APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

STUDY MATERIALS

Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

EIGHTH SEMESTER B.TECH DEGREE EXAMINATION, MAY 2019

Course Code: CS402

Course Name: DATA MINING AND WAREHOUSING

Max. Marks: 100 Duration: 3 Hours

PART A

Answer all questions, each carries 4 marks. Marks

1		How is data mining related to business intelligence?	(4)
2		Differentiate between OLTP and OLAP.	(4)
3		Why do we need data transformation? What are the different ways of data transformation?	(4)
4		An airport security screening station wants to determine if passengers are criminals or not. To do this, the faces of passengers are scanned and kept in a database. Is this a classification or prediction task? Justify	(4)
5		Where do we use Linear regression? Explain linear regression.	(4)
6		What is the significance of tree pruning in decision tree algorithms?	(4)
7		What are the two measures used for rule interestingness?	(4)
8		Given two objects represented by the tuples (22,1,42,10) and (20,0,36,8) Compute the Manhattan distance between the two objects.	(4)
9		How density based clustering varies from other methods?	(4)
10		Differentiate web content mining and web structure mining.	(4)

PART B

Answer any two full questions, each carries 9 marks.

11	a)	Explain various stages in knowledge discovery process with neat diagram	(5)
	b)	Use the two methods below to normalize the following group of data: 1000,2000,3000,5000,9000 i)min-max normalization by setting min=0 and max=1 ii) z-score normalization	(4)
12		Suppose that a data warehouse for University consists of four dimensions date, spectator, location and game and two measures count and charge, where charge is the fare that a spectator pays when watching a game on the given date. Spectator may be students , adults or seniors ,with each category having its own charge rate	

a) Draw a star scheme for the data warehouse. (6)

b) Starting with the basic cuboid [date,spectator,location,game] ,what specific OLAP operation should be performed in order to list the total charge paid by student spectators at GM_PLACE in 2010. (3)

13 Summarize the various pre-processing activities involved in data mining (9)

PART C

Answer any two full questions, each carries 9 marks.

14 Based on the following data determine the gender of a person having height 6 ft., weight 130 lbs. and foot size 8 in. (use Naive Bayes algorithm). (9)

person	height (feet)	weight (lbs)	foot size (inches)
male	6.00	180	10
male	6.00	180	10
male	5.50	170	8
male	6.00	170	10
female	5.00	130	8
female	5.50	150	6
female	5.00	130	6
female	6.00	150	8

15 (9)

The "Restaurant A" sells burger with optional flavours: Pepper, Ginger and Chilly. Every day this week you have tried a burger (A to E) and kept a record of which you liked. Using Hamming distance, show how the 3NN classifier with majority voting would classify
{pepper = false, ginger =true, chilly = true}

	Pepper	Ginger	Chilly	liked
A	true	true	true	false
B	true	false	flase	true
C	false	true	true	false
D	false	true	false	true
E	true	false	false	true

16 a) How C4.5 differs from ID3 algorithm? (3)

b) How does backpropagation algorithm works? (6)

PART D

Answer any two full questions, each carries 12 marks.

17 Consider the transaction database given below. Set minimum support count as 2 and minimum confidence threshold as 70%

Transaction ID	List of Item_Ids
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

a) Find the frequent itemset using FP Growth Algorithm. (8)

b) Generate strong association rules. (4)

18 a) Explain BIRCH Clustering Method. (8)

b) What are the advantages of BIRCH compared to other clustering method. (4)

19 a) Explain k-means partition algorithm. What is the drawback of K-means? (6)

b) Term frequency matrix given in the table shows the frequency of terms per document. Calculate the TF-IDF value for the term T4 in document 3. (6)

Document/term	T1	T2	T3	T4	T5	T6
D1	5	9	4	0	5	6
D2	0	8	5	3	10	8
D3	3	5	6	6	5	0
D4	4	6	7	8	4	4

****

1. ***How is data mining related to business intelligence?***

Ans. Data mining is an integral part of business intelligence when it comes to cleansing, standardizing, and utilizing business data. It also contributes to your ability to use that data to make accurate and dependable predictions that can allow you to operate at higher level than simply relying on the historical data, and guessing a future they need and use business intelligence and analytics to determine why it is important.

Since time human have been collecting information, and in recent decades growth of technology sector has also caused disproportionate increase in the volume of information data. Therefore it requires more sophisticated and complex data storage. Due to this boom organizations implemented big data to analyse, discover and understand the info beyond this system. Big data is the computerised processing of large amounts of information.

It is therefore necessary that the response speed of the system would be as quick as possible, in order to obtain the right information at the right time. Big data also analyse and classify them into different categories.

of smart city appears. In it, the innovative use of data helps to provide better and more inventive services to improve people's lives. And it is estimated that about 3000 million people live currently in cities.

