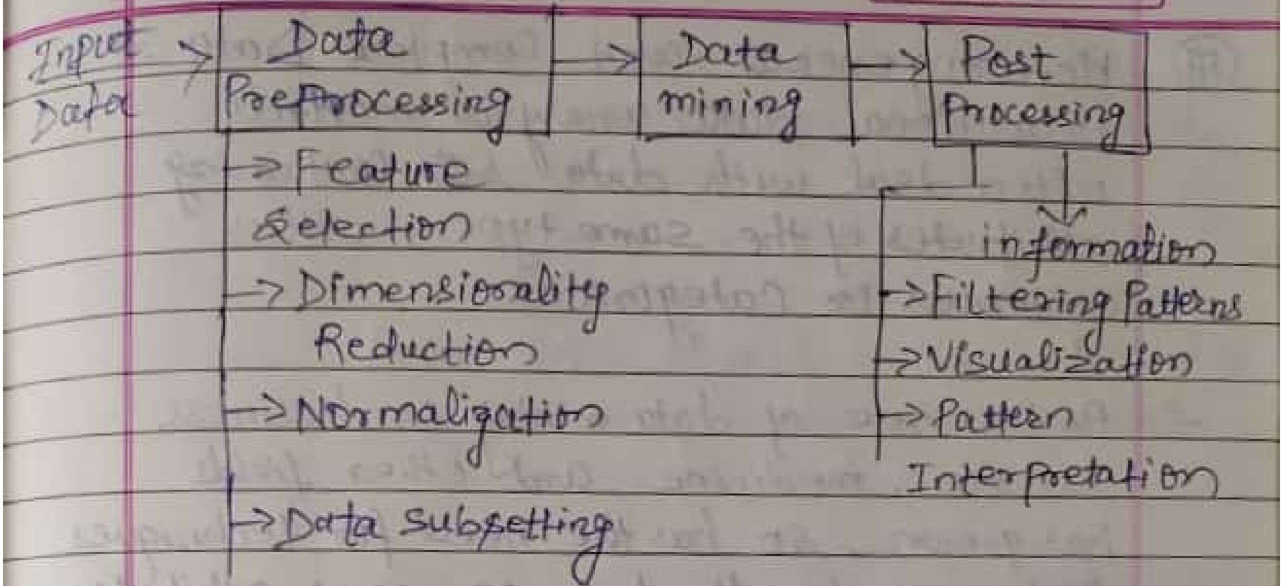


* Motivating challenges

(i) Scalability → Because of advances in data generation and collection, data sets with sizes of gigabytes, terabytes, or even petabytes are becoming common.

→ If data mining algorithms are to handle these massive data sets, then they must be scalable.

(ii) High Dimensionality :→ It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago.

→ Data sets with temporal or spatial components also tend to have high dimensionality.

Data mining Tasks:

→ Data mining tasks are generally divided into two categories

(1) Predictive tasks :→ The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes.

(2) Descriptive tasks :→ The objective is to derive patterns that summarize the underlying relationship in data.

* Four core data mining tasks :→

(i) Association analysis ~~is~~ is used to discover patterns that describes strongly associated features in the data.

(ii) Cluster analysis :→ seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other

(iii) Anomaly detection: is the task of identifying observations whose characteristics are significantly different from the rest of the data.

(iv) Predictive modeling: refers to the tasks of building a model for the target variable as a function of the variables.

(iii) Heterogeneous and Complex Data :->
Traditional data analysis methods often deal with data sets containing attributes of the same type, either continuous or categorical.

-> As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes.

(iv) Data Ownership and Distribution:
-> Sometimes, the data needed for an analysis is not stored in one location or owned by one organisation.
-> Instead, the data is geographically distributed among resources belonging to multiple entities.

(v) Non-traditional Analysis: The traditional statistical approach is based on a hypothesize and test paradigm.
-> An experiment is designed to gather the data, and then the data is analyzed with respect to the hypothesis.

③ $F_{k-1} \times F_{k-1}$ method :→

$F_k = F_{k-1} \times F_{k-1}$

(k-2) item set must be common.
↳ 1 item should be common among two.

eg. {Bread, milk}, {Bread, Diaper}

k-2 = {Bread, Diaper, Milk}

eg {Milk, Bread} {Diaper, Bread}

$F_k = F_{k-1} \times F_1$

F_3 = {Bread, Milk} {Diaper}

F_3 = {Bread, Milk, Diaper} (x)

F_2 = {Diaper, Milk} -> infrequent

lexiography order

abc
bcd
abd
acd
bda
bdc
bcd

Transition → Frequent Itemset
↓
Association rules

K - Frequent Itemset
X → X & Y
X → X - Y

For K items
$2^K - 2$ rules can be generated.

