Data Mining Unit 5 Handwritten

Data Mining And Analytics (SRM Institute of Science and Technology)

Data Mining

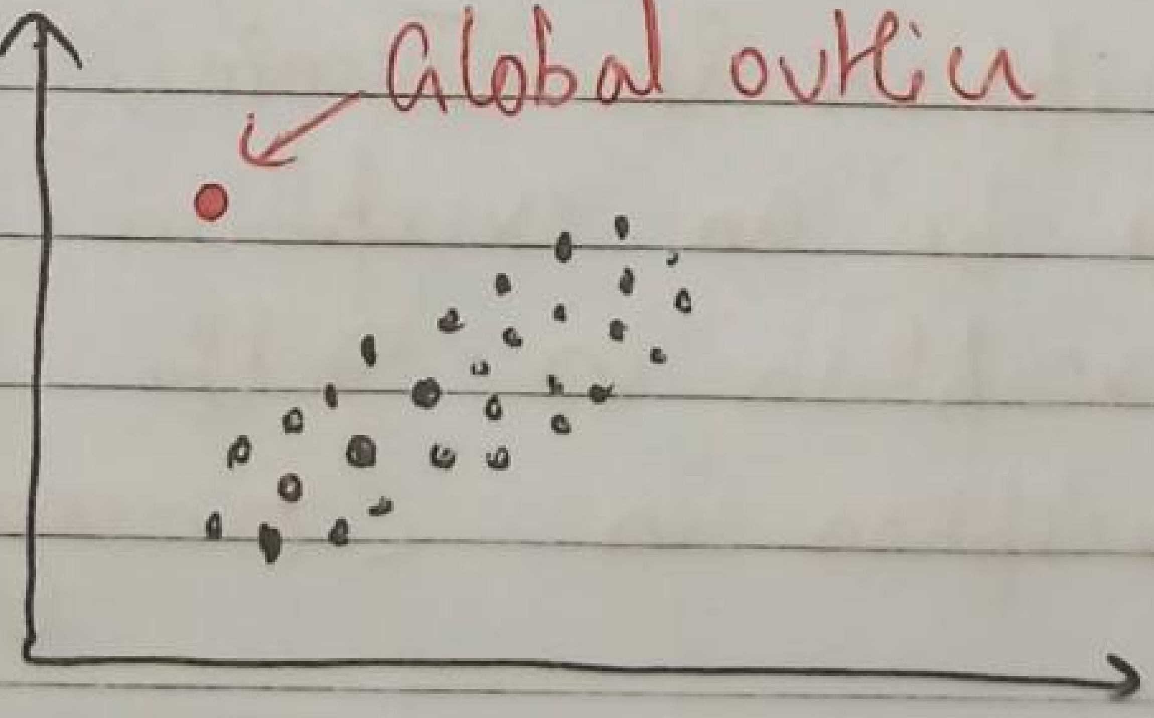
Unit - 5

Outlier :- An outlier is an data object that deviates ~~of~~ significantly from the rest of the data objects and behave in a different manner.

