No:

Data Mining
Part 01

Date: 21/08/2023

- Data mining is the process of discovering patterns, trends, correlations, and insights from large sets of data. It involves using various techniques and algorithms to extract valueble information and knowledge from raw data, often with the goal of making informed decisions, predicting future trends, or identifying hidden relationships.

Steps

Data Collections

- Gathering the relevant data from various sources, such as databases, websites and more

Data Prepreceссing

- Cleaning and preparing the data for analysis.

Exploratary Data Analysis

- Understanding the data by visualizing and summarizing its key characteristics.

Feature selection / Extraction

- Choosing the most relevent features from the data that will be used in the analysis.

Choosing Data Mining Techniques

- Common techniques include clustering, classification, regression, association rule mining, and more.

Applying Data Mining Algorithms

- decision trees, NNs, clustering, and more.

Interpreting results

- Analyzing the results of the algorithms to extract meaningful insights.

Validation and Evaluation

- Assessing the quality of the discovered patterns and models.

Atlas

No: Date:

Deployment

- Implementing the insights and models into world applications.

Continuous Monitoring and Maintenance

- As new data becomes available, the models need to be updated or adjusted to ensure they remain accurate and relevent.

Why Data mining

- With the rapid growth of digital info., data mining helps extract meaningful and relevant information from vast amount of data.
- DM techniques can identify hidden patterns, correlations and relationships within the data.
- Organizations use DM to gain insights into customer behavior, preferences, and bying patterns.
- DM aids in identifying potential risks and fraud patterns

DM Applications

Prediction and description

- Prediction involves using historical data to make informed forecasts about future outcomes. It is about building models that can predict an unknown or future value based on the pattern and relationships identified in the available data.

Customer profiling, Fraud detection, Customer segmentation

CRISP DM Process

Business Understanding

- In this phase, the project team works closely with stackholders to understand the business objectives, requirements and contraints. Define the problem or opportunity that the DM project aims to address.

Atlas

No: Date: / /

Data Understanding

- Tasks include ~~theta~~ data collection, data description, data exploration, and initial data quality assessment

Data preparation

- data cleaning, data transformation, data integration

Modeling

- Actual DM techniques are applied to the prepared data set. The performance of different models is evaluated to identify the most suitable one.

DM Tasks

Exploratory Data Analysis

- Data inspection, Data visualization, Summary statistics, Handling missing data, Identifying outliers, pattern recognition, Data transformation,

05/09/2023

Multi-Tiered Architecture

- if we noticed that monthly sales will be calculated frequently, we can calculate monthly sales and store it in the data warehouse. (Materialization)

- Full materialization- do summarization for all the variables and all the levels

10/09/2023

Data Preprocessing

- Today's real-world databases are highly susceptible to noicy, missing and inconsistent data due to their typically huge size, and their likely origin from multiple heterogenous sources.
  - Fields that are obsolete or redundant
  - Missing values
  - Outliers
  - Data in a form not suitable for data mining models

No: Date: __/__/____

Predictive modeling - This is a composition of classification, Regression, Time series Analysis

descriptive modeling - unsupervised learning or clustering

Association rule mining / Market Basket Analysis

- The aim of association rule mining is to determine which items are purchased together frequently, so that they may be grouped together on store shelves or the information may be used for cross-selling.

- In the content of retail and sales, these rules can provide valueble insights into customers behavior and preferences.
  - ↳ Determining items purchased together
  - ↳ Grouping items on store shelves
  - ↳ cross-selling opertunities - If customers often buy smart phones, association rule might reveal that they also tent to purchase phone cases, screen protectors.
  - ↳ personalization - Online retailers might suggest related products based on a customer's browsing and purchase history.
  - ↳ inventory management - Retailers can ensure that they have sufficient stock of related items and avoid stockouts or overstock situations.

- process for modeling and viewing data in multiple dimensions called OLAP.
- OLTP - Day-to-Day operations: purchasing, banking, registration..

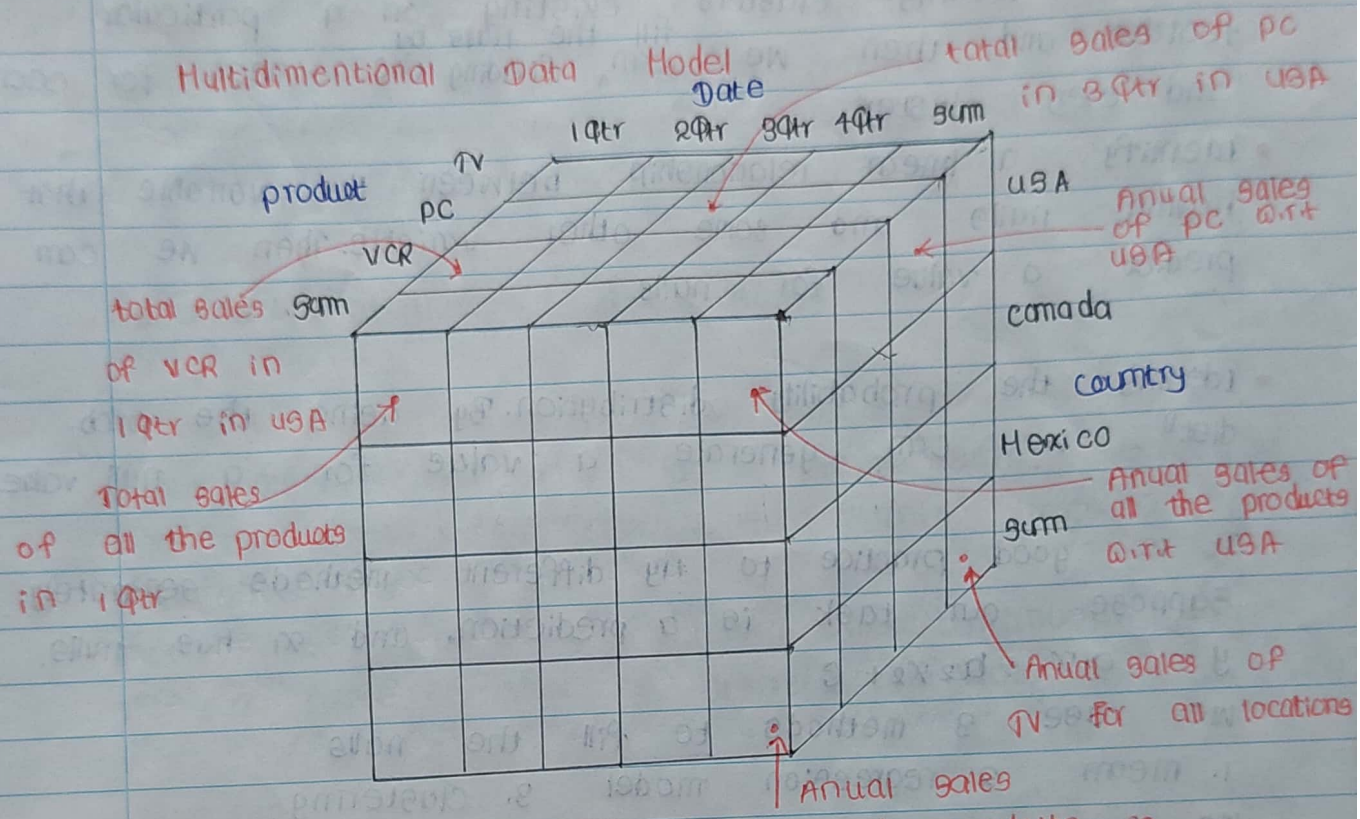
Ad-hoc queries are dynamic queries that are not predefined or stored as a part of a structured query language (SQL) program. These queries are typically used for one-time or infrequent data retrieval and analysis tasks.