eg. ~~Se~~ Rule generation. ~~Iitem set~~
{a, b, c}
{~~a~~ 3 - Frequent Item set

{a, b, c}
$2^K - 2$ rules $= 2^3 - 2 = 6$

a → b
a → c
ab → c
ac → b
bc → a
b → c

{a, b} {c} {a, b} → {c}

* Confidence - Based Praning :-

* Compact representation of frequent item

1. Maximal frequent Itemset.
2. closed frequent Itemset.

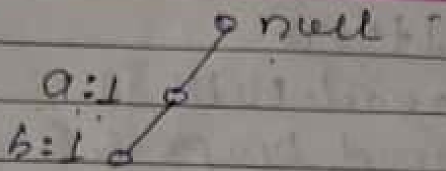
→ All frequent Itemset immediate supersets should be infrequent

{ad, ace, bcde}

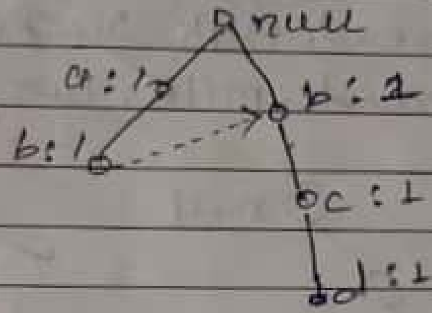
→ which is also closed Item set and satisfy ~~support~~ minimum support.

Model overfitting &
problem in resubstitution method
Pessimistic error

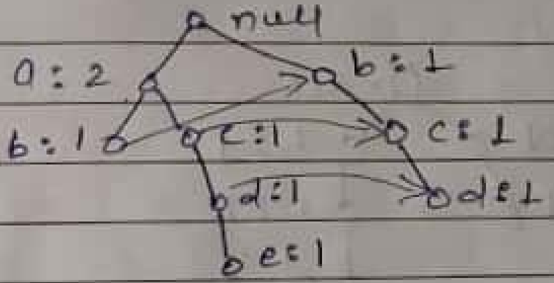
T=1



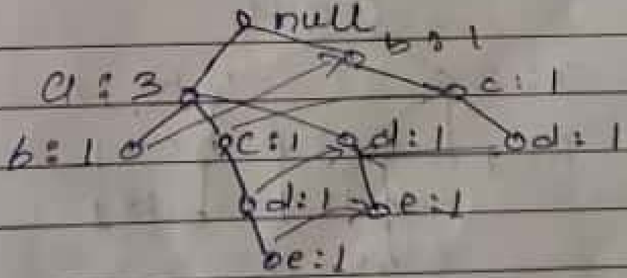
T=2



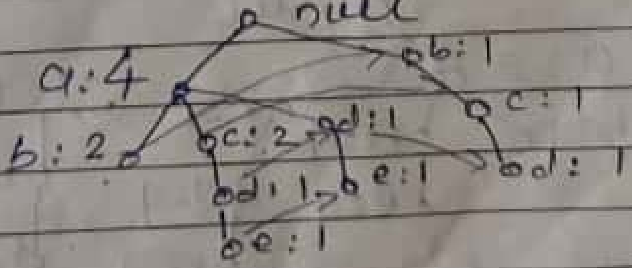
T=3



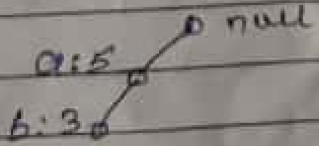
T=4



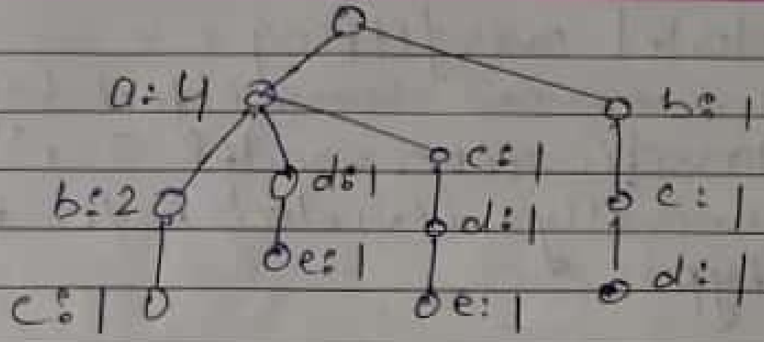
T=5



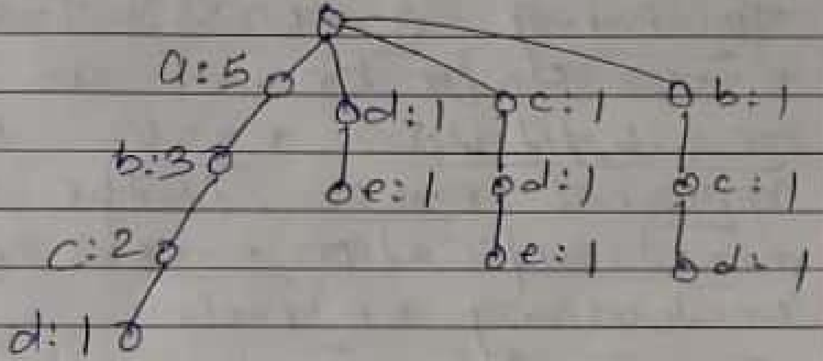
T=6.



