19CST203 DATA ANALYTICS

UNIT I- INTRODUCTION

1. What is Big Data?

Big data is a field that treats ways to analyse, systematically extract information from, or otherwise, deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

2. List out the best practices of Big Data Analytics.

1. Start at the End
2. Build an Analytical Culture.
3. Re-Engineer Data Systems for Analytics
4. Focus on Useful Data Islands.
5. Iterate often.

3. Write down the characteristics of Big Data Applications.

a) Data Throttling

b) Computation- restricted throttling

c) Large Data Volumes

d) Significant Data Variety

e) Benefits from Data parallelization

4. Write down the four computing resources of Big Data Storage.

a) Processing Capability

b) Memory

c) Storage

d) Network

5. What is HDFS?

Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

6. What is MapReduce?

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

7. What is YARN?

YARN is an Apache Hadoop technology and stands for Yet Another Resource Negotiator. YARN is a large-scale, distributed operating system for big data applications. YARN is a software rewrite that is capable of decoupling MapReduce's resource management and scheduling capabilities from the data processing component.

8. What is Map Reduce Programming Model?

MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. The model is a specialization of the split-apply-combine strategy for data analysis.

9. What are the characteristics of big data?

Big data can be described by the following characteristics
Volume - The quantity of data generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.
Variety -The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.
Velocity -In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.
Variability- Inconsistency of the data set can hamper processes to handle and manage it.
Veracity-The data quality of captured data can vary greatly, affecting the accurate analysis

10. What is Big Data Platform?

- Big Data Platform is integrated IT solution for Big Data management which combines several software systems, software tools and hardware to provide easy to use tools system to enterprises.

- It is a single one-stop solution for all Big Data needs of an enterprise irrespective of size and data volume. Big Data Platform is enterprise class IT solution for developing, deploying and managing Big Data.

Unit – II

GETTING INSIGHTS FROM DATA, DATA QUALITY AND PREPROCESSING

1. What is data pre-processing?

Data pre-processing is the process of transforming raw data into a useful, understandable format. Real-world or raw data usually has inconsistent formatting, human errors, and can also be incomplete. Data pre-processing resolves such issues and makes datasets are complete and more efficient to perform data analysis.

2. Why is data pre-processing required?

As you know, a database is a collection of data points. Data points are also called observations, data samples, events, and records.

3. List out the Characteristics of quality data.

- **Accuracy:** As the name suggests, accuracy means that the information is correct. Outdated information, typos, and redundancies can affect a dataset's accuracy.
- **Consistency:** The data should have no contradictions. Inconsistent data may give you different answers to the same question.
- **Completeness:** The dataset shouldn't have incomplete fields or lack empty fields. This characteristic allows data scientists to perform accurate analyses as they have access to a complete picture of the situation the data describes.
- **Validity:** A dataset is considered valid if the data samples appear in the correct format, are within a specified range, and are of the right type. Invalid datasets are hard to organize and analyse.
- **Timeliness:** Data should be collected as soon as the event it represents occurs. As time passes, every dataset becomes less accurate and useful as it doesn't represent the current reality. Therefore, the topicality and relevance of data is a critical data quality characteristic.

4. Write down the four stages of data pre-processing.

There are four stages of data processing:

- cleaning
- integration
- reduction
- transformation

5. What is Data cleaning?

Data cleaning or cleansing is the process of cleaning datasets by accounting for missing values, removing outliers, correcting inconsistent data points, and smoothing noisy data. In essence, the motive behind data cleaning is to offer complete and accurate samples for machine learning models.

6. What is Missing values?

The problem of missing data values is quite common. It may happen during data collection or due to some specific data validation rule. In such cases, you need to collect additional data samples or look for additional datasets.

7. Define Noisy data.

A large amount of meaningless data is called noise. More precisely, it's the random variance in a measured variable or data having incorrect attribute values. Noise includes duplicate or semi-duplicates of data points, data segments of no value for a specific research process, or unwanted information fields.

8. Define Data integration.

Since data is collected from various sources, data integration is a crucial part of data preparation. Integration may lead to several inconsistent and redundant data points, ultimately leading to models with inferior accuracy.

Here are some approaches to integrate data:

- Data consolidation
- Data virtualization
- Data propagation

9. What is Data virtualization?

In this approach, an interface provides a unified and real-time view of data from multiple sources. In other words, data can be viewed from a single point of view.

10. What is Data reduction?

As the name suggests, data reduction is used to reduce the amount of data and thereby reduce the costs associated with data mining or data analysis.

It offers a condensed representation of the dataset. Although this step reduces the volume, it maintains the integrity of the original data. This data pre-processing step is especially crucial when working with big data as the amount of data involved would be gigantic.

11. Define Dimensionality reduction.

Dimensionality reduction, also known as dimension reduction, reduces the number of features or input variables in a dataset.

The number of features or input variables of a dataset is called its dimensionality. The higher the number of features, the more troublesome it is to visualize the training dataset and create a predictive model.

The following are some ways to perform dimensionality reduction:

- Principal component analysis (PCA)
- High correlation filter
- Missing values ratio
- Low variance filter
- Random forest

UNIT III CLUSTERING
PART A

1. Define clustering.

Clustering is the process of grouping of similar data or data points together in a same group or cluster.

