FLIP ROBO TECHNOLOGIES

Worksheet Set – 1

Batch DS2301

Intern: Vishnukanth

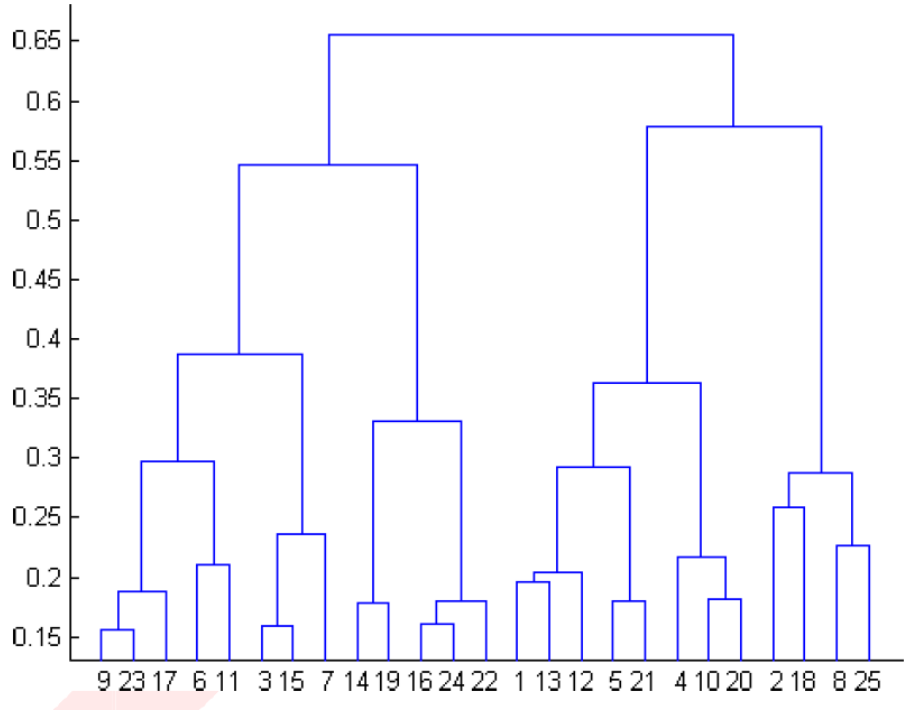
Mail ID: vishnukh25@gmail.com

- ✓ **Machine Learning**
- ✓ **SQL**
- ✓ **Statistics**

Machine Learning

Worksheet – 1

Q1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



Option B - **4**

Q2. In which of the following cases will K-Means clustering fail to give good results?
1. Data points with outliers
2. Data points with different densities
3. Data points with round shapes
4. Data points with non-convex shapes

Options:
a) 1 and 2
b) 2 and 3
c) 2 and 4
d) 1, 2 and 4

Option D - **1, 2 and 4**

Q3. The most important part of is selecting the variables on which clustering is based.
Option D - **formulating the clustering problem**

Q4. The most commonly used measure of similarity is the or its square.

Option A - **Euclidean distance**

Q5. is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.
Option B - **Divisive clustering**

Q6. Which of the following is required by K-means clustering?
a) Defined distance metric
b) Number of clusters
c) Initial guess as to cluster centroids
d) All answers are correct
Option D – **All answers are correct**

Q7. The goal of clustering is to-
Option A - **Divide the data points into groups**

Q8. Clustering is a-
Option B - **Unsupervised learning**

Q9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?
a) K- Means clustering
b) Hierarchical clustering
c) Diverse clustering
d) All of the above
Option D – **All of the above**

Q10. Which version of the clustering algorithm is most sensitive to outliers?
Option A - **K-means clustering algorithm**

Q11. Which of the following is a bad characteristic of a dataset for clustering analysis-
a) Data points with outliers
b) Data points with different densities
c) Data points with non-convex shapes
d) All of the above
Option D – All of the above

Q12. For clustering, we do not require-
Option B - **Unlabeled data**

Q13. How is cluster analysis calculated?

SPSS offers three methods for the cluster analysis: K-Means Cluster, Hierarchical Cluster, and Two-Step Cluster.

K-means cluster is a method to quickly cluster large data sets. The researcher defines the number of clusters in advance. This is useful to test different models with a different assumed number of clusters.