2. ***Differentiate between OLTP and OLAP?***

BASIS	**OLTP (OS)**	**OLAP (Data Warehouse)**
Source of data	Operational data; OLTPs are the original source of the data	Consolidation data; OLAP data comes from various OLTP databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving , decision support
What the data reveals	A snapshot of on-going business processes	Multi-dimensional views of various kinds of business activities
Inserts and updates	Short and fast inserts and updates by end users	Periodic long running batch jobs refresh the data
Queries	Relatively standardized and simple queries returning relatively few records	Often complex queries involving aggregations

3. ***Why do we need data transformation? What are the different ways of data transformation?***

Ans. Data transformation is required to correct the detected discrepancies. So it uses various commercial tools such as data migration and ETL tools. Data migration tools allow simple transformations and ETL tools helps transformation using GUI.

The most common data transformation techniques include:

a. **Smoothing:** It removes noise from data.
b. **Aggregation:** It summarises data and constructs data cubes.
c. **Generalisation:**It is also known as concept hierarchy.
d. **Attribute/feature construction:** It composes new attributes from the given ones.
e. **Normalisation:** It scales the data within small, specified range. The most dominant normalisation techniques are
    i. **Min-max normalisation:**Linear transformation is applied on the data

$$v' = \frac{v - \min A}{\max A - \min A}(new\_maxA - new\_minA) + new\_minA$$

    ii. **Z-score normalisation:** Attribute A is normalised w.r.t average value and standard deviation

$$v' = \frac{v - mean\ A}{stdA}$$

    iii. **Decimal scaling normalisation:** The values of attribute A are normalise by shifting their decimal part

$$v' = v/10^j$$

where, j=smallest integer that satisfies max(|v'|)<1

4. ***An airport security screening station wants to determine if passengers are criminals or not. To do this, the faces of passengers are scanned and kept in a database. Is this a classification or prediction task? Justify.***

Ans. The task is a classification as classification predicts categorical (discrete, unordered). Classification is a data mining technique used to predict group membership for data instances.

5. ***Where do we use linear regression? Explain linear regression.***

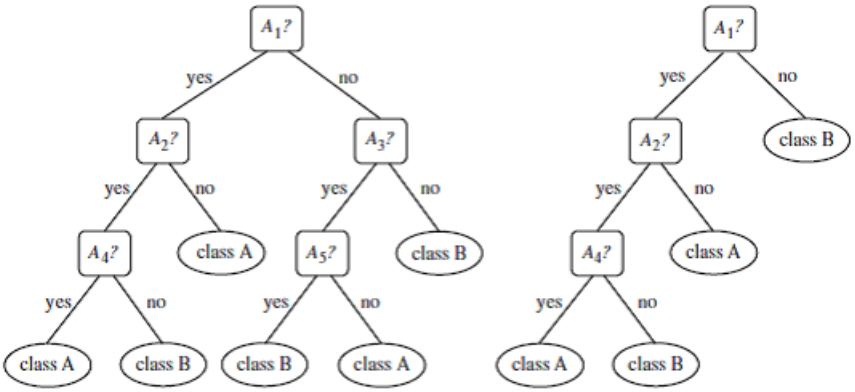
Ans. Linear regression is the most common regression model. Regression is a widely used statistical methodology for numeric prediction. And prediction is a data mining technique used to predict or appropriately guess the missing values for attributes which are either missing or corrupt.

Linear regression is the measure of the average relationship between two or more variables in terms of the original units of data. The simplest form of, linear regression uses the formula of a straight line (y=mx + b) and determines the appropriate values for m and b to predict the values of y based upon given values of x.

Linear models are approximate representations of real relations between variables, since data are rarely ideally linear. This model is not adequate for non-linear models, even though it is a widely accepted regression model.

6. ***What is the significance of tree pruning in decision tree algorithms?***

Ans. When a decision tree is built, many of the branches will reflect its anomalies in the training data due to noise or outliers. The pruning methods address this problem of over fitting the data. Such methods typically use statistical measures to remove the least reliable branches. An unpruned and pruned decision tree is shown below



Pruned trees tend to be smaller and less complex, and thus, easier to comprehend. They are usually faster and better at correctly classifying independent test data than unpruned trees.

7. ***What are the two measures used for rule interestingness?***

Ans. Two measures used for rule interestingness:

a. **Φ -coefficient:** The phi square is a measure of correlation between two categorical variables in a 2×2 table. Its value can range from 0 (no relation between factors) to 1 (perfect relation between the two factors). This measure is analogous to Pearson's product – moment correlation coefficient for continuous variables. The Φ coefficient is a statement of effect size in χ^2 that is independent of sample size.
b. **Goodman Kruskal's λ-coefficient :** Also known as the index of predictive association. It is the expected probability that in a randomly chosen case, the antecedent will lead to incorrect class label . it is based on rationale that if two variables are highly dependent on each other, than the error in predicting one of them would be small whenever the value of other variable is known. It is used to capture the amount of reduction in prediction error.

8. ***Given two objects represented by tuples (22,1,42,and 10) and (20,0,36,8). Compute the Manhattan distance between the two objects.***

Ans. d (i,k)= $\sum_{n=0}^{n}[|xi_n\text{-}xk_n|]$

d=[|22-20|+|1-0|+|42-36|+|10-8|] d=(2+1+6+2) d=11

9. ***How density based clustering varies from other clustering methods?***

Ans. This finds clusters of arbitrary shape. It grows the clusters with as many points as possible till some threshold is met. The e-neighbourhood of a point is used to find dense regions in the database. Density based algorithms typically regard clusters as dense regions of objects in the data space that are separated by regions of low density. The main ide of density- based approach is o find regions of high density and low density, with high density regions being separated from low density regions. These approaches can make it easy to discover arbitrary clusters.