- A data warehouse organized around major subjects such as customer, supplier, product, sales etc.

No: Date: / /

Time-variant : A data warehouse is designed to capture and store historical data and provide historical perspective on the information it contains.

- Usually - modelled by multidimensional database structure. It provides multidimentional view of data



- Data warehouse - Fact constellation schema is used
  Data mart - Star schema or Snowflake schema is used frequently.

- measures of ~~data~~ central tendency include mean, median, mode, and midrange, while measures of data dispersion include quartiles, interquartile range (IQR), and variance.

11/09/2023

Data Cleaning

Missing Data

- We can identify clusters existing on a particular feature and then we can fill the nulls by finding the mean for each and every cluster.
- Identify a linear relationship between the variable that having nulls and some other variable. Then we can predict a value for nulls.

- Identify the probability distribution. By using the prob. dist^n, we can generate a value for a null value.

It is good practice to try different methods separately.
Suppose our task is a prediction, and x_1 has nulls.
$y = a_1x_1 + b_2x_2 + c$
We use 3 methods to fill the nulls
1. mean 2. regression model 3. clustering
We have to try each and every method separately and check the accuracy of the model.

Handling Redundancy

- Object identification- we can use a similarity measure to identify an entity.

Ex. in one place a person has one name and in another place the same person has a different name. In these kind of situations, we can use other variables like age, id etc to detect whether this person is same or not.

No :

Date: / /

Decimal Scaling

- In a particular column, suppose the largest value is 980, 000 000. We have to get the largest value just before 1. [0, 1]

$\frac{x}{10^9} \Rightarrow 0.98$

Another value - 12 000 000 → 0.012

Data smoothing

- DOWN round to the nearest integer.

18/09/2023

Attribute Subset Selection

- Redundant attributes- multicollinearity

data discretization

- Data discretization is a data preprocessing technique use to convert continuous data into discrete or categorical data.
- process of grouping or binning these continuous values into a finite number of discrete intervals or categories

Sampling

- sample should be representative.

Data Cube Aggregation

lossless compression - if we compress an image then it should be the same as the original image. That is it will not destroy the information.

Equi-frequency binning

0, 4, 12, 16, 16, 18, 24, 26, 28

$\frac{12+16}{2} = 14$

$\frac{18+24}{2} = 21$

Atlas

No: Date: __/__/____

- Entropy-based binning- take different combinations and calculate information gain.

ChiMerge

- chi-square test

$H_0: q_1 = q_2$. if we accept H_0, then we can combine those 2 intervals.

$E_{11} = \frac{R_1 \times C_1}{N}$ ← column total; row total; N ← total # of obs.

- if we get E as 0, then we can take it as 0.1 from the table we can directly get the observed value.

Ex.

x	y	discrete_x	
0	p	Bin1	[-0.026, 9.33)
4	p	Bin1	[9.33, 18.66)
12	p	Bin2	[18.66, 28)
16	n	Bin2	
16	n	Bin2	
18	p	Bin2	
24	n	Bin3	
26	n	Bin3	
28	n	Bin3	

25/09/2023

Association Rule Mining

- Association rule mining is a data mining technique used to discover interesting relationships or patterns within large datasets. It focuses on finding associations or correlations among items in a dataset.

- An item refers to a unique element or object in a dataset.

Atlas

No :

Date: __/__/____

- A transaction is a collection of items that are bought or used together.
- support is a measure of how frequently a specific item or combination of items appears in the dataset. High support indicates that the itemset is popular or common in the dataset.

- confidence measures the strength of an association rule between 2 itemsets. High confidence suggets that if itemset A is present in a transaction, there is a high likelihood that itemset B will also be present in the same transaction.

- Association rule mining is a valueble technique for discovering hidden patterns and relationships in large datasets, which can be used for decision-making, marketing strategies, and improving business processes.

- $X = Y$
  if x in the basket, there is a high probability that item Y can be in the same basket.

  $n = (X, y)$, # of transactions that x, y occur together
  $N =$ Total # of transactions

  support $(X, y) = P(X, y) = \frac{n}{N}$ ← support frequency.

- confidence $= P(Y|X) = \frac{n(X \cup Y)}{n(X)}$

- consider 3 items (A, B, C)
  rules:
  $A \Rightarrow B$ $B \Rightarrow C$ $C \Rightarrow A$
  $B \Rightarrow A$ $C \Rightarrow B$ $A \Rightarrow C$
  Rules that satisfies both a minimum support or threshold and a minimum confidence threshold are called strong

No: Date: __/__/____

- Total # of rules for m # of items, $= \sum_{k=2}^{m} {}^{m}C_k (2^k - 2)$

Example

$A \Rightarrow C$

$n = 2 \quad N = 4$

support $(A, C) = \frac{2}{4} = 0.5 = 50\%$.

$P(C|A) = \frac{n(A \cup C)}{n(A)} = \frac{2}{3} = 0.6667 = 66.67\%$.

we will consider this as a valid rule.

$C \Rightarrow A$

$n = 2 \quad N = 4$ support $(A, C) = \frac{2}{4} = 0.5 = 50\%$.

$P(A|C) = \frac{n(A \cup C)}{n(C)} = \frac{2}{2} = 1 = 100\%$.

{A}	= 111 (3)	{A, B}	= 1 (1)	{A, B, C}	= 1 (1)
{B}	= 11 (2)	{A, C}	= 11 (2)	{B, E, F}	= 1 (1)
{C}	= 11 (2)	{B, C}	= 1 (1)		
{D}	= 1 (1)	{A, D}	= 1 (1)		
{E}	= 1 (1)	{B, E}	= 1 (1)		
{F}	= 1 (1)	{B, F}	= 1 (1)		
		{F, E}	= 1 (1)		

out of all combinations, {A}, {B}, {C}, {A, C} are the frequent item sets.

$P(C|A) = \frac{n(A \cup C)}{n(A)} = \frac{n(A \cup C)/N}{n(A)/N} = \frac{\text{support}(A, C)}{\text{support}(A)}$

- Let take frequent set {A, C}. Consider subsets {A}, {C}. If {A}, {C} is a frequent set, then {A}, {C} both are frequent set. – A priori principal

if {A} is a non freq. set, then {A, C} is also a non freq. set.

No: Date: / /

{A}	= 75%	{A}	= 75%
{B}	= 50%	{B}	= 50%
{C}	= 50%	{C}	= 50%
{D}	= 25%	{E}	= 50%
{E}	= 50%		

Candidate item set → frequent item sets.

Example

C_1 = one item candidate set

sup = 50%

$\frac{x}{4} \times 100 = 50 \Rightarrow x = 2$

if frequency < 2 is a non freq. set

$L_2 \rightarrow C_3$

Consider the 1st item of any 2 sets.

{1, 3}, {2, 3}. If they are same, combine them.

{2, 3}, {2, 5} ⇒ {2, 3, 5}

prune step

{1, 2, 4} ⇒ {1, 2}, {1, 4}, {2, 4}

↑ non freq. set. ∴ {1, 2, 4} is also a non-freq. set.

Generating Candidate Item sets $(C_1 C_k)$

p = {$A_1, A_2, \ldots, A_n, p$}
q = {$A_1, A_2, \ldots, A_n, q$}
} → {$A_1, A_2, A_3, \ldots, A_n, p, q$} ($A_1 \ldots A_n$ common) $p < q$

candidate sets.

{abcd}, {acde}

↑ ab is common

• c is prior than d.

{abcd} ⇒ {abc}, {abd}, {acd}, {bcd}

{acde} ⇒ {acd}, {ace}, {ade}, {cde} {ade} is not in L_3

No: Date: __/__/____

∴ {a c d e} is a non-frequent set.

