STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.

a) True

b) False

Ans.1: a) True

2. 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?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Ans.2: a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data

b) Modeling bounded count data

c) Modeling contingency tables

d) All of the mentioned

Ans.3: b) Modeling bounded count data

4. 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

Ans.4: d) All of the mentioned

5. ______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans.5: c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

Ans.6: b) False

7. 1. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Ans.7: b) Hypothesis

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.

a) 0

b) 5

c) 1

d) 10

Ans.8: a) 0

9. Which of the following statement is incorrect with respect to outliers?

a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

Ans.9: c) Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans.10:

Normal distribution is a symmetrical distribution of data around the mean of the data. The curve when plotted looks like a bell.

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

Ans.11:

If the percentage of missing values is low, then I consider removing the rows with missing values, but if missing data is not small in percentage or removing data is creating a bias, then we can consider imputing these missing values using various methods, depending upon the type of problem.

I recommend KNN imputation as a technique to impute missing values in the data. This method is simple to understand and implement.

12. What is A/B testing?

Ans.12:

A/B testing is a test done by creating two versions, A and B, of a product and then monitoring which product version is turning in more profit, and/or the change in behaviour of consumers with respect to certain characteristics of each product version.

13. Is mean imputation of missing data acceptable practice?

Ans.13:

Mean imputation is not an acceptable method because it does not consider the correlation between the target variable and the features, this will cause a negative effect on accuracy of model. Also imputing mean values in place of missing values will decrease the actual variance of the feature and thus, can introduce bias in the data.

14. What is linear regression in statistics?

Ans.14:

Linear regression is the most basic form of a regression analysis. It is an approach to modelling the relationships between a dependent variable and a group of independent variables.

15. What are the various branches of statistics?

Ans.15:

There are two different branches of statistics:

1. **Descriptive Statistics:**
   It focuses on collecting, summarizing and presenting a set of data.
2. **Inferential Statistics:**
   It analyses sample of data to draw conclusion about population.

WORKSHEET 1 SQL

Q1 and Q2 have one or more correct answer. Choose all the correct option to answer your question.

1. Which of the following is/are DDL commands in SQL?
A) Create B) Update C) Delete D) ALTER

Ans.1: A) Create & D) ALTER

2. Which of the following is/are DML commands in SQL?
A) Update B) Delete C) Select D) Drop

Ans.2: A) Update & B) Delete

Q3 to Q10 have only one correct answer. Choose the correct option to answer your question.

3. Full form of SQL is:
A) Strut querying language B) Structured Query Language C) Simple Query Language D) None of them

Ans.3: B) Structured Query Language

4. Full form of DDL is:
A) Descriptive Designed Language B) Data Definition Language C) Data Descriptive Language
D) None of the above.

Ans.4: B) Data Definition Language

5. DML is:
A) Data Manipulation Language B) Data Management Language C) Data Modeling Language
D) None of these

Ans.5: A) Data Manipulation Language

6. Which of the following statements can be used to create a table with column B int type and C float type?
A) Table A (B int, C float) B) Create A (b int, C float) C) Create Table A (B int,C float)
D) All of them

Ans.6: C) Create Table A (B int,C float)

7. Which of the following statements can be used to add a column D (float type) to the table A created above?
A) Table A (D float) B) Alter Table A ADD COLUMN D float C) Table A(B int, C float, D float)
D) None of them

Ans.7: B) Alter Table A ADD COLUMN D float

8. Which of the following statements can be used to drop the column added in the above question?
A) Table A Drop D B) Alter Table A Drop Column D C) Delete D from A
D) None of them

Ans.8: B) Alter Table A Drop Column D

9. 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?
A) Table A (D float int) B) Alter Table A Alter Column D int C) Alter Table A D float int
D) Alter table A Column D float to int

Ans.9: B) Alter Table A Alter Column D int

10. 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?
A) Alter Table A Add Constraint Primary Key B B) Alter table (B primary key)
C) Alter Table A Add Primary key B D) None of them

Ans.10: C) Alter Table A Add Primary key B

<u>Q11 to Q15 are subjective answer type questions, Answer them briefly.</u>

11. What is data-warehouse?

Ans.11:

Data Warehousing is process for collecting and managing data from various sources to provide meaningful business insights.

12. What is the difference between OLTP VS OLAP?

Ans.12:

Differences between OLTP and OLAP:

1) **OLTP is an online database modifying system, whereas OLAP is an online database query management system.**
2) **OLTP is characterized with large number of short online transactions, whereas OLAP is characterized with large volume of data and its analysis.**
3) **OLTP has the transactions as source of data, whereas OLAP uses various OLTP databases as source of data.**
4) **OLTP is designed for daily real time business operations, whereas OLAP is designed for analysis of business measures.**

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

Ans.13:

The major characteristics of data-warehouse are:

1) **Subject-oriented:**

   Focuses on analysis to help make business decisions.
2) **Time-variant:**

   Data is maintained and updated frequently with every specific time interval.
3) **Non-volatile:**

   Data in the data-warehouse is permanent.
4) **Integrated:**

   Data is integrated from various data sources.

