Chapter 1

½ Markers

Q.1 What is data science?

- Collection of techniques used to extract value from data. Essential tool for any organization that collects, stores and process data as part of its operations.
- Relies on finding useful patterns, connections and relationships in data.

Q.2 Define the term Data Analytics?

- It is science of extracting meaningful, valuable information from raw data.

Q.3 What is the purpose of Diagnostic analysis?

- Analytics which examine data to answer the question, why did it happen?
- It focuses on the process and causes, key factors and unseen patterns.

Q.4 Enlist the types of Data Analytics.

1) Description Analytics
2) Diagnostic Analytics
3) Predictive Analytics
4) Prescriptive Analytics

Q.5 Define exploratory analysis

- Refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations.

Q.6 Define linear model.

- Linear models are governed by equations that weigh each feature variable by a coefficient reflecting its importance and sum up these values to produce a score.

¾ Markers

Q.1 With the help of diagram describe life cycle of Data Analytics.

a. Data Discovery: 1st Phase
   - The data science team learn and investigate the problem.
   - Develop context and understanding.
   - Come to know about data sources needed and available for the project.
   - The team formulates initial hypothesis that can be later tested with data.

b. Data Preparation: 2nd Phase
   - Steps to explore, preprocess, and condition data prior to modeling and analysis.
   - It requires the presence of an analytic sandbox, the team execute, load, and transform, to get data into the sandbox.

- Data preparation tasks are likely to be performed multiple times and not in predefined order.

c. Model Planning: 3rd Phase
   - Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
   - In this phase, data science team develop data sets for training, testing, and production purposes.
   - Team builds and executes models based on the work done in the model planning phase.

d. Model Building: 4th Phase
   - Team develops datasets for testing, training, and production purposes.
   - Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.

e. Communicate Results: 5th Phase
   - After executing model team need to compare outcomes of modeling to criteria established for success and failure.
   - Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
   - Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

f. Operationalize: 6th Phase
   - The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
   - This approach enables team to learn about performance and related constraints of the model in production environment on small scale, and make adjustments before full deployment.
   - The team delivers final reports, briefings, codes.

Q.2 Differentiate between data analysis and Data Analytics.

S.No.	Data Analytics	Data Analysis
1.	It is described as a traditional form or generic form of analytics.	It is described as a particularized form of analytics.
2.	It includes several stages like the collection of data and then the inspection of business data is done.	To process data, firstly raw data is defined in a meaningful manner, then data cleaning and conversion are done to get meaningful information from raw data.
3.	It supports decision making by analyzing enterprise data.	It analyzes the data by focusing on insights into business data.
4.	It uses various tools to process data such as Tableau, Python, Excel, etc.	It uses different tools to analyze data such as Rapid Miner, Open Refine, Node XL, KNIME, etc.
5.	Descriptive analysis cannot be performed on this.	A Descriptive analysis can be performed on this.
6.	One can find anonymous relations with the help of this.	One cannot find anonymous relations with the help of this.
7.	It does not deal with inferential analysis.	It supports inferential analysis.

Q.3 What are the types of Data Analytics? Describe two of them in detail.

- Descriptive Analysis: What happened?
- Diagnostic Analysis: Why did it happened?
- Predictive Analysis: What will happen?
- Prescriptive Analysis: How can we make it happen?

Descriptive Analysis:

- Examines the raw data or content to answer the question What happened.
- Descriptive analytics looks at data and analyze past event for insight as to how to approach future events.
- It looks at the past performance and understands the performance by mining historical data to understand the cause of success or failure in the past.
- Almost all management reporting such as sales, marketing, operations, and finance uses this type of analysis.

Predictive Analysis:

- Prescriptive Analytics automatically synthesize big data, mathematical science, business rule, and machine learning to make a prediction and then suggests a decision option to take advantage of the prediction.
- Prescriptive analytics goes beyond predicting future outcomes by also suggesting action benefit from the predictions and showing the decision maker the implication of each decision option. Prescriptive Analytics not only anticipates what will happen.

Q.4 What is linear and nonlinear model?

- These are governed by equations that weigh each feature variable by a coefficient reflecting its importance and sum up these values to produce a score.
- Powerful machine learning techniques can be used to identify the best possible coefficients to fit training data.
- The world is not linear. Richer mathematical descriptions include higher order exponential, logarithms and polynomials.
- These permit mathematical models that fit training data much more tightly than linear function can.
- It is much harder to find the best possible coefficients to fit a non-linear model.

Q.5 What is confusion matrix? How to use it in Data Analytics? Explain diagrammatically.

- A confusion matrix contains information about actual and predicted classifications done by a classifier. The performance of such systems is commonly evaluated using the data in the matrix.
- A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.
- A confusion matrix is also known as an error matrix.
- A confusion matrix is a technique for summarizing the performance of a classification algorithm.
- A confusion matrix is nothing but a table with two dimensions viz. Actual, Predicted and furthermore, both the dimensions have ‘True Positives’, ‘True Negative’, ‘False Positive’, ‘False

Negative'.

		Predicted Class: Positive	Predicted Class: Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
Actual Class	Negative	False Positive (FP) Type I Error	True Negative (TN)

- The rows in a confusion matrix represent the actual class while the columns represent the predicted class.

Q.6 Define accuracy, precision, recall and f- score.

- Accuracy: Accuracy is how close or far off a given set of measurements are to their true value.
- Precision is how close or dispersed the measurements are to each other.
- Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.
- F-Measure provides a single score that balances both the concerns of precision and recall in one number.

Q.7 What is ROC curve? How to implement it? Explain with example.

