Name: Lakhan Kumawat
Roll No: 1906055
Branch: CSE-1

Course Title: Data Mining & Warehousing
Course Code: CS6403

Page No.
Date: 11 03 2022

Solution 1.0)

Supervised learning: Supervised learning is the Data mining task of inferring a function from labeled training data. The training data consists of a set of training examples. In Supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A Supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

Semi-Supervised learning: Semi-Supervised learning is of great interest in machine learning and data mining because it can use readily available unlabeled data to improve Supervised learning tasks when the labeled data are scarce or expensive. Semi-Supervised learning also shows potential as a quantitative tool to understand human category learning, where most of input is self-evidently unlabeled. The success of Semi-Supervised learning depends critically on some underlying assumptions. We emphasize the assumptions made by each model and give counter examples when appropriate to demonstrate the limitations of the different models.

Name: Lakhan Kumawat
Roll No: 1906055
Branch: CSE-1

Course Title: Data Mining & Warehousing
Course Code: CS6403

Page No.
Date: 11 | 03 | 2022

Solution 1) b)

Handling the missing values in data cleaning process:

1. Ignore the data row:

→ This is usually done when the class label is missing, or many attributes are missing from the row. However, you'll obviously get poor performance if the percentage of such rows is high.

2. Use a global constant to fill in for missing values.

→ Decide on a new global constant value, like "unknown", "N/A" or minus infinity, that will be used to fill all the missing values. This technique is used because sometimes it just doesn't make sense to try and predict the missing value.

3. Use attribute Mean:

→ Replace missing values of an attribute with the mean value for that attribute in database.

4. Use attribute mean for all samples:

→ Instead of using the mean of a certain attribute calculated by looking at all the rows in a database, we can limit the calculations to the relevant class to make the value more relevant to the row we're looking at.

Name: Lakhan Kumawat
Roll No: 1906055
Branch: CSE-1

Course Title: Data Mining & Warehousing
Course Code: CS6403

Page No.
Date 11 03 2022

Solution 1) b) Continue...

5. Use data mining algorithm to predict the most probable value

→ The value can be determined using regression, inference based tools using Bayesian formalism, decision trees, clustering algorithms.

Name: Kumawat Lavhan
Roll No: 1906055
Branch: CSE-1

Course Title: Data Mining & Warehousing
Course Code: CS6403

Page No.
Date 11 03 2022

Solution 2) a)

Data Transformation is a technique used to convert the raw data into a suitable format that efficiently eases data mining and retrieves strategic information. Data transformation includes data cleaning techniques and a data reduction technique to convert data into appropriate form

Data Transformation Techniques:

There are several data transformation techniques that can help structure and clean up the data before analysis or storage in a data warehouse.

1. Data Smoothing:

→ Data Smoothing is a process that is used to remove noise from the dataset using some algorithms. It allows for highlighting important features present in the dataset. It helps in predicting the patterns.

2. Attribute Construction:

→ In the attribute Construction method, the new attributes consult the existing attributes to construct a new data set that eases data mining.

3. Data Aggregation:

→ Data Collection or aggregation is the method of storing and presenting data in summary format.

Name: Lavkan Kumawat
Roll No: 1906055
Branch: CSE-1

Course Title: Data Mining & Warehousing
Course Code: CS6403

Page No.
Date 11 03 2022

→ The data may be obtained from multiple data sources into a data analysis description.

4. Data Normalization.

→ Normalizing the data refers to scaling the data values to a much smaller range such as [-1,1] or [0.0, 1.0]

We have three different methods for normalization.
1. Min-max normalization.
2. Z-Score normalization.
3. Decimal Scaling.

5. Data Discretization.

→ This is a process of converting continuous data into a set of data intervals. Continuous attribute values are substituted by small interval labels.