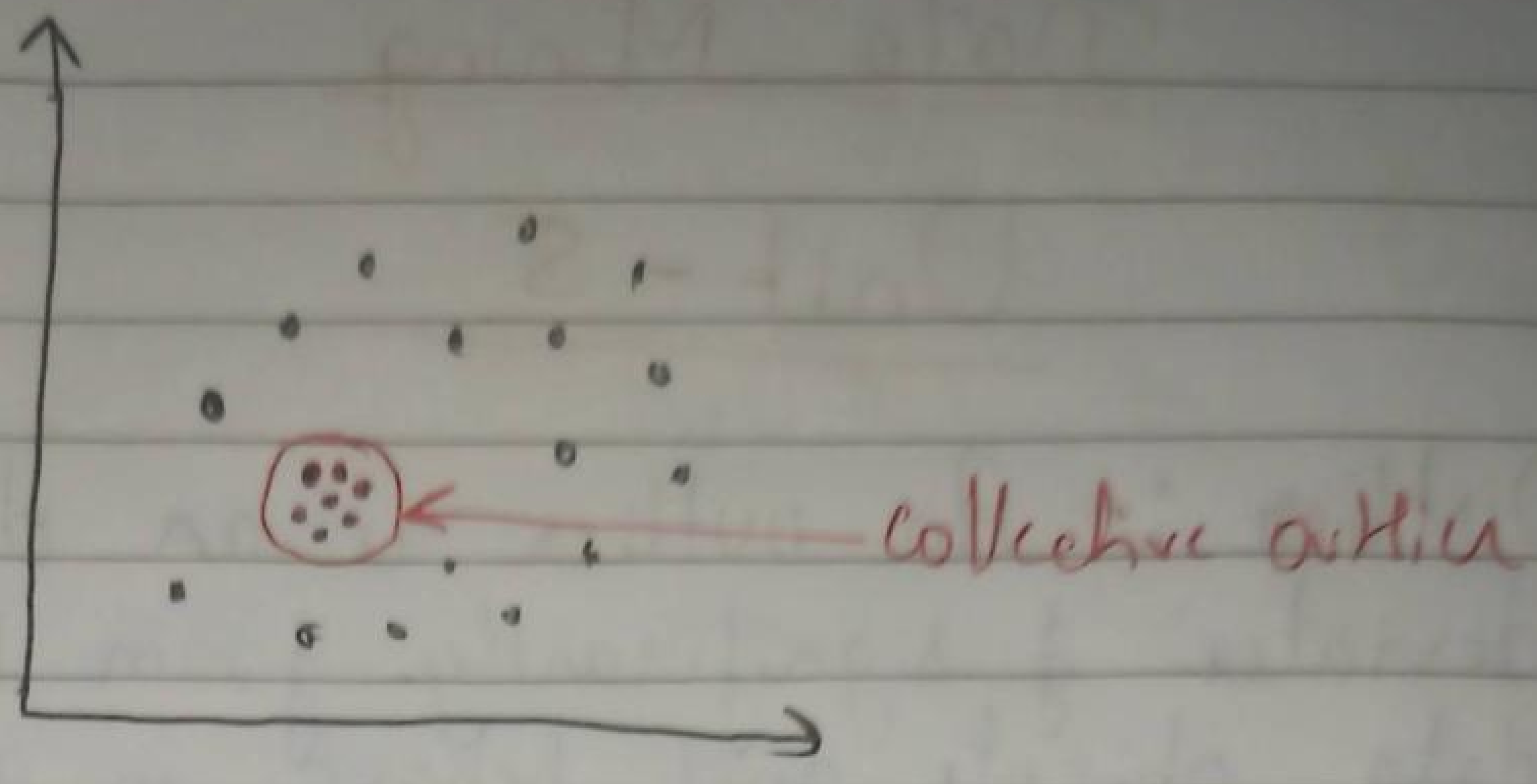
They are caused by measurement or execution error. The analysis of data outlier is reffered to as outlier analysis or outlier mining.

Types of Outlier

1) Global Outlier :- The outlier that deviates from complete database is called Global outlier. It is also known as Point anamolies.



2) Collective Outlier :- As name suggest, if in a given dataset some of the data points, as a whole, deviates from rest of dataset, they may be termed as collective outliers. Here, if we see one point it might not be an outlier but upon seeing as a collection of point it will behave as an outlier.



3) Contextual Outlier :- It is also known as Conditional outlier.
If in a given data set a data object deviates significantly from the other data points based on a specific context or condition only, then that data object is said to be an contextual outlier.

Ex → Today's ~~a~~ temperature is 28°C.
↑
It depend on various parameter (context) like time, location. for example 28°C may not be exception in Summer but in Dec & Jan it might be an exception. Also it may not be exception in Delhi But it may be ~~ex~~ exception in Srinagar.

Therefore a context must be specified in order to identify contextual outlier.
The attributes used ~~for~~ as parameters are of two types :-

i) Contextual attribute → time, location
ii) Behavioral attribute

Challenges of Outlier Detection-:

Outlier detection is useful in many applications yet faces many challenges such as

Modeling normal objects and outliers effectively

- Outlier detection quality highly depends on the modeling of normal (non outlier) objects and outliers. Often, building a comprehensive model for data normality is very challenging, if not impossible.
- This is partly because it is hard to enumerate all possible normal behaviors in an application.
- The border between data normality and abnormality (outliers) is often not clear cut. Instead, there can be a wide range of gray area.
- Consequently, while some outlier detection methods assign to each object in the input data set a label of either "normal" or "outlier," other methods assign to each object a score measuring the "outlier-ness" of the object.

Application-specific outlier detection

- Technically, choosing the similarity/distance measure and the relationship model to describe data objects is critical in outlier detection. Unfortunately, such choices are often application-dependent.
- Different applications may have very different requirements.
- For example, in clinic data analysis, a small deviation may be important enough to justify an outlier. In contrast, in marketing analysis, objects are often subject to larger fluctuations, and consequently a substantially larger deviation is needed to justify an outlier.