- Roc curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
- It is created by plotting TPR against FPR at various threshold settings.
- Each point on ROC curve represents a particular classifier threshold, defined by its false positive and false negative rates.

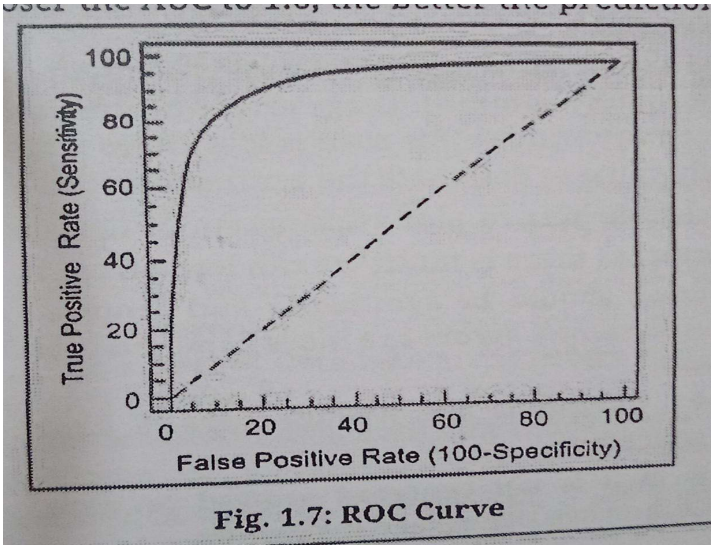


Fig. 1.7: ROC Curve

Chapter 2

½ Markers

Q.1 Define machine learning.

- Machine learning is a field of computer science that studies algorithms and techniques for automatic solutions to Complex problems.
- The basic idea of machine learning is to allow machines to independently learn from past data and solve problems.
- Machine learning is a subset of AI.
- Machine learning is the practice of getting machines to make decisions without being programmed.
- Machine learning is categorised into
    a. Supervised
    b. Unsupervised
    c. Reinforcement
- Application: image recognition and speech recognition.

Q.2 Define deep learning.

- It is a subset of machine learning. based on and neural network which mimics the working of the human brain.
- The objective is to build a neural network that automatically Discovers patterns for predictions.
- Deep learning algorithm depends on high-performance computing national power.
- It is used to solve high Complex problems.
- Application: self-driving Cars.

Q.3 List types of machine learning.

- Types of machine learning
    a. Supervised learning
    b. Unsupervised learning
    c. Semi-supervised learning
    d. Reinforcement learning

Q.4 Define classification and regression.

- Go to question 3 of 3 /4 marks

Q.5 State any two uses of machine learning.

1. Speech Recognition

2. Image Recognition
3. Automatic Language Translation
4. Product Recommendation

Q.6 Define neural networks.

- Neural Networks are such types of networks where each layer can perform complex operations such as representation and abstraction that make sense of images, sound and text.

Q.7 List any two applications of AI.

1. Robotics
2. Finance
3. Natural Language Processing
4. Speech Recognition

Q.8 Define supervised machine learning.

- Supervised learning is a learning technique that can only be applied to labelled data. In supervised learning, the computer is provided with example inputs that are labelled with their desired output.

Q.9 What is clustering?

- **Clustering** is basically a collection of objects on the basis of similarity and dissimilarity between data points. Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar.

Q.10 What is the purpose of the Apriori algorithm?

- The Apriori algorithm is used for mining frequent itemsets and devising association rules from a transactional database. The parameters “support” and “confidence” are used. Support refers to items' frequency of occurrence; confidence is a conditional probability.

Q.11 Differentiate between supervised and unsupervised machine learning.

Supervised Learning	Unsupervised learning
Supervised learning model predicts the output.	Unsupervised learning models find the hidden patterns in data.
It can be categorized in Classification and Regression problems	It can be classified in Clustering and Associations problems.

It is under the supervision	It is not under the supervision
It produces an accurate result.	It may give less accurate results as compared to supervised learning.

Q.12 Define regression analysis.

- Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables.

Q.13 What is logistic regression?

- Logistic regression means binary logistic regression having binary target variables; It is a supervised learning classification algorithm used to predict the probability of a target variable.

Q.14 Define polynomial regression.

- Polynomial regression is a regression algorithm that models the relationship between a dependently and independent variable as nth polynomial.

Q.15 List ensemble techniques.

1. Bagging ensemble technique
2. Boosting ensemble technique
3. Random forests ensemble technique

Q.16 Enlist types of clustering.

1. Overlapping Clusters
2. Exclusive or strict partitioning clusters
3. Fuzzy probabilistic clusters
4. Hierarchical clusters

¾ Markers

Q.1 Write a short note on learning models for algorithms.

- A Learning algorithm is a set of instructions used in machine learning.
- The math and logic that supports the learning algorithms can update itself.
- Learning supervised and unsupervised machine learning.
- Where supervised learning algorithms required training data to be labelled and, unsupervised learning algorithms look for patterns.

Q.2 What are the types of machine learning? Compare them.

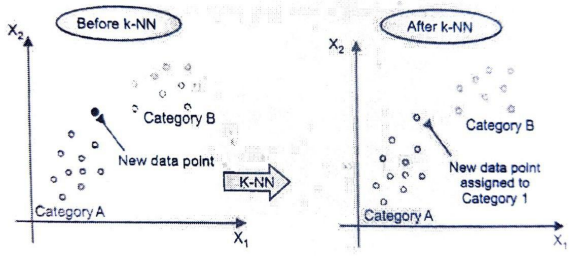
Supervised	unsupervised	semi-supervised
Trained using labelled data.	Trained using unlabeled data.	Combines a small amount of labelled and a large amount of unlabeled data.
Both input and output variables are provided to predict the output.	Only input variables are provided.	Input variables are provided whereas a few outputs are provided.
Highly accurate and trustworthy.	Comparatively less accurate.	unsup< semi<super
Net supervision to train the model.	No supervision is needed.	Requires minimal supervision.
Consumes a lot of time.	Very sensitive.	Very stable algorithm.