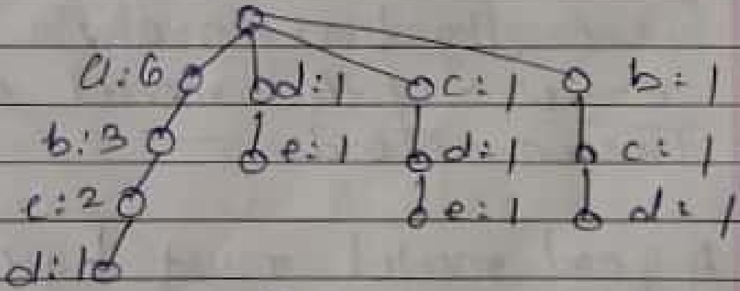
T5 =



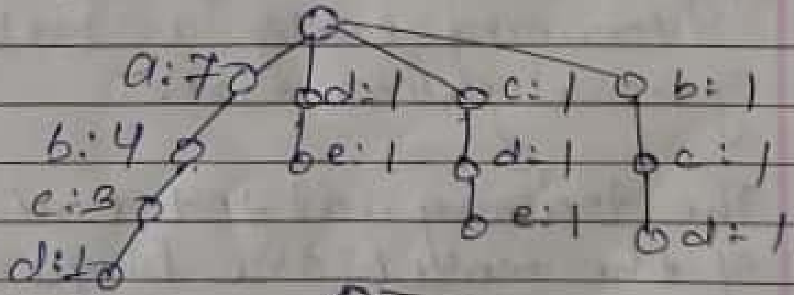
T6 =



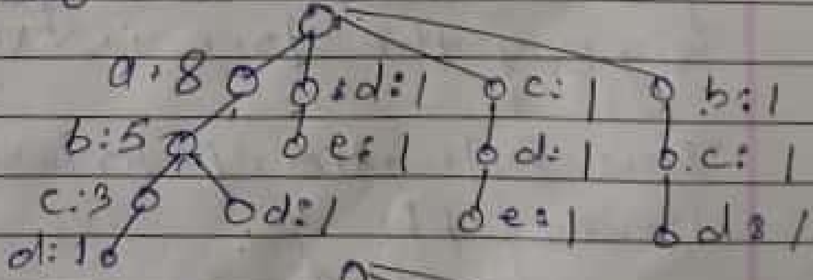
T7 =



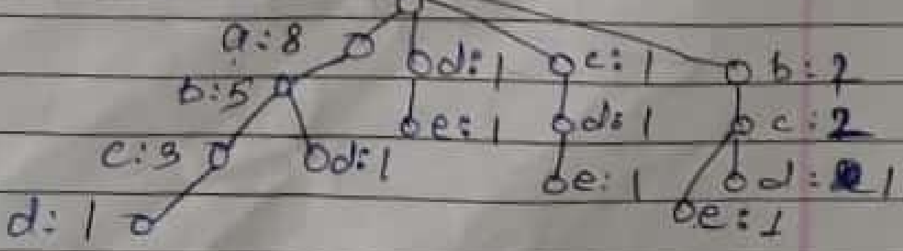
T8 =



T9 =



T10



Model overfitting:
The errors committed by a classification model are generally divided into two types:

(i) Training errors: Training error, also known as resubstitution error or apparent error, is the no of misclassification errors committed on training records.

(ii) Generalization error is expected error of the model on previously unseen records.

A good model must have low training error as well as low generalization error.

→ The training and test error rates of the model are large when the size of the tree is very small.

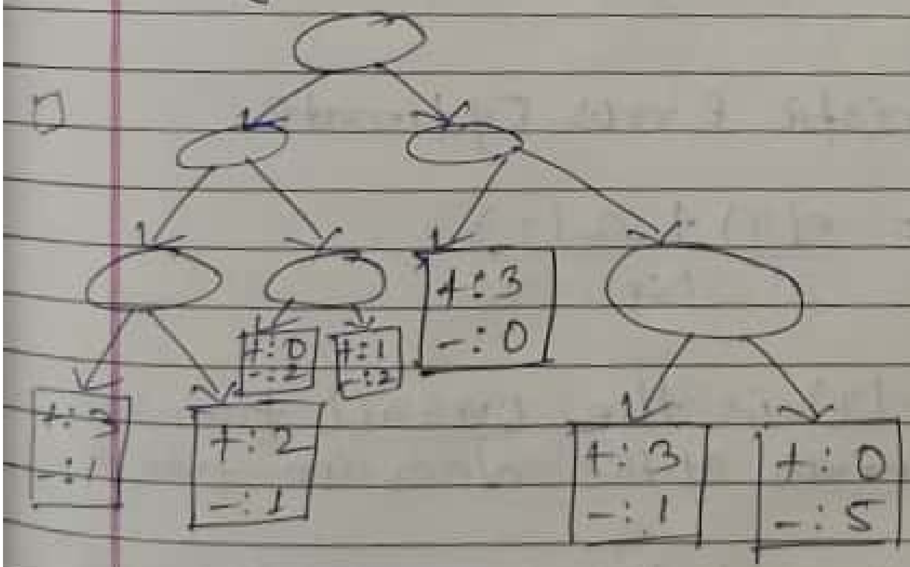
This situation is known as model underfitting

-> As the no of nodes in the decision tree increases, the tree will have fewer training and test errors

-> However, Once the tree becomes too large, its test error rate begins to increase even though its training error rate continues to decrease

This phenomenon is known as model overfitting.