2. List the categories of clustering methods.

a) Partitioning methods
b) Hierarchical methods
c) Density based methods
d) Grid based methods
e) Model based methods

3. Explain various steps in clustering process.

- Find groups of similar data items
- Statistical techniques require some definition of "distance" (e.g. between travel profiles) while conceptual techniques use background concepts and logical descriptions Uses:
- Demographic analysis Technologies:
- Self-Organizing Maps
- Probability Densities
- Conceptual Clustering

4. What are the requirements of cluster analysis?

The basic requirements of cluster analysis are

- Scalability
- Ability to deal with different types of attributes
- Ability to deal with noisy data
- Minimal requirements for domain knowledge to determine input parameters
- Constraint based clustering
- Interpretability and usability

5. What is Cluster Analysis?

- **Cluster analysis: A task that does**

– Grouping a set of data objects into clusters.
– Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

- A cluster analysis is the process of analyzing the various clusters to organize the different objects into meaningful and descriptive objects.

6. What is a "decision tree"?

It is a flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions.

Decision tree is a predictive model. Each branch of the tree is a classification question and leaves of the tree are partition of the dataset with their classification.

7. Define the concept of classification.

Two step process

a) A model is built describing a predefined set of data classes or concepts. The model is constructed by analysing database tuples described by attributes. b) The model is used for classification.

8. What are Bayesian Classifiers?

- Bayesian Classifiers are statistical classifiers.
- They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class.

9. How will you solve a classification problem using Decision Tree?

- Decision Tree Induction:
- Construct a decision tree using training data.
- For each ti € D apply the decision tree to determine its class ti-tuple D-Database

10. Define k-means clustering.

Given a collection of objects each with n measurable attributes, k-means is an analytical technique that, for a chosen value of k, identifies k clusters of objects based on the objects' proximity to the center of the k groups. The center is determined as the arithmetic average (mean) of each cluster's n-dimensional vector of attributes.

UNIT IV CLASSIFICATION

PART A

1. What is classification?

Classification is:

– the data mining process of
– finding a model (or function) that
– describes and distinguishes data classes or concepts,
– for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

• That is, predicts categorical class labels (discrete or nominal).
• Classifies the data (constructs a model) based on the training set.
• It predicts group membership for data instances.

2. What are the types of classification algorithm with examples

1. Binary Classification algorithm
2. Predictive performance measures
3. Distance based learning algorithms
4. Probabilistic classification algorithms

3. Differentiate Distance based algorithm and probability algorithm

Distance based algorithm is simplest approach to predicting the class of a new object is to see how similar this object is to other, previously labeled, objects. Probability algorithm estimate the probability of an object belonging to each class, given the information we have.

4. What is a "decision tree"?

It is a flow-chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions.
Decision tree is a predictive model. Each branch of the tree is a classification question and leaves of the tree are partition of the dataset with their classification.

5. Define the concept of classification.

Two-step process
a) A model is built describing a predefined set of data classes or concepts. The model is constructed by analysing database tuples described by attributes. b) The model is used for classification.

6. What are Bayesian Classifiers?

• Bayesian Classifiers are statistical classifiers.

- They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class.

7. How will you solve a classification problem using Decision Tree?

- Decision Tree Induction:
- Construct a decision tree using training data.
- For each ti € D apply the decision tree to determine its class ti-tuple D-Database

8. Define k-means clustering.

Given a collection of objects each with n measurable attributes, k-means is an analytical technique that, for a chosen value of k, identifies k clusters of objects based on the objects' proximity to the center of the k groups. The center is determined as the arithmetic average (mean) of each cluster's n-dimensional vector of attributes.

9. Write the advantages of case-based reasoning

Development is easier, systems learn by acquiring new cases through use, and this makes maintenance easier. CBR also enables the reasoner to propose solutions to problems quickly.

10. Distinguish between Linear and non-linear regression

Linear means something related to a line. All the linear equations are used to construct a line. A non-linear equation is such which does not form a straight line. It looks like a curve in a graph and has a variable slope value.

UNIT V REGRESSION AND APPLICATIONS

PART-A

1. What are Recommenders?

• Recommenders are instances of personalization software.

• Personalization concerns adapting to the individual needs, interests, and preferences of each user.

Includes:

- Recommending
- Filtering
- Predicting (e.g., form or calendar appt. completion)

From a business perspective, it is viewed as part of Customer Relationship Management (CRM).

2. What is Dimensionality Reduction?

Dimension Reduction refers to:

- The process of converting a set of data having vast dimensions into data with lesser dimensions ensuring that it conveys similar information concisely.
- These techniques are typically used while solving machine learning problems to obtain better features for a classification or regression task.

3. List out the problems on using Recommendation systems

• Inconclusive user feedback forms

• Finding users to take the feedback surveys

• Weak Algorithms

• Poor results

• Poor Data

• Lack of Data

• Privacy Control (May NOT explicitly collaborate with recipients)

4. List out the types of Recommender Systems.

- Content
- Collaborative
- Knowledge

5. What is Association Mining?

Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.

6. What is the Purpose of Apriori Algorithm?

Apriori algorithm is an influential algorithm for mining frequent item sets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties**.**

7. List out the applications of Association rules.

- Basket data analysis,
- cross-marketing,
- catalogue design,
- loss-leader analysis,
- clustering,
- classification

8. Define support and confidence in Association rule mining.

Support S is the percentage of transactions in D that contain AUB.

Confidence c is the percentage of transactions in D containing A that also contain B.

Support (A=>B) = P(AUB)

Confidence (A=>B) =P(B/A)

9. What is Association rule?

Association rule finds interesting association or correlation relationships among a large set of data items, which is used for decision-making processes. Association rules analyses buying patterns that are frequently associated or purchased together.

10. Describe the method of generating frequent item sets without candidate generation.

Frequent-pattern growth (or FP Growth) adopts divide-and-conquer strategy.

Steps:

- Compress the database representing frequent items into a frequent pattern tree or FP tree
- Divide the compressed database into a set of conditional databases
- Mine each conditional database separately.