A common way is to divide the high-dimensional space into density-based grid units. Units containing relatively high densities are the cluster centres and the boundaries between clusters fall in regions of low density units.

Different types of density-based methods are:

a. DBSCAN
b. OPTICS
c. DENCLUE

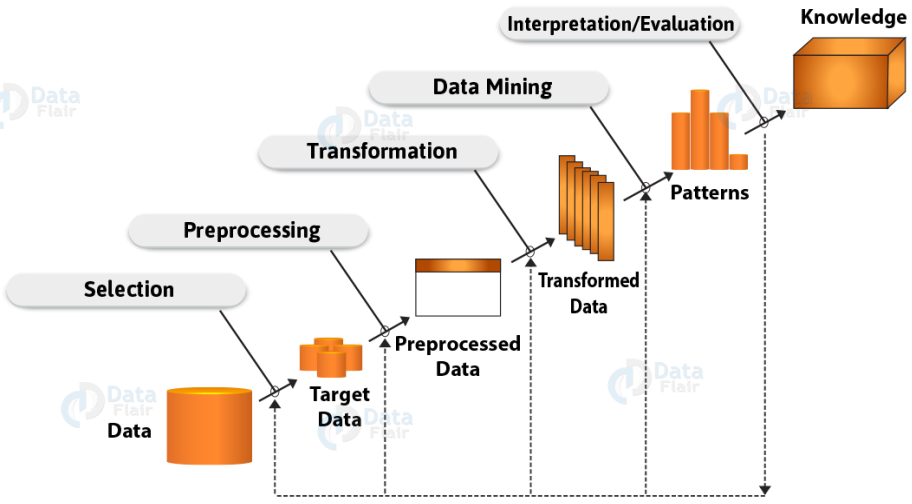
10. Differentiate between web content mining and web structure mining?

Ans.

Web content mining	**Web structure mining**
Web content mining is the task of extracting knowledge from the content of documents on world wide web	Web structure mining is the process of extracting knowledge from the link
It targets the knowledge discovery, in which the main objects are traditional collections of multimedia documents	It focuses on analysis of the link structure if the web and one of its purposes is to identify more preferable documents
Mainly focuses on structure of inner document	While in here mining tries to discover the link structure of the hyperlinks at the inner document level
Due to heterogeneity and lack of structure of web data, automated discovery of targeted or unexpected knowledge information still present many research problems	The challenge for web structure mining is to deal with the structure of the hyperlinks within the web itself link analysis is an old area of research
Agent based approach and database approach are the two approaches in web content mining	Link classification, link based cluster analysis, link types, link strength, link cardinality

11. A) Explain various stages in knowledge discovery process with neat diagram.

Ans.



a. Data Integration First of all the data is collected and integrated from all the different sources.

b. Data Selection

Generally, we may not all the data we have collected in the first step. Also, in this step, we select only those data which we think useful for data mining.

c. Data cleaning

Generally, the data we have collected is not clean and may contain errors, missing values, noisy or inconsistent data. Therefore we need to apply different techniques to get rid of such anomalies.

d. Data Transformation

Basically, the data even after cleaning is not ready for mining. Also, we need to transform them into forms appropriate for mining. Thus, the techniques used to do this are smoothing, aggregation, normalization etc.

e. Data Mining

As now in this step, we are ready to apply data mining techniques on the data. Basically, it is to discover the interesting patterns. Hence, clustering and association analysis are among the many different techniques present. Also, we used for data mining.

f. Pattern Evaluation and Knowledge Presentation

Generally, this step includes visualization, transformation, removing redundant patterns from the patterns we generated.

g. Decisions / Use of Discovered Knowledge

This step is beneficial to us. Also, it helps to use the knowledge acquired to take better decisions.

B) ***Use the two methods below to normalize the following group of data: 1000, 2000,3000,5000,9000***

i) min-max normalization by setting min=0 and max=1

ii) z-score normalization

Ans. i)

Data(v)
1000
2000
3000
5000
9000

Minimum value = 1000 and Maximum value = 9000

Newmax = 1 and Newmin = 0

$$v' = \frac{v - \min A}{\max A - \min A}(\text{new\_maxA-new\_minA}) + \text{new\_minA}$$

for data (v) 1000,

$$v' = \frac{1000-1000}{9000-1000}(1-0) + 0, \qquad v' = 0$$

for v = 2000

$$v' = \frac{2000-1000}{9000-1000}(1-0) + 0, \qquad v' = \frac{1000}{8000}\times 1, \qquad v' = 0.125$$

for v = 3000

$$v' = \frac{3000-1000}{9000-1000}(1-0) + 0, \qquad v' = \frac{2000}{8000}\times 1, \qquad v' = 0.25$$

for v = 5000

$$v' = \frac{5000-1000}{9000-1000}(1-0) + 0, \qquad v' = \frac{4000}{8000}\times 1, \qquad v' = 0.5$$

for v = 9000

$$v' = \frac{9000-1000}{9000-1000}(1-0) + 0, \qquad v' = \frac{8000}{8000}\times 1, \qquad v' = 1$$

