STATISTICS WORKSHEET 1

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

a) True

b) False

**Answer. 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

**Answer. 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

**Answer. 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

**Answer. d) All of the mentioned**

5. \_\_\_\_\_\_ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

**Answer. c) Poisson**

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

a) True

b) False

**Answer. b) False**

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

**Answer. b) Hypothesis Testing**

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

**Answer. 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

**Answer. c) Outliers cannot conform to the regression relationship**

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

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

**Answer**. Normal Distribution also called the Gaussian distribution is a continuous probability distribution that is symmetric about the mean. In a normal distribution the area under the curve represents the probability and it sums up to one. The graph of the probability density looks like a bell for a normal distribution. Mostly the continuous data values aggregate around the mean. The tails in the normal distribution curve are asymptotic which means that they tend to approach but never meet the x-axis. Mean, median and mode in a normal distribution are same.

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

**Answer**. Missing data is an inevitable part of the data analysis process. It is a huge problem in data analysis as it results in biased findings. Hence, missing data must be appropriately addressed. Several ways are there to handle the missing data the common practice of all is the deletion process.

1. Deletion Method- Two common practised deletion techniques are listwise deletion and pair wise deletion. Listwise deletion omits all the observation of the missing data and analyse the rest of the data. This is a good practice if only a few of the observations have missing data but it would give a biased result if we have a large number of missing data. Pairwise deletion only eliminates the data only if the particular data-point vital for testing is missing. Pairwise deletion preserves the data better than listwise deletion as listwise deletes the observation regardless of their importance to the testing.
2. Imputation- Imputation is the process of replacing the missing data with the estimated values. Instead of dropping the missing values, it replaces the missing data with a value estimated with the help of available information. There are two common imputation techniques- average imputation and common-point imputation. Average imputation fills out the missing data with the average value of variables. Common-point imputation is replacing the missing values with the middle observation or the most commonly occurring observation. One can use mean, median or mode to replace the missing values after clearly checking which method fits best for the situation.

12. What is A/B testing?

**Answer**. It is a randomized controlled experiment. It is a way to compare two variants of a variable to check which of the two variant is effective in controlled environment. For example, a company wants to check whether sale for a product would increase if they change the appearance of the product cover. The company divides the product into two variants- A with no changes and B with some changes in the appearance of the product. Now on the basis of customer of response for the two variants company tries to decide which has a better sale. This is A/B testing. It is a hypothesis testing methodology for making decisions that estimates population parameters based on sample statistics. The population here refers to the customer buying the product and the sample is the number of customers that participated in the test.

13. Is mean imputation of missing data acceptable practice?

**Answer**. Mean imputation is a good practice if the missing data is completely random as it will not bias the estimation. Mean Imputation is generally not acceptable for the missing values that are not random as it may lead to inconsistent bias. Furthermore, this method adds no new information but only increases the sample size hence decreasing the estimator’s standard error. As there will be low standard errors when we have imputed the data. With low standard errors we will have a low p values, we will end up making a Type-1 error without realizing. Also, Mean imputation does not preserve the relationship of the variables. So, mean imputation is not a general acceptable imputation technique.

14. What is linear regression in statistics?

**Answer**. Linear regression is a way to model linear relationship between the scalar response (or the dependent variable) and one or more explanatory variables (independent variables). Simple linear regression plots one independent variable against one dependent variable. Multiple linear regression is plotting of more than one independent variable with a dependent variable. A linear regression is the result of a regression analysis between the variables where the relationship between the variable can be described with a straight line. A curved line is produced in a non-linear regression.

15. What are the various branches of statistics?

**Answer**. The two main branches of statistics are-

1. Descriptive Statistics involves the process of collection, summarizing and presentation of data. This is the initial work in a statistical analysis process. It is utmost important to collect the data keeping in mind the experiment so as to target the right groups and also to avoid biases that are very commonly found in the data.
2. Inferential Statistics as the name suggests is the process of analysing the sample data to draw conclusions about the population. Mostly the prediction about the future and the generalizations about the behaviour of the population by studying a sample population comes under inferential statistics.

SQL WORKSHEET 1

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

**Answer. A) Create and D) Alter**

2. Which of the following is/are DML commands in SQL?

A) Update

B) Delete

C) Select

D) Drop

**Answer. A) Update, B) Delete and C) Select**

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

**Answer. 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.

**Answer. B) Data Definition Language**

5. DML is:

A) Data Manipulation Language

B) Data Management Language

C) Data Modeling Language

D) None of these

**Answer. A) Data Manipulation Language**

6. Which of the following statements can be used to create a table with column B int type and C floattype?

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

**Answer. 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

**Answer. 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

**Answer. 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

**Answer. 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

**Answer. A) Alter Table A Add Constraint Primary Key B**

Q11 to Q15 are subjective answer type questions, Answer them briefly.

11. What is data-warehouse?

**Answer.** Data warehouse is a place where valuable historical and commutative data of an organization is stored for analytical purpose. It is the large amount of data from different fields of a company that helps a company in decision making and forecasting.