Handling noise in outlier detection

- As mentioned earlier, outliers are different from noise. It is also well known that the quality of real data sets tends to be poor.
- Noise often unavoidably exists in data collected in many applications.
- Noise may be present as deviations in attribute values or even as missing values.
- Low data quality and the presence of noise bring a huge challenge to outlier detection.
- They can distort the data, blurring the distinction between normal objects and outliers.

Understandability

- To meet the understandability requirement, an outlier detection method has to provide some justification of the detection.
- For example, a statistical method can be used to justify the degree to which an object may be an outlier based on the likelihood that the object was generated by the same mechanism that generated the majority of the data.
- The smaller the likelihood, the more unlikely the object was generated by the same mechanism, and the more likely the object is an outlier.

Outlier Detection Methods :-

i) Supervised Method :-

This method model data normality and abnormality. The task is to learn a classifier that can recognize outliers. The sample is used for training and testing.

Basically we can have 2 ways to deal with it. First is to label normal object and mark other object not matching the ~~other~~ model as outlier. Secondly, we can model the outlier and any object not matching them can be marked as normal object.

Challenges in this Model :-

(i) Two classes are imbalanced. Therefore methods for handling imbalanced class may be used such as oversampling Outlier to increase their distribution in training set used to construct ~~a~~ classifier. The lack of sample of outlier can limit the capability of classifiers built as such.

~~II~~

2) Unsupervised Method :-

In some application, ~~ob~~ scenarios, objected labeled as normal or outlier are not available. Thus an unsupervised method has to be used.

Unsupervised outlier detection method assumes the normal object as somewhat clustered. In other words, it ~~except~~ expect normal object follow a pattern more frequently than outlier.

Unsupervised Methods-:

- In some application scenarios, objects labeled as “normal” or “outlier” are not available. Thus, an unsupervised learning method has to be used.
- Unsupervised outlier detection methods make an implicit assumption: The normal objects are somewhat “clustered.” In other words, an unsupervised outlier detection method expects that normal objects follow a pattern far more frequently than outliers.
- Normal objects do not have to fall into one group sharing high similarity. Instead, they can form multiple groups, where each group has distinct features. However, an outlier is expected to occur far away in feature space from any of those groups of normal objects.
- For instance, in some intrusion detection and computer virus detection problems, normal activities are very diverse and many do not fall into high-quality clusters. In such scenarios, unsupervised methods may have a high false positive rate—they may mislabel many normal objects as outliers (intrusions or viruses in these applications), and let many actual outliers go undetected.
- Due to the high similarity between intrusions and viruses (i.e., they have to attack key resources in the target systems), modeling outliers using supervised methods may be far more effective.
- Many clustering methods can be adapted to act as unsupervised outlier detection methods.
- The central idea is to find clusters first, and then the data objects not belonging to any cluster are detected as outliers. However, such methods suffer from two issues. First, a data object not belonging to any cluster may be noise instead of an outlier. Second, it is often costly to find clusters first and then find outliers.
- It is usually assumed that there are far fewer outliers than normal objects.
- The latest unsupervised outlier detection methods develop various smart ideas to tackle outliers directly without explicitly and completely finding clusters.

Semi-Supervised Methods-:

- In many applications, although obtaining some labeled examples is feasible, the number of such labeled examples is often small.
- We may encounter cases where only a small set of the normal and/or outlier objects are labeled, but most of the data are unlabeled. Semi-supervised outlier detection methods were developed to tackle such scenarios.
- For example, when some labeled normal objects are available, we can use them, together with unlabeled objects that are close by, to train a model for normal objects. The model of normal objects then can be used to detect outliers—those objects not fitting the model of normal objects are classified as outliers.
- If only some labeled outliers are available, semi-supervised outlier detection is trickier. A small number of labeled outliers are unlikely to represent all the possible outliers.
- Therefore, building a model for outliers based on only a few labeled outliers is unlikely to be effective. To improve the quality of outlier detection, we can get help from models for normal objects learned from unsupervised methods.

Statistical → Box Plot → already Studied

Univariate Outlier Detection :-

1) Maximum Likelihood :-

Q. Suppose a city temp value in last 10 years are

24, 28.9, 28.9, 29, 29.1, 29.1, 29.2, 29.2, 29.3, 29.4

Sol →

$$\mu = \frac{24+28.9+28.9+29+29.1+29.1+29.2+29.2+29.3+29.4}{10}$$

$$= 28.61$$

$$\sigma^2 = 2.29$$

$$\sigma = \sqrt{2.29} = 1.51$$

Most deviating value is 24.

