Data Analytics (BE 2015)

Paper Solution

Class: BE Computer Academic Year: 2018-19 SEM:- I
Exam: INSEM Marks:30

Q1a) What is Big Data? Explain characteristics of Big Data [4M]

Ans:

'Big Data' is also a **data** but with a **huge size**. 'Big Data' is a term used to describe collection of data that is huge in size and yet growing exponentially with time.In short, such a data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

Characteristics of 'Big Data'

(i)Volume **–** The name 'Big Data' itself is related to a size which is enormous. Size of data plays very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with 'Big Data'.

(ii)Variety **–** The next aspect of 'Big Data' is its **variety**.

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Now days, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. is also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analysing data.

(iii)Velocity **–** The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

Q1 b) Explain different phases of Data Analytics life cycle[6M]

Ans:

Phase 1—Discovery:

In Phase 1, the team learns the business domain, includingrelevant history such as whether the organization or business unit has attemptedsimilar projects in the past from which they can learn. The team assesses theresources available to support the project in terms of people, technology, time, anddata. Important activities in this phase include framing the business problem as ananalytics challenge that can be addressed in subsequent phases and formulating initialhypotheses (IHs) to test and begin learning the data.

Phase 2—Data preparation:

Phase 2 requires the presence of an analytic sandbox,in which the team can work with data and perform analytics for the duration of theproject. The team needs to execute extract, load, and transform (ELT) or extract,transform and load (ETL) to get data into the sandbox. The ELT and ETL aresometimes abbreviated as ETLT. Data should be transformed in the ETLT process sothe team can work with it and analyze it. In this phase, the team also needs tofamiliarize itself with the data thoroughly and take steps to condition the data

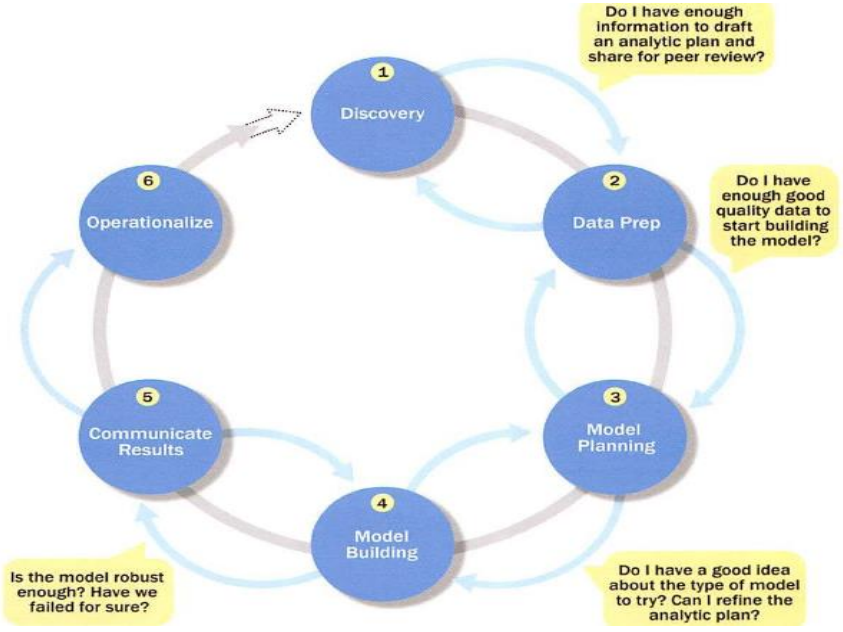


Fig: Data Analytics Life Cycle

Phase 3—Model planning:

Phase 3 is model planning, where the team determinesthe methods, techniques, and workflow it intends to follow for the subsequent modelbuilding phase. The team explores the data to learn about the relationships betweenvariables and subsequently selects key variables and the most suitable models.

Phase 4—Model building:

In Phase 4, the team develops datasets for testing,training, and production purposes. In addition, in this phase the team builds andexecutes models based on the work done in the model planning phase. The team alsoconsiders whether its existing tools will suffice for running the models, or if it willneed a more robust environment for executing models and workflows (for example,fast hardware and parallel processing, if applicable).