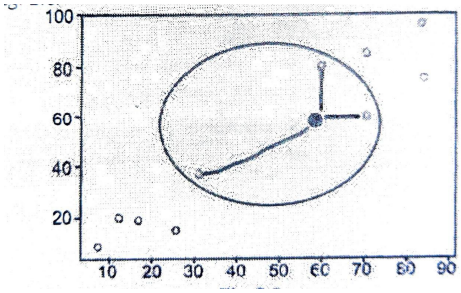
Q.3 What is supervised learning? How it works? State its advantages and disadvantages.

- It is basically: learn from the past and predict the future.
- It can be only e applied to labelled data.
- The machines are trained with labelled data under supervision.
- Working:
    a. Labelled input data is provided to the Machines
    b. Output variables of fast iterations are provided.
    c. Under supervision machine learning the algorithm and predicts the output.
- Divided into two classes:
    a. Classification: To predict output labels based on what model has learnt in training.
    b. Regression: credits output labels which are continuing numeric values and the output depends upon what model has been learnt in the training.
- Advantages:
    a. More accurate.
    b. Solves real-world problems.
- Disadvantages:
    a. Consumes a lot of time
    b. Not for Complex tasks.

Q.4 What is k-nn? How it works? Explain diagrammatically and state its advantages and disadvantages.

- Supervised ML algorithm used for classification and regression predictive problems.
- But mostly used for classification problems.
- This algorithm stores all the available data and classifies new data points based on their similarity.
- Here the function is only approximated locally and all computation is deferred until function evaluation.
- Uses ‘feature similarity’ to predict the values of new data points.





- Advantages:
    1. Simple and easy to implement.
    2. Versatile algorithm.
    3. Useful for non-linear data.
- Disadvantages:
    1. Becomes slower as the number of independent variables increases.
    2. Expensive algorithm cause it stores all training data.
    3. Requires high memory usage.

Q.5 What is a decision tree? How it works? List advantages and disadvantages.

- A decision tree is a graphical representation for getting all the possible solutions to a problem based on given conditions.
- Used for classification and regression.
- Internal nodes represent the features of a dataset, branches represent the decision rule and leaves represent the outcome.
- It is called a decision tree because it starts with the root node which expands on further branches and constructs a tree-like structure.
- Decision tree working:
    - For predicting the class of the given dataset, the algorithm starts from the root node of the tree.
    - It compares the values of the root attribute with the record attribute and based on the comparison it follows the branch and jumps to the next node.

    - For the next node, it does the same.
- Advantages:
    1. Simple to understand.
    2. Handles both numerical and categorical data.
    3. Works well with large data.
    4. Fast and accurate.
- Disadvantages:
    1. A small change in the training data can result in a larger change in the tree and consequently in the final predictions.
    2. Performance is not good if there are lots of uncorrelated variables in the data set.
    3. Making huge trees with a huge number of branches is complex.

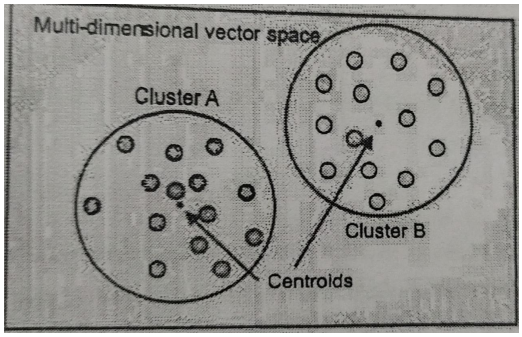
Q. 6 Explain the support vector machine with the help of a diagram.

- Support Vector Machines(SVM) are powerful yet flexible supervised machine learning algorithms.
- Used for classification and regression.
- The goal is to create the best line or decision boundary that can segregate n-dimensional space into classes.
- The objective is to find a hyperplane in an n-dimensional space that distinctly classifies the data points.
- SVM working:
    - It is a representation of different classes in a hyperplane in a multidimensional space.
    - Hyperplane will be generated iteratively by SVM.
- Advantages:
    1. Great accuracy.
    2. Works well with high dimensional space.
    3. Effective in cases where the dimension is greater than the number of samples.
- Disadvantages:
    1. High training time hence in practice not suitable for large datasets.
    2. Does not perform well when the data set has more noise.

Q.7 Explain the k-means clustering algorithm with the help of a diagram.

- The k-means clustering is a simple and popular clustering algorithm that originated in signal processing.
- The goal of the k-means algorithm is to partition examples from a data set into k clusters.

- Each example is a numerical vector that allows the distance between vectors to be
  calculated as a Euclidean distance.
- The simple example below visualizes the partitioning of data into k = 2 clusters,
- where, the Euclidean distance between examples is smallest to the centre of the cluster, which indicates its membership.



.

- We begin by randomly assigning each example from the data set into a cluster, calculate the centroid of the clusters as the mean of all member examples
- Then iterate the data set to determine whether an example is closer to the member cluster or the alternate cluster (given that k = 2).
- If the member is closer to the alternate cluster, the example is moved to the new cluster and its centroid recalculated. This process continues until no example moves to the alternate cluster.
- As illustrated, k-means partitions the example data set into k clusters without any understanding of the features within the example vectors (that is, without supervision).

Q.8 What is association rule mining? Describe with the example.