Generating Rules from Frequent Itemset

I = {A, B, C, D}

let S = {A, B, C} sup(S) ≥ sup(I)

{A, B, C, E} ⇒ {D}

Consider combination. D in right side, that is like {D, E}

Then {A, B, C} is mail. E comes from the left.

support ↑ confidence ↓

26/09/2023

Improving the Efficiency of Apriori

- we have to scan the transaction table several times

Hash-Based Techniques

- Reduce the # of transactions to be scanned.
- We put the item sets into different bins.

- consider {1, 3}

h(x, y) = ⌈(1×10) + (3×1)⌉ % 7 = 13 % 7 = 6

In this way we can find the position for each and every item set.

- After building the 1st hash table, again ~~th~~check the transaction table. If we identify any non-frequent item from the previous step, remove the non-freq. item set from the transactional table. (or whole transaction)

Ex. {4} is a non-frequent item. ∴ {1, 3, 4} is also a non-freq. set.

3-item hash fn.

{2, 3, 5} → ⌈(2×100) + (3×10) + (5×1)⌉ % 7 = 4

{1, 2, 3, 5} → {1, 2, 3}, {1, 3, 5}, {1, 2, 5}, {2, 3, 5}

No: Date: / /

- Once we removed an item from the transaction table, then it will no longer available

- All the bins assign with 1 or 0. (According to the minimum support count). If the assign value is 1, then we can take item sets from that bin a candidate sets. (based on the Bucket count)

C_2

{1,4}, {3,5}, {2,3}, {3,5}, {2,3}, {3,4}
 ‖ ‖ ‖ ‖ ‖ ‖
 1 2 2 3 2 1

Now we can generate L_2.

L_2

{3,5}, {2,3}, {2,5}, {1,3}

C_3

{2,3,5}, {1,2,3}

- remember: When getting these count, also check whether there are non-freq. item sets or not. In this step, it is based on C_2.

Frequent Pattern Growth (FP-Growth) Algorithm.

Ex.

Transaction ID	Items
T1	{E, K, H, N, O, Y}
T2	{D, E, K, N, O, Y}
T3	{A, E, K, H}
T4	{C, K, H, U, Y}
T5	{C, E, I, K, O, O}

Let the minimum support be 3

Item	Freq.
A	1
C	2
D	1
E	4
I	1
K	5
H	3
N	2
O	3
U	1
Y	3

Atlas

No: Date: __/__/____

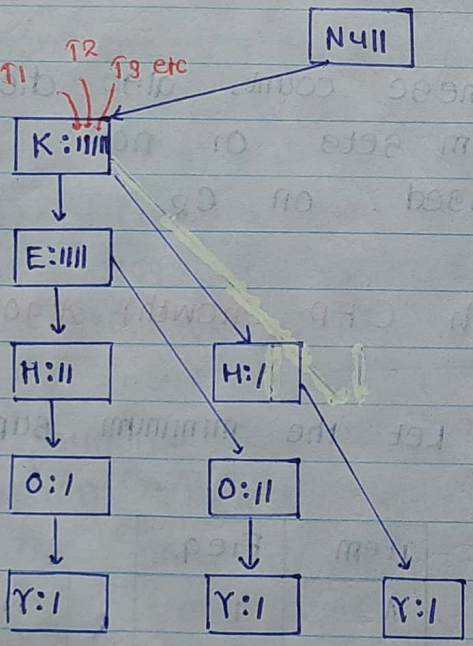
- Frequen pattern set (L) - contain all the elements whose freq. is greater than or equal to the minimum support.

  L = { K : 5, E : 4, H : 3, O : 3, Y : 3 }

- ordered - item set

Transaction ID	Items	Ordered-Item set
T1	{E, K, H, N, O, Y}	{K, E, H, O, Y}
T2	{D, E, K, N, O, Y}	{K, E, O, Y}
T3	{A, E, K, H}	{K, E, H}
T4	{C, K, H, U, Y}	{K, H, Y}
T5	{C, E, I, K, O, O}	{K, E, O}

- Now all the Ordered-item sets are inserted into a Trie Data structure.
- Consider 1st Transaction.



- Conditional pattern base is computed which is path labels of all the paths which lead to any node of the given item in the frequent pattern tree.

Items	Conditional Pattern Base
Y	{{K,E,M,O:1}, {K,E,O:1}, {K,M:1}}
O	{{K,E,M:1}, {K,E:2}}
M	{{K,E:2}, {K:1}}
E	{{K:4}}
K	

- Conditional Frequent Pattern Tree

It is done by taking the set of elements which is common in all the paths in the Conditional Pattern Base of that item.

Items	Conditional P.B	C. F. Pattern Tree
Y	{{K,E,M,O:1}, {K,E,O:1} {K,M:1}}	{K:3} (1+1+1)
O	{{K,E,M:1}, {K,E:2}}	{K,E:3} (2+1)
M	{{K,E:2}, {K:1}}	{K:3} (2+1)
E	{{K:4}}	{K:4}
K		

Items	Frequent Pattern Generated
Y	{<K,Y:3>}
O	{<K,O:3>, <E,O:3>, <K,E,O:3>}
M	{<K,M:3>}
E	{<E,K:4>}
K	

- Frequent pattern rules

L = {K, Y}

1) $K \longrightarrow Y$
2) $Y \longrightarrow K$

$K, Y \longrightarrow \Phi$

1) $Conf = \frac{3}{5} = 0.6 = 60\%$.

2) $conf = \frac{3}{3} = 1 = 100\%$.

No: Date: / /

Ex 2

TID	Items
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I1, I2, I3, I5
9	I1, I2, I3

L1

Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Items	Sup. count
I2	7
I1	6
I3	6
I4	2
I5	2

Association rules can be generated as follows,

- For each frequent item set l, generate all non-empty subsets of l.
- For every non-empty subsets s of l, output the rule '$s \Rightarrow (l - s)$' if $\frac{\text{support\_count}(l)}{\text{support\_count}(s)} \geq$ min conf,

where min_conf is the minimum confidence threshold.

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}

Item	Frequency
a	8
b	7
c	6
d	5
e	3

if we have non-freq. items we can remove

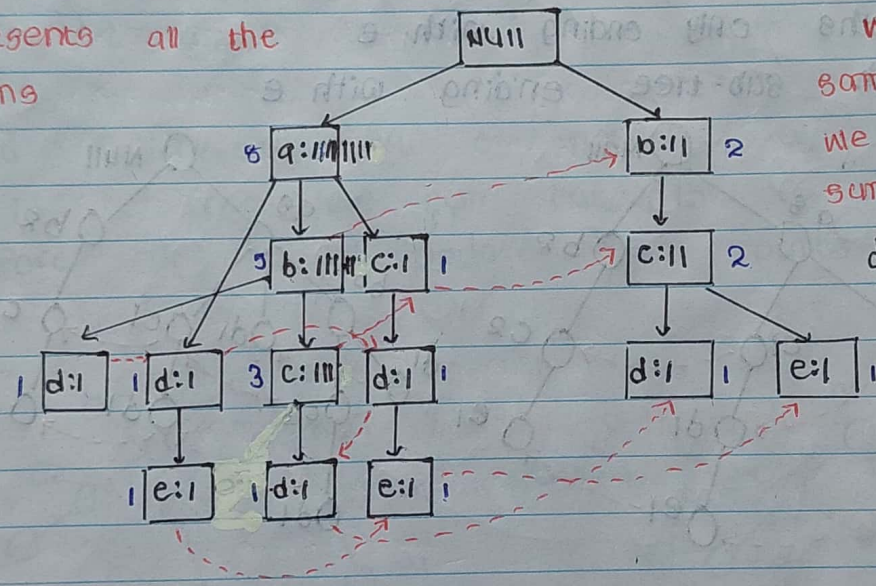
minimum support = 2

L = Frequent pattern set (descending order)
= {a:8, b:7, c:6, d:5, e:3}

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}

We can share the common items.

