k-th nearest neighbor
  - $N_k(o) = \{o' | dist(o', o) \leq dist_k(o)\}$
  - Note that $N_k(o)$ can contain more than k data objects.



- **Local density**: the average distance to o in $N_k(o)$
  - It is sensitive to small distance

If the local density of o is significantly lower than its nearest neighbors, it is an outlier.

Summary: Proximity-Based Methods

- **Pros**
  - Understandable to humans
  - Non-parametric. No assumptions on the data distribution
  - Flexible to different proximity measurements
- **Cons**
  - Computation cost can be high especially in high-dimensional space
  - Not suitable for collective outliers detection

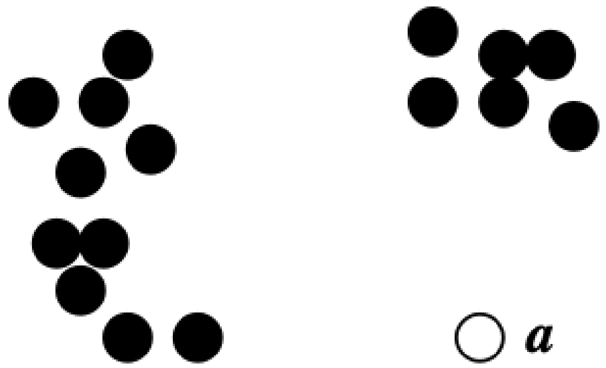
CLUSTERING-BASED METHODS

Clustering-Based Outlier Detection

- Clustering-based methods examine the relationship between data objects and clusters
  1. If a data object doesn't belong to any cluster, it is an outlier
  2. If a data object is far from its nearest cluster, it is an outlier
  3. If a data object belongs to a small or sparse cluster, all objects in that cluster are outliers

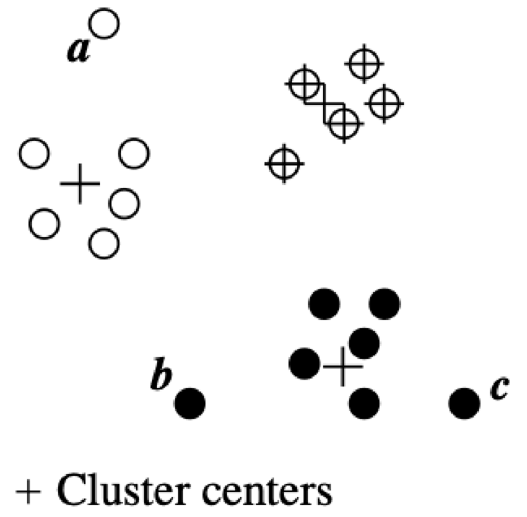
Case 1: Not Belong to Any Cluster

- Using clustering methods like DBSCAN, some data points may not belong to any cluster.
- These unclustered points are considered outliers.
  - Consider organizing a library. Most books fit into well-defined categories, such as "Fiction" or "Science.“
  - However, a rare, unrelated book that doesn’t belong to any category (like a handwritten manuscript) would be an outlier.



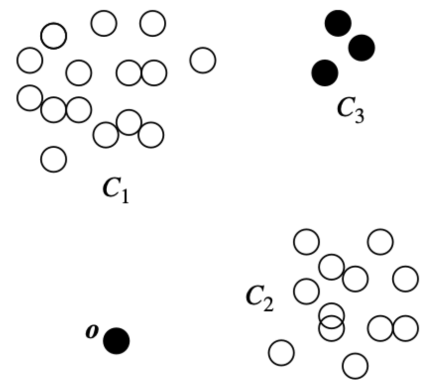
Case 2: Far Away from Nearest Cluster

- k-means clustering is sensitive to outliers, as points far from cluster centers may distort the clustering process.
- We can find the data objects who is far away from the cluster center as outliers
  - Unusual activity (e.g., rare login attempts or abnormal patterns) often appears far from established clusters of normal behavior.