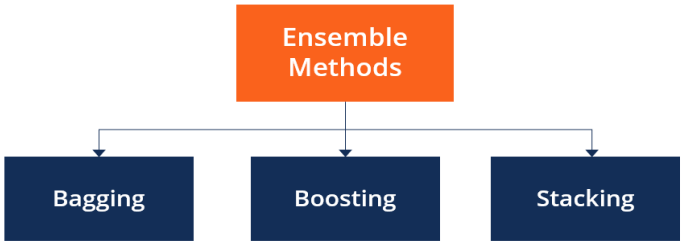
- Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.
- Another useful unsupervised ML method is Association which is used to analyze large datasets.
- **The main application of association rule mining:**
  a. **Basket Data Analysis** is to analyze the association of purchased items in a single basket or single purchase, for example, peanut butter and jelly are often bought together because a lot of people like to make PBandJ sandwiches.
  b. **Cross Marketing** is to work with other businesses that complement your own, not competitors. For example, vehicle dealerships and

manufacturers have cross-marketing campaigns with oil and gas companies for obvious reasons.

c. **Catalog Design** the selection of items in a business catalogue are often designed to complement each other so that buying one item will lead to buying another. So these items often complement or are very related.

Q.9 Write a short note on ensemble techniques.

- Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. This has boosted the popularity of ensemble methods in machine learning

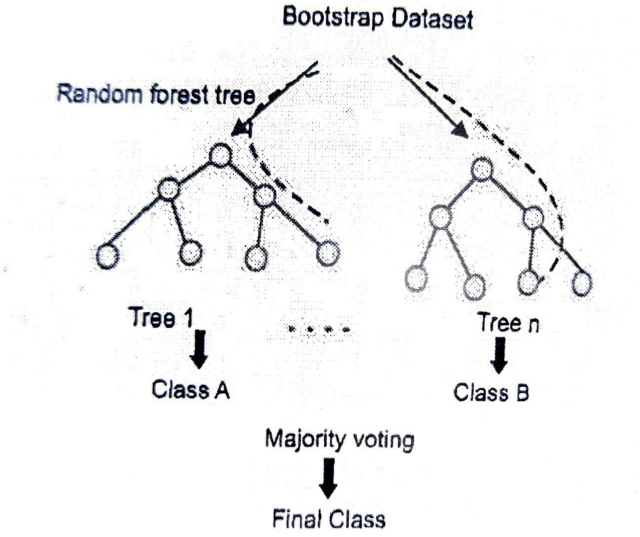


1. **Bagging**, the short form for bootstrap aggregating, is mainly applied in classification and regression. It increases the accuracy of models through decision trees, which reduces variance to a large extent. The reduction of variance increases accuracy, eliminating overfitting, which is a challenge to many predictive models. Bagging is classified into two types, i.e., bootstrapping and aggregation.
    a. **Bootstrapping** is a sampling technique where samples are derived from the whole population (set) using the replacement procedure. The sampling with replacement method helps make the selection procedure randomized.
    b. **Aggregation** in bagging is done to incorporate all possible outcomes of the prediction and randomize the outcome. Without aggregation, predictions will not be accurate because all outcomes are not put into consideration.
2. **Boosting** is an ensemble technique that learns from previous predictor mistakes to make better predictions in the future. The technique combines several weak base learners to form one strong learner, thus significantly improving the predictability of models. Boosting works by arranging weak learners in a sequence.

Q.10 What is a random forest? Describe diagrammatically.

- One of the most commonly used classifications in supervised machine learning is Random Forest Classification.

- It can be used both for classification and regression problems. As the name indicates it is a forest of decision trees.
- Decision trees are generated with randomly drawn instances from the training set. However, the final class of classification of the input instances is based on majority voting.



Q.11 Differentiate between supervised, unsupervised, semi-supervised and reinforcement machine learning.

Supervised	Unsupervised	Semi-supervised	Reinforcement
Both input and output data variables are provided on the basis of which the output could be predicted and the probability of its correctness is higher.	Only input variables are provided and no output variables are available due to which the outcome or resultant learning is dependent on one intellectual observation.	It is neither fully supervised nor fully unsupervised. It basically falls between the two i.e. supervised and unsupervised learning methods.	It is a feedback-based ML technique in which an agent learns to behave in an environment by performing actions and seeing the result of actions.
These are trained using labelled data.	These are trained using unlabelled data.	It combines a small amount of labelled data with a large amount of unlabelled data during training.	It is an agent-based goal-seeking technique whire an AI agent tries to determine the

			best action to take in a given environment-depending on a reward.
It predicts the output.	Finds hidden patterns in data.	Improves predictive power of models.	Predicts the output of categorical dependent variables.
Produces accurate results.	The less accurate result compared to supervised.		

Chapter 3

½ Markers

Q.1 Define data mining.

- Extracting for mining knowledge from the massive amounts of data sets.
- Advantages:
    a. A quick process that helps new users
    b. Help in decision making
    c. Helps in predicting future
- Disadvantages:
    a. Violates the privacy of the user.
    b. High implementation cost

Q.2 Define frequent patterns.

- Frequent patterns:
    a. Patterns that occur frequently in data.
    b. Users can apply a threshold to get a specifically occurred pattern.
    c. A pattern can be an item set or subsequences.

Q.3 Define support and confidence.

- Support: The support of rule x->y is defined as the proportion of transactions in the data set which contain the item set x as well as y. So,
  Support (x->y) = number of transactions contains x & y / Total number of transactions
  Whereas,
  Support (x) = number of transactions contains the item set x / Total number of transactions
- Confidence: The confidence of rule x->y is defined as
  Confidence (x->y) = Support (x->y) / Support (x)

Q.4 What is the purpose of FP growth algorithm?