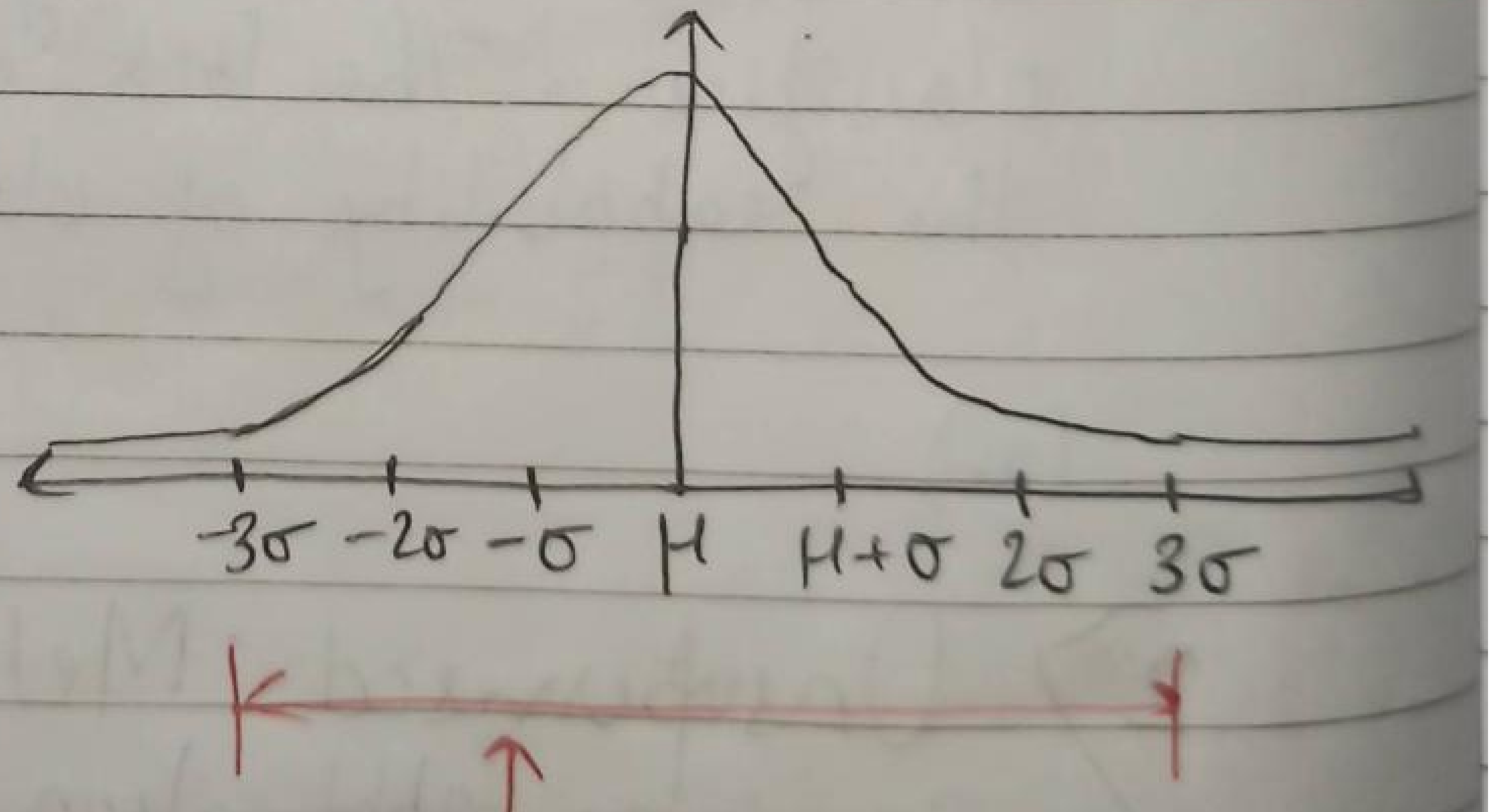
28.61
− 24.00
4.61

1.51
× 3
4.53

4.61 > 4.53

∴ 24 lie behind 3σ

⇒ Outlier



99% data is contained in this μ ± 3σ region

Likehood junction :-

$$L(\mu, \sigma^2) = \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

2) Grubb test :- (Maximaum Normal Residue) :-

$$Z\text{-}score = \frac{|x - \bar{x}|}{s}$$

$$Z > \frac{N-1}{N}\sqrt{\frac{t^2_{(\alpha/2N,\, N-2)}}{N-2+t^2_{(\alpha/2N,\, N-2)}}}$$

from t-table

If $Z_{cal} > Z_{table}$

$\Rightarrow$ that point is an outlier.