Case 3: Outliers in Small Clusters

- The above two cases focus on detecting individual outliers
- Cluster-based local outlier factor (*CBLOF*)
  - Clusters are divided into large and small
    - based on # objects they cover
  - For an object o in large cluster: $CBLOF = \#\, objects \times sim(o, C)$
    - C is the cluster o lies in
  - For an object o in small cluster: $CBLOF = \#\, objects \times sim(o, C)$
    - C is the nearest large cluster of o



Summary: Clustering-Based Methods

- **Pros**
  - Unsupervised. Suitable for any type of data
  - Clusters can be regarded as a summary of data and help other tasks.
  - Detection process is fast as # cluster is typically small
- **Cons**
  - Effectiveness is limited since the labels are missing.

MINING CONTEXTUAL AND COLLECTIVE OUTLIERS

Contextual Outlier Detection

- The attributes of data objects are divided into two groups
  - Context attribute: e.g. longitude, latitude, time, etc.
  - Behavioral attribute: e.g. temperature

- How to analyze the corresponding contextual information?
  - In some scenarios, the contexts can not be clearly identified

Extending Conventional Outlier Detection

- When the contexts can be clearly identified
  - Identify the contexts of data objects using contextual attributes
  - Apply a conventional outlier detection

Example: Is 28℃ an outlier for Hong Kong in April?

- First find all data objects whose "City" equals "Hong Kong" and "Month" equals "April"
- Apply a conventional outlier detection on these selected data objects

City	Month	Temperature
Hong Kong	April	28.7℃
Hong Kong	April	26℃
Tokyo	March	12℃
……	……	……

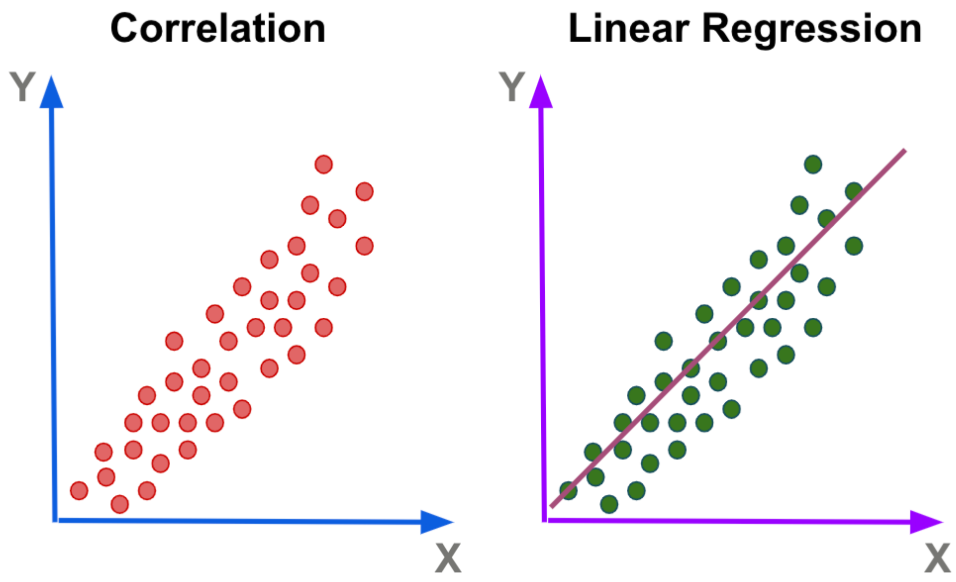
Modeling Normal Behavior *wrt* Contexts

- When the contexts cannot be clearly identified
  - E.g. finding an abnormal purchase *wrt* to the browser log
  - There is no straightforward way to determine how much of a customer's browsing history should be considered

- Use a predictive model to predict the purchase based on the browser log
  - If an actual purchase is significantly different from the prediction, it can be considered as an outlier。