14. What is Star-Schema??

Ans.14:

Star-Schema is a type of Data-warehouse schema, where a table, known as fact-table, is at the center of the schema and surrounded by various associated dimension tables, such that the structure resembles a star. The fact-table contains keys to every dimension table. The dimension tables are not joined to each other and are joined individually with the fact table using a foreign key.

15. What do you mean by SETL?

Ans.15:

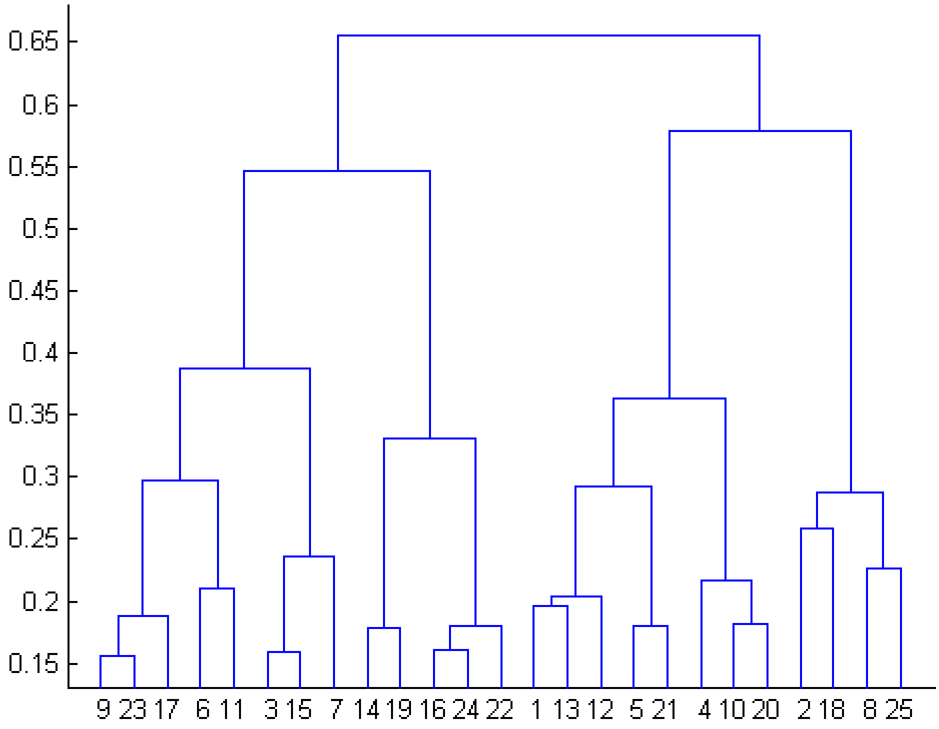
SETL stands for Semantic Extract-Transform-Load framework for semantic data warehouses.

SETL builds on Semantic Web (SW) standards and tools and supports developers by offering a number of powerful modules, classes, and methods for (dimensional and semantic) Data Warehouse constructs and tasks. Thus, it supports semantic data sources in addition to traditional data sources, semantic integration, and creating or publishing a semantic (multidimensional) Data Warehouse in terms of a knowledge base. A comprehensive experimental evaluation comparing SETL to a solution made with traditional tools (requiring much more hand-coding) on a concrete use case, shows that SETL provides better programmer productivity, knowledge base quality, and performance.

Machine Learning

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

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



a) 2
b) 4
c) 6
d) 8

Ans.1: b) 4

2. 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

Ans.2: d) 1,2 and 4

3. The most important part of is selecting the variables on which clustering is based.
a) interpreting and profiling clusters
b) selecting a clustering procedure
c) assessing the validity of clustering
d) formulating the clustering problem

Ans.3: d) formulating the clustering problem

4. The most commonly used measure of similarity is the or its square.
a) Euclidean distance
b) city-block distance
c) Chebyshev’s distance
d) Manhattan distance

Ans.4: a) Euclidean distance

5. 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.
a) Non-hierarchical clustering
b) Divisive clustering
c) Agglomerative clustering
d) K-means clustering

Ans.5: c) Agglomerative clustering

6. 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

Ans.6: d) All answers are correct

7. The goal of clustering is to-
a) Divide the data points into groups
b) Classify the data point into different classes
c) Predict the output values of input data points
d) All of the above

Ans.7: d) All of the above

8. Clustering is a-
a) Supervised learning
b) Unsupervised learning
c) Reinforcement learning
d) None

Ans.8: b) Unsupervised learning

9. 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

Ans.9: a) K- Means clustering

10. Which version of the clustering algorithm is most sensitive to outliers?
a) K-means clustering algorithm
b) K-modes clustering algorithm
c) K-medians clustering algorithm
d) None

Ans.10: a) K-means clustering algorithm

11. 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

Ans.11: d) All of the above

12. For clustering, we do not require-
a) Labeled data
b) Unlabeled data
c) Numerical data
d) Categorical data

Ans.12: a) Labeled data

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.