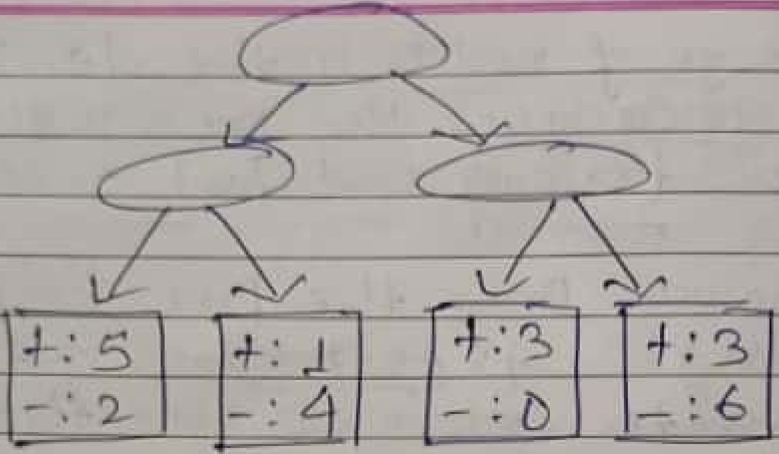
* Using Resubstitution Estimate :



$$e(T_L) = \frac{1+1+0+1+0+1+0}{24} = \frac{4}{24} = ~~0.25~~ = 0.16$$

↓

error rate in left tree.



15
9
24

$$e(T_R) = \frac{2+1+0+3}{24} = \frac{6}{24} = 0.25$$

Based on resubstitution estimate left tree is considered better than the right tree

* Pessimistic Error Estimate:

