MACHINE LEARNING – WORKSHEET 1

(CLUSTERING)

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. a. K- Means clustering
10. a. K-means clustering algorithm
11. b. All of the above
12. a. Labeled data
13. Cluster analysis can be carried on in two ways, that is either using an unsupervised technique or using a know method of binning. For example, we can cluster customers directly on the basis of their income or age and it does not require any mathematical interpretation. However, on the other hand using various distance measures (Euclidean, Manhattan, Hamming etc) we can try to find similar points grouped together, spread across in a multi-dimensional space. We can also use various proximity methods as well as clustering methods (KMeans, KMedoids, Hierarchical, Fuzzy) to carry on the analysis.
14. Cluster quality is usually measured using the WSSE or within sum of squared errors for plotting the elbow graphs to know the right   
      
    Also, silhouette score is used to find the quality of the cluster on mathematical terms which is a function of inter-cluster and intra-cluster distances.  
      
    Apart from the mathematical terms clusters are often visualized using cluster analysis like plotting various variables to see if there is a distinct division of them in the various clusters. An example will be after clustering on customer demographics data, we may see how the factors like age, employment, income are distributed in various clusters. A well-defined distribution and explainablilty of the clusters is another measure for evaluation.
15. Broadly speaking, clustering can be divided into :  
    **Hard Clustering**: Every data point either belongs to a cluster fully or not at all.  
    **Soft Clustering**: Each data point does not belongs to a cluster fully and a probability or likelihood of that data point to be in those clusters is assigned.  
      
    Apart from that we can subdivide the types as:  
    **Connectivity models**: Algorithms are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. They can either be Hierarchical or decisive i.e assigning all points in different clusters and then joining them together else, start with a giant cluster and then allocate all points in seperate clusters.  
    **Centroid models**: These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is the most popular one in this category.  
    **Distribution models**: These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution. Expectation-Maximization is a well know algorithm in this class.  
    **Density Models**: These models search the data space for areas of varied density of data points in the data space. DBSCAN is a good example.

MACHINE LEARNING – WORKSHEET 2

(CLUSTERING)

1. d. 2 and 3
2. e. 1, 2 and 4
3. a. True
4. a. 1 only
5. b. 1
6. b. No
7. a. Yes
8. d. All of the above
9. \_
10. a. K-means clustering algorithm
11. f. All of the above
12. e. All of the above
13. The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. However, if we introduce enough clusters i.e. enough clusters, then theoretically, the extreme value/values will be assigned to a separate cluster and will not mess up with the clustering of the remaining points.
14. Various clustering algorithms like density-based, distribution models with higher levels of complexity are often expensive in run-time. On the other hand, K-means is easier to run, faster to implement and have easy explainabailty and is almost a great solution for pre-clustering, which reduce the space into disjoint smaller sub-spaces based on centroids, where other complex clustering algorithms can be applied.
15. The K-means clustering is a non-deterministic algorithm. When we run it several times on the same data, we get different results. The non-deterministic nature of K-Means is because of its random selection of data points as initial centroids. However, using seeds (ensuring that the same value is passed in the randomized centroid identification), it can be made kind-of deterministic. However, if the initial choice was biased, we may not get a generalized model. So, even if there is provision for seed, it is recommended to run the K-means in its non-deterministic form several times before coming to a conclusion.

WORKSHEET 1

SQL

1. a, d. Create, Alter
2. a,b. Update, Delete
3. b. Structured Query Language
4. b. Data Definition Language
5. a. Data Manipulation Language
6. c. Create Table A (B int,C float)
7. b. Alter Table A ADD COLUMN D float
8. b. Alter Table A Drop Column D
9. b. Alter Table A Alter Column D int
10. c. Alter Table A Add Primary Key B
11. A data warehouse is a system used for reporting and data analysis. It is considered a core component of business intelligence. They are central repositories of integrated data from one or more disparate sources. It stores current and historical data in one single place and are used for creating analytical reports.  
      