Multivariat Data Outlier Detection :-

1) χ^2- test :-

If x^2 value is greater, then there is an outlier

2) Mahalanobis distance :-

object Mean vector

$$MDist(O, \bar{O}) = (O - \bar{O})^T S^{-1} (O - \bar{O})$$

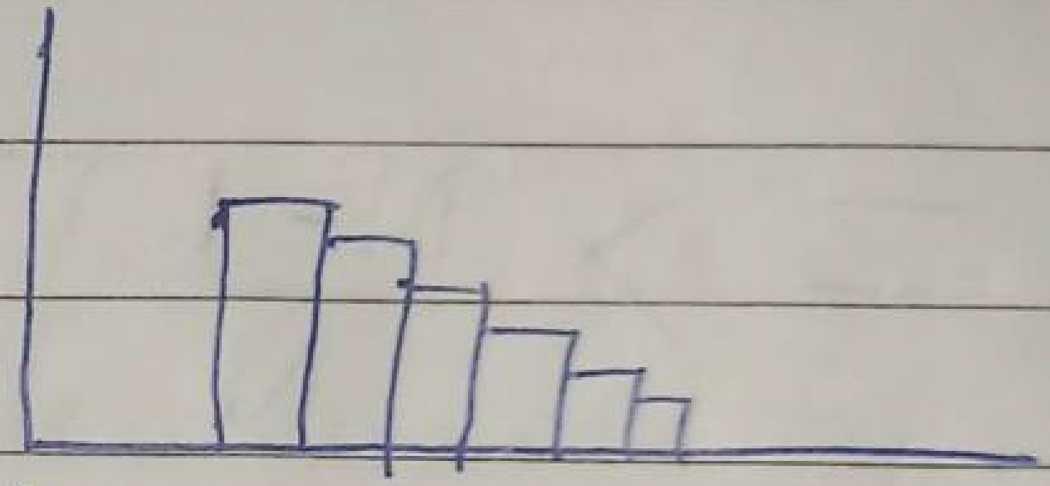
Covariance Matrix

Steps :-

1. Comput mean vector from Multivariat dataset.
2. For each 0, calculat MDis(0, $\bar{0}$)
3. Detect outlier in transformed univariat dataset
4. If MDist(0, $\bar{0}$) is determined to be outlier then 0 is regard as outlier as well.