This represents all the transactions



We linked the same items.

We have to get sum. Ex: e = 3 (count) (1+1+1)

d = 5 (1+1+1+1+1)

c = (3+2 = 5)

a = 8

b = (5+2 = 7)

C:1 C:1

Item	Conditional P.B	C.P.T
e	{a, d:1} {a, c, d:1} {b, c:1}	{a:2, d:2, c:2, b:1}
d	{a, b, c:1} {b, c:1} {a, b:1} {a:1} {a, c:1}	{a:4, b:2, c:2} {b:1, c:1}
c	{a, b:3} {a:1} {b:2}	{a:4, b:3} {b:2}
b	{a:5}	{a:5}
a	ϕ	

↑ only items which count ≥ min support

No: Date: / /

item	Frequent Itemsets
e	{e}, {a,e}, {d,e}, {c,e}, {a,d,e}
d	{d}, {a,d}, {b,d}, {c,d}, {a,b,d}, {b,c,d}, {a,c,d}
c	{c}, {a,c}, {b,c}, {a,b,c}
b	{b}, {a,b}
a	{a}

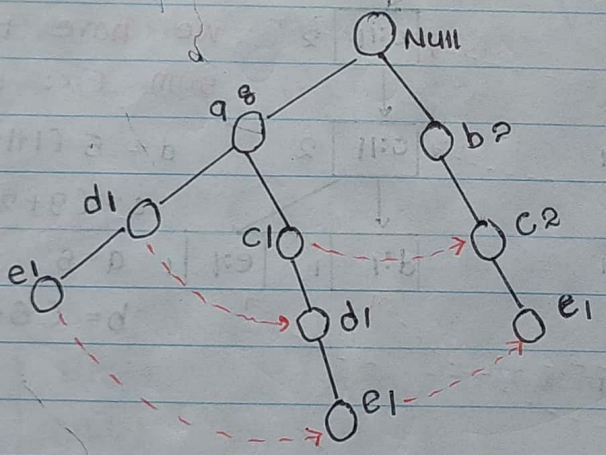
Consider {c,e}

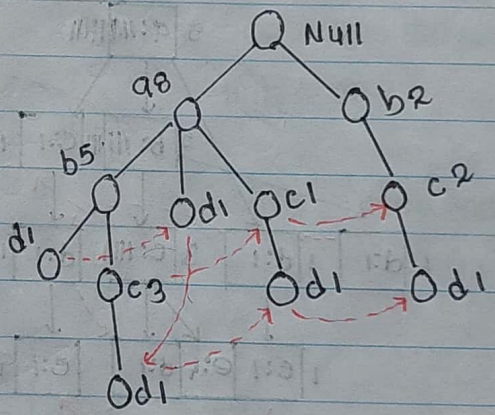
02/10/2023

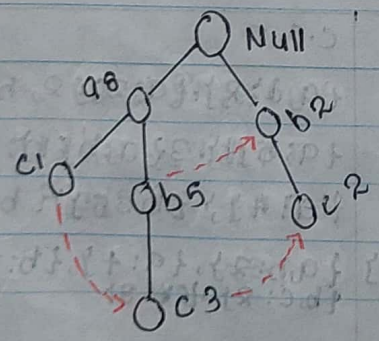
Example

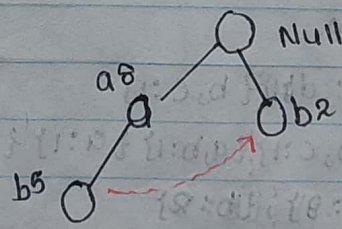
We start with least frequent item. That is e.
Extract paths only ending with e.

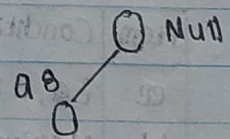
- prefix path sub-tree ending with e











- Consider sub-tree of 'e'
count of e is 3 ≤ 2 ∴ {e} is a freq. item set
Then combination of e can be a freq. item (2 items)

Item	Freq. item
a	
b	
c	
d	
e	{e}

e → {e,d}, {c,e}, {b,e}, {a,e}

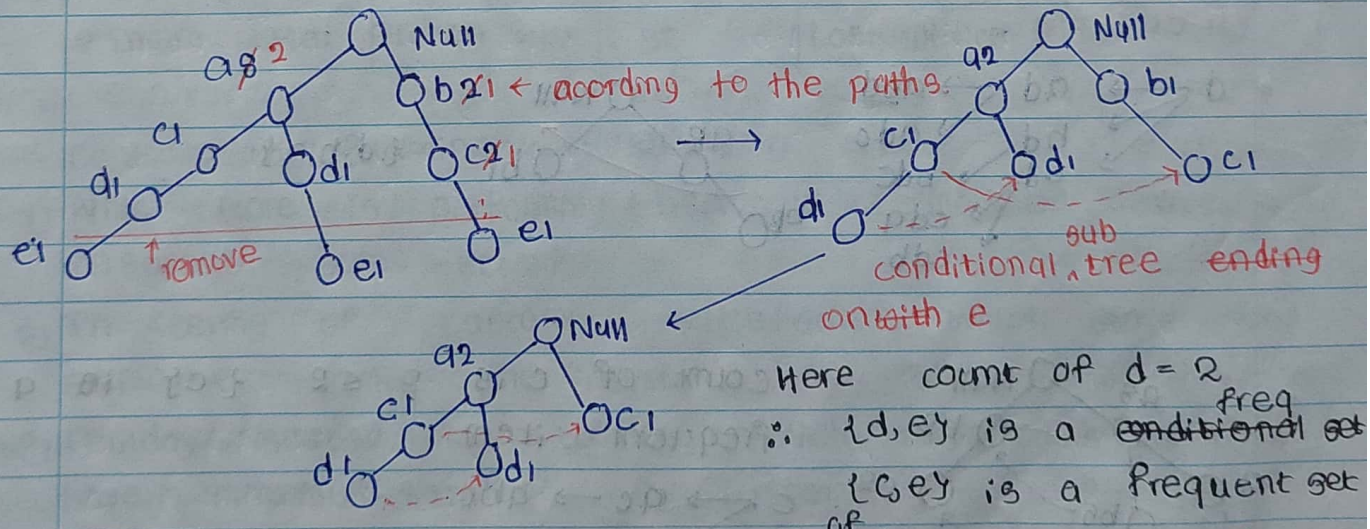
d → {c,d}, {b,d}, {a,d}

c → {b,c}, {a,c}

b → {a,b} {a}

we can use the sub-tree of e to find these item set are freq.