Data(v)	Data after min-max normalisation
1000	0
2000	0.125
3000	0.25
5000	0.5
9000	1

ii) z-score normalization, $s=\sqrt{\frac{\sum(x-\overline{x})^2}{n}}$

n = 5

mean $\overline{x} = \frac{1000+2000+3000+5000+9000}{5}$, $\overline{x}$ = 4000

$$s=\sqrt{\frac{(1000-4000)^2+(2000-4000)^2+(3000-4000)^2+(5000-4000)^2+(9000-4000)^2}{5}}$$

$$s=\sqrt{\frac{9000000+4000000+1000000+1000000+16000000}{5}} \qquad s=\sqrt{62000000} \qquad s=2489.97$$

$Z_{score}=\frac{x-N}{\sigma}$ (N=mean, x=data value, σ=s)

For 1000

$$Z_{score}=\frac{1000-4000}{2489.97}, \qquad Z_{score}= -1.20$$

For 2000

$Z_{score}=\frac{2000-4000}{2489.97}$, Z_{score}= -0.80

For 3000

$Z_{score}=\frac{3000-4000}{2489.97}$, Z_{score}= -0.40

For 5000

$Z_{score}=\frac{5000-4000}{2489.97}$, Z_{score}= -0.40

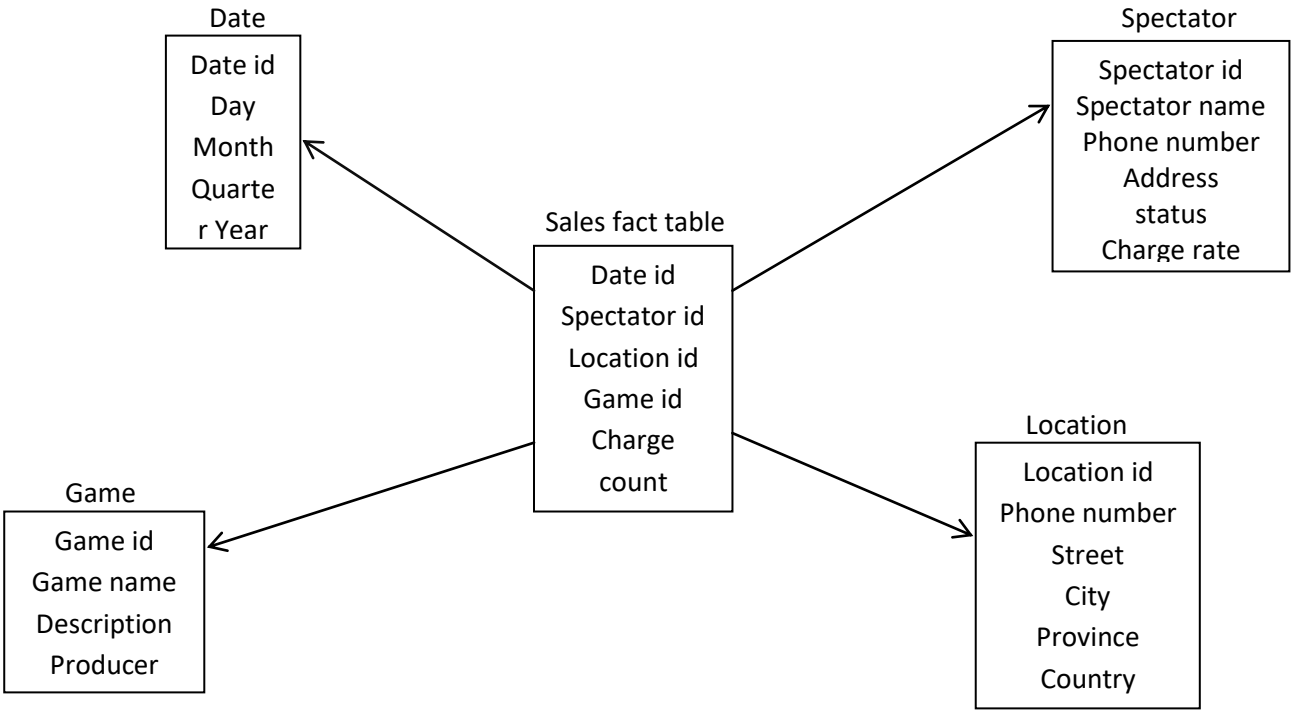
For 9000

$Z_{score}=\frac{9000-4000}{2489.97}$, Z_{score}= -2.00

Data(v)	Data after Z normalisation
1000	-1.20
2000	-0.80
3000	-0.40
5000	-0.40
9000	-2.00

12. Suppose that a data warehouse for University consists of four dimensions date, spectator, location and game and two measures count and charge, where charge is the fare that a spectator pays when watching a game on the given date. Spectator may be students , adults or seniors ,with each category having its own charge rate

a. Draw a star scheme for the data ware house.



b. Starting with the basic cuboid [date, spectator, location, game] ,what specific OLAP operation should be performed in order to list the total charge paid by student spectators at GM_PLACE in 2010.