Non-Parametric Methods :-

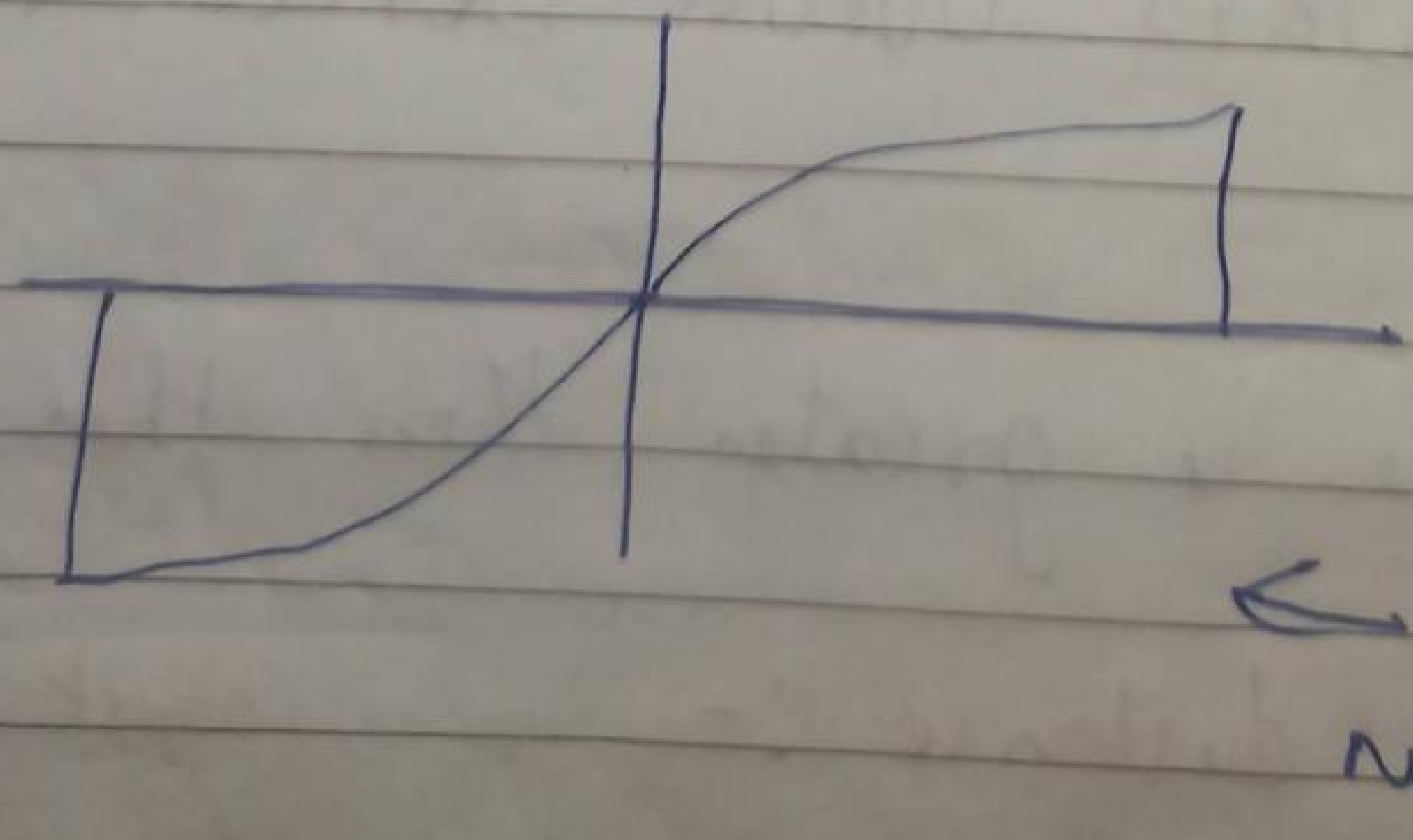
Histogram :-



1. Creat bins & put data inside Bins.
2. If data is place in this then it is not outlier else it is an outlier.

Drawback :- Bins should be correct else will result in wrong output.

Kernel estimation function :-



← Area of curve should not be 0.

Proximity-Based Methods :-

→ We define some threshold if it exceeds that threshold, then that is outlier.

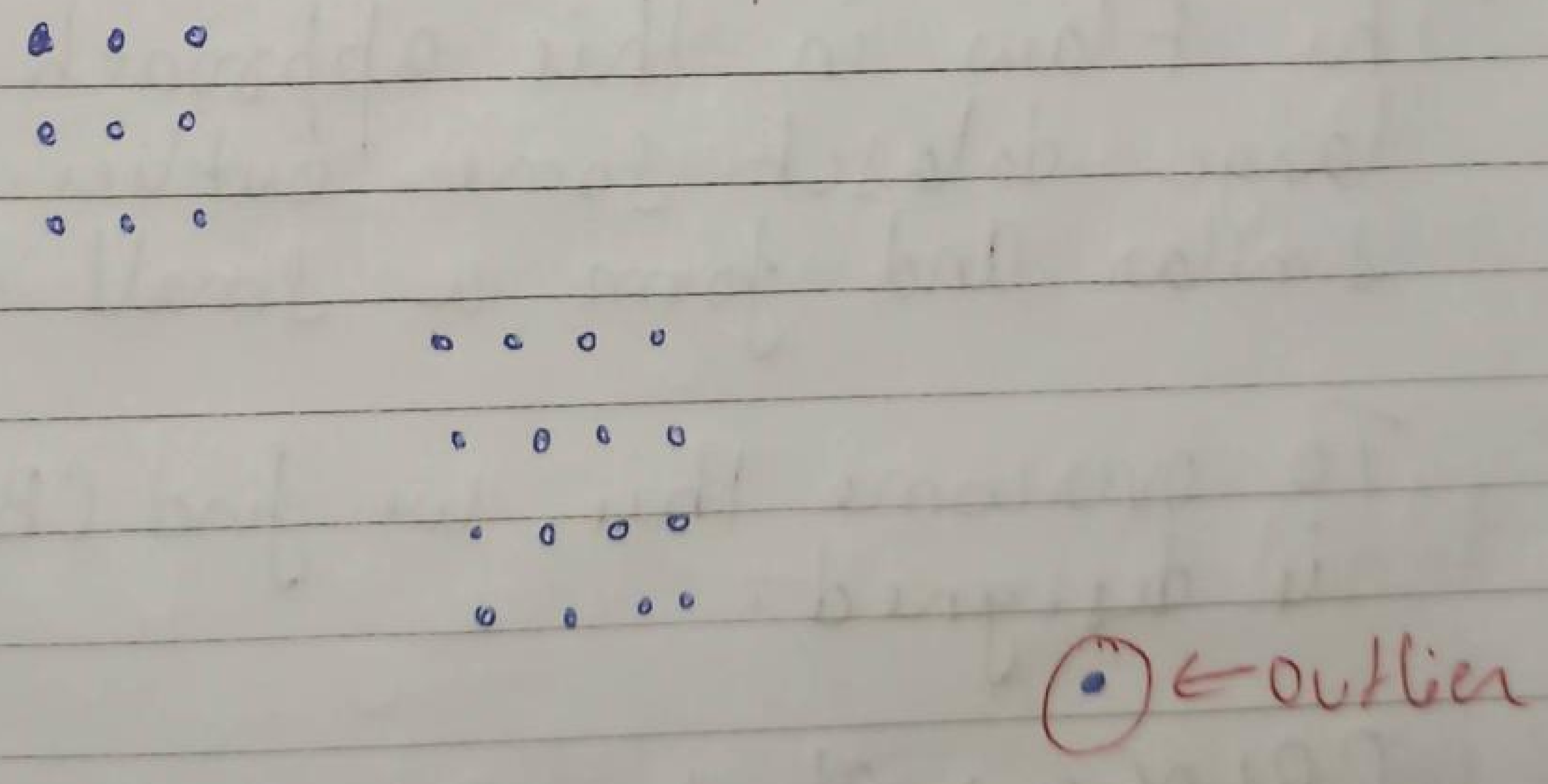
→ Example → Mobile phone sensor, while bringing screen to ears, light go off.

