Number of Book II-nd-used

34890

B.S. 2071

TRIBHUVAN UNIVERSITY

No. 38

INSTITUTE OF SCIENCE & TECHNOLOGY

EXAMINATION BOARD

KIRTIPUR

Book I

Students are required to write their answers on both sides & a margin of about 1¼ inches should be left on each page.

Master's Level

Year/Part II

Centre SMS TU

Roll No. 06

Roll No. In Words Six

Arpan Saphota

Reg. No. .....

Subject Advanced Data Mining

Paper .....

Script English

Date 2081/01/31

INSTRUCTIONS TO CANDIDATES

[1] Each Candidate will write legibly on the title page his or her ROLL NO. REGISTERED NO. AND THE SCRIPT in which answers are written but not his or her name and the name of the campus from which he or she appears. This should be done in each answer - book used before beginning to write inside.

[2] No loose papers will be provided for scribbling and no other paper should be brought in for this purpose. Any candidate found with loose paper in his or her possession WILL BE EXPELLED. All work must be done in the book provided and pages MUST NOT BE TORN OUT. The book provided CANNOT BE REPLACED BY ANOTHER but, if necessary, an additional book will be given. All work intended for the examiner must be written ON BOTH SIDES of the paper. Anything cancelled will not be looked into. Should a torn leaf be discovered inside an answer book, it should not be removed but crossed out and folded after bringing it to the notice of the invigilator.

[3] Candidate is forbidden to write answers or anything else on the question - paper.

[4] No Candidate will be allowed to leave the room until one hour has passed from the time when the papers are distributed.

[5] Candidate who uses two or more answer-books should see, before handing over to the invigilator, that they are properly stitched together.

INSTRUCTIONS TO THE EXAMINER

1. Aggregate marks of each question should be placed in the given appropriate box.
2. Marks for parts of a qusetion should be totalled under the question inside the answer- book.

MARKS OBTAINED

1st Q		11th Q	
2nd Q		12th Q	
3rd Q		13th Q	
4th Q		14 th Q	
5th Q		15th Q	
6th Q		16th Q	
7th Q		17th Q	
8th Q		18th Q	
9th Q		19th Q	
10thQ		20th Q	

TOTAL

Signature of Scrutiny Board

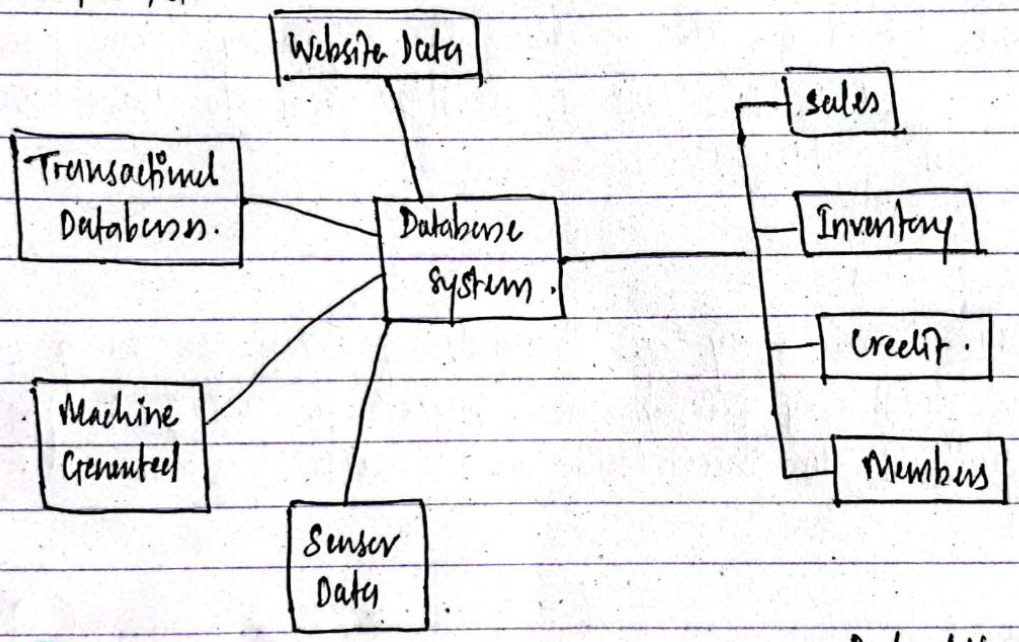
Signature of Examiner

Group 'A'

1. Warehousing is the process of storing subject oriented, time variant, integrated, non-volatile data into a repository for Data Mining, Data Analysis purposes.