$$e_g(T) = \frac{e(T) + \Omega(T)}{N_t}$$

~~e(T)~~ e(T) is the overall training ~~or~~ error of the decision tree

$\Omega(T)$ penalty term associated with ~~ne~~ each node t_i

N_t = No of training records

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

(c) {Milk, Diapers} → {Beer}

Support Count = 2

total no of transaction = 5

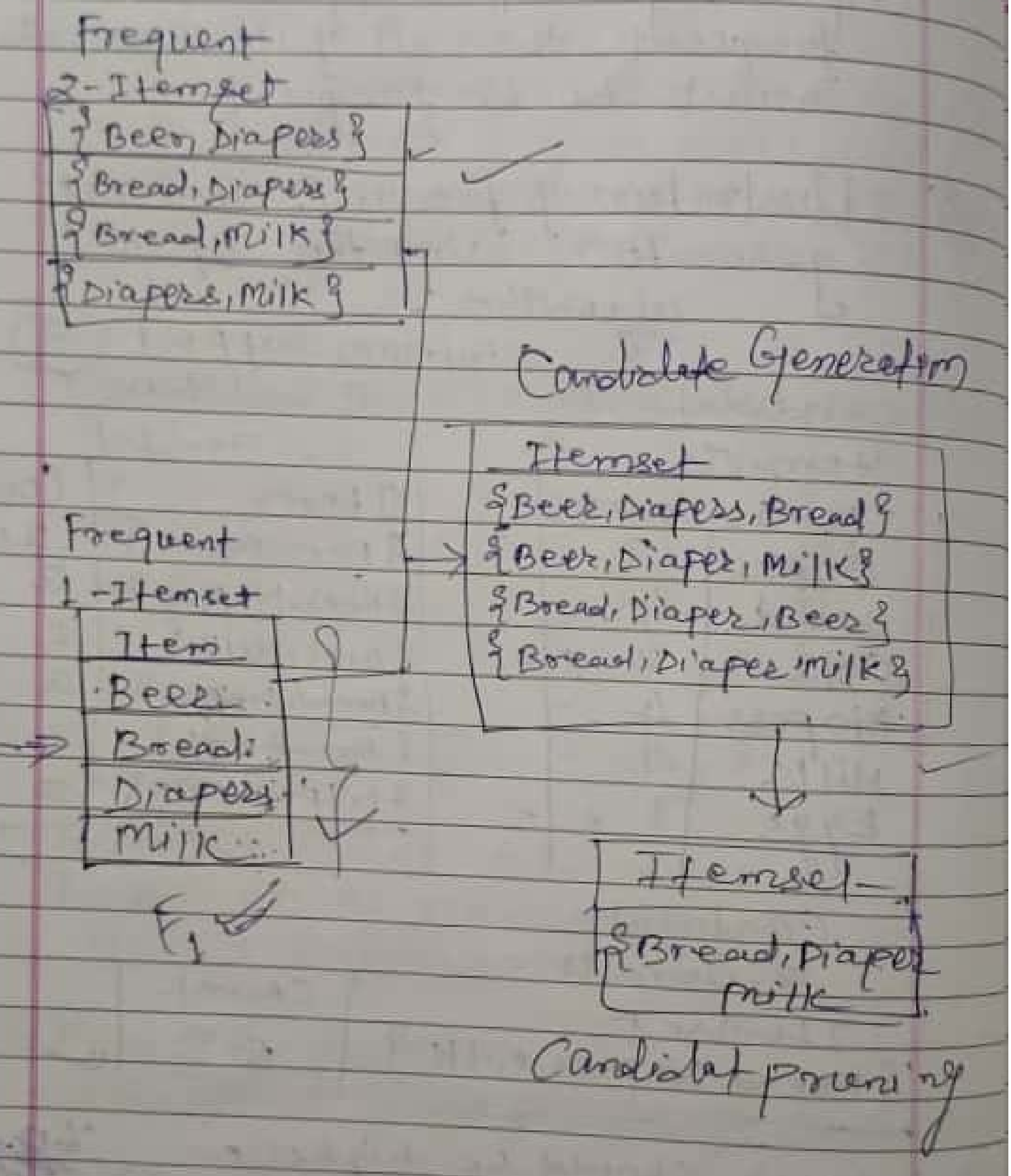
rule's support = 2/5 = 0.4

Rule's Confidence =

$$\frac{\text{Support Count for \{Milk, Diapers, Beer\}}}{\text{Support Count for \{Milk, Diapers\}}}$$