Ans. **Specific OLAP operation to be performed area**

1) Roll up on date from date id to year.
2) Roll up on game from game id to all.
3) Roll up on location from location id to location name.
4) Roll up on spectator from spectator id to status.
5) Dice with status = "students", location name = "GM place" and year = "2010".

13. ***Summarise the various pre-processing activities involved in data mining.***

Ans. **Various pre-processing activities involved in data mining**

1) **Data cleaning:** Data can have many irrelevant and missing data. To handle this part, data cleaning is done. It involves handling of missing and noisy data.
    i. **Missing data:** This situation occurs when some data is missing in the data. It handled in various methods. Some of the ways are
        a. **Ignore the tuple:** This method is applicable when given dataset is quite large and there are multiple data is missing within a single tuple.
        b. **Filling the missing tuple:** There are various ways to do this method, you may chose in between them, and it includes manual input of missing values, by means of attribute or most probable value.
    ii. **Noisy data:** Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. it can be handled in following ways
        a. **Binning method:** This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segment is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
        b. **Regression:** Here data can be made smooth by fitting it to a regression function the regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
        c. **Clustering:** This approach groups the similar data in a cluster. The outliers may be undetected or will fall outside the clusters.
2) **Data transformation:** This step is taken in order to transform the data appropriate forms suitable for mining process. This involves following ways:
    i. **Normalization:** It is done in order to scale the data values in a specified range (0.0 to 1.0 or -1.0 to 1.0)
    ii. **Attribute selection:** In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

iii. **Discretization:** This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

iv. **Concept hierarchy generation:** Here attributes are converted from lower level to higher level in hierarchy. For example – the attribute "city" can be converted to "country".

3) **Data reduction:** Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. It aims to increase the storage efficiency and reduce data storage and analysis costs.
The various steps to data reduction are:

i. **Data cube aggregation:** Aggregation operation is applied to data for the construction of data cube.

ii. **Attribute subset selection:**The highly relevant attributes should be used; rest of all can be discarded. For performing attribute selection, one can use level of significance and p-value of attribute. The attribute having p-value greater than significance value can be discarded.

iii. **Numerosity reduction:**This enables to store the model of data instead of whole data, for example: Regression models.

iv. **Dimensionality reduction:** This reduces the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: wavelet transforms and PCA (Principal Component analysis.

14. Based on the following data determine the gender of a person having height 6 ft., weight 130 lbs. and foot size 8 in. (use Naive Bayes algorithm).

Person	Height(ft)	Weight(lbs)	Foot size(inches)
Male	6.00	180	10
Male	6.00	180	10
Male	5.50	170	8
Male	6.00	170	10
Female	5.00	130	8
Female	5.50	150	6
Female	5.00	130	6
Female	6.00	150	8

Ans.**Naïve Bayes algorithm**

The classifier created from the training set using a Gaussian distribution assumption would be:

Sex	Mean(height)	Variance(height)	Mean(weight)	Variance(weight)	Mean(foot size)	Variance (foot size)
Male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
Female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667+00

Let's say we have equiprobable classes so P (male) = P (female) = 0.5

There was no identified reason for making this assumption so it may have been a bad idea. If we determine P(c) based on frequency in the training set, we happen to get the same answer.

Below is a sample to be classified as a male or female.

Sex	Height(ft)	Weight(lbs)	Foot size(inch)
Sample	6	130	8

We wish to determine which posterior is greater, male or female.

For the classification as male, the posterior is given by:

Posterior (male) = $\frac{P(male)p(height|male)p(weight|male)p(footsize|male)}{Evidence}$

For the classification as female, the posterior is given by:

Posterior (female) = $\frac{P(female)p(height|female)p(weight|female)p(footsize|female)}{Evidence}$

The evidence may be calculated since the sum of posteriors equals one.

Evidence = $P(male)p(height|male)p(weight|male)p(footsize|male) + P(female)p(height|female)p(weight|female)p(footsize|female)$

The evidence may be ignored since it is a positive constant. We now determine sex of the sample.

P(male) = 0.5

P(height|male) = $\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789$

Where μ = 5.855 and σ = 3.50333e – 02 are the parameters of normal distribution which have been previously determined from the training set. Note that a value greater than 1 I OK here – it is probability density rather than probability, because height is a continuous variable.

P (weight | male) = 5.9881e – 06

P (foot size | male) = 1.3112e – 3

Posterior numerator (male) = their product = 6.1984e – 09

P (female) = 0.5

P (height | female) = 2.2346e – 1

P (weight | female) = 1.6789e - 2

P (foot size | female) = 2.8669e - 1

Posterior numerator (male) = their product = 5.377e – 04

Since posterior numerator is greater in female case we predict the sample is a female.