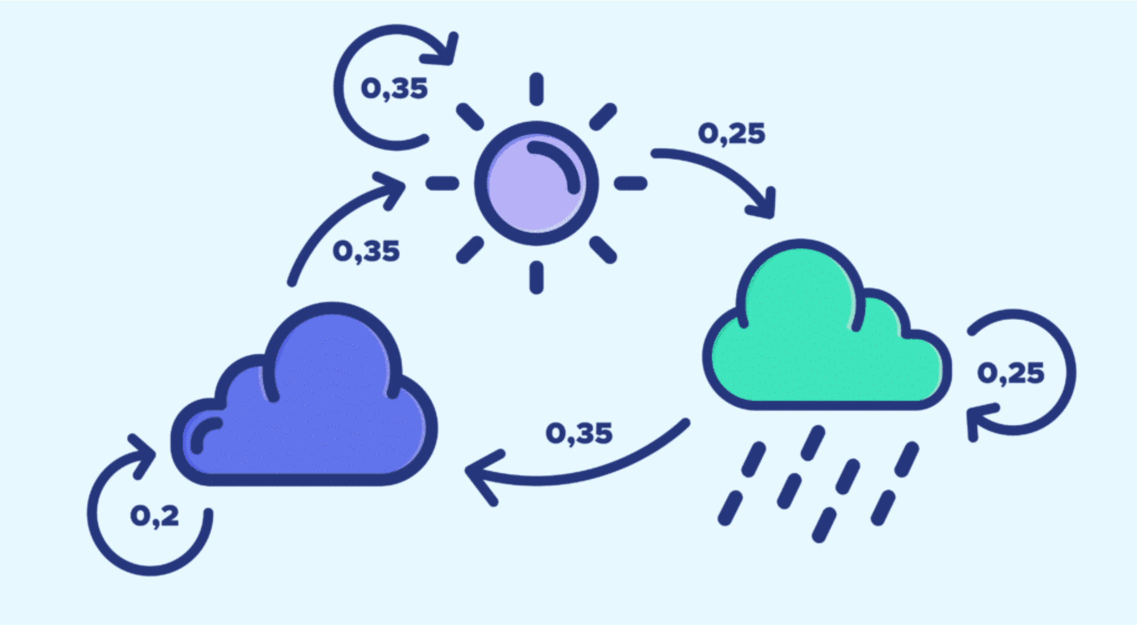
Modeling Behavior – Regression Analysis

- Regression analysis reveals the correlation between data objects
  - Assume the purchase highly depends on the browser log
  - Learn a regression model to predict purchase behavior based on browser log
  - Find outliers if actual purchases deviate significantly from prediction



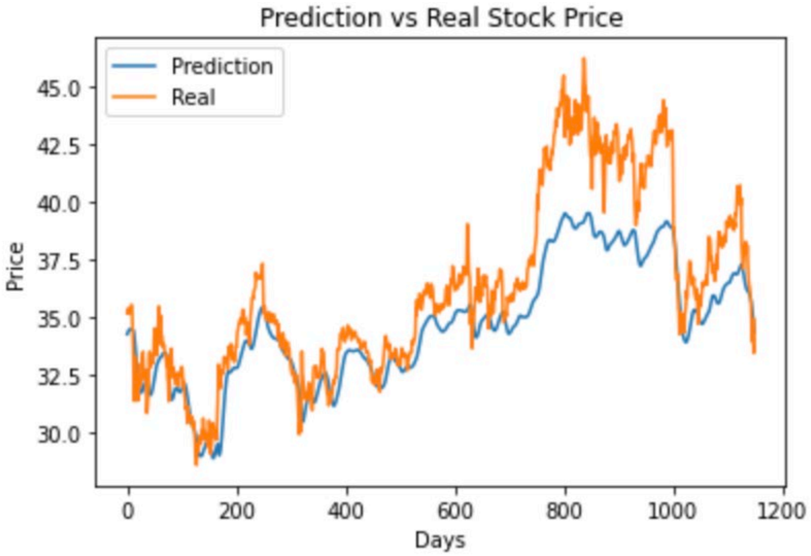
Modeling Behavior – Markov Models

- Markov property: $P(X_n|X_{n-1},\dots,X_1) = P(X_n|X_{n-1},\dots,X_{n-k})$
  - The number k is called the order
- Based on Markov property, we can learn a Markov model to represent the transition probability from one product to another, or from one product to the purchase