Phase 5—Communicate results:

In Phase 5, the team, in collaboration with majorstakeholders, determines if the results of the project are a success or a failure basedon the criteria developed in Phase 1. The team should identify key findings, quantifythe business value, and develop a narrative to summarize and convey findings tostakeholders.

Phase 6—Operationalize:

In Phase 6, the team delivers final reports, briefings,code, and technical documents. In addition, the team may run a pilot project toimplement the models in a production environment.

Q2 a) Explain current analytical structure with suitable diagram [6M]

Ans:

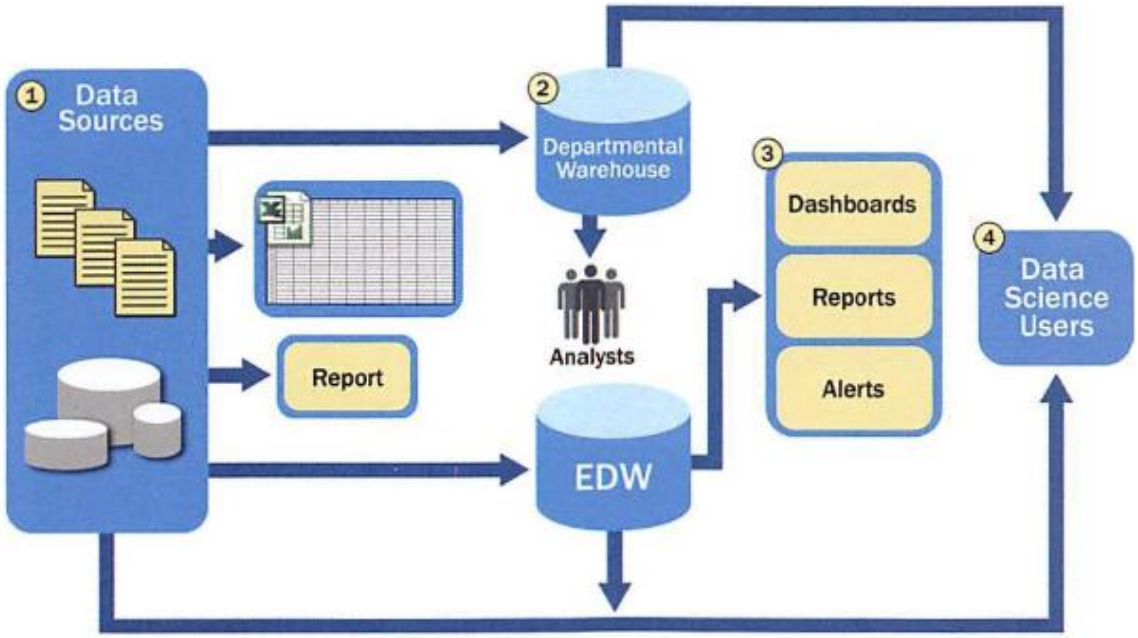


Fig: Current Analytical architecture

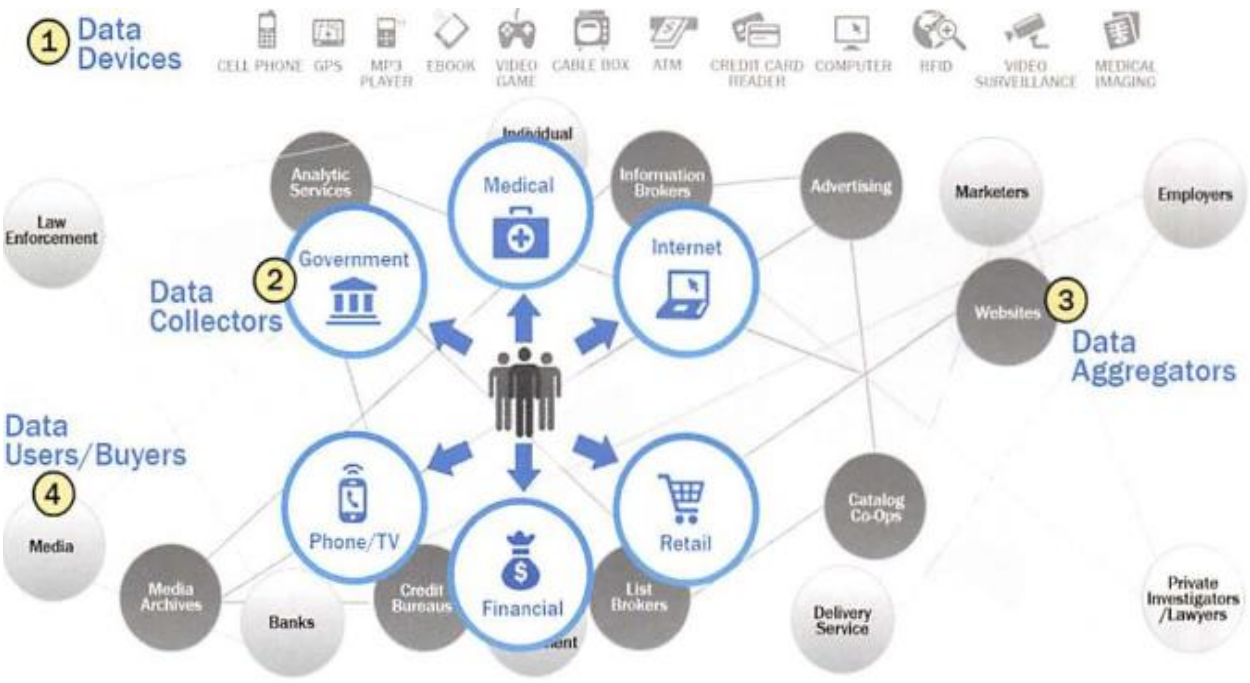
1. For data sources to be loaded into the data warehouse, data needs to be wellunderstood, structured, and normalized with the appropriate data type definitions.Although this kind of centralization enables security, backup, and failover of highlycritical data, it also means that data typically must go through significantpreprocessing and checkpoints before it can enter this sort of controlled environment,which does not lend itself to data exploration and iterative analytics.
2. As a result of this level of control on the EDW, additional local systems may emergein the form of departmental warehouses and local data marts that business userscreate to accommodate their need for flexible analysis. These local data marts maynot have the same constraints for security and structure as the main EDW and allowusers to do some level of more in-depth analysis. However, these one-off systemsreside in isolation, often are not synchronized or integrated with other data stores, andmay not be backed up.