15. ***The "Restaurant A" sells burger with optional flavours: Pepper, Ginger and Chilly. Every day this week you have tried a burger (A to E) and kept a record of which you liked. Using Hamming distance, show how the 3NN classifier with majority voting would classify {pepper = false, ginger =true, chilly = true}***

	Pepper	**Ginger**	**Chilly**	**liked**
A	True	True	True	False
B	True	False	False	True
C	False	True	True	False
D	False	True	False	True
E	True	False	False	True

Ans. Assume the true = 1 and false = 0, so the above table can be written as

	Pepper	Ginger	Chilly	Liked
A	1	1	1	0
B	1	0	0	0
C	0	1	1	0
D	0	1	0	1
E	1	0	0	1

And the given majority voting, (pepper = 0, ginger = 1, chilly = 1)

Now we can calculate the hamming distance for A,B,C,D and E

	Pepper	Ginger	Chilly
A	1	1	1
X	0	1	1
	DIFF	SAME	DIFF
HD		1	FALSE

	Pepper	Ginger	Chilly
B	1	0	0
X	0	1	1
	DIFF	DIFF	DIFF
HD		3	TRUE

	Pepper	Ginger	Chilly
C	0	1	1
X	0	1	1
	SAME	SAME	SAME
HD		0	FALSE

	Pepper	Ginger	Chilly
D	0	1	0
X	0	1	1
	SAME	SAME	DIFF
HD		1	TRUE

	Pepper	Ginger	Chilly
E	1	0	0
X	0	1	1
	DIFF	DIFF	DIFF
HD		3	TRUE

3NN classifier with majority voting would classify {pepper = false, ginger = true, chilly = true} will belong to true class so this is linked on majority of days.

16. **a)** ***How C4.5 differs from ID3 algorithm?***
    b) ***How does back propagation algorithm work?***

Ans.

a. The main difference about C4.5 from ID3 is that, C4.5 is just an extension of ID3. Apart from ID3 it is a statistical classifier. It is better than ID3 algorithm because it deals with both continuous and discrete attributes and also with the missing values and pruning trees after construction. It can be easily implemented, builds models that can be easily interpreted, and can deal with noise and missing value attributes.

b. Backpropagation learns by iteratively processing a data set oftraining tuples, comparing the network's prediction for each tuple with the actual known target value.The target value may be the known class label of the training tuple (for classification problems) or acontinuous value (for numeric prediction). For each training tuple, the weights are modified so as tominimize the mean-squared error between the network's prediction and the actual target value. Thesemodifications are made in the "backwards" direction (i.e., from the output layer) through each hiddenlayer down to the first hidden layer (hence the name back propagation). Although it is not guaranteed, ingeneral the weights will eventually converge, and the learning process stops. The algorithm issummarized in figure below. The steps involved are expressed in terms of inputs, outputs, and errors, andmay seem awkward if this is your first look at neural network learning. However, once you becomefamiliar with the process, you will see that each step is inherently simple. The steps are described next.

17. ***Consider the transaction database given below. Set minimum support count as 2 and minimum confidence threshold as 70%.***

Transaction id	**List of items ids**
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2,I3

a. Find the frequent item set using FP growth Algorithm.

b. Generate strong association rules.

Ans.

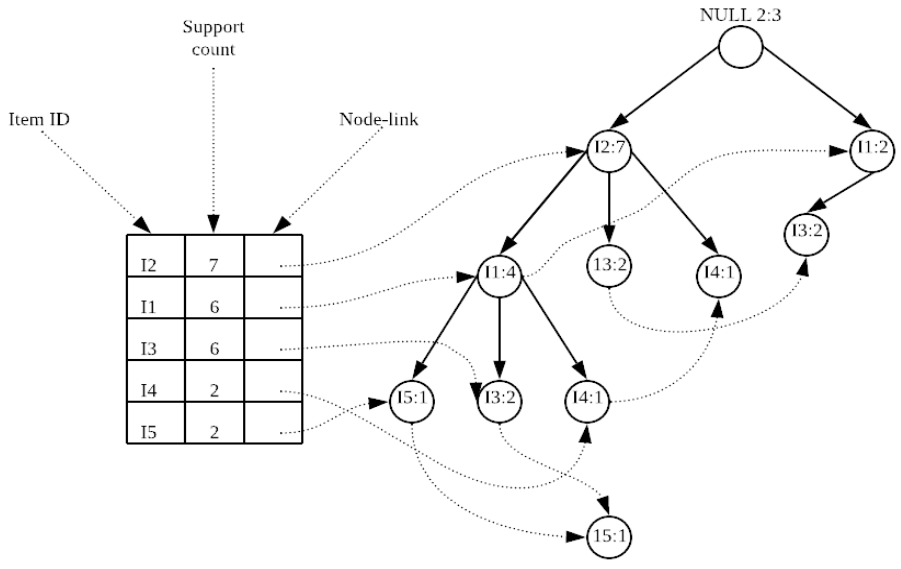
a) Support count = 2

List of Item_Ids	Support count	Frequency pattern	Items
I1	6	I2	I4
I2	7	I1	I5
I5	2	I3	I3
I4	2	I5	I1
I3	6	I4	I2

Transaction id	List of Items_Id	Ordered item set
T100	I1, I2, I5	I2, I1, I5
T200	I2, I4	I2, I4
T300	I2, I3	I2, I3
T400	I1, I2, I4	I2, I1, I4
T500	I1, I3	I1, I3
T600	I2, I3	I2, I3
T700	I1, I3	I1, I3
T800	I1, I2, I3, I5	I1, I2, I3, I5
T900	I1, I2,I3	I2, I1, I3