6. Data Generalization.

→ It converts low-level data attributes to high level data attributes using concept hierarchy, which is useful to get a clearer picture of data. e.g. age data can be in the form of (20, 30) in a dataset. It is transformed into a higher conceptual level into a categorical values (young, old)

Name : Lakhan Kumawat
Roll No: 1906055
Branch: CSE-1

Course Title: Data Mining & Warehousing
Course Code: CS6403

Date 11 03 2022

Solution 2) b)

Given: Group of data: 200, 300, 400, 600, 1000

min_A = minimum of all data values = 200 (these)
max_A = maximum of all data values = 1000 (these)

Given,

$new\_min_A = 0$, $new\_max_A = 1$

We have, For Min-Max Normalization

$$v' = \frac{v - min_A}{max_A - min_A} (new\_max_A - new\_min_A)$$

Putting values of parameters,

Normalized value, $v' = \frac{v - 200}{1000 - 200} \cdot (1 - 0) = \frac{v - 200}{800}$

For each value in data.

i) For 200: $v' = \frac{200 - 200}{800} = 0$

ii) For 300: $v' = \frac{300 - 200}{800} = \frac{100}{800} = 0.125$

iii) For 400: $v' = \frac{400 - 200}{800} = \frac{200}{800} = 0.25$

iv) For 600: $v' = \frac{600 - 200}{800} = \frac{400}{800} = 0.5$

v) For 1000: $v' = \frac{1000 - 200}{800} = \frac{800}{800} = 1$

Solution 2) b) Continue...

Original Data	200	300	400	600	1000
Normalized Data	0	0.125	0.25	0.5	1

Name : Lakhan Kumawat
Roll No: 1906055
Branch: CSE-1

Course title: Data Mining & Wasehousing
Course Code: CS6403

Date 11 03 2022

Solution 3) OLAP stands for On-Line Analytical Processing

OLAP implement the multidimensional analysis of bussiness information. Based on multidimensional data model, it allows user to query multidimensional data. OLAP data bases are divided into one or more cubes and these cubes are known as Hyper-cubes.

At the core of OLAP lies OLAP, which addresses the complex queries easily to give information. It is a multidimensional matrix, where information is stored in facts

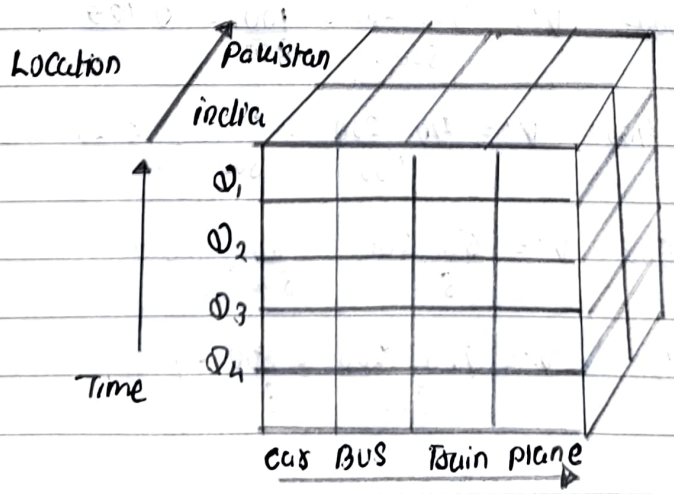
* The OLAP operations are:-

a) Roll up.

It is also just opposite of drill-down operation. It performs aggregation on the OLAP cube. It can be done by.

- Climbing up in Concept hierarchy
- Reducing the dimensions

In figure below rollup operation is performed by climbing up in the Concept hierarchy.