- find {e,d} is freq. or not
e is a freq., we can cut the e from the subtree. before cut the 'e' update the counts of each node



Here count of d = 2

∴ {d,e} is a freq

{c,e} is a frequent set

freq of {b,e} = 1 remove b

{a,e} freq. item

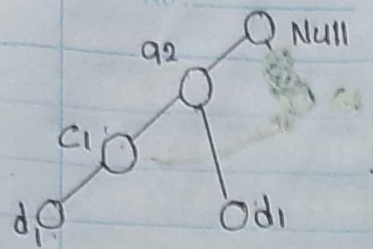
Now {d,e} is a freq. item

de → cde, bde, ade

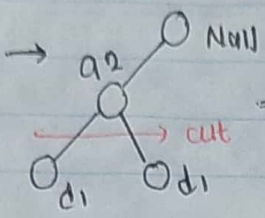
these can be freq. items.

some for ae, ce

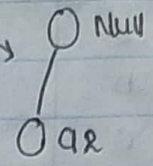
↳ we have to get prefix path sub tree ending with de



{c, d, e} = 1

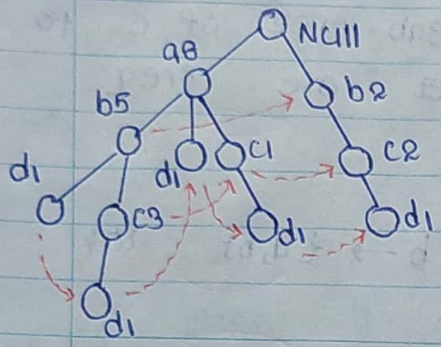


{a, d, e} = freq. ↔

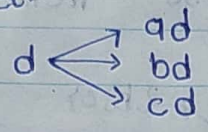


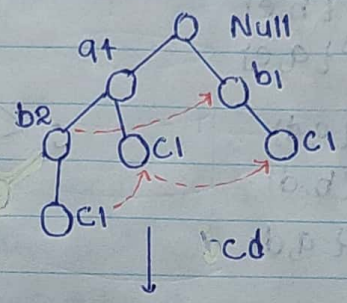
⇒ conditional sub tree on 'de'

- sub-tree of d

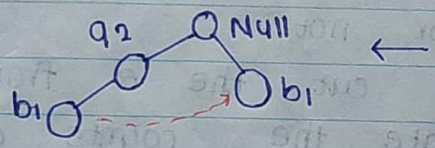


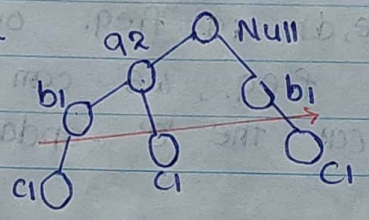
count of d is 4. ≤ 2 ∴ {d} is a frequent set.



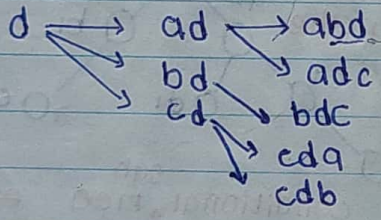


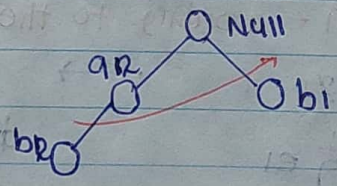
{d, a}
{d, b}
{d, c}



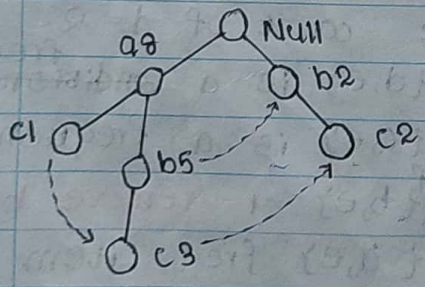


{b, c, d} , {a, c, d}

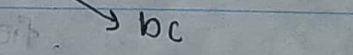


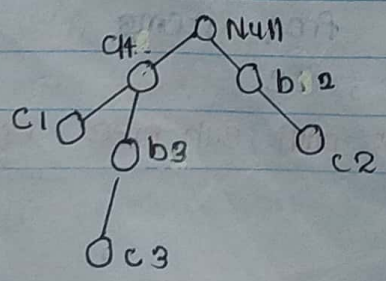


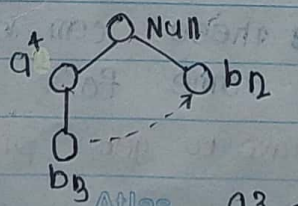
{d, b, a}



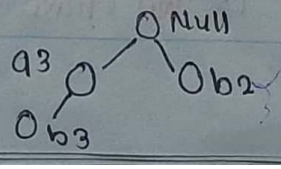
count of c is 3 ≤ 2 {c} is a frequent item.

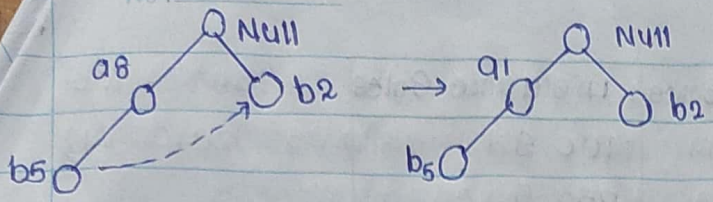






{b, c} , {a, c}
{a, b, c}





{b} is a freq. set

08/10/2023

OLAP vs OLTP

- Consider the following 2 data stores of 2 different companies Sales Data.

Store Name	Cust Name	Device	Price	Sales Data
3 Jay St, NY	John Smith	iPhone11	1100 $	20 Jan 2018
3 Jay St, NY	Brad Pitt	Pixel 4	850 $	15 Sep 2018
3 Jay St, NY	Maria David	Vivo Z5i	450 $	5 Jan 2020
2 Indira Nagar	Mohan S	iPhone10	80,000 RS	2 Jan 2020

Insurance Data

Store Name	Cust Name	Period	Plan	Price
3 Jay St, NY	John Smith	2 yr	Device Damage	50 $
3 Jay St, NY	Brad Pitt	5 yr	Screen Protection	30 $
3 Jay St, NY	Maria David	6 months	Lost Device	20 $
2 Indira Nagar	Mohan S	1 yr	Lost Device	17,500 RS

- Consider the following questions.

1) Whic store is performing best in terms of device and insurance sales total?
2) In terms of customer satisfaction which store and employee ranks the best?
3) Holiday season is comming, which region is going to have maximum trafic of customers?

Aggregation

Store Name	Device Sales	Insurance Sales
3 Jay St, NY	2350 $	100 $
2 Indira Nagar	80,000 RS	17,500 RS

can not compare, because of the difference currency.

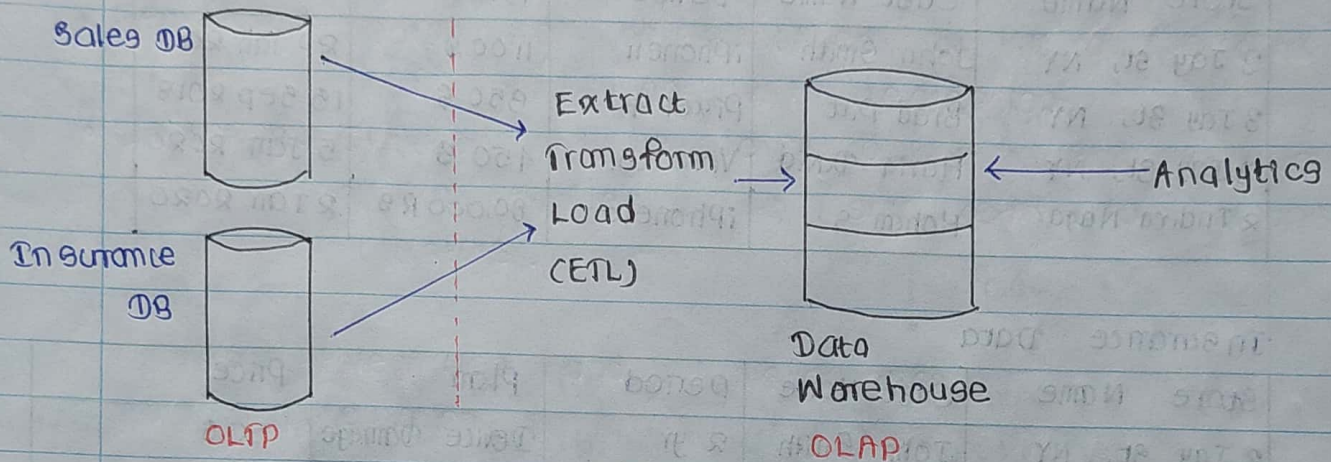
No: Date: __/__/____

Normalization

Store Name	Device Sales	Insurance Sales
3 Jay st, NY	2350 $	100 $
2 Indira Nagar	1142 $	250 $

We can store this data in another database called Data Warehouse.