$$= \frac{2}{3} = 0.67$$

$F_{k-1} \times F_1$ ✓ $F_1 \times F_{k-1}$



Frequent
2-Itemset

{Beer, Diapers}
{Bread, Diapers}
{Bread, Milk}
{Diapers, Milk}

Frequent
1-Itemset

Item
Beer
Bread
Diapers
Milk

F_1

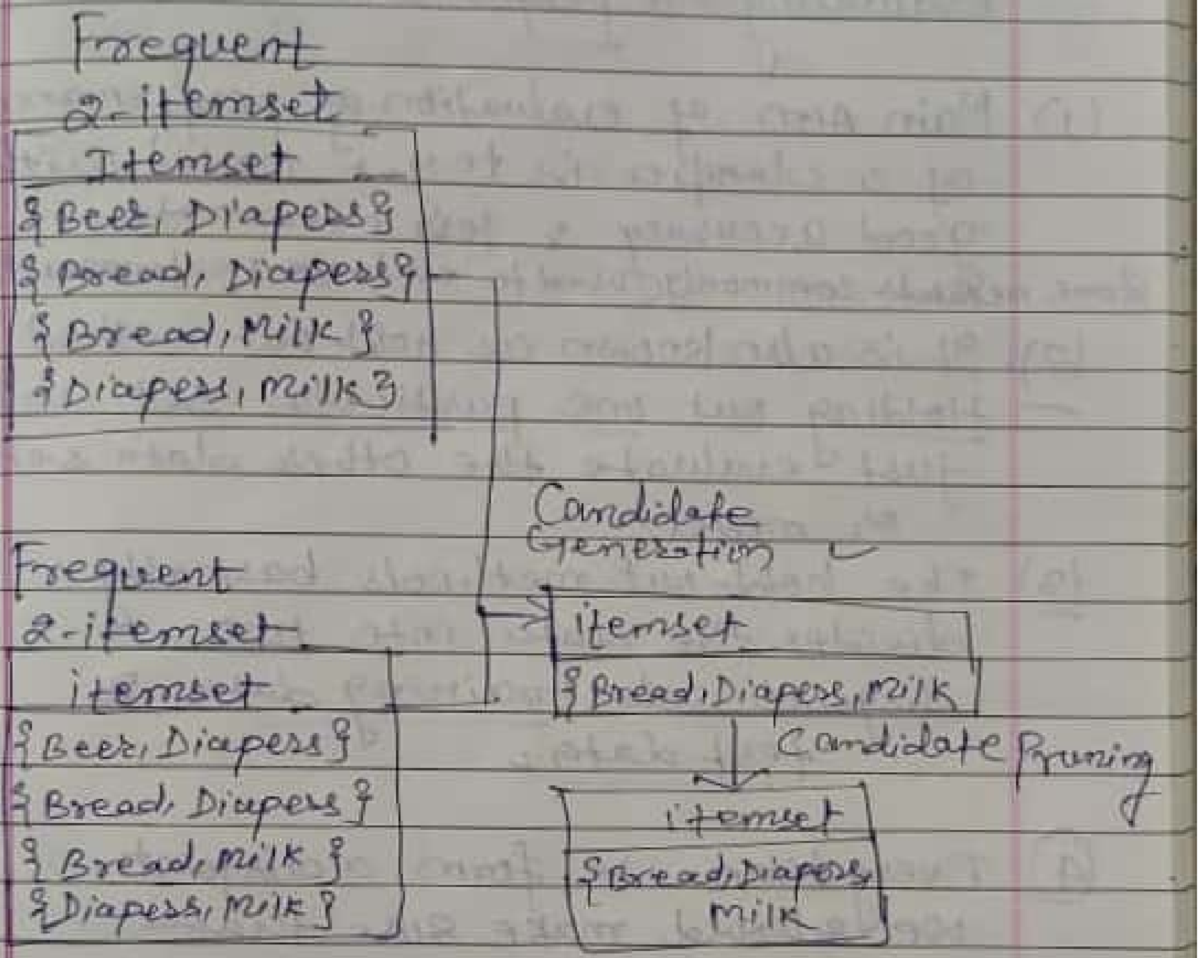
Candidate Generation

Itemset
{Beer, Diapers, Bread}
{Beer, Diaper, Milk}
{Bread, Diaper, Beer}
{Bread, Diaper, Milk}

Itemset
{Bread, Diaper, Milk}

Candidat pruning

Fk-1 x Fk-1 :



Frequent 2-itemset

Itemset
{Beer, Diapers}
{Bread, Diapers}
{Bread, Milk}
{Diapers, Milk}

Frequent 2-itemset

itemset
{Beer, Diapers}
{Bread, Diapers}
{Bread, Milk}
{Diapers, Milk}

Candidate Generation

itemset
{Bread, Diapers, Milk}

Candidate Pruning

itemset
{Bread, Diapers, Milk}

Evaluating the performance of a classifier

(1) Main Aim of evaluation of performance of a classifier is to get model with good accuracy & less error rate.

Some methods commonly used to evaluate performance of classifier

(2). It is also known as hold-out method.
— Holding out one particular set and just evaluate the other data set or model.

(3) The hold-out methods basically divides the data into two disjoint sets training data & test data.

(4) Every time we form a disjoint we should make sure that training data is more and test data is less

* Bootstrap: In bootstrap approach, The training records are sampled with replacement i.e, a record already chosen for training ~~records~~ is put back into the Original pool of records so that it is equally likely to be redrawn.

Random subsampling:

Hold-out method will be repeated for several times to improve the estimation of classification

$\boxed{acc_{sub} = \sum_{i=1} acc_i/k}$

Iteration-I

ID	X	Y	class	
20	0	0	+	→ testdata
21	0	1	-	— ~~test~~ training data
22	1	0	+	
23	1	1	+	

ID	X	Y	class	
20	0	0	+	— training data
21	0	1	-	
22	1	0	+	
23	1	1	+	— test date

(1) Cross validations:

Alternative approach of Random subsampling.

→ In this approach ~~is~~ each record is used same no of time for training and exactly one for testing

Iteration1

ID	X	Y	class
20	0	0	+
21	0	1	+
22	1	0	−
23	1	1	+

ID 20 & 21 are considered to be test set and ID 22 & 23 are training set.

Iteration 2

ID	X	Y	class
20	0	0	+
21	0	1	+
22	1	0	−
23	1	1	+

ID 22 & 23 are test data.
ID 20 & 21 are training data.

The Advantage of this approach is that it avoids generating overly complex subtrees that overfit the training data.

* Post-Pruning: In this approach, the decision tree is initially grown to its maximum size.

-> This is followed by a tree-pruning step, which proceeds to trim the fully grown tree in a bottom-up fashion.

-> Trimming can be done by replacing a subtree with(1) ~~a new leaf node whose class label is determined~~

-> the most frequently used branch of the Subtree.

-> Post-Pruning tends to give better results than prepruning because it makes pruning decision based on fully grown tree, unlike prepruning, which can suffer from premature termination of the tree-growing process.

BIG BOOK
DATE: / /
PAGE NO.

7/6/22

Cluster Analysis:

-> Grouping of objects that are meaningful, useful or both.

Cluster for understanding.
1. Biology
2. Information Retrieval
3. Business

Cluster utility:
-> cluster is used for further data analysis or pre-processing.

-> Each cluster will have cluster prototype that is a representative of that cluster.

1. Summarization: - Dealing with the group rather than dealing with individuals.

2. Compression: Vector quantization

3. Efficiently discovering Nearest Neighbor.

Cluster Analysis:->
-> Clustering the data involves grouping of data items that have Similarity within the same group & dissimilarity across the group.

cluster Algorithm:

1. K-means : Prototype based cluster and it performs single level partitioning of data points.

K -> how many clusters do you want.
we need to classify the starting point of clusters.

-> we need to specify seed points(centroids)

10 data points
~~10 data~~ 3 clusters

C1	C2	C3
3	4	3

There is no more change in clusters if centroid remains same across the iterations.

Algo:

1. Select K points as initial centroid
2. repeat.
3. Form K-clusters by ~~assigning the~~ assigning the point to closest centroid
4. Recompute the centroid for each clusters.
5. Until centroid do not change.