The sources of data for building a data warehouse are:

(i) Transactional Database.
(ii) Machines
(iii) Sensors
(iv) People generated
(v) Websites, etc.



Data Warehouse.

3

fig: Sources of Data Warehouse.

2. The limitation of OLTP that are solved in OLAP so that OLAP can be used for data mining are:

(i) Database size

the size of OLTP is in GB to higher GB, so there is limited size which get improved in OLAP so TB's of data can be stored in OLAP for Data Mining.

(ii) Access in Read and Write.

In OLTP there is both Read and write access which makes the database in inconsistent mode while performing high processing operation in OLTP, so OLAP is mostly Read so there won't be any such problems.

(iii) Number of Users

In OLTP there may be thousands of users accessing OLTP but in OLAP only few hundred people are using it for the core operations so this limitation of OLTP has overcomed by OLAP.

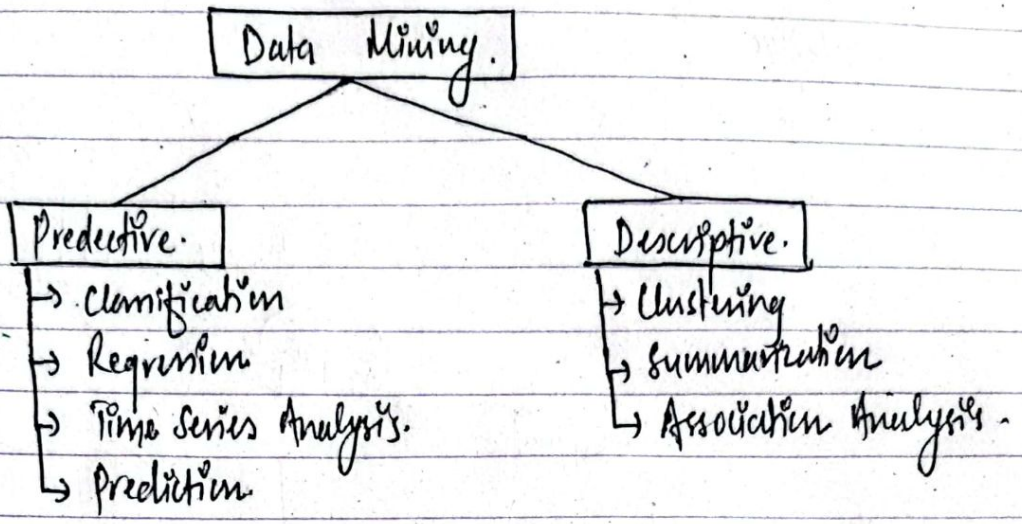
(iv) Processing Power.

The processing power of OLTP is slower than OLAP, so this is also improved in OLAP.

2

(v) Many other like: Infrastructure Setup, Hardware Requirements, etc

3. The different techniques that can be applied for data mining are.



The use cases of the above data mining techniques are.

(i) Classification → Email Spam or Non-Spam identification
→ Tumer cell classification.

(ii) Regression → Predicting Age with Heights

(iii) Time Series Analysis → Weather forecasting
→ Stock Market prediction.

25

(iv) Clustering → Outlier detection
→ Intrusion detection

(v) Association Analysis → Market Basket Analysis.
→ Inventory Management in shopping Markets.
→ Recommendation of items

4. KDD stands for knowledge Discovery from Data which is a data mining process which is carried out to get the valid patterns which then interpreted and evaluated to obtain the knowledge.