- FP-growth algorithm is used to mine the complete set of frequent itemsets. It creates a Frequent Patter(FP) tree to complete a large dataset.
- In FP tree nodes, frequent items are arranged in such a manner that more frequently occurring nodes have better chances of sharing nodes than the less frequently occurring ones.

Q.5 Define outlier analysis.

- Outlier analysis is the process of identifying outliers, or abnormal observations, in a dataset. Also known as outlier detection.

Q.6 List applications of outlier analysis.

- Outlier analysis has numerous applications in a wide variety of domains, such as the financial industry, quality control, fault diagnosis, intrusion detection, Web analytics, and medical diagnosis.

Q.7 List frequent itemset mining methods.

1. Apriori Algorithm
2. FP Growth

Q.8 What is the purpose of the Apriori algorithm?

- It is a widely used algorithm for generating frequent itemsets and for learning association rules.

¾ Markers

Q.1 Explain the usage of market basket analysis for example.

- Data mining techniques are used by retailers to increase sales by better understanding customers purchasing patterns.
- Analyzes large data sets such as purchasing history.
  Example:
  buys(x:"computer") -> buys(x:"antivirus")
  [support= 1% , confidence = 50%]

  support is the percentage of a transaction by all transactions. i.e 1% of all transaction analysis shows a computer and antivirus were bought together. Confidence is if a customer buys a computer there is 50% chance of him buying an antivirus.

  $$\text{Support } (x \to y) = \frac{\text{number of transactions contains x \& y}}{\text{Total number of transaction}}$$

  $$\text{Support}(x) = \frac{\text{Number of transactions containing x}}{\text{Total no. of transactions}}$$

  $$\text{Confidence}(x \to y) = \frac{\text{support}(x \to y)}{\text{support}(x)}$$

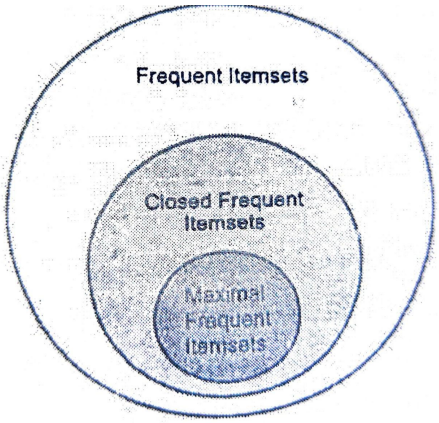
  Where x and y are events/items.

Q.2 Explain the Apriori algorithm in detail.

- Widely used an algorithm to generate frequent itemset.
- Uses prior knowledge of frequent itemset properties.
- Classic algorithm for learning Association rules.
- It is easy to use and execute and very simple to mine frequent itemsets.
- Based on antimonotonicity property.
- If an item is infrequent then all its supersets must also be infrequent.
- If {a} is infrequent then all its supersets {a,b}, {a,b,c} will be infrequent. Page 3.20 example crow.

Q.3 What are frequent itemsets, closed itemsets and association rules? Describe in detail.

- A frequent itemset typically refers to a set of items that frequently appear together in a transactional data set.
- Maximal frequent itemset is a frequent item set for which none of its immediate supersets is frequent.
- In Closed Frequent Itemset none of its immediate supersets has the same support as that of the itemset.
- As mentioned earlier closed and maximal frequent itemsets are subsets of a frequent itemset.
- The analytical process that finds frequent itemsets and associations from data sets is called frequent pattern mining or association rule mining.



Q.4 How to mind the following: a. Frequent patterns b. Associations c. Correlations

- Frequent pattern mining searches for recurring relationships in a given data set. It is a data mining technique with the objective of extracting frequent itemsets from a data set.
- Association analysis is one of the functions of data mining which discovers the association relationships among huge amounts of data.
- Association rule mining finds interesting associations and/or correlation relationships among large sets of such data items.

Q.5 What are structured and unstructured data? Distinguish between them.

- Structured data is highly specific and is stored in a predefined format, whereas unstructured data is a conglomeration of many varied types of data that are stored in their native formats.

Structured	Unstructured
Self-service access	Requires data science expertise
Only select data types	Many varied types conglomerated
Commonly stored in data warehouses	Commonly stored in data lakes
Predefined format	Native format
It is based on a relational database.	It is based on character and binary data.
It is easy to search	Searching for unstructured data is more difficult.

Q.6 With the help of an example describe the FP growth algorithm.

- FP growth algorithm is used to mine the complete set of a frequent itemset. It creates an FP tree to compress a large dataset.
- In FP tree nodes, frequent items are arranged in such a manner that more frequently occurring nodes have better chances of sharing nodes than the less frequently occurring ones.
- FP-growth algorithm preserves whole information for frequent pattern mining. FP-growth algorithm is as follows:
    a. Construct Condition Pattern Base for Each Node in the FP-Tree.
    b. Construct a Conditional FP-Tree from each Conditional Pattern-Base.
    c. Recursively Mine Conditional FP-trees and Grow Frequency Patterns obtained so Far.

Example 1: Consider following table,

Database D

Transaction ID	Item
T_1	I_1, I_2, I_3, I_5
T_2	I_2, I_3, I_4, I_5
T_3	I_1, I_4, I_5
T_4	I_1, I_2, I_3, I_4, I_5

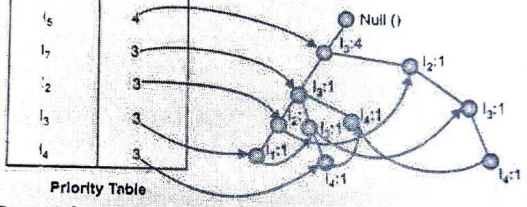
Scan Database →

Item	Count
I_1	3
I_2	3
I_3	3
I_4	3
I_5	5

After that, all the itemset arrange in a sequential order or give the priority and construct FP-tree. And new consider the items as a suffix for the database D and arrange the items or Transaction according to priority.

