MACHINE LEARNING

Assignment 2

1) Movie Recommendation systems are an example of: i) Classification ii) Clustering iii) Regression Options: a) 2 Only b) 1 and 2 c) 1 and 3 d) 2 and 3
Answer = a) 2 Only

2) Sentiment Analysis is an example of: i) Regression ii) Classification iii) Clustering iv) Reinforcement Options: a) 1 Only b) 1 and 2 c) 1 and 3 d) 1, 2 and 4
Answer = d) 1, 2 and 4

3) Can decision trees be used for performing clustering? a) True b) False
Answer = a) True

4) Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points: i) Capping and flooring of variables ii) Removal of outliers Options: a) 1 only b) 2 only c) 1 and 2 d) None of the above
Answer = a) 1 only

5) . What is the minimum no. of variables/ features required to perform clustering? a) 0 b) 1 c) 2 d) 3
Answer = b) 1

6) For two runs of K-Mean clustering is it expected to get same clustering results? a) Yes b) No
Answer = b) No

7) Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means? a) Yes b) No c) Can't say d) None of these
Answer = a) Yes

8) Which of the following can act as possible termination conditions in K-Means? i) For a fixed number of iterations. ii) Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum. iii) Centroids do not change between successive iterations. iv) Terminate when RSS falls below a threshold. Options: a) 1, 3 and 4 b) 1, 2 and 3 c) 1, 2 and 4 d) All of the above
Answer = d) All of the above

9) Which of the following algorithms is most sensitive to outliers? a) K-means clustering algorithm b) K-medians clustering algorithm c) K-modes clustering algorithm d) K-medoids clustering algorithm
Answer = a) K-means clustering algorithm

10) How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning): i) Creating different models for

different cluster groups. ii) Creating an input feature for cluster ids as an ordinal variable. iii) Creating an input feature for cluster centroids as a continuous variable. iv) Creating an input feature for cluster size as a continuous variable. Options: a) 1 only b) 2 only c) 3 and 4 d) All of the above
Answer = d) All of the above

11) What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset? a) Proximity function used b) of data points used c) of variables used d) All of the above
Answer = d) All of the above

12) Is K sensitive to outliers?
Answer = Yes because it is easily influenced by extreme values. Since a single mislabelled example dramatically changes class boundaries

13) Why is K means better?
Answer = It is relatively simple to implement, its easily adapts to new example, guarantee convergence, k-means is one of the simplest algorithms which uses unsupervised learning method to solve known clustering issues. It works really well with large datasets

14) Is K means a deterministic algorithm?
Answer = NO. K-Means is a non-deterministic algorithm. This means that a compiler cannot solve the problem in polynomial time and doesn’t clearly know the next step. This is because some problems have a great degree of randomness to them. K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters

SQL
Assignment 2

1) Which of the following constraint requires that there should not be duplicate entries? A) No Duplicity B) Different C) Null D) Unique
   Answer = D) Unique
2) Which of the following constraint allows null values in a column? A) Primary key B) Empty Value C) Null D) None of them
   Answer = C) Null
3) Which of the following statements are true regarding Primary Key? A) Each entry in the primary key uniquely identifies each entry or row in the table B) There can be duplicate values in a primary key column C) There can be null values in Primary key D) None of the above.
   Answer = A) Each entry in the primary key uniquely identifies each entry or row in the table
4) Which of the following statements are true regarding Unique Key? A) There should not be any duplicate entries B) Null values are not allowed C) Multiple columns can make a single unique key together D) All of the above
   Answer = D) All of the above
5) Which of the following is/are example of referential constraint? A) Not Null B) Foreign Key C) Referential key D) All of them
   Answer = C) Referential key
6) How many foreign keys are there in the Supplier table? A) 0 B) 3 C) 2 D) 1
   Answer = A) 0
7) The type of relationship between Supplier table and Product table is: A) one to many B) many to one C) one to one D) many to many
   Answer = C) one to one
8) The type of relationship between Order table and Headquarter table is: A) one to many B) many to one ASSIGNMENT C) one to one D) many to many
   Answer = D) many to many
9) Which of the following is a foreign key in Delivery table? A) delivery id B) supplier id C) delivery date D) None of them
   Answer = D) None of them
10) The number of foreign keys in order details is: A) 0 B) 1 C) 3 D) 2
    Answer = C) 3
11) The type of relationship between Order Detail table and Product table is: A) one to many B) many to one C) one to one D) many to many
    Answer = D) many to many
12) DDL statements perform operation on which of the following database objects? A) Rows of table B) Columns of table C) Table D) None of them
    Answer = D) None of them
13) Which of the following statement is used to enter rows in a table? A) Insert in to B) Update C) Enter into D) Set Row
    Answer = A) Insert
14) Which of the following is/are entity constraints in SQL? A) Duplicate B) Unique C) Primary Key D) Null
    Answer = B) Unique C) Primary Key