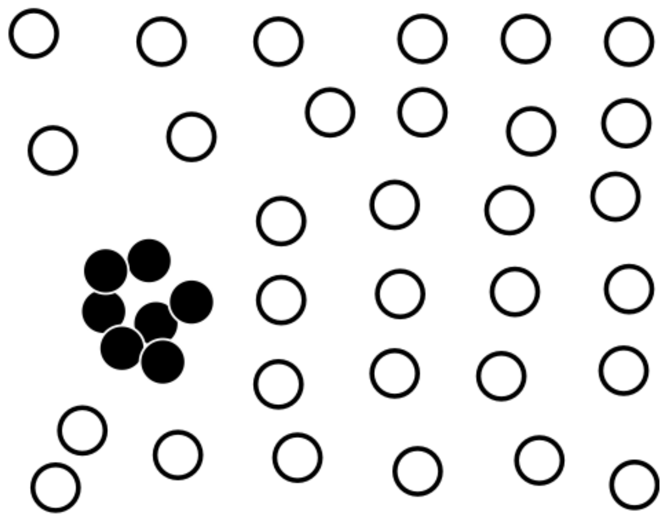
Modeling Behavior – LSTM

- To handle the fixed-order issue of Markov models, we can use a recurrent neural network (RNN) called LSTM
- Long Short-Term Memory (LSTM) network aims to provide a short-term memory for RNN that can last thousands of timesteps
  - LSTM is good at time-series data prediction



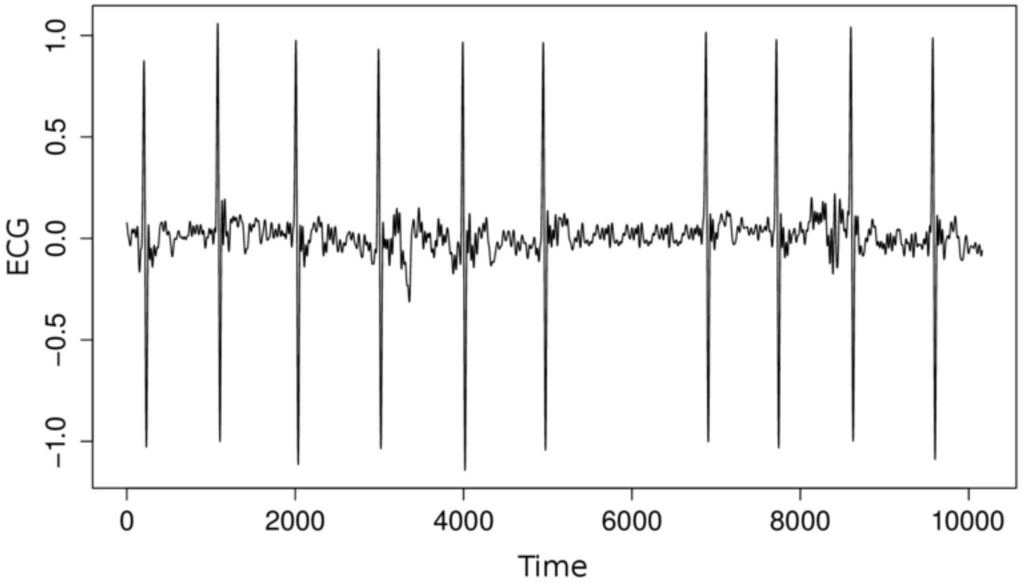
Collective Outlier Detection

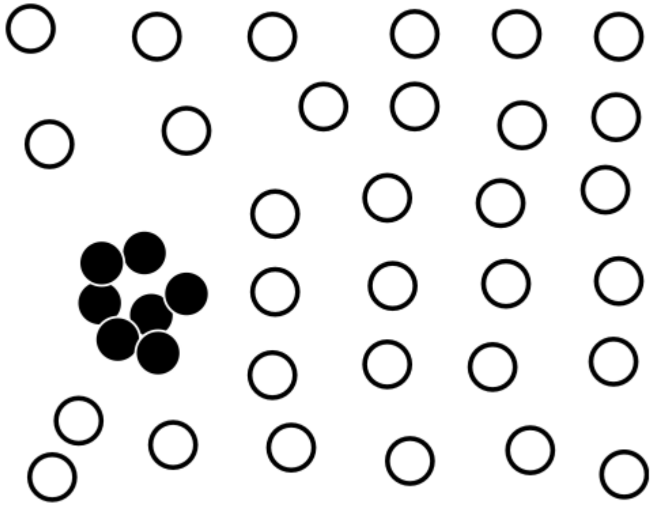
- A group of data objects as a whole deviating significant from the entire dataset
- Question: How can we identify the abnormal behavior of a group of data objects
  - We need to examine the structure of the dataset



Collective Outlier Detection

- Structure
  - Temporal data: sub-sequences
  - Spatial data: local areas
  - Graph and network data: subgraphs