Hierarchical cluster is the most common method. It generates a series of models with cluster solutions from 1 (all cases in one cluster) to n (each case is an individual cluster). Hierarchical cluster also works with variables as opposed to cases; it can cluster variables together in a manner somewhat similar to factor analysis. In addition, hierarchical cluster analysis can handle nominal, ordinal, and scale data; however it is not recommended to mix different levels of measurement.

Two-step cluster analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

The hierarchical cluster analysis follows three basic steps: 1) calculate the distances, 2) link the clusters, and 3) choose a solution by selecting the right number of clusters. The Dendrogram will graphically show how the clusters are merged and allows us to identify what the appropriate number of clusters is.

Q14. How is cluster quality measured?

Cluster quality is measured generally using Dissimilarity/Similarity metric but other methods are also available to measure its quality which include:

- Dissimilarity/Similarity metric
- Cluster completeness
- Ragbag
- Small cluster preservation

To measure the quality of a clustering, the average silhouette coefficient value of all objects in the data set.

Q15. What is cluster analysis and its types?

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

- K Means Clustering
- Density-Based Clustering
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- OPTICS (Ordering Points to Identify Clustering Structure)
- HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)
- Hierarchical Clustering
- Fuzzy Clustering
- Partitioning Clustering

- PAM (Partitioning Around Medoids)
- Grid-Based Clustering

Structured Querying Language

Worksheet – 1

Q1. Which of the following is/are DDL commands in SQL?
Options A and D - **Create and Alter**

Q2. Which of the following is/are DML commands in SQL?
Options A and B - **Update and Delete**

Q3. Full form of SQL is **Structured Querying Language** [Option B]
Q4. Full form of DDL is **Data Definition Language** [Option B]
Q5. Full form of DML is **Data Manipulation Language** [Option A]

Q6. Which of the following statements can be used to create a table with column B int type and C float type?
Option C **- Create Table A (B int, C float)**

Q7. Which of the following statements can be used to add a column D (float type) to the table A created above?
Option B - **Alter Table A ADD COLUMN D float**

Q8. Which of the following statements can be used to drop the column added in the above question?
Option B **- Alter Table A Drop Column D**

Q9. Which of the following statements can be used to change the data type (from float to int) of the column D of table A created in above questions?
Option B **- Alter Table A Alter Column D int**

Q10. Suppose we want to make Column B of Table A as primary key of the table. By which of the following statements we can do it?
Option A - **Alter Table A Add Constraint Primary Key B**

Q11. What is data-warehouse?

Data Warehouse is a relational database that is designed for query and analysis for the purpose of decision-making and business insights. Data Warehouse is the core of Business Intelligence.

In short, Data Warehouse is a central repository of information collected from various sources which can be analysed to make informed decisions.

Q12. What is the difference between OLTP VS OLAP?

Category	OLTP	OLAP
Expansion	Online Transaction Processing	Online Analytical Processing
Definition	It is known as an online database modifying system	It is known as an online database query management system
Function	It is responsible for collecting, storing and processing data from transactions in real-time	It is responsible for analysing historical data from OLTP system using queries
Purpose	It serves the purpose to insert, update, delete information from the database	It serves the purpose to extract the data/information for analysis and decision making
Processing time	It is comparatively fast in processing because of simple queries operating on 5% of the data	It is relatively slow because of complex queries involving large amount of data
Type of users	This data is managed by clerks and managers usually	This data is generally managed by CEO, MD, GM and other top tier management
Operation	Read and write are performed frequently	Usually read and rarely write operation are performed
Application	It is application oriented used for business operations/tasks	It is subject oriented used for Data Mining, Analytics and others
Productivity	Improves the users' productivity	Improves the efficiency of business analysis

