ASSIGNMENT- 1
Machine Learning

1. b) 4
2. d) 1, 2 and 4
3. d) formulating the clustering problem
4. a) Euclidean distance
5. b) Divisive clustering
6. d) All answers are correct
7. a) Divide the data points into groups
8. b) Unsupervised learning
9. d) All of the above
10. a) K-means clustering algorithm
11. d) All of the above
12. a) Labelled data

13. How is cluster analysis calculated?

Ans:- Cluster analysis is basically identifying the similarities and dissimilarities between the data points and grouping them accordingly. Cluster analysis is generally carried out by calculating the distance between two points from centroid, squaring them and summing up them.

14. How is cluster quality measured?

Ans:- If all the data objects in the cluster are highly similar then the cluster has high quality. We can measure the quality of clustering by using the Dissimilarity/Similarity metric in most situations. There are other techniques also like measuring cluster completeness. This is associated with the properties of data points. If two data points have similar properties they must be grouped in a single cluster to increase cluster completeness.

15. What is cluster analysis and its types?

Ans:- Cluster analysis is basically identifying the similarities and dissimilarities between the data points and grouping them accordingly. There are mainly three types of cluster analysis based on approaches used , they are

1. K means Clustering

2. Hierarchical clustering

3. Density based clustering

STATISTICS WORKSHEET-1

1. a) True
2. a) Central Limit Theorem
3. b) Modelling bounded count data
4. d) All of the mentioned
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a)0
9. c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?
Ans:- Normal Distribution is a distribution where data is distributed symmetrical to mean and many data points lies near mean. The normal distribution has a mean zero and standard deviation 1.and it forms bell curve in graphical presentation.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans:- Missing data can be handled in multiple ways and can be ignored based on the problem statement and data under consideration. Most commonly it is handled by varies imputation techniques. 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.
Generally mean imputation would be recommended.

12. What is A/B testing?
Ans :- From software testing perspective A/B testing is a method where two versions of softwares are tested with different random sets of user and can be checked which version creates good impact.

13. Is mean imputation of missing data acceptable practice?
Ans:- Yes

14. What is linear regression in statistics?
Ans:- Linear regression is the technique of establishing the relation between dependant and independent variables and the value of dependant variable is predicted based on independent variables value.

15. What are the various branches of statistics?
Ans:- 1.Descriptive statistics
2.Inferential statistics