Item	Count
I_5	4
I_1	3
I_2	3
I_3	3
I_4	3

Priority Table



After that FP-tree of priority table, insert this priority table except item I_5 because it contains higher priority and mining the above FP-tree as summarized in Table 3.7.

Table 3.7: FP-Tree of Priority Table

Item	Conditional Pattern Base FP-Tree	Conditional Generated	Frequent Pattern
I_4	$(I_1, I_2, I_3 : 1), (I_2, I_3 : 1), (I_1 : 1)$	$(I_2, I_3 : 2), (I_1 : 2)$	$(I_2, I_3, I_4 : 1)$ $(I_1, I_4 : 2)$
I_3	$(I_1, I_2 : 2)$ $(I_2 : 1)$	$(I_1, I_2 : 2), (I_2 : 3)$	$(I_1, I_2, I_3 : 3)$ $(I_2, I_3) : 3$
I_2	$(I_1 : 2)$	$(I_1 : 2)$	$(I_1, I_2 : 2)$
I_1	$(I_5 : 3)$	$(I_5 : 3)$	$(I_5, I_1 : 3)$

Chapter 4

½ Markers

Q.1 Define text Analytics.

- Text analysis consist of extracting, analysing and interpreting the textual elements over social media, like comments.
- Textual messages include textual posts, tweets, comments, status updates, blog posts, etc.
- Data collected from textual analysis is used in business analysis for knowing opinion or sentiment regarding a particular product, topic or individual.

Q.2 Define social network.

- Social Network is a type of complex network that can be described as a social structure made of social actors or users, their inter-connections and interactions.
- These networks are useful for studying relationships between individuals, groups, social units and societies.

Q.3 What is tokenization?

- First part of any NLP process is breaking a piece of text into consistent parts (or words) this process is called tokenisation.
- This extracts unwanted elements from a text document, converts all letters to lowercase and removes the punctuation marks.

Q.4 What is the purpose of n-grams?

- An n-gram means a sequence of n words. An n-gram is a piece of text containing M words that can be broken into a collection of M - n + 1 n-grams.
- n-grams are used for a variety of things. Some examples include auto-completion of sentences, auto spell check and to a certain extent, we can check for grammar in a given sentence.

Q.5 List challenges for social media.

- Unstructured Data
- High Volume and Velocity of Data
- Diversity of Data
- Organizational Level Issues

Q.6 Define NLP.

- Natural Language Processing (NLP) is concerned with the development of computational models of aspects of human language processing.
- NLP is the convergence between linguistics, computer science and AI.

Q.7 List example of stop words.

- “the”, “a”, “an”, “so”, “what” are a few stop words.

Q.8 Define the term stemming and lemmatization.

- Stemming and lemmatization are methods used by search engines and chatbots to analyse the meaning behind a word. Stemming uses the stem of the word, while lemmatization uses the context in which the word is being used.

Q.9 Define community detection.

- Community detection is a social media network which can be used in machine learning to detect groups with similar properties and extract groups for various reasons.

Q.10 What is the purpose of Social Network Analysis?

- It mainly involves studying the relationship between media and users, organizations, user communities, users from a particular demographic group and so on. It is the process of investigating social structure through the use of social media networks.

Q.11 List applications of NLP.

- Automatic text summarization is a technique which creates a short, accurate summary of longer text documents
- Machine translation is basically a process of translating one source language or text into another language

Q.12 Define link prediction.

- Link prediction is the problem of predicting the existence of a link between two entities in a social network.

Q.13 What is trend Analytics?

- Idea of this analysis is that ML which observing data of a given time period and this data can be used to predict the trend of future.

¾ Markers

Q.1 What is social media Analytics? What is its purpose? List its benefits.

- Social Media Analytics is the process of collecting, tracking and analysing data from social networks.
- Social media analytics serves in business analysis. it also helps in knowing sentiments and opinions about certain products, etc.
- Benefits of social media analytics include:
  a. Monitoring, capturing and analysing of social media data can provide valuable information for decision making.
  b. Governments around the word have started to realise the potential of analysis in making timely and effective decisions.
  c. It gives the ability to track and analyse the growth of the community on the social media platform and the behaviour of the people using it.

Q.2 Describe layers of social media analytics with the help of diagram.

a. LAYER ONE: TEXT
   Social media text analytics deals with the extraction and analysis of business insights from textual elements of social media content, such as comments, tweets, blog posts, and Facebook status updates.

b. LAYER TWO: NETWORKS
   Social media network analytics extract, analyse, and interpret personal and professional social networks, for example, Facebook, Friendship Network, and Twitter.

c. LAYER THREE: ACTIONS
   Social media actions analytics deals with extracting, analysing, and interpreting the actions performed by social media users, including likes, dislikes, shares, mentions, and endorsement.

d. LAYER FOUR: MOBILE
   Mobile analytics is the next frontier in the social business landscape. Mobile analytics deals with measuring and optimizing user engagement with mobile applications (or apps for short).

e. LAYER FIVE: HYPERLINKS
   Hyperlink analytics is about extracting, analysing, and interpreting social media hyperlinks (e.g., in-links and out-links).

f. LAYER SIX: LOCATION
   Location analytics, also known as spatial analysis or geospatial analytics, is concerned with mining and mapping the locations of social media users, contents, and data.

g. LAYER SEVEN: SEARCH ENGINES
   Search engines analytics focuses on analysing historical search data for gaining a valuable insight into a range of areas, including trends analysis, keyword monitoring, search result and advertisement history, and advertisement spending statistics.