- Here, we performed basically 3 steps.
Data Extraction, Data Transformation, Data Loading
Now we can perform analysis based on the data warehouse



Disadvantages of using OLTP for Data mining

Performance Impact

- Data mining queries involve complex aggregations and scans of large datasets, which can place a significance burden on OLTP databases, leading to performance degradation for regular transactional operations.

Inefficient Data Retrieval

- OLTP databases are typically normalized. This normalized structure can make it inefficient and time consuming to retrieve the denormalized data often needed for data mining tasks.

- OLTP databases, are not designed for complex analytical functions required in data mining, such as clustering, regression analysis, or machine learning.

No :

Date: __/__/____

- Total number of cuboids = $\prod_{i=1}^{n}(L_i + 1)$
  L_i is the number of levels associated with dimension i
  1 is added to L_i to include the virtual top level, all.

- No materialization - do not precompute any of the nonbase cuboids
  Full materialization - precompute all of the cuboids.

Data Preprocessing

- We noticed that several of the attributes for various tuples have no recorded value. For our analysis, we would like to include information as to whether each item purchased was advertised as on sale, yet we discover that this info. has not been recorded.

The data we wish to analyze by data mining techniques are incomplete, noisy, and inconsistent.

Descriptive data summarization

- Can be used to identify the typical properties of our data and highlights which data values should be treated as noise or outliers.
  ↳ mean, median, mode, quantiles, IQR, variance

- Histograms, quantile plots, q-q plots, scatter plots are very helpful for the visual inspection of our data.

Data cleaning.

Missing values

- Ignore the tuple
- Fill in the missing value manually
- Use a global constant to fill in the missing value. (like 'Unknown')

No: Date: / /

- Use the attribute mean to fill in the missing value.
- Use the attribute mean for all samples belonging to the same class as the given tuple.
- Use the most probable value to fill in the missing value (regression, Bayesian formalism, decision tree etc.)

Noisy Data

- Binning - The sorted values are distributed into a number of buckets.
- Regression - Data can be smoothed by fitting the data to a fn, such as with regression.
- Clustering - Outliers may be detected by clustering, where similar values are organized into groups or 'clusters'.

Data Integration and Transformation

Data Integration

- Data analyst task will involve data integration, which combines data from multiple sources into a coherent data store, as in data warehousing.

Data Transformation

- In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following.

Smoothing - works to remove noice from the data. Such techniques include binning, regression, and clustering.

Aggregation - summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly or annual total amounts.

Generalization - low-level or primitive data are replaced by high-level concepts through the use of concept hierarchies.

No: Date: / /

Ex. categorical attribute, like street, can be generalized to high level concept, like city or country.
age → youth, middle age, senior.

Normalization- attributes data are scaled so as to fall within a small specified range, such as -1 to +1 or 0 to 1

Attribute construction- new attributes are constructed and added from the given set of attributes to help the mining process.

- Min-max normalization- $v' = \frac{v - min_A}{max_A - min_A}$

z-score normalization- $v' = \frac{v - \bar{A}}{\sigma_A}$

Data Reduction

- Data reduction techniques can be applied to obtained a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.

1. Data cube aggregration
2. Attribute subsets selection
3. Dimensionality reduction
4. Numerosity reduction
5. Discretization and coneept hierarchy generation.

Attribute subset selection

- Stepwise forward selection
- Stepwise backward elemination
- combination of forward selection and backward elemination
- Decision tree induction.

- If the original data can be reconstructed from the compressed data without any loss of infor., the data reduction is called lossless.

No: Date: __/__/____

If, instead, we can reconstruct only an approximation of the original data, then the data reduction is called lossy.

Data Discretization and Concept Hierarchy Generation

- Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals.
Replacing numerous values of a continuous attribute by a small number of intervals labels thereby reduces and simplifies the original data.

Binning

- Attribute values can be discretized by applying equal-width or equal-frequency binning, and then replacing each bin value by the bin mean or median.

Entropy-based Discretization

- Calculate the 'Entropy' for the target.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2(p_i)$$

$$= -\frac{7}{24} \times \log_2\left(\frac{7}{24}\right) + \left(-\frac{17}{24}\right) \log_2\left(\frac{17}{24}\right)$$

$$= 0.871$$

- Calculate 'Entropy' for the target given a bin

$$E(S, A) = \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$$

$$= \frac{|S_1|}{|S|} E(S_1) + \frac{|S_2|}{|S|} E(S_2)$$

S_1 ; $A \leq$ split point S_2 ; $A >$ split point

No: Date: __/__/____

A = temperature.

$$E(S, A) = \frac{|S_1|}{|S|} E(S_1) + \frac{|S_2|}{|S|} E(S_2)$$

$S_1 = A \leq 60$ $S_2 = A > 80$

$$E(S_1) = -\left(\frac{3}{3}\right) \log_2(3/3) - (0/3) \log_2(0) = 0$$

$$E(S_2) = -\left(\frac{4}{21}\right) \log_2(4/21) - (17/21) \log_2(17/21)$$

$$= -\frac{4}{21} \times (-2.3923) - \frac{17}{21} (-0.30495)$$

$$= 0.702$$

$$E(S, A) = \left(\frac{3}{24} \times 0\right) + \left(\frac{21}{24} \times 0.7\right) = 0.615$$

Let

9/10/2023

- $E(D) = -\sum_{i=1}^{c} p_i \log_2(p_i)$ Entropy

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times info(D_j)$$

P-338

Date: __/__/____

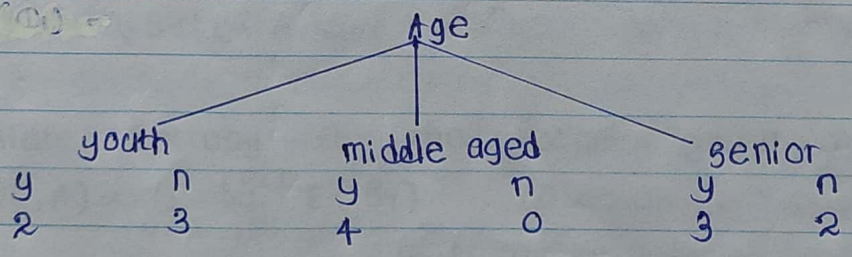
	Age	income	student	credit-rating	buys_computer
1	youth	high	no	fair	no
2	youth	high	no	Excellent	no
3	middle_aged	high	no	f	yes
4	senior	medium	no	f	yes
5	senior	low	yes	f	yes
6	senior	low	yes	ex	no
7	middle-aged	low	yes	ex	yes
8	youth	medium	no	f	no
9	youth	low	yes	f	yes
10	senior	medium	yes	f	yes
11	youth	medium	yes	ex	yes
12	middle-aged	medium	no	ex	yes
13	middle_aged	high	yes	f	yes
14	senior	medium	no	ex	no

$$E(D) = -\frac{9}{14}\log_2(9/14) - \frac{5}{14}\log_2(5/14)$$ yes = 9 no = 5

$$= 0.94$$ //

- First consider 'age'

$$info_A(D) = \frac{|D1|}{|D|} info(D_1) + \frac{|D2|}{|D|} info(D_2) + \frac{|D3|}{|D|} info(D_3)$$



$$info(D_1) = -\frac{2}{5}\log_2(2/5) - \frac{3}{5}\log_2(3/5) = 0.971$$

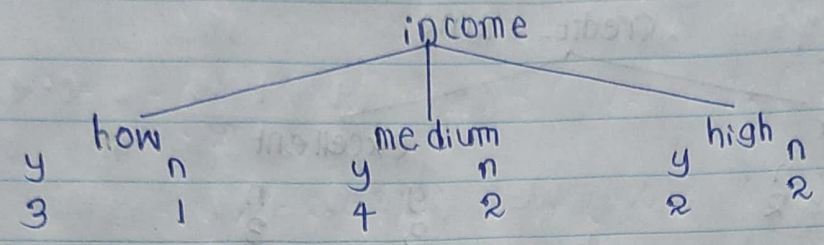
$$info(D_2) = -\frac{4}{4}\log_2(4/4) - \frac{0}{4}\log_2(0/4) = 0$$

$$info(D_3) = -\frac{3}{5}\log_2(3/5) - \frac{2}{5}\log_2(2/5) = 0.971$$

$info_A(D) = \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971 = 0.694$

info gain (Age) = 0.94 − 0.694 = 0.246 //.