3. Once in the data warehouse, data is read by additional applications across theenterprise for BI and reporting purposes. These are high-priority operationalprocesses getting critical data feeds from the data warehouses and repositories.
4. At the end of this workflow, analysts get data provisioned for their downstreamanalytics. Because users generally are not allowed to run custom or intensiveanalytics on production databases, analysts create data extracts from the EDW toanalyze data offline in R or other local analytical tools. Many times these tools arelimited to in-memory analytics on desktops analyzing samples of data, rather than theentire population of a dataset. Because these analyses are based on data extracts, theyreside in a separate location, and the results of the analysis—and any insights on thequality of the data or anomalies—rarely are fed back into the main data repository.

Q2 b) Explain Big Data Ecosystem.[4M]

Ans:

Organizations and data collectors are realizing that the data they can gather from individuals contains intrinsic value and, as a result, a new economy is emerging. As the new ecosystem takes shape, there are four main groups of players within this interconnected web.

1. **Data devices**
    i. Gather data from multiple locations and continuously generate new data about this data. For each gigabyte of new data created, an additional petabyte of data is created about that data.
    ii. For example, playing an online video game, Smartphones data, Retail shopping loyalty cards data



2. **Data collectors**

   i. Include sample entities that collect data from the device and users.
   ii. For example, Retail stores tracking the path a customer

3. **Data aggregators – make sense of data**

   iii. They transform and package the data as products to sell to list brokers for specific ad campaigns.