* Different types of clustering

* Hierarchical V/s Partitional

* Exclusive V/s Overlapping v/s Fuzzy

* Probabilistic

* Complete V/s Partial.

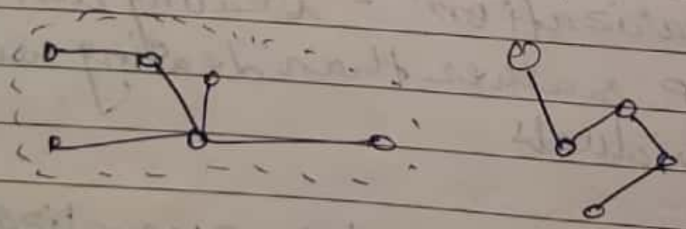
8/10/22

Different types of clusters.

1. Well separated

2. Prototype based cluster

3. Graph based cluster: ~~Connectd Component.~~
Connected Component G



4. Contiguity based cluster

*: Shared Property:

Support Counting:
Support Counting is the Process of determining the frequency of Occurence for every candidate itemset.

Confidence:

Support: support represents the popularity of that product of all the production transactions.

Support of the product is calculated as the ratio of the no of transactions includes that product and the total no of transactions.

$$\frac{\text{No of transactions includes that Product}}{\text{Total no of transactions.}}$$

Confidence can be interpreted as the likelihood of purchasing both the products A and B.

Confidence is Calculated as the no of transaction that include both A&B divided by the no of transaction only product A.

Bootstrap:

-> 9n bootstrap training records are sampled with replacement. i.e, a record already chosen for training is put back & into the original pool of records so that it is equally likely to be redrawn.

-> There are several variations to the bootstrap sampling approach in terms of how the overall accuracy of the classifier is computed.

$$\boxed{acc_{boot} = \frac{1}{b}\sum_{i=1}^{b}(0.632 \times \epsilon_i + 0.368 \times acc_s)}$$

Prepruning: 9n this approach, the tree-growing algorithm is halted before generating a fully grown tree that perfectly fits the entire training data.

-> To do this, a more restrictive stopping condition must be used.

-> eg. stop expanding a leaf node when the observed gain in impurity measures falls below a certain threshold.

→ If the original pool of records already chosen for training

→ If the original data has N records, it can be shown that, on average, a bootstrap sample of size N contains about 63.2% of the records in the original data.

* Random Subsampling :→

→ The holdout method can be repeated several time to improve the estimation of a classifier's performance.

This approach is known as a random subsampling.

$$acc_{sub} = \sum_{i=1}^{K} \frac{acc_i}{K}$$

* Cross validation : An alternative to random subsampling is cross-validation.

→ In this approach each record is used the same no of times for training and exactly once for testing.

Apriori principle: If an itemset is frequent, then all of its subsets must also be frequent.

⑥ Illustration of frequent itemset generation using the Apriori algorithm:

Minimum support = 3

Candidate-1 Itemsets

Item	Count
Beer	3
Bread	4
Cola	2 x
Diapers	4
Milk	4
Eggs	1 x

Candidate 2-Itemset

Item	Count
{Beer, Bread}	2 x x
{Beer, Diapers}	3
{Beer, Milk}	2 x
{Bread, Diapers}	3
{Bread, Milk}	3
{Diapers, Milk}	3