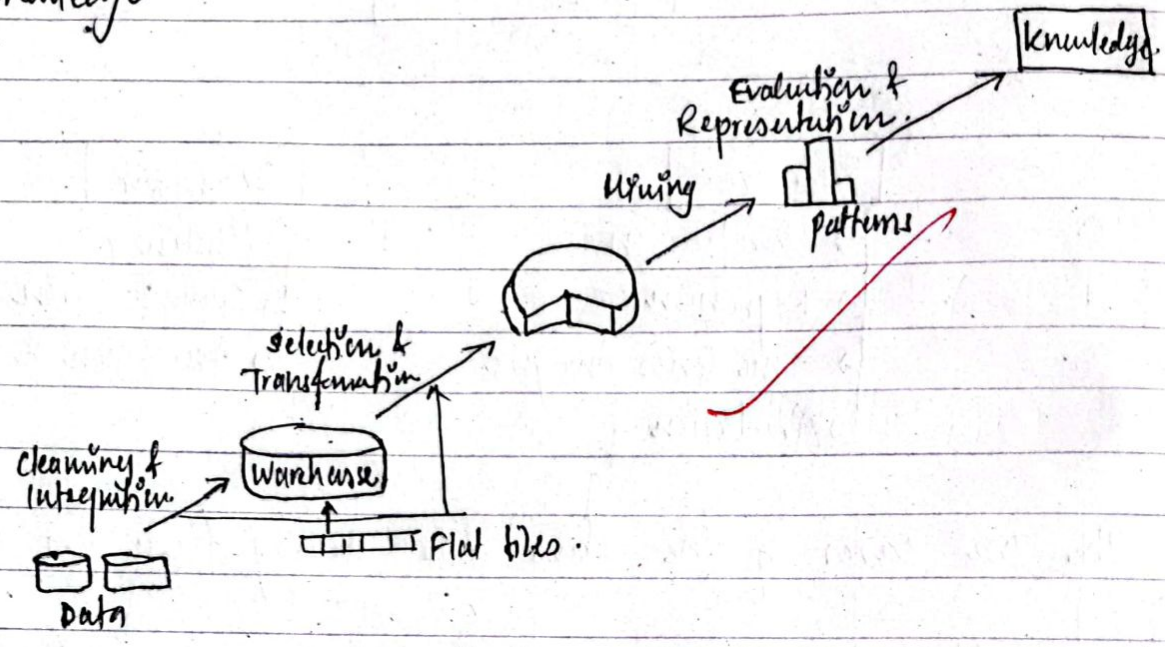


fig: KDD process.

The possible tasks that can be done in the data-preprocessing phase are:-

(i) Dealing with missing values.
(ii) Dealing with Duplicates and outliers.
(iii) Encoding the categorical features.
(iv) Splitting the data into Train and Test set.
(v) Dealing with Imbalanced data.

3

Example

5. The different steps of the kNN algorithm are:-

(1) Compute the distance between the training set and the given query point which needs to be predicted.

(2) Compare the distance obtained between each records in the training set.

Here, we can use any distance calculation metric like.

Euclidean distance = $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

Manhattan distance = $|x_2-x_1|+|y_2-y_1|$

(3) In order to make prediction, we need the value of k. The best values of k lies between 5 to 10,

Sum the computed distance get k smallest distances results from the training set

(4). Analyze the majority of the class that is being labelled as output for those k smallest distance result..

(5) The final output of the prediction will be based on the k smallest results being classified in majority output.

25 This result be the output of the query in kNN Algorithm.

In this way the kNN Algorithm works.

Group 'B'

6. Apriori Algorithm.

It is an association analysis algorithm which is used to mine the association rules & the frequent itemset. This algorithm works in two phases. i.e.

(i). Frequent itemset generation.

(ii) Generating the association rules.

In this algorithm, the database needs to be scanned multiple times so this is not computationally good algorithm.

Let us consider one use case to illustrate Apriori Algorithm for association rule mining.

Consider a transaction:

TID.	Items
1	A, B, C
2	A, C
3	A, D
4.	B, E, F

Support = 50%.
Confidence = 50%.

Here, we have given the support and confidence as

$$\text{Support} = \frac{50}{100} \times 4 = 2$$

$$\text{Confidence} = \frac{50}{100} \times 4 = 2.$$

Now,

Lets get the candidate items from above given Dataset 1.