4. **Data users and buyers**

   iv. These groups directly benefit from the data collected and aggregated by others within the data value chain.
   v. For Example, People want to determine public sentiments toward a candidate by analyzing related blogs and online comments

Q3 a) What is Clustering? Explain k-means clustering algorithm with usecases [6M]
Ans:

Clustering:

In general, clustering is the use of unsupervised techniques for grouping similar objects. In machine learning, unsupervised refers to the problem of finding hidden structure within unlabeled data. Clustering techniques are unsupervised in the sense that the data scientist does not determine, in advance, the labels to apply to the clusters.

k-means Clustering:

- Given a collection of objects each with n measurable attributes and a chosen value k that is the number of clusters, the algorithm identifies the k clusters of objects based on the objects proximity to the centers of the k groups.
- The algorithm is iterative with the centers adjusted to the mean of each cluster's n-dimensional vector of attributes

k-means clustering algorithm

- Choose the value of k and the initial guesses for the centroids
- Compute the distance from each data point to each centroid, and assign each point to the closest centroid
- Compute the centroid of each newly defined cluster from step 2
- Repeat steps 2 and 3 until the algorithm converges (no changes occur)

Use cases of k-means clustering

- **Image Processing:**

Video is one example of the growing volumes of unstructured data being collected. Withineach frame of a video, k-means analysis can be used to identify objects in the video. Foreach frame, the task is to determine which pixels are most similar to each other. Theattributes of each pixel can include brightness, color, and location, the x and y coordinatesin the frame.

- **Medical**

Patient attributes such as age, height, weight, systolic and diastolic blood pressures,cholesterol level, can identify naturally occurring clusters. Theseclusters could be used to target individuals for specific preventive measures or clinicaltrial participation.

- **Customer Segmentation**

Marketing and sales groups use k-means to better identify customers who have similarbehaviors and spending patterns.

Q3 b) Explain Hypothesis testing with example. [4M]

Ans:

Data Set	1	2
Mean	49.2	37.4
Standard deviation	7.463	7.301

Hypothesis Testing:

- Hypothesis testing refers to
  1. Making an assumption, called hypothesis, about a population parameter.
  2. Collecting sample data.
  3. Calculating a sample statistic.
  4. Using the sample statistic to evaluate the hypothesis
- Basic concept is to form an assertion and test it with data
- Common assumption is that *there is no difference between samples* (default assumption)
- Statisticians refer to this as the *null hypothesis (H_0)*
- The *alternative hypothesis (H_A)* is that *there is a difference between samples*

Example

Some hairs were found on the clothing of a victim at a crime scene.The five of the hairs were measured: 46, 57, 54, 51, 38µm.A suspect is the owner of a shop with similar brown hairs. A sample of the hairs has been taken and their widths measured: 31, 35, 50, 35, 36 µm.Is it possible that the hairs found on the victim were left by the suspect‟s ? Test at the %5 level.

- Ho (Null Hypothesis is) Both the data sets are same **µ1=µ2**
- Ha Alternate hypothesis is that µ1!= µ2

Steps to solve using t-test:

1. Calculate the mean and standard deviation for the data sets
2. Calculate the magnitude of the difference between the two means**49.2 – 37.4 = 11.8**
3. Calculate the standard error in the difference**4.669**
4. Calculate the value of *T*
   T = difference between the means / standard error in the difference **=(11.8/4.664)= 2.53**
5. Calculate the degrees of freedom**= $n_1 + n_2 – 2$ (5 + 5 - 2 = 8)**
6. Find the critical T* value for the particular significance you are working to from the table **(2.03)**
7. **As T< T* (critical value) then there is no significant difference between the two sets of data ,i.e. null hypothesis is Accepted.**

Q4 a) Explain any two of the following [4M]

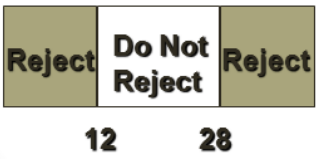
Ans:

1. Wilcoxon-ran-sum-test: If the populations cannot be assumed or transformed to follow a normal distribution, a nonparametric test can be used. The Wilcoxon rank-sum test is a nonparametric hypothesis test that checks whether two populations are identically distributed.

Example

- You're a production planner. You want to see if the operating rates for 2 factories is the same.
- For factory 1, the rates are :**71**, **82**, **77**, **92**, **88**.
- For factory 2, the rates are:**85**, **82**, **94**&**97**.
- Do the factory rates have the same **probability distributions** at the **.05** level?
- **H_0: Identical Distrib.**
- **Ha: Shifted Left or Right**
- ***n*1 = 4 *n*2 = 5**

Use table to find critical region using n1 and n2 values. So we got the Do Not reject region.