→ We apply loop for each data point & then decide.

Drawback :- • Complexity $O(n^2)$.

Statistically approach's Complexity → $O(n)$.

Density Based Outlier :-



Application of Outlier detection :-

1) Intrusion Detection System :-

1. training dataset is used to find pattern of normal data.
TCP connection data segments according to, say, dates
Frequent itemset that are in majority of segment is considered as pattern termed as base connection.

2. Connection in base connection are attack free such group are clustered in group.

3. Data point in original data set are compared with clustered mined in step 2. Any point that deemed outlier are termed as attack.

The Flaw in this approach is that in large dataset some outlier may be similar and form a small cluster.

To overcome this we find CBLOF Algorithm is designed.

CBLOF → Cluster Based Local Outer Factor

Steps :-

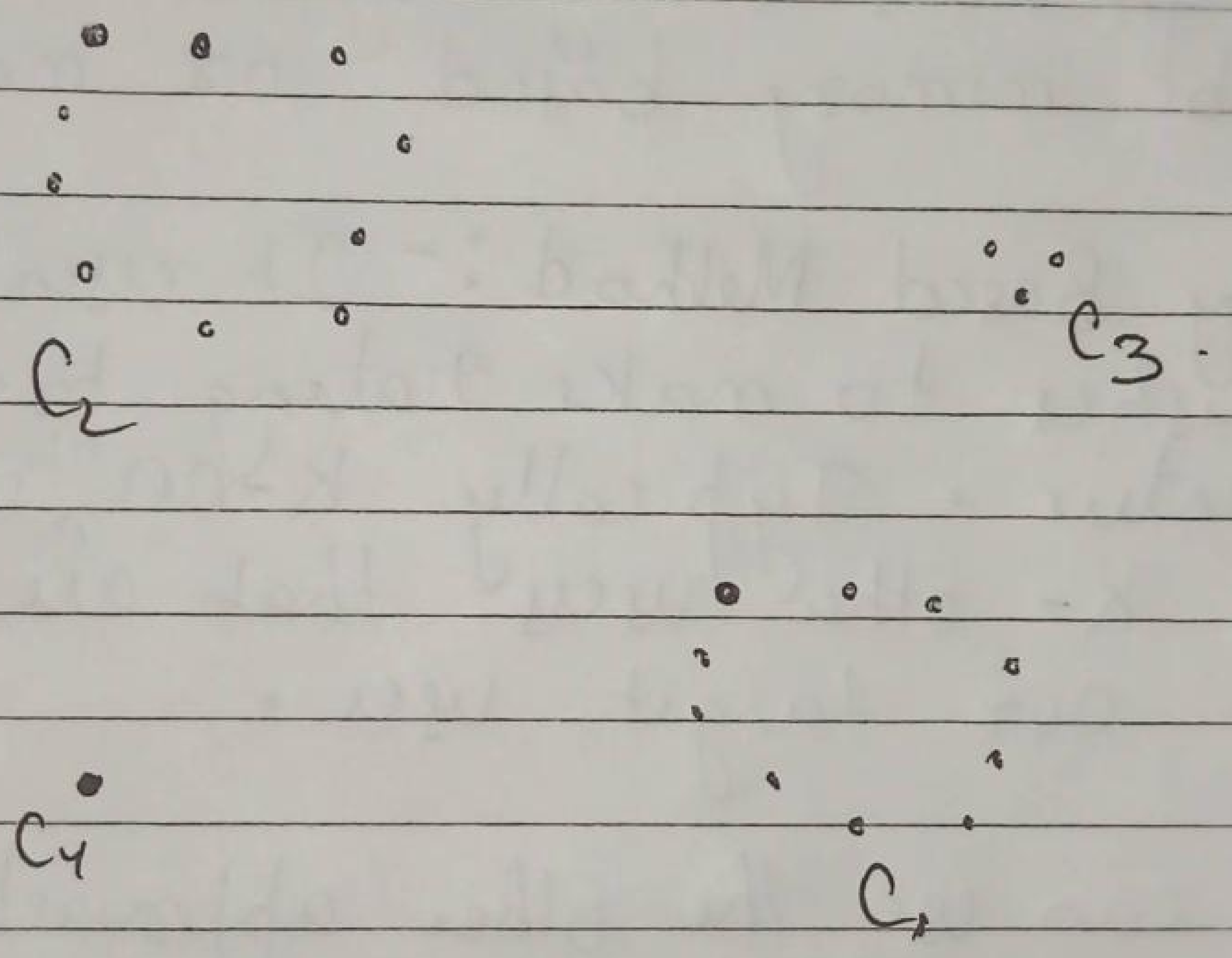
1. Find cluster in dataset & sort them in decreasing order.