15) Which of the following statements is an example of semantic Constraint? A) A blood group can contain one of the following values - A, B, AB and O. B) A blood group can only contain characters C) A blood group cannot have null values D) Two or more donors can have same blood group
Answer = C) A blood group cannot have null values

STATISTICS WORKSHEET-2

1) What represent a population parameter? A) SD B) mean C) both D) none
   Answer = D) none
2) What will be median of following set of scores (18,6,12,10,15)? A) 14 B) 18 C) 12 D) 10
   Answer = C) 12
3) What is standard deviation? A) An approximate indicator of how number vary from the mean B) A measure of variability C) The square root of the variance D) All of the above
   Answer = D) All of the above
4) The intervals should be ______ in a grouped frequency distribution A) Exhaustive B) Mutually exclusive C) Both of these D) None
   Answer = C) Both of these
5) What is the goal of descriptive statistics? A) Monitoring and manipulating a specific data B) Summarizing and explaining a specific set of data C) Analyzing and interpreting a set of data D) All of these
   Answer = B) Summarizing and explaining a specific set of data
6) A set of data organized in a participant by variables format is called A) Data junk B) Data set C) Data view D) Data dodging
   Answer = B) Data set
7) In multiple regression,_______ independent variables are used A) 2 or more B) 2 C) 1 D) 1 or more
   Answer = D) 1 or more
8) Which of the following is used when you want to visually examine the relationship between 2 quantitative variables? A) Line graph B) Scatterplot C) Bar graph D) Pie graph
   Answer = B) Scatterplot
9) Two or more groups means are compared by using A) analysis B) Data analysis C) Varied Variance analysis D) Analysis of variance
   Answer = D) Analysis of variance
10) _______is a raw score which has been transformed into standard deviation units? A) Z-score B) t-score C) e-score D) SDU score
    Answer = A) Z-score
11) _______is the value calculated when you want the arithmetic average? A) Median B) mode C) mean D) All
    Answer = C) mean
12) Find the mean of these set of number (4,6,7,9,2000000)? A) 4 B) 7 C) 7.5 D) 400005.2
    Answer = B) 7
13) _______ is a measure of central tendency that takes into account the magnitude of scores? A) Range B) Mode C) Median D) Mean
    Answer = D) Mean
14) ______ focuses on describing or explaining data whereas ______involves going beyond immediate data and making inferences A) Descriptive and inferences B) Mutually exclusive and mutually exhaustive properties C) Positive skew and negative skew D) Central tendency
    Answer = A) Descriptive and inferences
15) What is the formula for range? A) H+L B) L-H C) LXH D) H-L
    Answer = D) H-L