Give ranks to all rates of factories according to ascending order (smallest to largest) as shown below

Factory 1		Factory 2	
Rate	Rank	Rate	Rank
71	1	85	5
82	~~3~~ 3.5	82	~~4~~ 3.5
77	2	94	8
92	7	97	9
88	6	...	...
Rank Sum	19.5		25.5

Test Statistic: T2 = 5 + 3.5 + 8+ 9 = 25.5 (Smallest Sample). As 25.5 value comes in Do Not Reject region means our H_0 null hypothesis is accepted.

2. Type I and Type II Errors:

- Type I error refers to the situation when ***we reject the null hypothesis when it is true*** (H_0 is wrongly rejected). **Denoted by α**

- Type II error refers to the situation when ***we accept the null hypothesis when it is false.*** (H_0 is wrongly Accepted). **Denoted by β**

		Conclusion about null hypothesis from statistical test	
		Accept Null	Reject Null
Truth about null hypothesis in population	True	Correct	Type I error Observe difference when none exists
	False	Type II error Fail to observe difference when one exists	Correct

3. ANNOVA:

- A generalization of the hypothesis testing of the difference of two population means
- Good for analyzing more than two populations
- ANOVA tests if any of the population means differ from the other population means

Procedure:

- Find the mean for each of the groups.
- Find the overall mean (the mean of the groups combined).
- Find the Within Group Variation; the total deviation of each member's score from the Group Mean.
- Find the Between Group Variation: the deviation of each Group Mean from the Overall Mean.
- Find the F critical and F statistic: the ratio of Between Group Variation to Within Group Variation.
- F statistic < F critical accept Ho else reject H0 and accept Ha

Q 4b) use the data and group them using k-means clustering algorithm. Show calculations of centroid.[6M]

Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Ans:

- This is two dimensional data.

Step 1)

- We assume the value of k=2. And staring centroid values of k1=(170,70) and k2=(180,80).

Step 2)

- Now Compute the distance from each data point to each centroid, and assign each point to the closest centroid
- We use Euclidean distance measure : $\sqrt{(x1-x2)^2+(y1-y2)^2}$ to find distance between two data points.
- X is height and Y is weight.
- Calculation for data d1-
  - d(k1,d1) = $\sqrt{(170-185)^2+(70-72)^2}$ =15.3
  - d(k2,d1) = $\sqrt{(180-185)^2+(80-72)^2}$ =9.4
  - As distance of data d1 from centroid k2 is smaller than centroid k1 so d1 will be assigned to k2.
- Calculation for data d2-
  - d(k1,d2) = $\sqrt{(170-170)^2+(70-56)^2}$ =14
  - d(k2,d2) = $\sqrt{(180-170)^2+(80-56)^2}$ =17.2
  - As distance of data d2 from centroid k1 is smaller than centroid k2 so d2 will be assigned to k1.

- Calculation for data d3-
  - d(k1,d3) = $\sqrt{(170-168)^2+(70-60)^2}$ =10.19
  - d(k2,d3) = $\sqrt{(180-168)^2+(80-60)^2}$ =23.32
  - As distance of data d3 from centroid k1 is smaller than centroid k2 so d3 will be assigned to k1.
- Calculation for data d4-
  - d(k1,d4) = $\sqrt{(170-179)^2+(70-68)^2}$ =9.2
  - d(k2,d4) = $\sqrt{(180-179)^2+(80-68)^2}$ =12.04
  - As distance of data d4 from centroid k1 is smaller than centroid k2 so d4 will be assigned to k1.
- Calculation for data d5-
  - d(k1,d5) = $\sqrt{(170-182)^2+(70-72)^2}$ =12.16
  - d(k2,d5) = $\sqrt{(180-182)^2+(80-72)^2}$ =12.16
  - As distance of data d5 from centroid k1 is sameas centroid k2 so d5can be assigned to k1 or k2. We assign it k1.
- Calculation for data d6-
  - d(k1,d6) = $\sqrt{(170-188)^2+(70-77)^2}$ =9.2
  - d(k2,d6) = $\sqrt{(180-188)^2+(80-77)^2}$ =8.5
  - As distance of data d6 from centroid k2 is smaller than centroid k1 so d6 will be assigned to k2.
- Calculation for data d7-
  - d(k1,d7) = $\sqrt{(170-180)^2+(70-71)^2}$ =10.04
  - d(k2,d7) = $\sqrt{(180-180)^2+(80-71)^2}$ =9
  - As distance of data d7 from centroid k2is smaller than centroid k1 so d7 will be assigned to k2.
- Calculation for data d8-
  - d(k1,d8) = $\sqrt{(170-180)^2+(70-70)^2}$ =10
  - d(k2,d8) = $\sqrt{(180-180)^2+(80-70)^2}$ =10
  - As distance of data d8 from centroid k1 is same as centroid k2 so d8 can be assigned to k1 or k2. We assign it k1.
- Calculation for data d9-
  - d(k1,d9) = $\sqrt{(170-183)^2+(70-84)^2}$ =19.10
  - d(k2,d9) = $\sqrt{(180-183)^2+(80-84)^2}$ =5
  - As distance of data d9 from centroid k2 is smaller than centroid k1 so d9 will be assigned to k2.
- Calculation for data d10-
  - d(k1,d10) = $\sqrt{(170-180)^2+(70-88)^2}$ =20.59
  - d(k2,d10) = $\sqrt{(180-180)^2+(80-88)^2}$ =8
  - As distance of data d10 from centroid k2 is smaller than centroid k1 so d10 will be assigned to k2.
- Calculation for data d11-
  - d(k1,d11) = $\sqrt{(170-180)^2+(70-67)^2}$ =10.44

