MACHINE LEARNING

ASSIGNMENT – 1

1.(a) - 2

2. (a) - 1 and 2 i.e data points with outliers and data points with different densities.

3. (d) – formulating the clustering problem.

4. (a) – Euclidean distance.

5 .(b) – Divisive clustering.

6.(d) – All answers are correct.

7.(b) – Divide the data points into groups.

8.(b) – Unsupervised learning.

9.(a) – K means clustering.

10.(a) – K means clustering algorithm.

11.(d) – All of the above.

12.(a) – Labeled data.

13. Cluster analysis can be calculated by the following 3 important steps those are calculating the distance, linking the clusters, choosing the solution by the appropriate number of clusters.

14. The quality of cluster analysis can be seen be a number of metrices. An efficient cluster has minimal intra cluster distance and maximum inter cluster distance.

15. Cluster analysis is a data analysis technique that explores the naturally occurring groups within a data set known as clusters. Cluster analysis is an unsupervised learning method .There are three types of cluster analysis such as K-means , density based , distribution based.

<u>WORKSHEET 1 SQL</u>

<u>ASSIGNMENT</u>

1.(A) – Create and (D) – Alter.

2.(A) – Update and (B) – Delete.

3.(B) – Structured Query Language.

4.(B) – Data Definition Language.

5.(A) – Data Manipulation Language.

6.(C) – All of them.

7.(B) – Alter Table A ADD COLUMN D Float.

8.(D) – None of them.

9.(B) – Alter Table A Alter Column D int

10.(A) - Alter Table A Add Constraint primary key B.

11. A data warehouse is where data can be collected for mining purposes , usually with large storage capacity. Various organizations systems are in the data warehouse , where it can be fetched as per usage.

12. i) Online transaction processing (OLTP) captures and stores the real time data whereas Online analytical processing (OLAP) is a method of using queries to analyse historical data of OLTP.

ii) OLTP makes use of standard data base management system(DBMS) and OLAP makes use of a data warehouse.

13. Data warehouse have 4 important characteristics and these are i) Data warehouse is always a subject oriented as it delivers information about a theme instead of organization’s current operation.

ii) Data warehouse is integrated as it will integrate data from all sources.

iii) Time - variant.

iv) It is non volatile.

14. Star schema is a database organization structure optimized for use in data warehouse that uses a single large fact table to store transactional or measured data .

15. SETL is a very high level programming language based on the mathematical theory of sets.

STATISTICS WORKSHEET-1

1.(a) - True

2.(a) – Central Limit Theorem.

3.(c) – Modeling contingency table.

4.(a) – The exponent of normally distributed random variables follows what is called log normal distribution.

5.(c) - Poission

6.(b) - False

7.(b) – Hypothesis.

8.(a) - 0

9.(c) – Outliers cannot conform to the regression relationship.

10. Normal distribution is such type of distribution of which has mean=median=mode .It is a type of continuous probability distribution in which most data points cluster towards middle range. It has bell shaped.

11. Missing data can be handle by deleting the rows if it is less than 20% else it can be replaced by median or mode .

12. AB testing is a process of splitting the data in different ways to check which split is giving more accurate value like cross validation in python.

13. Mean imputation should not be used as it does not consider the correlations.

14. Linear Regression is used for predictive analysis. It is also used to describe data and to explain the relationship.

15. The main branches of statistics are data collection , descriptive statistics and inferential statistics.