MACHINE LEARNING
ASSIGNMENT – 3

1) Which of the following is an application of clustering? a. biological network analysis b. Market trend prediction c. Topic modelling d. All of the above
Answer = d. All of the above
2) On which data type, we cannot perform cluster analysis? a. Time series data b. Text data c. Multimedia data d. None
Answer = d. None
3) Netflix's movie recommendation system usesa. Supervised learning b. Unsupervised learning c. Reinforcement learning and Unsupervised learning d. All of the above
Answer = c. Reinforcement learning and Unsupervised learning
4) The final output of Hierarchical clustering isa. The number of cluster centroids b. The tree representing how close the data points are to each other c. A map defining the similar data points into individual groups d. All of the above
Answer = b. The tree representing how close the data points are to each other
5) Which of the step is not required for K-means clustering? a. A distance metric b. Initial number of clusters c. Initial guess as to cluster centroids d. None
Answer= d. None
6) Which is the following is wrong? a. k-means clustering is a vector quantization method b. k-means clustering tries to group n observations into k clusters c. k-nearest neighbour is same as k-means d. None
Answer = c. k-nearest neighbour is same as k-means
7) 7. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering? i. Single-link ii. Complete-link iii. Average-link Options: a.1 and 2 b. 1 and 3 c. 2 and 3 d. 1, 2 and 3
Answer = d. 1, 2 and 3
8) Which of the following are true? i. Clustering analysis is negatively affected by multicollinearity of features ii. Clustering analysis is negatively affected by heteroscedasticity Options: a. 1 only b. 2 only c. 1 and 2 d. None of them
Answer = a. 1 only
9) In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed? a. 2 b. 4 c. 3 d. 5
Answer = a. 2
10) For which of the following tasks might clustering be a suitable approach? a.Given sales data from a large number of products in a supermarket, estimate future sales for each of these products. b. Given a database of information about your users, automatically group them into different market segments. c. Predicting whether stock price of a company will increase tomorrow. d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy
Answer= b. Given a database of information about your users, automatically group them into different market segments
11) Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:
Answer = a)

12) Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering?

    Answer = b)

13) What is the importance of clustering?

    Answer = It’s a process that has enormous applicability. It can effectively address diverse problems and objectives from the simplest to most complex. It helps in understanding the natural grouping in a dataset. Main advantage of clustering solution is automatic recovery from failure.

14) How can I improve my clustering performance?

    Answer = Applying unsupervised feature learning to input data , by applying ICA blind source separation

STATISTICS WORKSHEET-3

1) Which of the following is the correct formula for total variation? a) Total Variation = Residual Variation – Regression Variation b) Total Variation = Residual Variation + Regression Variation c) Total Variation = Residual Variation * Regression Variation d) All of the mentioned

   Answer = b) Total Variation = Residual Variation + Regression Variation

2) Collection of exchangeable binary outcomes for the same covariate data are called outcomes. a) random b) direct c) binomial d) none of the mentioned

   Answer = c) binomial

3) How many outcomes are possible with Bernoulli trial? a) 2 b) 3 c) 4 d) None of the mentioned

   Answer = a) 2

4) If Ho is true and we reject it is called a) Type-I error b) Type-II error c) Standard error d) Sampling error

   Answer = a) Type-I error

5) Level of significance is also called: a) Power of the test b) Size of the test c) Level of confidence d) Confidence coefficient

   Answer = b) Size of the test

6) The chance of rejecting a true hypothesis decreases when sample size is: a) Decrease b) Increase c) Both of them d) None

   Answer = b) Increase

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

8) What is the purpose of multiple testing in statistical inference? a) Minimize errors b) Minimize false positives c) Minimize false negatives d) All of the mentioned

   Answer = d) All of the mentioned

9) . Normalized data are centred at and have units equal to standard deviations of the original data a) 0 b) 5 c) 1 d) 10

   Answer = a) 0

10) What Is Bayes' Theorem?

    Answer = It states that the conditional probability of an event, based on occurrence of another event is equal to the likelihood of second event given the first event multiplied by the probability of the first event

11) What is z-score?

    Answer = It indicates how much a given values differ from the standard deviation

12) What is t-test?

    Answer = It is a statistical test that compares the means of two samples. It is used in hypothesis testing with a null hypothesis.

13) What is percentile?

    Answer= value on a scale of 100 that indicates the percent of a distribution that is equal to or below it

14) What is ANOVA?

    Answer = Analysis of variance is a collection of statistical model and their associated estimation procedures used to analyse the difference among them

15) How can ANOVA help?

Answer = It is helpful for testing three or more variables. It is similar to multiple two sample t- tests.