C_1 =

Item	Support
{A}	3
{B}	2
{C}	2
{D}	1
{E}	1
{F}	1

Eliminating the items whose support < Min. threshold support i.e 50% = 2.

from C_1, we get L_1 as,

Item	Support
{A}	3
{B}	2
{C}	2

Now, again getting the two items in one transaction from L_1 as,

C_2

Item	Support
{A,B}	1
{B,C}	1
{A,C}	2

Similarly, L_2 becomes.

Item	Support
{A,C}	2

No more itemset can be generated after L_2

Thus, we obtain the association rule as,

rule	Support	Confidence	Confidence %
A → C	2	2/3 = 0.66	66.66%.
C → A	2	2/2 = 1	100 %.

Both the rules has confidence > Min threshold confidence, Hence,

A → C

C → A

are the required final rule from the given set of Transactions. In this way we can implement Apriori Algorithm for

Use case ?

Association rule mining.

This Algorithm is computationally costly so its improved version is FP-Growth Algorithm.

In order to improve the performance of Apriori Algorithm we can implement some following techniques.

(i) Sampling.
(ii) Partitioning.
(iii) Hash based itemset count.
(iv) Dynamic itemset count.
(v) Reducing Transactions.

---

7. Data Visualization Techniques.

Data Visualization is the technique which can be use to visually represent the data so that it will be easy for the data interpretation and understanding the data.

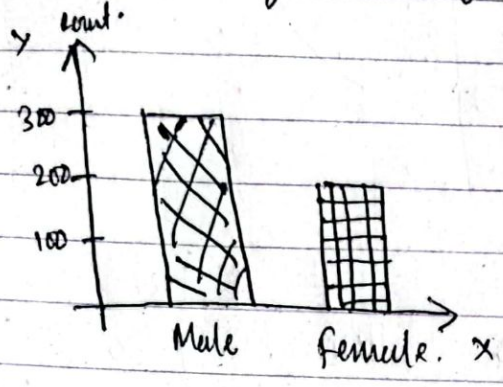
The Data Visualization techniques are:-

(i) Bar plot
(ii) Histogram
(iii) pie chart.
(iv) Scatter plot.
(v) Box and Wisker plot.
(vi) Heat Map

(i) Bar plot.

Bar plot is one of the visualization technique which represents the data in bars and give the frequency distribution of the data.

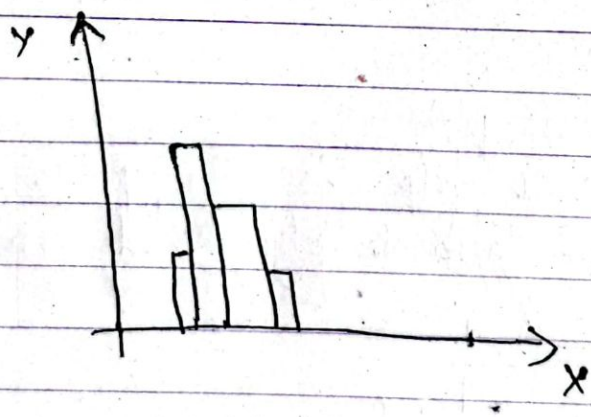
i.e.



(ii). Histogram.

Histogram is similar to bar plot but it is not like a bar structure but it gives the distribution of the data. In order to know the distribution, we use histogram to view it.

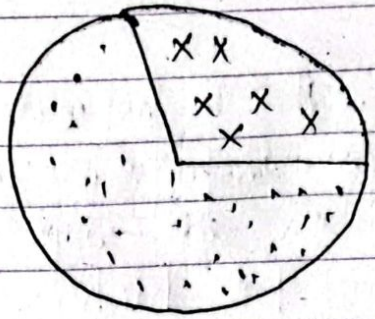
i.e.



(iii) Pie chart.

Pie chart is the circular pie shaped visualization technique which also represents the distribution of the data. With the help of pie-chart we can easily interpret the data.

i.e.