12. What is the difference between OLTP VS OLAP?

**Answer.**

|  |  |
| --- | --- |
| OLTP | OLAP |
| OLTP or Online Transaction Process is defined as the operational system that supports transaction-oriented applications in a three-tier system. | OLAP or Online Analytical Processing is a software tool which allows user to analyse and retrieve data. User can analyse data from multiple databases at one time. |
| The main objective is data processing. | The main objective is data analysis. |
| OLTP uses the traditional data base management system. | OLAP uses the data warehouse. |
| OLTP is a market-oriented process. | OLAP is a customer-oriented process. |
| Example of OLTP is ATM centre. | Any data-warehouse is an example of OLAP. |

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

**Answer.** Characteristics of data warehouse-

1. Subject-oriented: A data warehouse is subject-oriented as it analyses information based on a particular subject and not about the company’s current operations. The subject areas can be sales, marketing, HR etc.
2. Integrated: A data warehouse works in an integrated manner by establishing a common unit for the similar data from distinct databases. Integration helps in effective analysis of data.
3. Time-variant: Time-variant means we have time associated with the data in the data warehouse. We have historical data in the data warehouse which we can retrieve and analyse.
4. Non-Volatile: Once a data is entered in a data warehouse it cannot be altered.

14. What is Star-Schema?

**Answer.** A Star-schema is a relational database schema that contains a fact table and one or more dimensions. The entity-relationship diagram between the fact tables and the dimensions dimension resembles the shape of a start which is why it is called a Star-Schema.

15. What do you mean by SETL?

**Answer**. Also known as Set Language or Set Theory of a language, SETL is a high-level programming language that enables user to solve programming problems in efficient manner. It is based on the mathematical theory of sets. It is an interpreted language with syntax resembling loosely to C and Perl. In SETL every statement terminates with a semi-colon, variable sin SETL are not case sensitive and are automatically determined by their last assignment.

MACHINE LEARNING ASSIGNMENT – 1

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

**Answer. 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

**Answer. 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

**Answer. 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

**Answer. 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

**Answer. b) Divisive 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

**Answer. 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

**Answer. d) All of the above**

8. Clustering is a-

a) Supervised learning

b) Unsupervised learning

c) Reinforcement learning

d) None

**Answer. 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

**Answer. 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

**Answer. 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

**Answer. d) All of the above**

12. For clustering, we do not require-

a) Labeled data

b) Unlabeled data

c) Numerical data

d) Categorical data

**Answer. a) Labeled data**

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

13. How is cluster analysis calculated?

Answer. Cluster analysis is an iterative method which is calculated by finding appropriate number of centroids which have certain properties to which each observation belongs to exactly one of these clusters. These observations share similar properties to the other observations of the clusters and distinct properties to other clusters.

14. How is cluster quality measured?

Answer. In categorical problems, Cross tab helps us analyse better which data set should fall in which cluster and would help us validate the clustering.

Inertia measures clustering quality:

Good clustering has tight clusters and the data points are not spread out. To check the spread of data points inertia is used, lower the spread better the cluster. We can also consider the elbow method to validate a clustering. It is good to consider the elbow point of the inertia plot as the k value. To further validate the chosen k value, we can go for Silhouette value.

Silhouette method can also be used to validate clustering. Silhouette method uses distance metrics to validate the consistency within the data points of a cluster. Silhouette measures the similarity of an object to the cluster it belongs and its dissimilarity to the clusters it doesn’t belong. Silhouette value ranges from -1 to 1 where a value closer to +1 is the indication that the observation well belongs to the cluster it is in and poorly belongs to the other clusters. A high value of silhouette for most of the objects means the clustering is done in appropriate manner.

15. What is cluster analysis and its types?

Answer. Cluster Analysis or Clustering is the task of grouping similar data points in a similar group in a way that they can be characterized by their relevance with each other. This technique creates clusters that allows us to understand how data is related. Data points belonging to a cluster would have similar traits on the basis of which they are grouped. The data points in a cluster would have same properties. The aim of this analysis is to create group in which objects are similar to one another and distinct from the other objects of other groups. Greater the similarity between the objects of a group greater the group be distinguished from another group. Cluster analysis is most commonly used in business setting to segment customers or business activities. Broadly clustering can be divided into two parts:

Hard clustering: In hard clustering each data point completely belongs to a cluster i.e. each observation belongs to exactly one cluster.

Soft clustering: In soft clustering an observation is put in a cluster based on the likelihood of belonging to the cluster. An observation can belong to more than one cluster.

Types of cluster analysis algorithm:

1. Centroid Clustering: As the name suggests this algorithm works with closeness of data points to the centroid. In Centroid clustering we choose the number of clusters required at the end have to be mentioned beforehand. These are iterative clustering algorithm in which the notion of similarity is derived by the closeness of the data points to the centroid of the cluster. K-Means algorithm is a popular algorithm that falls into this category. We must have prior knowledge of dataset to be able to choose the number of clusters.
2. Density Clustering: This techniques clusters data points based on how densely populated they are. This technique works on the belief that denser the data points more related they are. Popular example of density clustering is DBSCAN and OPTICS.
3. Distribution Clustering: Distribution clustering identifies the probability that a data point belonging to the cluster belongs to the same distribution. Around each possible centroid the algorithm defines a probability distribution for each cluster, quantifying the probability of belonging on the distributions. These models often suffer from overfitting.
4. Connectivity Clustering: This technique sees each data point as its own cluster. This technique works with the consideration that the closer data points exhibit similar properties and are similar to each other than to those lying farther in data space. The iterative process of this technique is to continually incorporate a data point or group of data points with other data points until a big cluster is formed. Example of this algorithm is Hierarchical Clustering algorithm.