Q.3 What is social network? List any four examples of it. Explain two of them in short.

a. A social network can be defined as a dedicated website or other application which enables users to communicate with each other by posting information, comments, messages, images, etc.
b. ex- Facebook, Instagram, Twitter, Snapchat, Tumblr, etc.
c. Facebook: Facebook provides a platform, where people come to socialise, talk and share their views with each other. Facebook social networking sites are commonly used by large number of people to interact with their families and friends and also making business appearances or meeting online with other users.
d. Instagram: Instagram is a photo and video sharing platform owned by Facebook. The platform has provision over the photos and videos shared, it can either be viewed publicly or only among the followers. The concept of hashtags became popular over Instagram which can be used to group certain posts and comments under one hashtag.

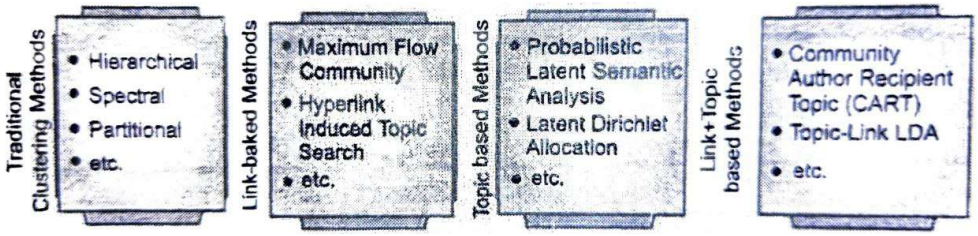
Q.4 What is link prediction? Explain with example.

- Link prediction is the problem of predicting the existence of a link between two entities in a social network.
- Prediction problem issue in social network analysis and mining.
- The link prediction problem is a common feature found on many social networking sites for possible friend suggestions as found on Facebook or Twitter.
- Allows a user to increase the personal or professional friends circle to broaden the social links and connections.
- This will increase the social networking activities as each user will be then connected to more users on the social network.

- The objective of link prediction is to identify pairs of nodes that will wither form a link or not in the future.
- Similarity-based approach and learning-based approach are the two types of approaches in link prediction.

Q.5 What is community detection? What are its different methods?

- Community detection techniques are useful for social media algorithms to discover users with common interests and keep them tightly connected.
- This technique can be used to discover manipulative groups inside a social network or a stock market.

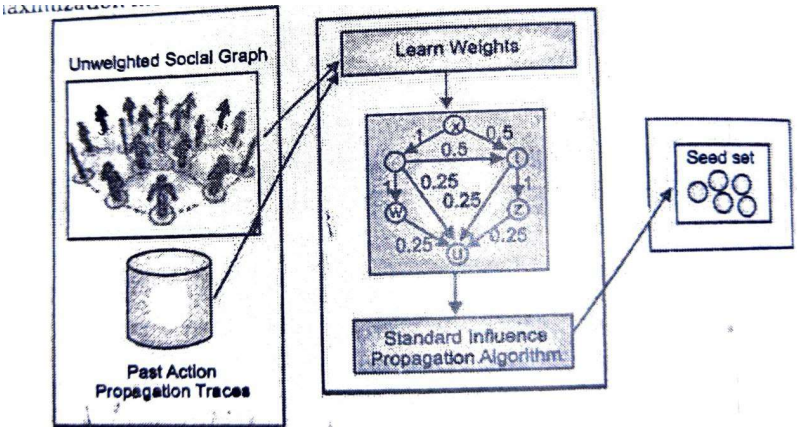


- A user belonging to the same community is expected to share similar tastes, likes and dislikes which helps in the prediction of what products a user is likely to buy.
- Categories of community detection:
  1. Traditional Clustering Methods:
     - Hierarchical Clustering: This method either gradually merges or splits the groups to create nested clusters.
     - Spectral Clustering: This method creates groups of communities by using the spectral properties of the similarity matrix.
     - Partitioning Clustering: This method divides all nodes into n-clusters, where the values of n are provided as a parameter well in advance.
  2. Link-based Clustering Methods:
     - Mainly explored is the strength of connections between nodes and not the basic semantics such as the common topic of interest of likings among nodes.
     - The two-standard link-based community detection methods are Hyperlink Induced Topic Search (HITS) and Maximum Flow Community (MFC).
  3. Topic-based Methods:
     - The topic-based community detection methods emphasise the generation of communities based on the common topic of interest.
     - Two standard topic-based community detection methods are probabilistic latent semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA).
  4. Topic-link Based Methods:
     - These community detection methods are hybrid as these approaches consider both the strength of connections between nodes and well-finding communities that are topically similar.
     - Two standard topic-link-based community detection methods are community-author-recipient-topic (CART) and topic-link LDA.

Q.6 Explain influence maximization diagrammatically.

- Influence maximization is the problem of finding a small subset of nodes in a social network that could maximize the spread of influence.
- The seed set of users can help other users to decide in choosing as to which movie to watch, which political party to follow, which product to buy, which community to join and so on.

- An effective tool being adapted by all companies and organisations.
- Viral marketing is a strategy that uses existing social networks to promote a product mainly on various social media platforms.
- The user's friends will again, in turn, recommend or publicize the same product to their friends and this helps in easy and simple product promotion.
- The main challenge in influence maximization is to generate the best influential users in the social media network.



Q.7 What is expert finding? How to find an expert? Explain with example.