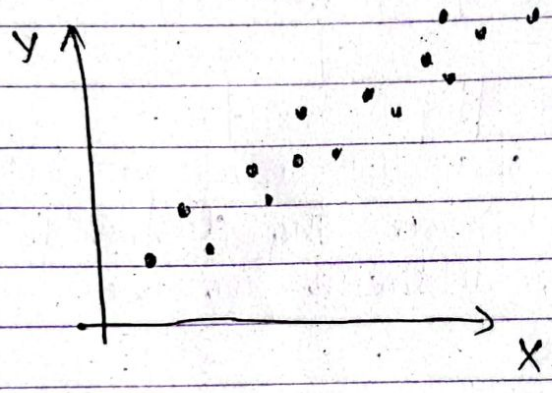
Index:

× × × ×	Female.
⁘	Male.

(iv) Scatter plot

Another important visualization technique is the scatter plot which gives the scatterness of the data.

i.e.

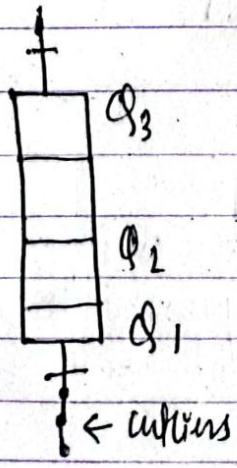


2D scatter plot

(v) Box and Whisker plot

This plot will help to identify the outliers in the data. This is statistical plot based on quartiles Q1, Q2 and Q3

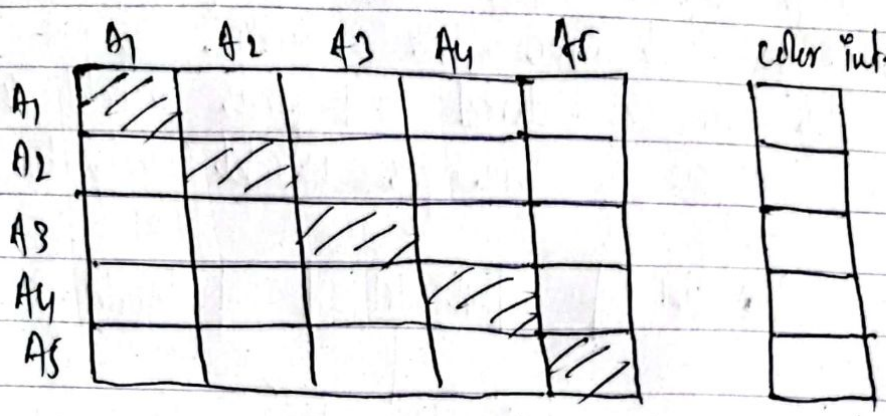
i.e.



(vi) Heat Map

Heat map will helps to visually represent the density of the data in a matrix form. It is normally used with color intensities.

i.e.



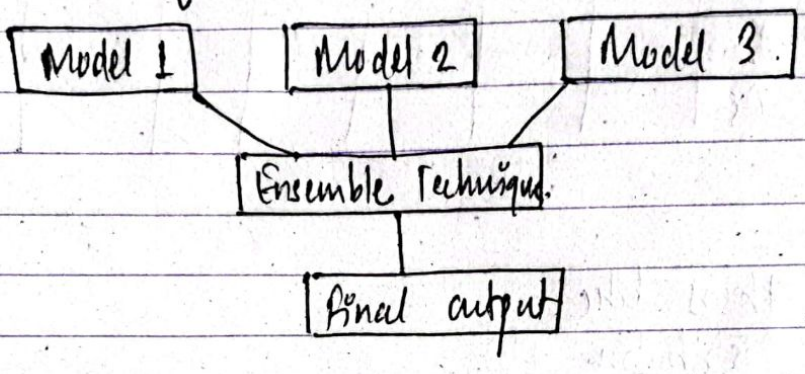
(vii) Many others like:

- Distribution plot
- Word cloud., etc.

All this visualization can be implemented from the python library matplotlib pyplot or also can be done in R, MATLAB.

8. Ensemble technique.

Ensemble technique is the advance technique which is used to improve the performance of the predictive models. In this technique the multiple output from different model are considered and final output is obtained by the analysis of multiple model output.