    The data stored in the warehouse is uploaded from the operational systems. The data may pass through an operational data store and may require data cleansing. Extract, Transform, Load (ETL) are the main approaches used to build a data warehouse systems.
12. Online Analytical Processing (OLAP), is a category of software tools which provide analysis of data for business decisions. OLAP systems allow users to analyze database information from multiple database systems at one time. Any data-warehouse system is an OLAP system.  
      
    Online transaction processing shortly known as OLTP supports transaction-oriented applications in a 3-tier architecture. OLTP administers day to day transaction of an organization. An example of OLTP system is ATM center such as joint account holders are transacting at the same time.
13. **Subject Oriented**A data warehouse never put emphasis only current operations. Instead, it focuses on demonstrating and analysis of data to make various decisions.  
    **Integrated**Integration means founding a shared entity to scale the all similar data from the different databases.   
    **Time Variant**  
    In this data is maintained via different intervals of time such as weekly, monthly, or annually etc. It founds various time limit which are structured between the large datasets and are held in online transaction process (OLTP).   
    **Non Volatile**The data resided in data warehouse is permanent. It also means that data is not erased or deleted when new data is inserted.
14. Star schema is widely used to develop or build a data warehouse and dimensional data marts. In Star Schema, that holds the quantitative data about a business is distributed in fact tables, and dimensions which are descriptive characteristics related to fact data.
15. In order to create better decisions for business analytics, organizations increasingly use external structured, semi-structured, and unstructured data in addition to the (mostly structured) internal data. Semantic Extract-Transform-Load (SETL) tools are advanced tools to aid the purpose of structuring the data and be in the core for Business Intelligence.

WORKSHEET2

SQL

1. d. Unique
2. d. None of them (UNIQUE can only do)
3. a. Each entry in the primary key uniquely identifies each entry or row in the table
4. a .There should not be any duplicate entries
5. b. Foreign Key
6. d. 1
7. a. one to many
8. c. one to one
9. b. supplier id
10. b. 1
11. a. one to many
12. c. Table
13. a. Insert into
14. b,c. Unique,Primary Key
15. a,c,d

Statistics – WORKSHEET 1

1. a. True
2. a. Central Limit Theorem
3. b. Modeling 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. Normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions with 68%, 95% and 99.7% of the data lying within 1, 2 and 3 standard deviation.
11. There are various ways to handle missing data like deletion of the values, for continuous variables, we can do mean/median and for categorical mode imputation. For time series data, we can do last/next observation carry forward/backward. However, these are on the basis of central tendency and on heavy assumptions. We can have models like Linear Regression for continuous variables and KNN for categorical variables and they tend to prove more worthy and hence recommended.
12. A/B testing is basically statistical hypothesis testing for making decisions that estimates population parameters based on sample statistics. The population refers to all the visitors coming to your website, while the sample refers to the number of visitors that participated in the test. After you made a change in the webpage, you launch your test to gather statistical evidence to accept or reject the claim (hypothesis) about your website visitors (the test increased conversions for a group of visitors on product pages). The final data shows you whether your hypothesis was correct, incorrect or inconclusive.
13. Mean imputation is a simple, easy to understand and computationally light method. If the data has no outliers, then at many times it tends to do a good job because of the approximation and introducing bias. Now again in cases where we tend to achieve higher accuracy, maybe we need a balanced variance and less bias and in those cases mean imputation will not suffice.
14. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. In cases of multiple linear regressions, we can have more then one explanatory variable and in those cases a hyper-plane is fitted rather than a straight line to understand the linear relationship among the dependant and independent variables.
15. The two main branches of statistics are   
      
    **Descriptive statistics**: It deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data.  
    **Inferential statistics**: It deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.

Statistics – WORKSHEET 2

1. c. both
2. c. 12
3. d. All of the above
4. b. Mutually exclusive
5. b. Summarizing and explaining a specific set of data
6. b. dataset
7. c. 1
8. b. Scatterplot
9. d. Analysis of variance
10. a. Z-score
11. c. mean
12. d. 400005.2
13. d. Mean
14. a. Descriptive and inferences
15. d. H-L