Name: Lakhan Kumawat
Roll No: 1906055
Branch: CSE-1

Course Title: Data Mining & Warehousing
Course Code: CS6403

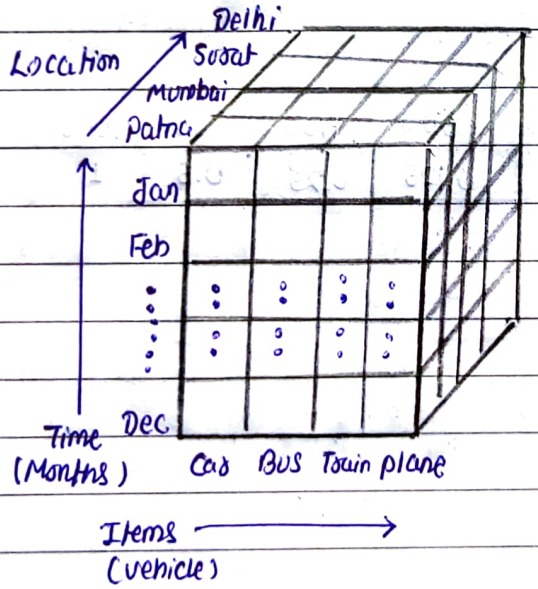
11 03 2022

b) Drill Down:-

In drill down operation, the less detailed data is converted into highly detailed data.

It can be done by:-
- Moving down in Concept hierarchy
- Adding a new dimension.

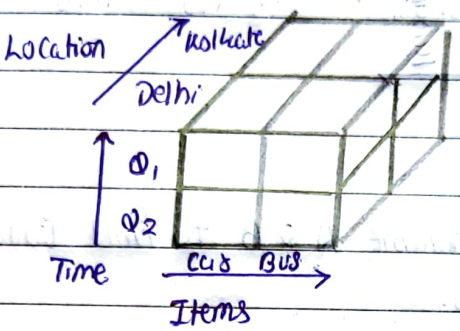
The drill down operation is performed by moving down in the concept hierarchy of Time dimension



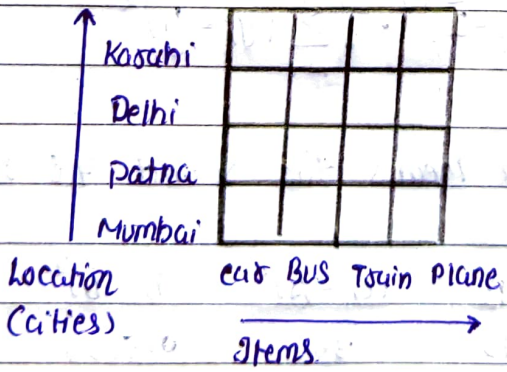
Lakhan Kumawat / 1906055
CS6403

Solutions) Continue...

C) Dice: It selects a sub-cube from the OLAP cube by selecting two or more dimensions. In the cube given in the overview section, a sub cube is selected by selecting following dimensions.



D) Slice : It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the overview section.

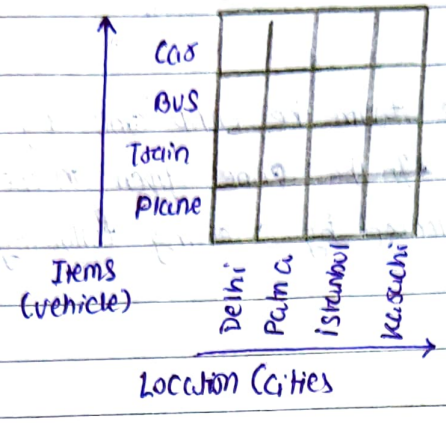


E) Pivot: It is also known as rotation operation as it rotates the current view to get a new view of representation.

In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.

Lakhan Kumawat / 1906055
CS6403.

Solution 3) Continue...



Solution 4) Given two variable A & B to find correlation between the two.

We will use pearson correlation which states that pearson correlation fact

$$\gamma_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Where $\bar{x}$ and $\bar{y}$ are means for x and y's respectively

So.