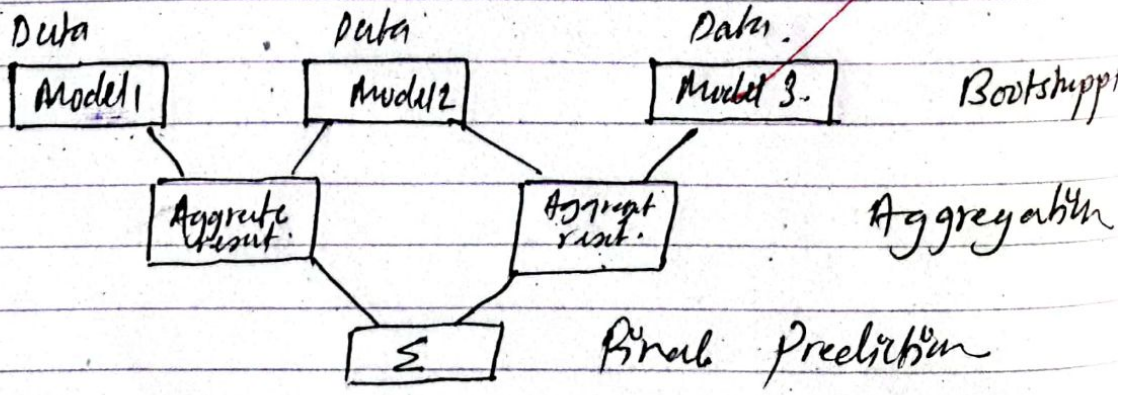
The ensemble techniques are:-

(i) Bagging
(ii) Boosting.
(iii) Stacking.

(5)

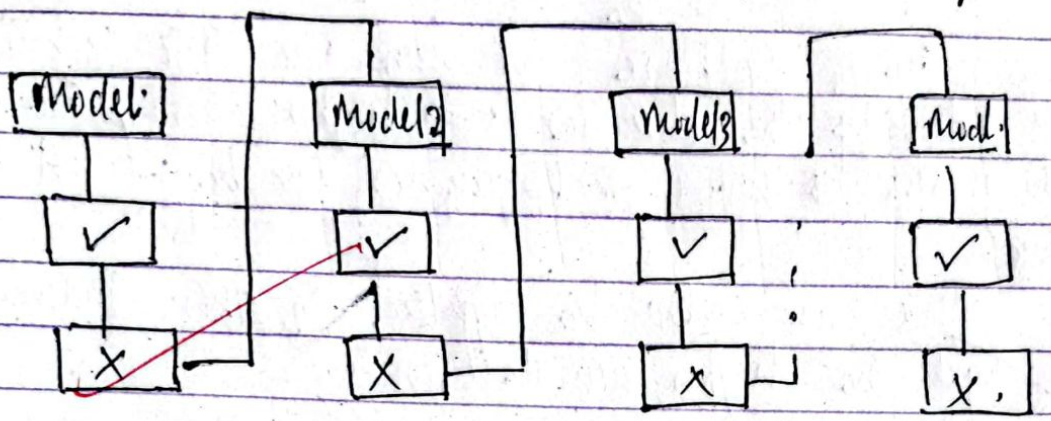
(i) Bagging.

In this Bagging technique, the model are bootstrapped and then Aggregated to obtain the final output.



(ii) Boosting:

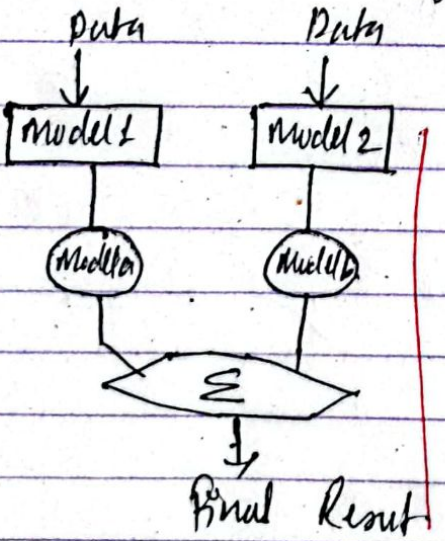
Boosting is done to give output of one model to another model as input and so on with multiple boosting models.



In this way the final output will be more accurate with boosting technique in the model.

(iii) Stacking.

It is another ensemble method which is in stack form and summarize the output of multiple model and give final output.