- d(k2,d11) = $\sqrt{(180-180)^2+(80-67)^2}$ =13
- As distance of data d11 from centroid k1 is smaller than centroid k2 so d11 will be assigned to k1.

- Calculation for data d12-
  - d(k1,d12) = $\sqrt{(170-177)^2+(70-76)^2}$ =9.2
  - d(k2,d12) = $\sqrt{(180-177)^2+(80-76)^2}$ =5
  - As distance of data d12 from centroid k2 is smaller than centroid k2 so d12 will be assigned to k2.

K1={(170,56),(168,60),(179,68),(182,72),(180,70),(180,67)}
K2={(185,72),(188,72),(180,71),(183,84),(180,88),(177,76)}

Step 3)

- Now we will calculate new centroid using following formula.
- (Xc,Yc)= $\frac{\sum_{i=1}^{m} Xi}{m}, \frac{\sum_{i=1}^{m} Yi}{m}$
- For **k1 (Xc,Yc)=**$\frac{170+168+179+182+180+180}{6}, \frac{56+60+68+72+70+67}{6}$ **= (176, 65)**
- For **k2 (Xc,Yc)=**$\frac{185+188+180+183+180+177}{6}, \frac{72+72+71+84+88+76}{6}$ **= (182, 77)**
- So, New centroids are k1=(176,65) k2=(182,77)

Step 4) Again repeat the step 2 and step 3 until convergence.

Q5 a) What is market basket analysis? Explain Apriori algorithm with Example.[6M]

Ans:

Market Basket Analysis is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy.

Association Rules are widely used to analyze retail basket or transaction data, and are intended to identify strong rules discovered in transaction data using measures of interestingness, based on the concept of strong rules. **Apriori algorithm** is used to find the Association rules from transaction.

Apriori algorithm Example

Assume that a large supermarket tracks sales data by stock-keeping unit (SKU) for each item: each item, such as "butter" or "bread", is identified by a numerical SKU. The supermarket has a database of transactions where each transaction is a set of SKUs that were bought together.

Let the database of transactions consist of following itemsets:

Itemsets
{1,2,3,4}
{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}
{2,4}

We will use Apriori to determine the frequent item sets of this database. To do this, we will say that an item set is frequent if it appears in at least 3 transactions of the database: the value **3 is the *support threshold*.**

The first step of Apriori is to count up the number of occurrences, called the support, of each member item separately. By scanning the database for the first time, we obtain the following result

Item	Support
{1}	3
{2}	6
{3}	4
{4}	5

All the itemsets of size 1 have a support of at least 3, so they are all frequent.

The next step is to generate a list of all pairs of the frequent items.

For example, regarding the pair {1,2}: the first table of Example 2 shows items 1 and 2 appearing together in three of the itemsets; therefore, we say item {1,2} has support of three.