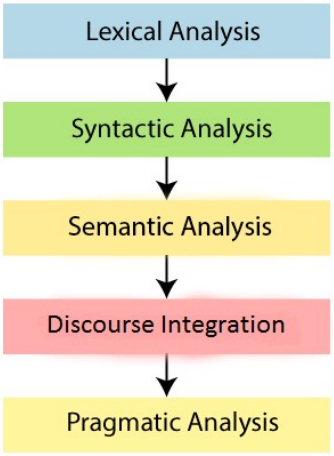
- The task of expert finding, as one of the most important research issues in social networks. The expert finding is aimed at identifying persons with relevant expertise or experience in given topic.
- Some social media networking sites provide a pool of experts for certain topics and discussions.
- The expert finding is the process of generating and grouping experts of a social network based on his/her expertise on certain topics.
- For example, refer top figure on page no. 4.23.

Q.8 Write a short note on prediction of trust and distrust.

- In some online social network services, users are allowed to label users with similar interests as "trust" to get the information they want and use "distrust" to label users with opposite interests to avoid browsing content they do not want to see.
- This remarkable growth in the number of users for social media network usage has indirectly raised a question of trust and distrust among the connected users.
- Trusted users in social media networks spread the right information and positive effects on a social network.
- However, the distrusted users in social media networks pose a threat to the social network as there is a likelihood that such users may cause a disturbance or threat in the near future.
- As every social media networking site wants to build a reputed network that can be fully trusted by users, it has become essential to trade such individual users that have a likelihood of conducting harmful or mischievous online activities in the near future.
- For maintaining the reputation of a social media network, it is important to partition users into two groups – trusted users and distrusted users of a social network.
- This will help in preventing any king of malicious activity occurring via the social media network and in turn, will bring an increase in the number of users joining the particular networking site based on the level of trust.

Q.9 What is NLP? Describe its phases with the help of diagram.

- NLP stands for Natural Language Processing, which is a part of Computer Science, Human language, and Artificial Intelligence. It is the technology that is used by machines to understand, analyse, manipulate, and interpret human's languages.



1. Lexical Analysis and Morphological: The first phase of NLP is the Lexical Analysis. This phase scans the source code as a stream of characters and converts it into meaningful lexemes. It divides the whole text into paragraphs, sentences, and words.
2. Syntactic Analysis (Parsing): Syntactic Analysis is used to check grammar, word arrangements, and shows the relationship among the words.
3. Semantic Analysis: Semantic analysis is concerned with the meaning representation. It mainly focuses on the literal meaning of words, phrases, and sentences.
4. Discourse Integration: Discourse Integration depends upon the sentences that precede it and also invokes the meaning of the sentences that follow it.
5. Pragmatic Analysis: Pragmatic is the fifth and last phase of NLP. It helps you to discover the intended effect by applying a set of rules that characterize cooperative dialogues.

Q.10 Explain n-gram with example?

- An n-gram means a sequence of n words. An n-gram is a piece of text containing M words can be broken into a collection of M-n+1 n-grams.

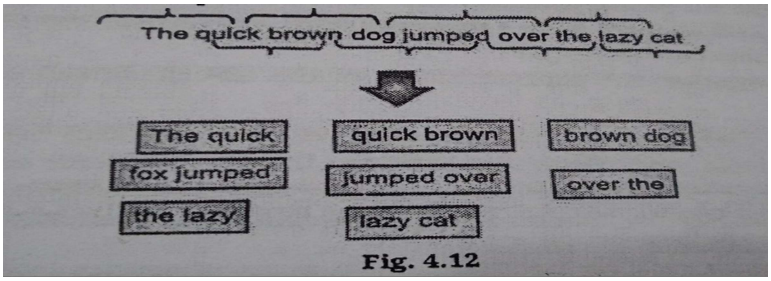


Fig. 4.12

Q.11 What is stemming and lemmatization? How they differ from each other?

- Stemming: It is the process of reduces the terms to their linguistic root to obtain index terms. Words found in text document can have different forms like organize, organization and organizing
- Lemmatization: It is just like stemming. But after lemmatization we will be getting a valid word which have same meaning. Lemmatization switches any kind of word to its base mode.

Stemming	Lemmatization
1.It is fast	Slower as compared to Stemming
2.Rule base approach	Dictionary ased approach
3.Accuracy is less	Accuracy is more
4.Example: “Studies” : “studi”	Example: “Studies” : “study”

Q.12 What is sentiment analysis? Explain with its classification?

- It is used to identify emotions conveyed by unstructured text. It basically means to find the emotion behind a piece of text or any mode of communication

  1. Document-level: Document is analysed as one unit and documents are classified on basis of positive or negative opinions
  2. Sentence-level: Document is partitioned in sentence and then classification is done on basis of positive or negative opinions.
  3. Aspect-level: Document is classified on basis of specific aspects of entities. Entities are identified and aspects of those entities are classified. Both positive and negative aspects are described

Q.13 Write a short note on challenges to social media Analytics.

1. Unstructured Data: It refers to information that either does not have a predefined data model. Unstructured data is not organized, it can have various forms i.e., textual, graphical, etc.
2. High volume and velocity of data: It refers to speed at which data is getting accumulated. Social media generates data every second and analysing a data that is high in volume and velocity is a real challenge.
3. Diversity in data: Users in social media belongs to various cultures, regions, and countries and the data generated are of various types and multilingual. Finding and capturing only important content from such diverse data is a challenging and time-consuming task.
4. Organizing level issues: Many organizing spends money to develop resource to collect, manage and analyse data but they do not understand how to ethically use social media analytics. They lack ethical data control practices.