13. How is cluster analysis calculated?
Ans.13:
Basic steps to perform cluster analysis:

1) **Formulating a problem**
2) **Selecting a distance measure**
3) **Selecting a clustering procedure**
4) **Deciding the number of clusters**
5) **Interpreting the profile clusters**
6) **Finally, assessing the validity of clustering.**

14. How is cluster quality measured?
Ans.14:
There are many ways to measure cluster quality. The three popular ways are listed and explained below:

1. **Silhouette Coefficient:**
   The Silhouette Coefficient is measured using below formula:

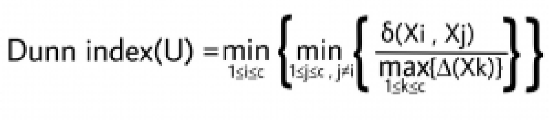
$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

where *a(i)* is the average distance of point *i* from all other points in its cluster and *b(i)* is the smallest average distance of *i* to all points in any other cluster. To clarify, *b(i)* is found by measuring the average distance of *i* from every point in cluster A, the average distance of i from every point in cluster B, and taking the smallest resulting value.

The Silhouette Coefficient tells us how well-assigned each individual point is. If *S(i)* is close to 0, it is right at the inflection point between two clusters. If it is closer to -1, then we would have been better off assigning it to the other cluster. If *S(i)* is close to 1, then the point is well-assigned and can be interpreted as belonging to an ‘appropriate’ cluster.

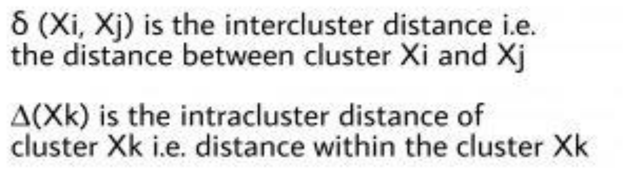
2. **Dunn index:**
   The Dunn index is an internal evaluation scheme, where the result is based on the clustered data itself. The aim is to identify sets of clusters that are compact, with a small variance between members of the cluster, and well separated, where the means of different clusters are sufficiently far apart, as compared to the within cluster variance.
   Higher the Dunn index value, better is the clustering. The number of clusters that maximizes Dunn index is taken as the optimal number of clusters k. It also has some drawbacks. As the number of clusters and dimensionality of the data increase, the computational cost also increases.

The Dunn index for c number of clusters is defined as:



$$\text{Dunn index}(U) = \min_{1\le i\le c}\left\{\min_{1\le j\le c,\, j\ne i}\left\{\frac{\delta(Xi\,,Xj)}{\max\limits_{1\le k\le c}\{\Delta(Xk)\}}\right\}\right\}$$

Where,



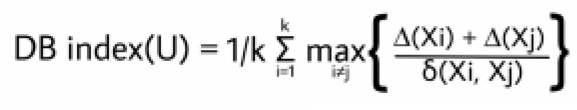
δ (Xi, Xj) is the intercluster distance i.e.
the distance between cluster Xi and Xj

Δ(Xk) is the intracluster distance of
cluster Xk i.e. distance within the cluster Xk

3. **DB index:**

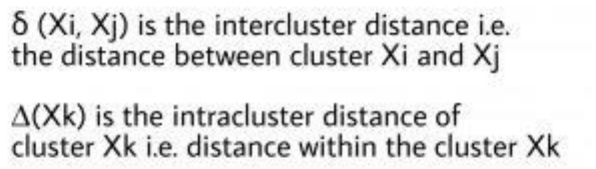
The Davies–Bouldin index is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset.
Lower the DB index value, better is the clustering. It also has a drawback. A good value reported by this method does not imply the best information retrieval.

The DB index for k number of clusters is defined as:



$$\text{DB index}(U) = 1/k \sum_{i=1}^{k} \max_{i\ne j}\left\{\frac{\Delta(Xi) + \Delta(Xj)}{\delta(Xi,\, Xj)}\right\}$$

Where,



δ (Xi, Xj) is the intercluster distance i.e.
the distance between cluster Xi and Xj

Δ(Xk) is the intracluster distance of
cluster Xk i.e. distance within the cluster Xk

15. What is cluster analysis and its types?
Ans.15:
Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting, is to segment customers or activities.

Four basic types of cluster analysis:

1. **Centroid Clustering.**
2. **Density Clustering.**
3. **Distribution Clustering.**
4. **Connectivity Clustering.**