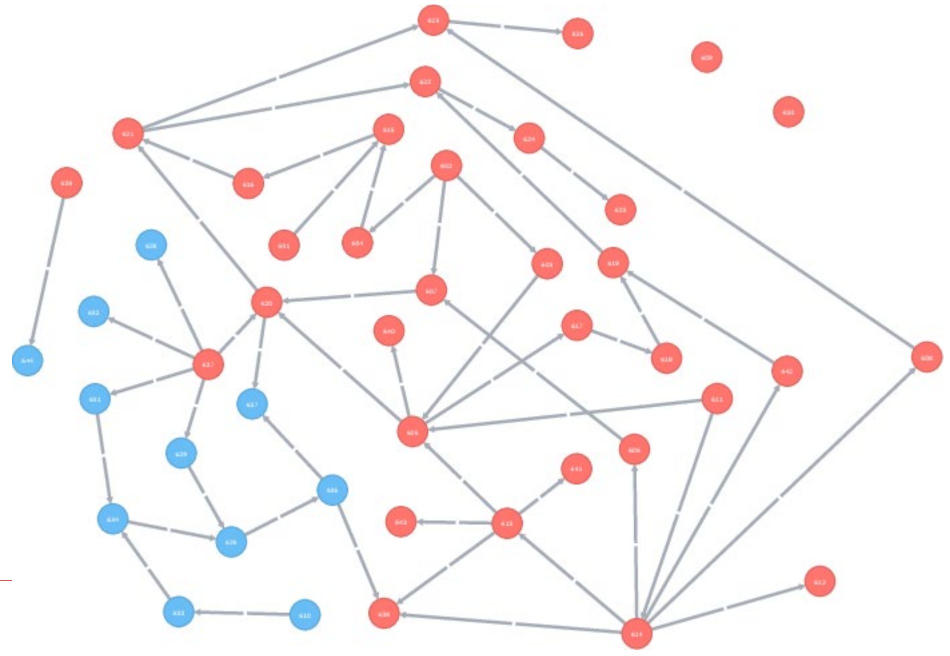
Extending Conventional Outlier Detection

- We first break the entire dataset into small "structure units"
  - Sub sequences, local areas, subgraphs, etc.

- Then we can conduct conventional outlier detection on these "structured objects"

Example: Graph outlier detection

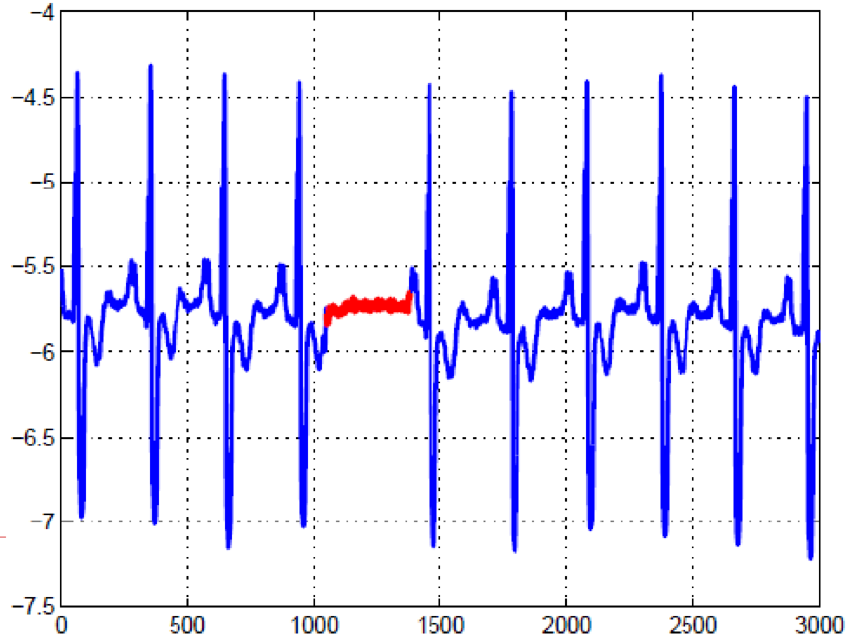
- Structure units: subgraphs with different number of nodes
- Typically, the number of subgraphs decreases while more nodes and edges are involved
- Find the outlier with unexpected #subgraph



Modeling the Normal Behavior

- In real scenarios, the number of structure units can be huge and it's impossible to examine all of them

- We can use predictive models (Markov models, LSTM, etc.) to model and predict normal data



OUTLIER IN HIGH-DIMENSIONAL SPACE

Curse of Dimensionality & High Dimensions

- **Data objects are sparse in high-dimensional space**
  - Hard to understand the structure of data: making it difficult to cluster, classify, or understand relationships.
  - Sensitive to noises: Small variations in high-dimensional data can drastically affect results due to the sparsity.