L1	I
ITEM SET	SUPORT COUNT
I2	7
I1	6
I3	6
I4	2
I5	2

Frequent pattern growth (fp-growth):



Step	Conditional pattern base	Conditional FP-tree	Frequent patterns generated
I5	{(I2,I1:1), (I2,I1,I3:1)}	|I2:2, I1:1|	{I2,I5:2},{I1,I5:2},{I2,I1,I5:2}
I4	{(I2,I1:1), (I2:1)}	|I2:2|	{I2,I4:2}
I3	{(I2,I1:2), (I2:2), (I1:2)}	|I2:4, I1:2| |I1:2|	{I2,I3:4}, {I1,I3:4}, {I2,I,I3:2}
I1	{(I2:4)}	|I2:4|	{I2,I1:4}

b) Generate strong association rules
The set of frequent items is sorted in the order of descending support count. This resulting set or list is denoted by L. So, we have L = {(I2:7), (I1:6), (I3:6), (I4:2), (I5:2)}.

An FP-tree is then constructed as follows:

1) First, create the root of the tree, labelled with "null". Scan database D a second time. The items in each transaction are processed in L order (i.e., sorted according to descending support count), and a branch is created for each transaction. For example, the scan of the first transaction, "T100:I1,I2,I5" which contains three items (I1,I2,I5 in L order), leads to the construction of the first branch of the tree with three nodes, (I2:1), (I1:1), and (I5:1), where I2 is linked as a child to the root, I1 is linked to I2, and I5 is linked to I1.
2) The second transaction, T200, contains the items I2 and I4 in L order, which would result in a branch where I2 is linked to the root and I4 is, linked I2, with the existing path for T100.
3) Therefore, we instead increment the count of the I2 node by 1, and create a new node, <I4:1>, which is linked as a child to <I2:2>. In general, when considering the branch to be added for a transaction, the count of each node along a common prefix is incremented by 1, and nodes for the items following the prefix are created and linked accordingly.
4) To facilitate tree traversal, an item header table is built so that each item points to its occurrences in the tree via a chain of node-links. The tree obtained after scanning all of the transactions is shown in figure with the associated node-links. In this way, the problem of mining frequent patterns in databases is transformed to that of mining the FP-tree.

The FP-tree is mined as follows:

1) Start from each length-1 pattern (as an initial suffix pattern), construct its conditional pattern base (a "sub-database" which consists of the set of prefix paths in the FP-tree co-occurring with the suffix pattern), then construct such a tree. The pattern growth is achieved by the concatenation of suffix pattern with the frequent patterns generated from a conditional FP-tree.
2) Mining of FP-tree is summarized in table and detailed as follows. We first consider I5, which is the last item in L, rather than the first. The reason for starting at the end of the list will become apparent as we explain the FP-tree mining process. I5 occur in two branches of the FP-tree of figure. (The occurrences of I5 can easily be found by following its chain node-links). The paths formed by these branches are (I2, I1, I5: 1) and (I1, I1, I3, I5: 1). Therefore, considering I5 as a suffix, its corresponding two prefix paths are (I2, I1: 1) and (I2, I1, I3: 1), which form its conditional pattern base.

Using this conditional pattern base as a transaction database, we build an I5-conditional FP-tree, which contains only a single path, (I2:2, I1:2); I3 is not included because its support count of 1 is less than the minimum support count. The single path generates all the combinations of frequent patterns: {(I2, I1:1), (I1, I5:2), (I2, I1, I5:2)}.

For I4, its two prefix paths form the conditional pattern base, {(I2, I1:1), (I2:1)}, which generates a single-node conditional FP-tree, |I2:2|, and derives one frequent pattern, {I2, I4:2}

18. a) Explain BIRCH Clustering Method.
b) What are the advantages of BIRCH compared to other clustering method?

Ans.

a. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is anintegrated hierarchical clustering method.It is based on the notation of CF (Clustering Feature) a CF Tree.CF tree is a height balanced tree that stores the clustering features for a hierarchical clustering.Cluster of data points is represented by a triple of numbers (N,LS,SS) Where

N= Number of items in the sub cluster

LS=Linear sum of the points

SS=sum of the squared of the points

It consists of two phases:
Phase 1: BIRCH scans the database to build an initial in-memory CF tree, which can be viewed as a multilevel compression of the data that tries to preserve the inherentclustering structure of the data.
Phase 2: BIRCH applies a (selected) clustering algorithm to cluster the leaf nodes of the CF-tree.
For Phase 1, the CF tree is built dynamically as objects are inserted. Thus, themethod is incremental. An object is inserted to the closest leaf entry (subcluster). Ifthe diameter of the subcluster stored in the leaf node after insertion is larger thanthe threshold value, then the leaf node and possibly other nodes are split. After theinsertion of the new object,information about it is passed towards the root of thetree. The size of the CF tree can be changed by modifying the threshold. If the size ofthe memory that is needed for storing the CF tree is larger than the size of the mainmemory, then a smaller threshold value can be specified and the CF tree is rebuilt.The rebuild process is performed by building a new tree from the leaf nodes of theold tree. Thus, the process of rebuilding the tree is done without the necessity ofrereading all of the objects or points. This is similar to the insertion and node split inthe construction of B+-trees. Therefore, for building the tree, data has to be read justonce. Some heuristics and methods have been introduced to deal with outliers and improve the quality of CF trees by additional scans of thedata.
After the CF tree is built, any clustering algorithm, such as a typical partitioningalgorithm, can be used with the CF tree in Phase 2.

b. A single scan of data set yields a basic good clustering, and one or more additionalscans can (optionally) be used to further improve the quality. The computationcomplexity of

thealgorithm is O(n), where n is the number of objects to be clustered.Experiments have shown the linear scalability of thealgorithm with respect to the number of objects, and good quality of clustering of thedata.