A	B	$a_i - \bar{A}$	$b_i - \bar{B}$	$(a_i - \bar{A})^2$	$(b_i - \bar{B})^2$	$(a_i - \bar{A})(b - \bar{B})$
35	0	-6	-0.75	36	0.5625	4.5
49	9	8	0.25	64	0.0625	2
25	9	-16	-3.75	256	14.0625	60
33	9	-8	0.25	64	0.0625	-2
65	11	24	2.25	576	5.0675	54
25	5	-16	-3.75	256	14.0625	60
45	11	4	2.25	16	5.06	9
51	12	10	3.25	100	10.56	32.5

Lakhan Kumawat | 1906055

CS6403.

Solution 4) Continue...

$\Sigma A = 320$

$\Sigma B = 70$

$\Sigma a_i - \bar{A} = 0$

$\Sigma b_i - \bar{B} = 0$

$\Sigma (a_i - \bar{A})^2 = 1368$

$\Sigma (b_i - \bar{B})^2 = 49.5$

$\Sigma (a_i - \bar{A})(b - \bar{B}) = 220$

$$\left[\bar{A} = \frac{328}{8} = 41 \right]$$

$$\left[\bar{B} = \frac{70}{8} = 8.75 \right]$$

Putting the values inside equation.

$$\gamma_{A+B} = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{\sqrt{\sum_{i=1}^{n} (a_i - \bar{A})^2} \cdot \sqrt{\sum_{i=1}^{n} (b_i - \bar{B})^2}}$$

$$\gamma_{A+B} = \frac{220}{\sqrt{1368 \times 49.5}}$$

$$= \frac{220}{\sqrt{260}} = \frac{220}{260}$$

Lakhan Kumawat 1906055
CS6403

Solution 4)

$$\gamma_{A,B} \approx 0.84$$

Since $\gamma_A \approx 0.84$ (positive value)

we can say that A & B are positively related.

Solution 5)

	Male	Female	total
Fiction	300	250	550
Non-Fiction	100	900	1000
	400	1150	1550

General Format: $$\chi^2 \text{ value} = \sum_{i=1}^{c} \sum_{y=1}^{x} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

where O_{ij} = observed frequency (actual count) of the joint event (A_i, B_y)

e_{ij} = the expected frequency of (A_i, B_y)

$$e_{ij} = \frac{\text{Count}(A = a_i) * (\text{Count}(B = b_j)}{N}$$

Where N → no. of data tuples

Solution 5) Continue...

Count (A = a_i) → No. of tuples value a_i for A.
Count (B = b_j) → No. of tuples value b_j for B.

$$e_{11} \text{ (male, fiction)} = \frac{\text{Count (male)} \times \text{Count (fiction)}}{N}$$

$$= \frac{400 \times 550}{1550} = 141.93$$

$$e \text{ (male, non fiction)} = \frac{\text{Count (male)} \times \text{Count (non fiction)}}{N}$$

$$= \frac{400 \times 1000}{1550} = 258.06$$

Similarly, for female

$$e \text{ (female, fiction)} = \frac{1150 \times 550}{1550}$$

$$= 408.06$$

$$e \text{ (female, nonfiction)} = \frac{1150 \times 1000}{1550}$$

$$= 741.93$$

Page No.
Date

According to x^2 formula,

$$x^2 = \frac{(300-141.93)^2}{141.93} + \frac{(100-258.06)^2}{258.06}$$

$$+ \frac{(250-408.06)^2}{408.06} + \frac{(900-741.93)^2}{741.93}$$

$$= \frac{(158.07)^2}{141.93} + \frac{(-158.06)^2}{258.06} + \frac{(-158.06)^2}{408.06} + \frac{(158.07)^2}{741.93}$$

$$= \frac{24986.12}{141.93} + \frac{24982.96}{258.06} + \frac{24982.96}{408.06} + \frac{24986.12}{741.93}$$

$$= 176.04 + 96.81 + 61.22 + 33.67$$

$$= 367.74$$

$[\, x^2 \approx 368 \,]$ Solution