Parameter α is used to identify large clusters if they have atleast α% of point, Otherwise they are small cluster.

2. to each datapoint assign CBLOF

$$CBLOF = |Cluster\ size| * Similarity\ b/w\ point\ \&\ cluster$$

Ex →



C_1	10	large
C_2	10	
C_3	3	
C_4	1	Small

for α = 3

find CBLOF & identify attack

2) Data Mining in Recomendation System :-

Data mining uses various methologies in statics & different algo like clustering, classification regression to exploit insight of present large dataset

→ user based :- We calculat person similarity measures collabrative cluster.
└→ feedback

→ Item-based :- Content based

A collabrative recommender system tries to predict the utility of items for user U based on previous rating by other users. It can be memory based or model based.

Memory Based Method :- It essentially uses heuristics to make rating prediction based on entire. Typically k-nn is used that finds k- other users that are more similar to our target user.

We can use the other approaches like Cosine similarity or correlation coefficient

Model-based :- Classification
↓
false positive false negative

Report	Actual	
T	P	→ True Positive
T	N	→ True Negative
F	P	→ False Positive
F	N	→ False Negative

Data mining for financial data analysis-:

- Most banks and financial institutions offer a wide variety of banking, investment, and credit services (the latter include business, mortgage, and automobile loans and credit cards). Some also offer insurance and stock investment services.
- Financial data collected in the banking and financial industry are often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining. Here we present some cases.
    - **Design and construction of data warehouses for multidimensional data analysis and data mining**: Like many other applications, data warehouses need to be constructed for banking and financial data. Multidimensional data analysis methods should be used to analyze the general properties of such data. For example, a company's financial officer may want to view the debt and revenue changes by month, region, and sector, and other factors, along with maximum, minimum, total, average, trend, deviation, and other statistical information. Data warehouses, data cubes (including advanced data cube concepts such as multi feature, discovery-driven, regression, and prediction data cubes), characterization and class comparisons, clustering, and outlier analysis will all play important roles in financial data analysis and mining.
    - **Loan payment prediction and customer credit policy analysis**: Loan payment prediction and customer credit analysis are critical to the business of a bank. Many factors can strongly or weakly influence loan payment performance and customer credit rating. Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones. For example, factors related to the risk of loan payments include loan-to-value ratio, term of the loan, debt ratio (total amount of monthly debt versus total monthly income), payment-to-income ratio, customer income level, education level, residence region, and credit history.
    - **Classification and clustering of customers for targeted marketing:** Classification and clustering methods can be used for customer group identification and targeted marketing. For example, we can use classification to identify the most crucial factors that may influence a customer's decision regarding banking.
    - **Detection of money laundering and other financial crimes:** To detect money laundering and other financial crimes, it is important to integrate information from multiple, heterogeneous databases (e.g., bank transaction databases and federal or state crime history databases), as long as they are potentially related to the study. Multiple data analysis tools can then be used to detect unusual patterns, such as large amounts of cash flow at certain periods, by certain groups of customers. Useful tools include data visualization tools (to display transaction activities using graphs by time and by groups of customers), linkage and information network analysis tools (to identify links among different customers and activities), classification tools (to filter unrelated attributes and rank the highly related ones), clustering tools (to group different cases), outlier analysis tools (to detect unusual amounts of fund transfers or other activities), and sequential pattern analysis tools (to characterize unusual access sequences). These tools may identify important relationships and patterns of activities and help investigators focus on suspicious cases for further detailed examination.