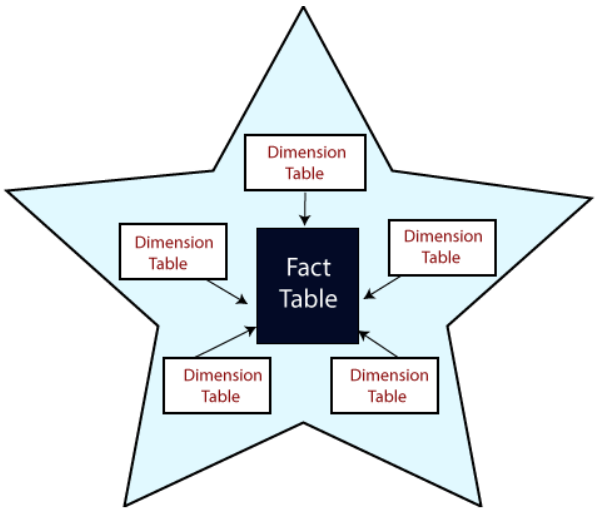
Q13. What are the various characteristics of data-warehouse?

The characteristics of data warehouse include:

- Subject oriented
- Time-variant
- Integrated
- Non-volatile

Q14. What is Star-Schema?

A star schema is the elementary form of a dimensional model, in which data are organized into facts and dimensions.
A fact is an event that is counted or measured, such as a sale or log in.
A dimension includes reference data about the fact, such as date, item, or customer.



Star Schema

A star schema is a relational schema where a relational schema whose design represents a multidimensional data model.
The star schema is the explicit data warehouse schema.
It is known as **star schema** because the entity-relationship diagram of this schemas simulates a star, with points, diverge from a central table. The center of the schema consists of a large fact table, and the points of the star are the dimension tables.

Q15. What do you mean by SETL?

SETL (SET Language) is a very high-level programming language based on the mathematical theory of sets. It was originally developed by (Jack) Jacob T. Schwartz at the New York University (NYU) Courant Institute of Mathematical Sciences in the late 1960s.

- SETL provides two basic aggregate data types: unordered sets, and sequences (the latter also called tuples).
- SETL provides quantified boolean expressions.
- SETL provides several iterators to produce a variety of loops over aggregate data structures.

STATISTICS

Worksheet – 1

Q1. Bernoulli random variables take (only) the values 1 and 0.
Option A - **True**

Q2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
Option A - **Central Limit Theorem**

Q3. Which of the following is incorrect with respect to use of Poisson distribution?
Option B - **Modeling bounded count data**

Q4. Point out the correct statement.
a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
c) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned
Option D - **All of the mentioned**

Q5. _____ random variables are used to model rates.
Option C - **Poisson**

Q6. Usually replacing the standard error by its estimated value does change the CLT.
Option B - **False**

Q7. Which of the following testing is concerned with making decisions using data?
Option B - **Hypothesis**

Q8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
Option A - **0**

Q9. Which of the following statement is incorrect with respect to outliers?
Option C - **Outliers cannot conform to the regression relationship**

Q10. What do you understand by the term Normal Distribution?

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
In graphical form, the normal distribution appears as a "bell curve". In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3. Normal distributions are symmetrical, but not all symmetrical distributions are normal.

Q11. How do you handle missing data? What imputation techniques do you recommend?

There are two ways to handle the missing the data either ignorance or imputation. By making no decision, the statistical programme will make the decision for the user. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

There are different imputation techniques which include:

- Mean Imputation
- Substitution
- Hot Deck Imputation
- Cold Deck Imputation
- Regression Imputation
- Stochastic Regression Imputation
- Interpolation and Extrapolation
- Single (or) Multiple Imputation

Among the above, multiple imputation eliminates many difficulties with missing data and when done correctly, it leads to unbiased parameter estimations and accurate standard errors.

Q12. What is A/B testing?

A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

Q13. Is mean imputation of missing data acceptable practice?

No, it is not acceptable. The process of replacing null values in a data collection with the data's mean is known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation. Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Q14. What is linear regression in statistics?

Simple linear regression is used to estimate the relationship between two quantitative variables. Linear regression is used to analyse:

- how strong the relationship is between two variables
- the value of the dependent variable at a certain value of the independent variable

Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Q15. What are the various branches of statistics?

The various branches of statistics include:

- Data Collection
- Descriptive Statistics
- Inferential Statistics

Descriptive statistics include:

- Central tendency – Mean, Median, Mode
- Dispersion of data – Standard Deviation, Variance, Percentile, IQR, Kurtosis, Skewness

Inferential Statistics include:

- ZScore
- Hypothesis testing - T-test, Z-test, Chi Square, Regression test, ANOVA test