- Now consider 'income'



$info_I(D) = \frac{|D_1|}{D} info(D_1) + \frac{|D_2|}{|D|} info(D_2) + \frac{|D_3|}{|D|} info(D_3)$

$info(D_1) = -\frac{3}{4} \log_2(3/4) - \frac{1}{4} \log_2(1/4) = 0.81$

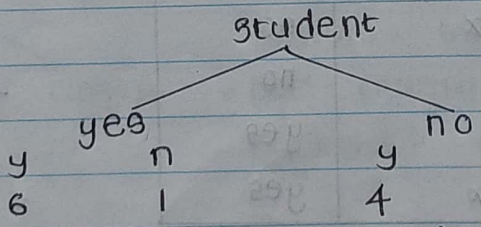
$info(D_2) = -\frac{4}{6} \log_2(4/6) - \frac{2}{6} \log_2(2/6) = 0.92$

$info(D_3) = -\frac{2}{4} \log_2(2/4) - \frac{2}{4} \log_2(2/4) = 1$

$info_I(D) = \frac{4}{14} \times 0.81 + \frac{6}{14} \times 0.92 + \frac{4}{14} \times 1 = 0.91$

information gain (Income) = 0.94 − 0.91 = 0.03

- Now take 'student'



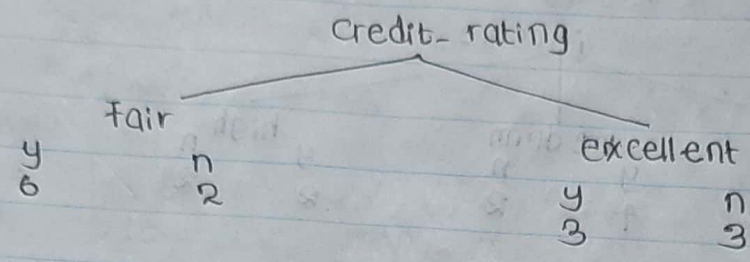
$info_S(D) = \frac{|D_1|}{|D|} info(D_1) + \frac{|D_2|}{|D|} info(D_2)$

$info(D_1) = -\frac{6}{7} \log_2(6/7) - \frac{1}{7} \log_2(1/7) = 0.59$

$info(D_2) = -\frac{1}{7} \log_2(1/7) - \frac{3}{7} \log_2(3/7) = 0.99$

$info_S(D) = \frac{7}{14} \times 0.59 + \frac{7}{14} \times 0.99 = 0.79$

info. gain (S) = 0.94 − 0.79 = 0.15 //.



$info_C(D) = \frac{|D_1|}{|D|} info(D_1) + \frac{|D_2|}{|D|} info(D_2)$

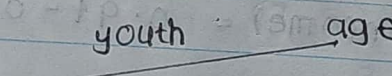
$info(D_1) = -\frac{6}{8}\log_2(6/8) - \frac{2}{8}\log_2(2/8) = 0.81$

$info(D_2) = -\frac{3}{6}\log_2(3/6) - \frac{3}{6}\log_2(3/6) = 1$

$info_C(D) = \frac{8}{14} \times 0.81 + \frac{6}{14} \times 1 = 0.89$

info. gain (C) = 0.94 − 0.89 = 0.05 //

- highest information gain belongs to 'age'.



income	student	credit_rating	class
high	no	f	no
high	no	ex	no
medium	no	f	no
low	yes	f	yes
medium	yes	ex	yes

$info(D) = -\frac{2}{5}\log_2(2/5) - \frac{3}{5}\log_2(3/5) = 0.971$

$info_I(D) = \frac{|D_1|}{|D|} info(D_1) + \frac{|D_2|}{|D|} info(D_2) + \frac{|D_3|}{|D|} info(D_3)$

$= \frac{1}{5}[-1 \times \log_2(1)] + \frac{2}{5}\left[-\frac{1}{2}\log_2(1/2) - \frac{1}{2}\log_2(1/2)\right] + \frac{2}{5}\left[-\frac{0}{2}\log_2(0/2) - \frac{2}{2}\log_2(2/2)\right]$

$= 0.4$

No:

Date: __/__/__

info gain (I) = 0.971 − 0.4 = 0.571 ||

- $\text{info}_S(D) = \frac{|D_1|}{|D|}\text{info}(D_1) + \frac{|D_2|}{|D|}\text{info}(D_2)$

$= \frac{2}{5}\left[-\frac{2}{2}\log_2(2/2) - 0\right] + \frac{3}{5}\left[-0 - \frac{3}{3}\log_2(3/3)\right]$

$= 0$

info gain (S) = 0.971 ||

- $\text{info}_C(D) = \frac{|D_1|}{|D|}\text{info}(D_1) + \frac{|D_2|}{|D|}\text{info}(D_2)$

$= \frac{3}{5}\left[-\frac{1}{3}\log_2(1/3) - \frac{2}{3}\log_2(2/3)\right] + \frac{2}{5}\left[-\frac{1}{2}\log_2(1/2) - \frac{1}{2}\log_2(1/2)\right]$

$= 0.951$

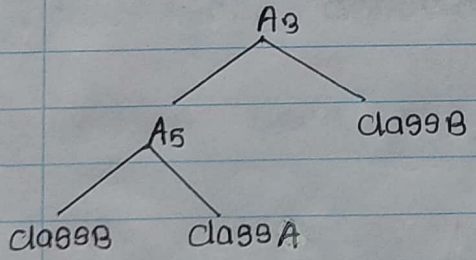
info. gain (C) = 0.971 − 0.951 = 0.02

highest information gain belongs to 'student'

10/10/2023

Tree Pruning and Scalability

- If we want to cut the A3, first consider the sub-tree.

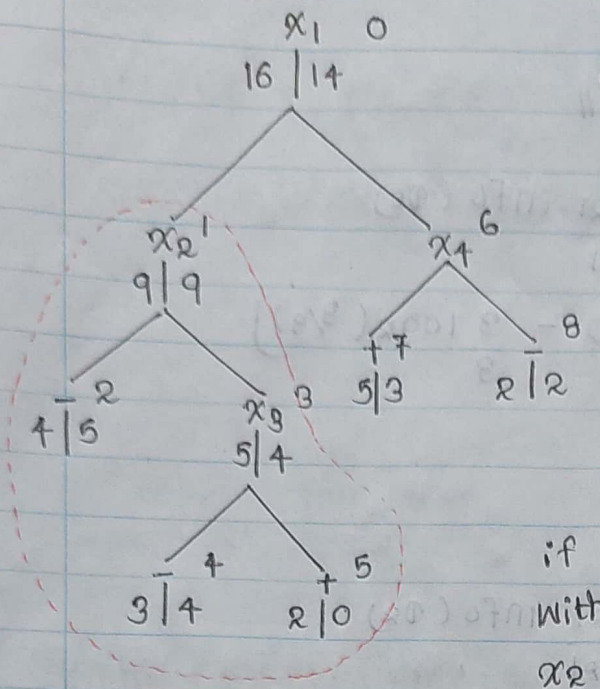


- pre-proning is the same as forward selection
- post-pruning is the same as backward elimination.