Item	Support
{1,2}	3
{1,3}	1
{1,4}	2
{2,3}	3
{2,4}	4
{3,4}	3

The pairs {1,2}, {2,3}, {2,4}, and {3,4} all meet or exceed the minimum support of 3, so they are frequent. The pairs {1,3} and {1,4} are not. Now, because {1,3} and {1,4} are not frequent, any larger set which contains {1,3} or {1,4} cannot be frequent. In this way, we can *prune* sets: we will now look for frequent triples in the database, but we can already exclude all the triples that contain one of these two pairs:

Item	Support
{2,3,4}	2

in the example, there are no frequent triplets. {2,3,4} is below the minimal threshold, and the other triplets were excluded because they were super sets of pairs that were already below the threshold.

We have thus determined the frequent sets of items in the database, and illustrated how some items were not counted because one of their subsets was already known to be below the threshold.

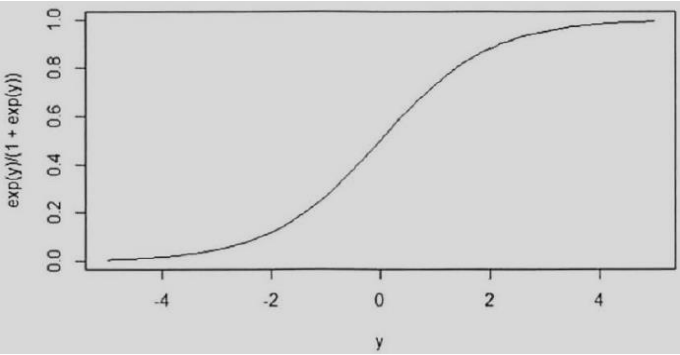
Q5 b)Explain logistic regression. Explain use cases of logistic regression.[4M]

Ans:

In logistic regression, the outcome variable is categorical, example two-valued outcomes like True/false, pass/fail, yes/no.Logical regression is based on the logistic function

$$f(y) = \frac{e^y}{1+e^y} \quad \text{for} -\infty < y < \infty$$

As y -> infinity, f(y)->1; and as y->-infinity, f(y)->0



Usecases of Logistic Regression:

- Medical
  - Probability of a patient's successful response to a specific medical treatment – input could include age, weight, etc.
- Finance
  - Probability an applicant defaults on a loan
- Marketing
  - Probability a wireless customer switches carriers (churns)
- Engineering
  - Probability a mechanical part malfunctions or fails

Q6a) Transactional data for an all electronics branch is as follows, find the frequent itemset and generate association rules with confidence values. [6M]

Ans:

Tid	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

Wewillconsiderminimumsupport level as 0.5. ie. 4. and minimum confidence 60%

The first step of Apriori is to count up the number of occurrences, called the support, of each member item separately. By scanning the database for the first time, we obtain the following result

Item	Support
I1	6
I2	7
I3	6
I4	2
I5	2

All the itemsets of size 1 have a support of at least 4, so i4 and i5 will not come in next step.

The next step is to generate a list of all pairs of the frequent items. For every pair we will calculate support and confidence of rule

Item	Rule	Support (A B)	Support A	Confidence= Support (A B)/ Support A
{I1,I2}	I1->I2	4	6	66%
{I1,I3}	I1->I3	4	6	66%
{I2,I3}	I2->I3	4	7	57%

Now for all three rules the support is same but the confidence if different. As confidence of {I2,I3} is less that minimum confidence we will remove the pair in next iteration step.

After pruning we have:

Item
{I1,I2}
{I1,I3}

Again we will make a pair

Item
{I1,I2,I3}

Q6b) What is regression? Explain any one type of regression in detail. [4M]

Ans:

Regression analysis attempts to explain the influence that input (independent) variables have on the outcome (dependent) variable

- Questions regression might answer
  - What is a person's expected income?
  - What is probability an applicant will default on a loan?
- Regression can find the input variables having the greatest statistical influence on the outcome
  - E.g. – if 10-year-old reading level predicts students' later success, then try to improve early age reading levels

There are two types of regression

1. Linear

2. Logistic

Linear Regression:

- Models the relationship between several input variables and a continuous outcome variable
  - Assumption is that the relationship is linear
  - Various transformations can be used to achieve a linear relationship