- **Visualization challenge:** Humans can’t intuitively visualize data beyond three dimensions.

Curse of Dimensionality & High Dimensions

- **Distance metrics:** The distance becomes meaningless!
  - In high dimensions, the difference between the nearest and farthest neighbors diminishes.
  - $\lim_{d \to \infty} E\left(\frac{dist_{max} - dist_{min}}{dist_{min}}\right) \to 0$

- **Combinatorial explosion**: High-dimensional data leads to an exponential increase in # feature combinations to analyze.
  - 30-dimensional space → $2^{30} \approx$ one billion possible combination!
  - Most attributes are irrelevant attributes

Challenges for High-Dimensional Outlier Detection

- Detecting outliers without saying why they are outliers is not very useful in high-dimensional space.
  - Many dimensions may have irrelevant or noisy features.
- Data sparsity: Noise dominates, making it hard to distinguish true outliers from noisy data.
- Subspaces and Scalability
  - Outliers often exist in specific subspaces, not in the full space.
  - Efficient exploration of subspaces is vital for meaningful outliers.

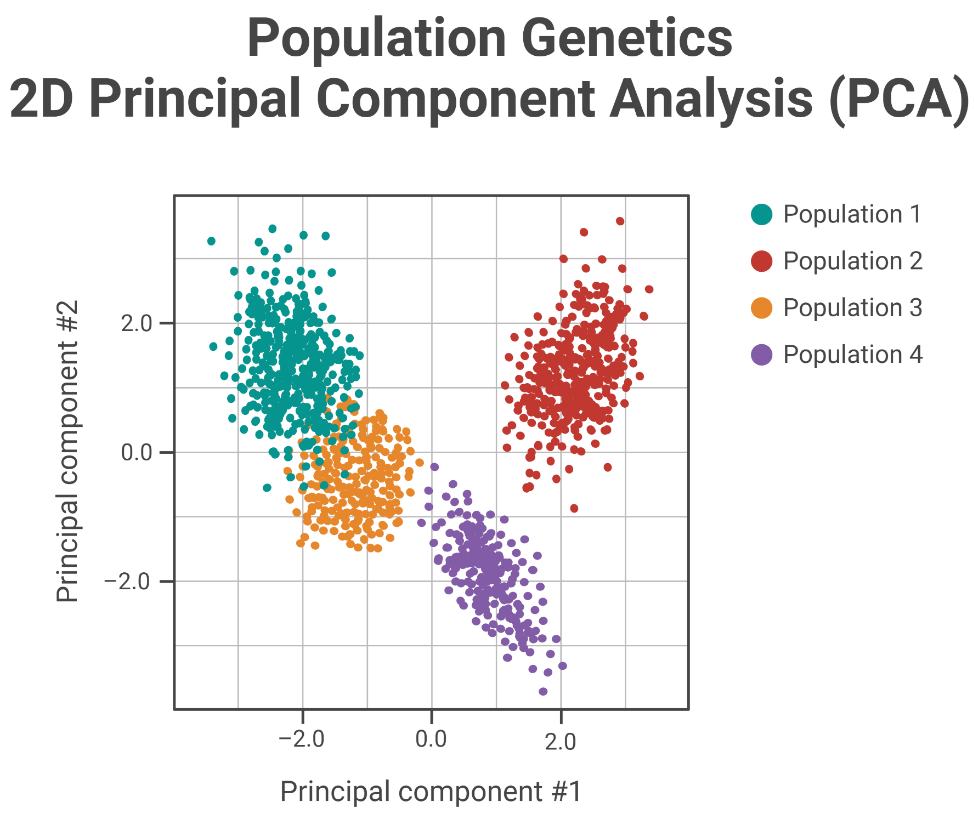
Extending Conventional Outlier Detection

- **Method 1**: Detect outliers in the full space (e.g., HilOut*)
  1. Find distance-based outliers, but use the ranks of distance instead of the absolute distance in outlier detection
  2. For each object o, find its k-nearest neighbors (nn denotes "nearest neighbors"): $nn_1(o), \ldots, nn_k(o)$
  3. Compute the weight of object o: $w(o) = \sum_{i=1}^{k} dist\left(o, nn_i(o)\right)$
  4. Rank all objects in weight-descending order
  5. Select the top-l objects as outliers (l: user-specified parameter)

*Angiulli, Fabrizio, and Clara Pizzuti. "Outlier mining in large high-dimensional data sets." *IEEE TKDE.* 17.2 (2005): 203-215.