$$q = \frac{N - n_c + 0.5}{N}$$

N: total examples; n_c: majority # of classes

Atlas



Let T_1 be the sub-stree

$$q(T_1) = \frac{\sum_{\ell \in leaf(T_1)} (N_\ell - N_{c,\ell}) + 0.5}{\sum_{\ell \in leaf(T_1)} N_\ell}$$

$$q(T_1) = \frac{(9-5) + (7-4) + (2-2) + 0.5}{18}$$

$$= \frac{4 + 3 + 0.5}{18} = 0.417$$

if we pruned this, replace it with a leaf node

$x_2 \Rightarrow$ we can replace with

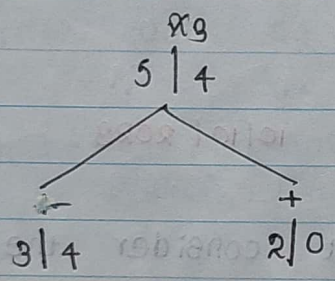
9 | 9 — or + ? # of − = 9

9 | 9 9 | 9 # of + = 9

$$q(v) = \frac{(18-9) + 0.5}{18} = 0.528$$

- $q(v) > q(T) \Rightarrow$ keep the sub-tree

Now consider



$$q(T_2) = \frac{(7-4) + (2-2) + 0.5}{9} = 0.389$$

$$q(v) = \frac{(9-5) + 0.5}{9} = 0.5$$

$q(v) > q(T_2) \Rightarrow$ Keep the sub-tree.

No: Data Mining 02 Date: / /

16/10/2023

Bayesian Classification

Bayer's Theorem

- $P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$

Hypothesis

$H : C = C_1$

- We can determine the prior probability in several ways.

Eg: If we know that there are 30% applications are safe and 70% are risky.

$P(C_i|X) = \frac{P(X|C_i)\,P(C_i)}{P(X)}$

$P(X)$ is constant for all classes. Hence, only $P(X|C_i)\,P(C_i)$

$P(X|C_i) = P(X_1, X_2, X_3, \cdots, X_n \mid C_i)$

$= P(X_1|C_i)\,P(X_2|C_i) \cdots P(X_n|C_i)$

- $P(X_k|C_i) = N(X_k - 0.5 \le X_k \le X_k + 0.5, \mu_{C_i}, \sigma_{C_i})$

Example

$X = (\text{age} = \text{youth}, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit\_rating} = \text{fair})$

Let C_1 = buys a computer C_2 = do not buy a computer

$P(C_1|X) = ?$ $P(C_2|X) = ?$

$P(C_1|X) = \frac{P(X|C_1)\,P(C_1)}{P(X)}$

$= \frac{P(X_1 = \text{youth}, X_2 = \text{medium}, X_3 = \text{yes}, X_4 = \text{fair} \mid C_1)\,P(C_1)}{P(X)}$

$P(X) = P(X_1 \cap X_2 \cap X_3 \cap X_4)$

$P(C_1) = \frac{9}{14}$ $P(C_2) = \frac{5}{14}$

$P(X_1|C_1) = \frac{2}{9}$ $P(X_2|C_1) = \frac{4}{9}$ $P(X_3|C_1) = \frac{6}{9}$ $P(X_4|C_1) = \frac{6}{9}$

$P(X|C_1)P(C_1) = \frac{2}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14} = 0.0282$

Atlas

$$P(C_2|X) = \frac{P(X|C_2)\,P(C_2)}{P(X)}$$

$$= \frac{P(X_1|C_2)\,P(X_2|C_2)\,P(X_3|C_2)\,P(X_4|C_2)\,P(C_2)}{P(X)}$$

$P(X_1|C_2) = \frac{3}{5}$ $P(X_2|C_2) = \frac{2}{5}$ $P(X_3|C_2) = \frac{1}{5}$ $P(X_4|C_2) = \frac{2}{5}$

$$P(X|C_2)\,P(C_2) = \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14} = \frac{12 \times 5}{(5)^3 \times 14} = 0.0343$$

- $P(X|C_1)\,P(C_1) < P(X|C_2)\,P(C_2)$.

Therefore, the given observation belongs to the class C_2//

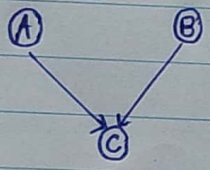
Bayesian Belief Networks (BBN)

- $P(X|Y,Z) = P(X|Z)$ means X, Y, Z are independent

if A is independent and B depends on A, then

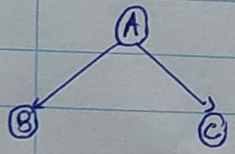
$P(X_A, X_B) = P(B|A)\,P(A)$

↑ joint probability

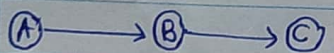


A - independent, B - independent, C - has 2 parents

$P(A,B,C) = P(C|A,B)\,P(A)\,P(B)$



$P(A,B,C) = P(B|A)\,P(C|A)\,P(A)$



$P(A,B,C) = P(C|B)\,P(B|A)\,P(A)$

The Alarm Example

B = a burglary occurs

E = an earthquakes occurs

A = alarm rigning

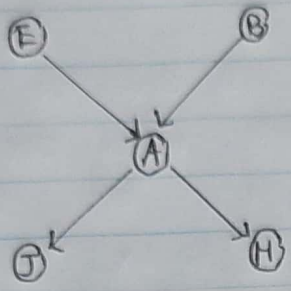
J = John calls

H = Mary calls

No: Date: / /

Find the probability of John and Mary report the alarm and neither earthquakes nor burglary occur.

A = Y, B = N, E = N, J = Y, H = Y



$P(A, B', E', J, H) = ?$

$= P(J|A) \times P(H|A) \times P(A|B', E') \times P(E') P(B')$

$= 0.9 \times 0.7 \times 0.001 \times 0.998 \times 0.999$

$= 0.00063$ //

• $P(J, A, B, E) = P(J|A) \times P(A|B, E) \times P(B) P(E)$

$P(J, A, B, E') = P(J|A) \times P(A|B', E') \times P(B) P(E')$

$P(J, A, B', E) = P(J|A) P(A|B', E) P(B') P(E)$

$P(J, A, B', E') = P(J|A) P(A|B', E') P(B') P(E')$

$P(J, A, B, E) = 0.9 \times 0.95 \times 0.001 \times 0.002 = 1.71 \times 10^{-6}$

$P(J, A, B, E') = 0.9 \times 0.94 \times 0.001 \times 0.998 = 8.448 \times 10^{-4}$

$P(J, A, B', E) = 0.9 \times 0.29 \times 0.999 \times 0.002 = 5.2148 \times 10^{-4}$

$P(J, A, B', E') = 0.9 \times 0.001 \times 0.999 \times 0.998 = 8.973 \times 10^{-4}$

$P(J, A', B, E) = P(J|A') P(A'|B, E) P(B) P(E)$

$= 0.05 \times 0.05 \times 0.001 \times 0.002 = 5 \times 10^{-9}$

$P(J, A', B, E') = P(J|A') P(A'|B, E') P(B) P(E')$

$= 0.05 \times 0.06 \times 0.001 \times 0.998 = 2.994 \times 10^{-6}$

$P(J, A', B', E) = P(J|A') P(A'|B', E) P(B') P(E)$

$= 0.05 \times 0.71 \times 0.999 \times 0.002 = 7.0929 \times 10^{-5}$

$P(J, A', B', E') = P(J|A') P(A'|B', E') P(B') P(E')$

$= 0.05 \times 0.999 \times 0.999 \times 0.998 = 0.0498$

$= 521.38763 \times 10^{-4} = 0.0521$ //