Here classifier and Regression model both can be used in Stack.

In this ensemble technique, we obtain the refined output from multiple model output. This ultimately helps to improve the accuracy of the model.

In Healthcare, we can use this technique to identify the diseases with higher accuracy in short diagnosis so that critical diseases can be cured in time.

This can also be applied in other industries to get the higher accurate results.

P.T.O.

e No.:

Code No.:

Copy No. 47125

TRIBHUVAN UNIVERSITY
INSTITUTE OF SCIENCE & TECHNOLOGY
EXAMINATION BOARD
KIRTIPUR

Book II

dents are required to write their answers on BOTH SIDES and a margin of ut 1¼ inches should be left on each page. Particulars below must be filled mmediately after the receipt of this book. This book must be stitched to the inal book before delivering the papers.e.

GNATURE OF INVIGILATOR

ROLL NO. 06.

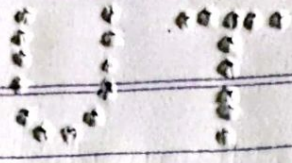
EXAMINATION .....

SUBJECT ADM.

PAPER II

DATE 2081/01/31

P.T.O.



9. Outlier

Outlier are the those values which are different than the normal values. These are the extreme values which does not fall in the normal region that the regular data falls.

Example.

Age = { 10, 15, 16, 19, 89, 13 }

↓

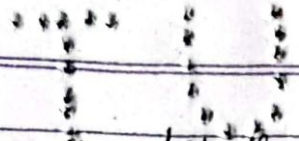
outlier in this data.

clustering can be used for outlier detection.

Outliers behaves like unusual behaviour in the data, so we can easily detect that unusual behaviour with the help of clustering algorithms. We have following clustering algorithms which can be applied to detect the outliers in the data.

(i) K-Means Clustering
(ii) Hierarchical clustering.
(iii) DBSCAN clustering.

let us discuss on k-means clustering for the outlier detection.



We perform the following steps to detect the outliers using k-means clustering Algorithm.

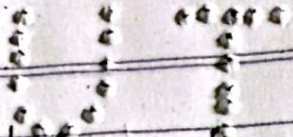
(i) First of all we have to get the number of clusters from the training data.

(ii) Perform the k-means clustering Algorithm for the given training dataset:
- Compute the distance between the initial cluster and other clusters
- Iteratively perform the clustering until we get the data grouped in the desired clusters.

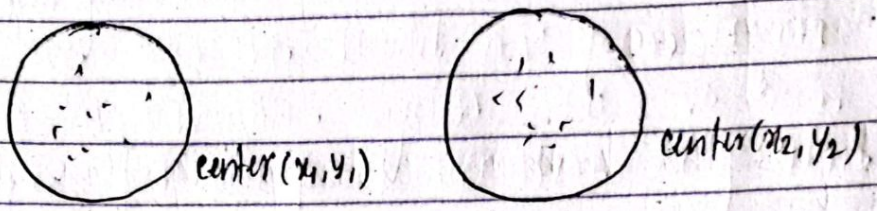
In order to stop the iteration of clustering algorithm, we can check for the data points grouped in a same cluster or different cluster in each iteration.
If the next cluster are being the same repeated then we can stop the clustering Algorithm.

(iii) Once the clusters are formed, then we can evaluate the clusters with the evaluation matrices like
- Silhuette Index. (SI)
- Davis Bouldin Index (DBI)
- CHI

Now, we are ready to perform the outlier detection technique in the formed clusters.



Compute the distance between the clusters and the outlier value.
Let we have choosen k = 2 then



Cluster 1 Cluster 2.

5.

Outlier = (a, b)

We compute the distance between the

cluster 1 center (x_1, y_1) and outlier (a, b) = [redacted] m

and

cluster 2 center (x_2, y_2) and outlier (a, b) = n.