Extending Conventional Outlier Detection

- **Method 2**: Dimensionality reduction
  - PCA-based Heuristic Approach: Principal components with low variance are preferred because:
    - Normal objects tend to cluster closely in these selected dimensions
    - Outliers are more likely to deviate significantly from the majority.



Finding Outliers in Subspaces

- Detecting outliers in full-dimensional spaces is hard to interpret.

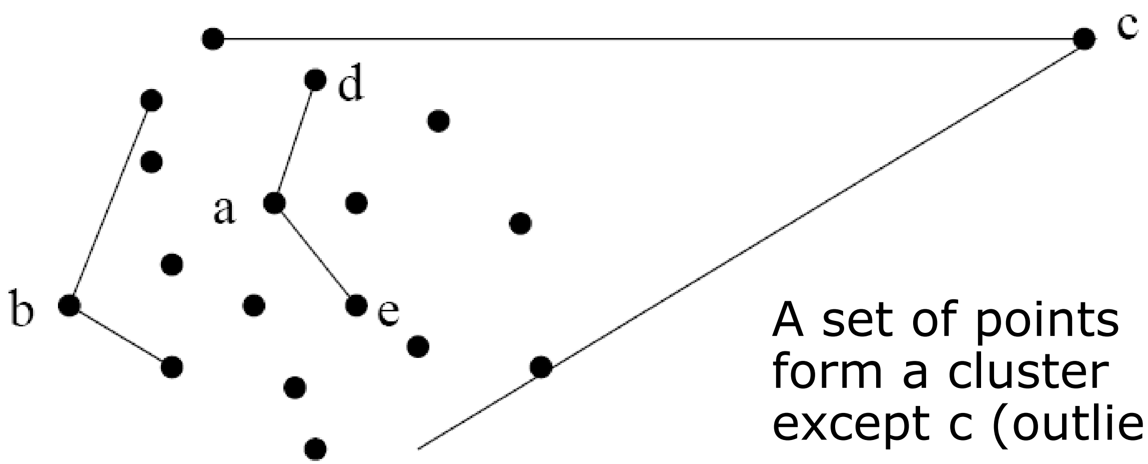
- Find outliers in much lower dimensional subspaces
  - Understand why the object is an outlier
  - Quantify to what extent it deviates from normal behavior
  - Example: Find "outlier customers" in certain subspace
    - **average transaction amount >> avg.**
    - **purchase frequency << avg.**

Finding Outliers in Subspaces: Example

- A grid-based subspace outlier detection method to identify an area with significantly lower density than the average
  - Project data onto various subspaces
  - Discretize the data into a grid with equal-depth regions
    - For each dimension, create ϕ partitions → each contains $f = \frac{1}{\phi}$ of total data
  - Search for regions that are significantly sparse
    - For an r-dimensional subspace, the expected # objects in a cell is $f^r n$
    - Calculate a sparsity score $S(C) = \frac{n(C) - f^r n}{\sqrt{f^r(1-f^r)n}}$, where $S(C) < 0$ means a sparse cell C that may contain outliers

Modeling High-Dimensional Outliers

- Develop new models for high-dimensional outliers directly
  - Avoid proximity measures and adopt new heuristics that do not deteriorate in high-dimensional data



A set of points form a cluster except c (outlier)

Summary

- Types of outliers: Global, Contextual, Collective
- Detection methods: Supervised/Unsupervised/Semi-supervised
  - Statistical methods: Parametric / Nonparametric
  - Proximity-based methods: Distance- / Density-based
  - Clustering-based methods: Three cases
  - High-dimensional outlier detection

Email: fengmei.jin@polyu.edu.hk

Office: PQ747

THANK YOU!