Candidate 3-itemsets

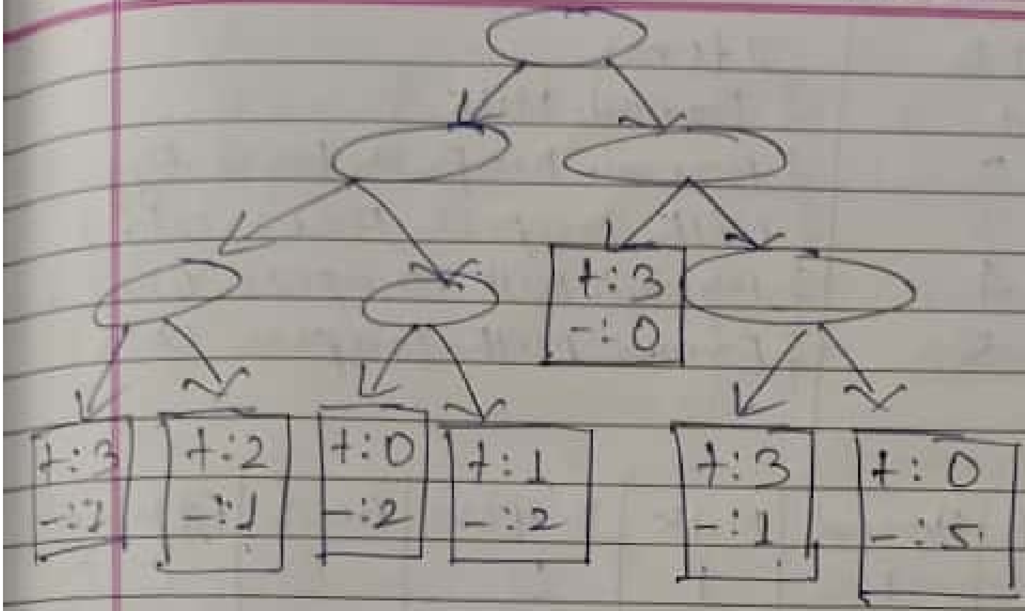
Itemset	Count
{Bread, Diapers, Milk}	2

→ Freqush should be higher among all 4 items

$\frac{60}{100} \times 5$

Bread
milk
Diaper
Beer

2/5 = 40 % < T
Not possible



$p = \boxed{0.5}$

$$eg(T) = \frac{e(T) + \Omega(T)}{N_t}$$

$$= \frac{4 + 0.5 \times 7}{24} = \frac{4 + 3.5}{24}$$

$$\frac{7.5}{24} = 0.3125$$

$$eg(T_R) = \frac{e(T) + \Omega(T)}{N_t}$$

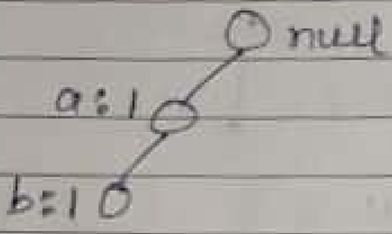
$$= \frac{6 + 0.5 \times 4}{24} = \frac{6 + 2.0}{24}$$

$$= \frac{8.0}{24} = 0.333$$

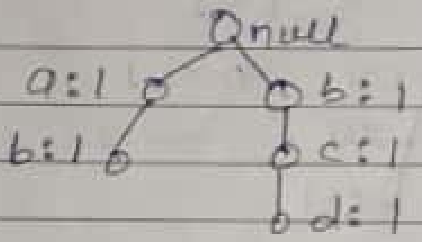
.5
4
2

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{milk, Diapers, Beer, Eggs}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, milk, Diapers, Cola}

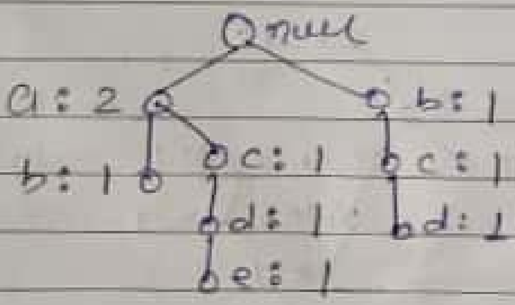
T_1 =



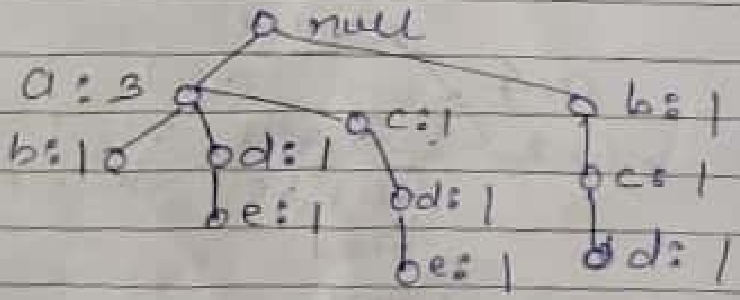
T2 =



T3 =



T4 =



20/5/20

Frequent Pattern tree

FP Growth Tree

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

a - 8
b - 7
c - 6
d - 5
e - 3

Support Count ⇒

1. Arrange the items in transaction according to the Support Count in decreasing fashion.

Theory

① Evaluate the performance of classifies ?

② Rule based classifies.
Rule evaluation -> Problem.
Statistical
laplace ✓
m estimate ✓
boyce ✓

1. Avoid too many Candidate generations
2. Candidate should be complete
3. Candidate itemset should not generated more than once.

(1.) Brute - force Method

{a, b, c, d}

a	ab	abc	abcd
b	ac	abd	
c	ad	acd	
d	bc	bcd	
	bd		
	cd		

(2) $F_{K-1} \times F_1$ method

$F_1 \Rightarrow$ 1 - Frequent Item set
$F_K \Rightarrow$ K - Frequent Item set

$F_K = F_{K-1} \times F_1$
$F_3 = F_2 \times F_1$
$F_5 = F_4 \times F_1$

e.g. F2-Itemset F_1 - Itemset $F_3 = f_2 \times F_1$

F2-Itemset
a, b
b, c
a, c
b, d

F_1 - Itemset
a
b
c
d

$F_3 = f_2 \times F_1$
a, b, c
b, c, a
a, b, d
a, c, d
b, d, a
b, d, c
b, c, d

Chapter=1

Ⓠ What is Data Mining?

→ Data mining is the process of automatically discovering useful information in large data respositories.

→ Datamining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown.

→ They also provide capabilities to predict the outcome of a future observation

→ Data mining is an integral part of knowledge discovery in databases (KDD), which is the overall process of converting raw data into useful information.

→ The input data can be stored in a variety of formats (flat files, spreadsheets, or relational tables) and may reside in a centralized data repository or be distributed across multiple sites.