[redacted]

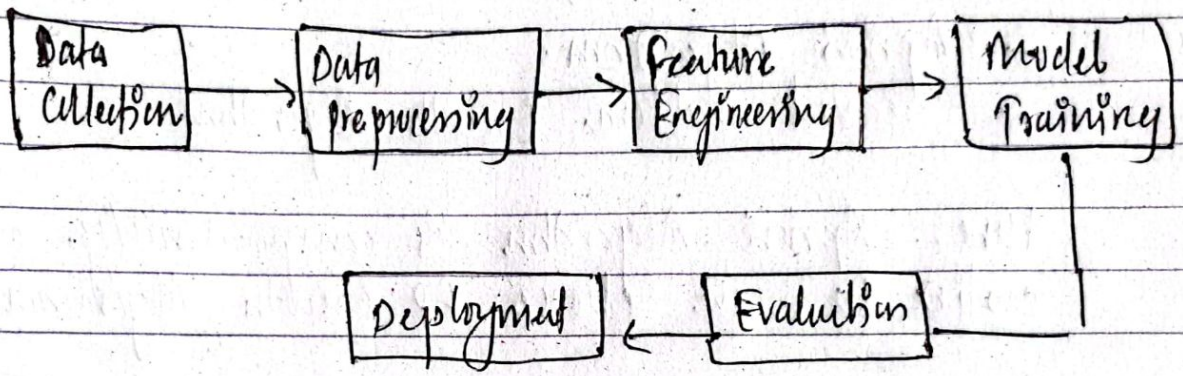
And here we can compare the distance and decide whether it is outlier or not.

Also we can get the minimum threshold value so that we can easily decide whether it is outlier or not

In this way we can perform outlier detection using clustering algorithm.

10. We can build Association based recommendation system

Association is the technique which helps to identify the next favorable items from the existing itemset. In order to build a recommendation system, we follow the given pipeline



(i) Data collection

Here, we collect the data which is our domain of focus i.e. if we want to give recommendation on ~~retail~~ market then we collect the retail data

(ii) Data preprocessing.

The collected data are preprocessed, unnecessary information, fields, duplicate values are removed and imputations are done.

(iii) Feature Engineering

The necessary features that are required for the

1. model training are kept and rest other which increases the model complexity are removed.

(iv) Model Traing

Here, we train the model, our focus is on Association based recommendation system, so we train the one of the following Association algorithm

- Apriori Algorithm
- Frequent Pattern Growth Algorithm

Here, Apriori Algorithm is computationally complex and costly so, we choose FP Growth Algorithm

In FP Growth Algorithm we perform the following steps:

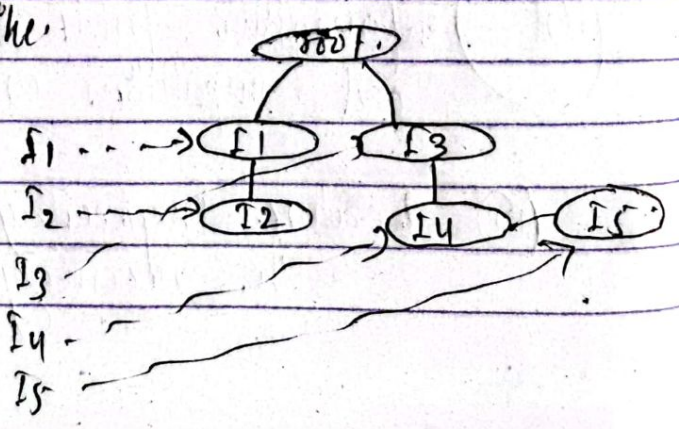
1. Construct the FP tree with our processed data
2. Generate the pattern from the FP tree.

Here, the model will only scan for the data in database just 2 times only so FP is improved version of Apriori Algorithm

Our FP tree looks like.

ordered items:

I_1, I_2, I_3, I_4, I_5



Finally, our FP tree is ready which will give the pattern of the itemset.

(v) Evaluation.

The mined model is then evaluted with evalution metrics. Here the evalution metrics will be

→ Support
→ Confidence

(vi) Deployment

Finally we deploy our recommendation system into production environment for the product use.

In this way we can build the Association based recommendation system.

Let us consider small rule that is globally popular in the retail market

{Milk, Diaper} ⟶ {Beer}.

This rule is mined with the Association Algorithm. In Retail store we can recommend the {Beer} by managing the Inventory near to Milk & Diaper. In this way we can implement Association in the recommendation system.