19. ***a) Explain k-means partition algorithm. What is the drawback of K-means?***
b) Term frequency matrix given in the table shows the frequency of terms per document. Calculate the TF-IDF value for the term T4 in document 3.

Document/term	T1	T2	T3	T4	T5	T6
D1	5	9	4	0	5	6
D2	0	8	5	3	10	8
D3	3	5	6	6	5	0
D4	4	6	7	8	4	4

Ans.

a. Input
D = {t1, t2, ... , tn} //set of elements
K //number of desired clusters

Output
K //set of clusters

Convergence criteria: A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity in elements in different clusters is achieved simultaneously.

Cluster mean of K_i = {t_{i1}, t_{i2},t_{i3},. . . . . .,t_{im}} is $m_i = \frac{1}{m}\sum_{j=1}^{m} t(i,j)$

K-means algorithm

i. assign initial values for means m1, m2,……, mk;
ii. repeat
iii. assign each item t1 to the cluster which has the closest
iv. mean; calculate new mean for each cluster;
v. until convergence criteria is met;

Drawbacks of k-means algorithm:

- **Lack of robustness.** As the sample mean and variance are very sensitive estimate against outliers. So-called breakdown point is zero, which means that one gross error may distort the estimate completely. The obvious consequent is that the k-means problem formulation is highly non-robust as well.
- **Unknown number of clusters.** Since the algorithm is a kind ”flat” or ”non-hierarchical” method , it does not provide any information about the number of clusters.
- Empty clusters. The Forgy’s batch version may lead to empty clusters on unsuccessful initialization.

- **Only spherical clusters.** K-means presumes the symmetric Gaussian shape for clusterdensityfunctions. Fromthisitfollowsthatalargeamountofcleandata is usually needed for successful clustering.
- **Handling of nominal values.** Thesamplemeanisnotdefinedfornominalvalues.
- **Sensitivity to initial configuration.** Since the basic algorithms are local search heuristics and K-means cost function is non-convex, it is very sensitive to the initial configuration and the obtained partition is often only suboptimal (not the globally best partition).

b. TF-IDF(T4 in D3) = 8*log(4/3)

Document/term	T1	T2	T3	T4	T5	T6
D1						
D2						
D3				1.74		
D4						

try it now

A KTU STUDENTS PLATFORM

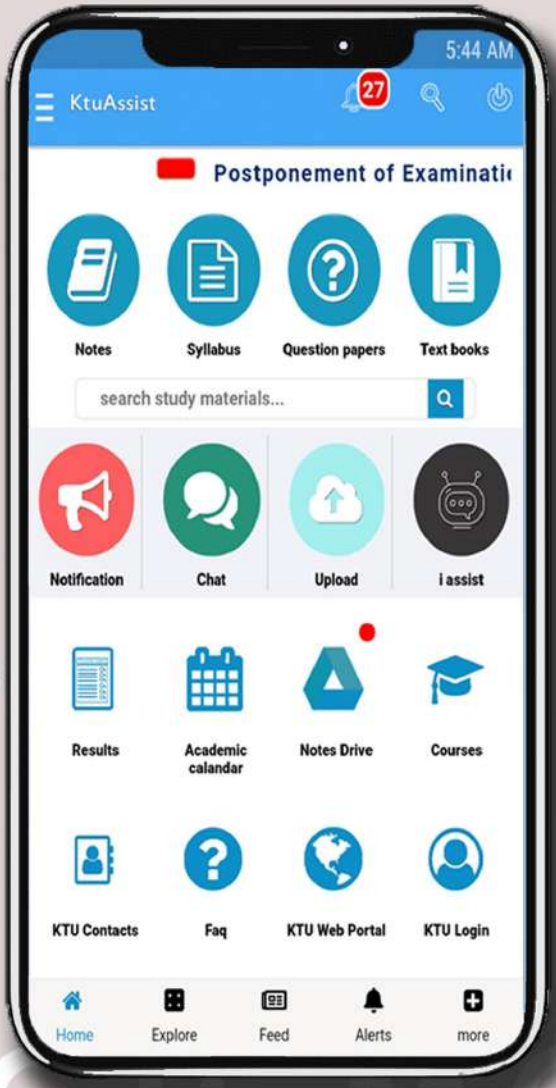
SYLLABUS

NOTES

TEXT BOOKS

QUESTION PAPERS

KTU NOTIFICATION



DOWNLOAD IT FROM GOOGLE PLAY

CHAT

FAQ

LOGIN

CALENDAR

MUCH MORE

DOWNLOAD APP

ktuassist.in

instagram.com/ktu_assist

facebook.com/ktuassist