BRAINWARE UNIVERSITY

Term End Examination 2020 - 21

Programme – Bachelor of Technology in Computer Science & Engineering

Course Name – Data Analytics

Course Code - BCSE701

Semester / Year - Semester VII

Time allotted : 85 Minutes Full Marks : 70

[The figure in the margin indicates full marks. Candidates are required to give their answers in their own words as far as practicable.]

Group-A

(Multiple Choice Type Question) 1 x 70=70

1. *(Answer any Seventy)*

(i) Which of the following would be more appropriate to be replaced with question mark in the following figure?

a) data analysis b) data science

c) descriptive analytics d) none of the mentioned

(ii) Which of the following approach should be used to ask data analysis question?

a) find only one solution for particular problem b) find out the question which is to be answered

c) find out answer from dataset without asking question d) none of the mentioned

(iii) Which of the following design term is perfectly applicable to the below figure?

a) correlation b) cofounding

c) causation d) none of the mentioned

(iv) The goal of _____ is to focus on summarizing and explaining a specific set of data

a) inferential statistics b) descriptive statistics

c) none of these d) all of these

(v) Which of the following represents the fiftieth percentile, or the middle point in a set of numbers arranged in order of magnitude?

a) Mode
b) median
c) mean
d) variance

(vi) Which of the following mentioned standard probability density functions is applicable to discrete random variables?

a) gaussian distribution
b) poisson distribution
c) rayleigh distribution
d) exponential distribution

(vii) What is the mean of this set of numbers: 4, 6, 7, 9, 2000000?

a) 7.5
b) 400005.2
c) 7
d) 4

(viii) When do the conditional density functions get converted into the marginally density functions?

a) only if random variables exhibit statistical dependency
b) only if random variables exhibit statistical independency
c) only if random variables exhibit deviation from its mean value
d) if random variables do not exhibit deviation from its mean value

(ix) The expected value of a discrete random variable ‘x’ is given by _____

a) P(x)
b) ? P(x)
c) ? x P(x)
d) 1

(x) If P(x) = 0.5 and x = 4, then E(x) =?

a) 1
b) 0.5
c) 4
d) 2

(xi) A fair six-sided die is rolled twice. What is the probability of getting 2 on the first roll and not getting 4 on the second roll?

a) 1/36
b) 1/18
c) 5/36
d) 1/6

(xii) Some test scores follow a normal distribution with a mean of 18 and a standard deviation of 6. What proportion of test takers have scored between 18 and 24?

a) 0.2
b) 0.22
c) 0.34
d) none of these

(xiii) Weight (Y) is regressed on height (X) of 40 adults. The height range in the data is 50-100 and the regression line is Y = 100+0.1X with R^2= 0.12. Which of the conclusions below does not necessarily follow?

a) the data suggests a weak relationship between X and Y
b) an adult with an X-value of 60 has an estimated Y-value of 106
c) an adult with an X-value of 80 has an estimated Y-value of 108
d) an adult with an X-value of 90 has an estimated Y-value of 109

(xiv) In the regression equation Y = 75.65 + 0.50X, the intercept is

a) 0.5
b) 75.65000000000001
c) 1
d) indeterminable

(xv) In a study, subjects are randomly assigned to one of three groups: control, experimental A, or experimental B. After treatment, the mean scores for the three groups are compared. The appropriate statistical test for comparing these means is:

a) the analysis of variance
b) the correlation coefficien
c) chi square
d) the t-test

(xvi) Assume that there is no overlap between the box and whisker plots for three drug treatments where each drug was administered to 35 individuals. The box plots for these data:

a) represent evidence against the null hypothesis of ANOVA
b) provide no evidence for, or against, the null hypothesis of ANOVA

c) represent evidence for the null hypothesis of ANOVA

d) None of these

(xvii) What is the function of a post-test in ANOVA?

a) describe those groups that have reliable differences between group means

b) set the critical value for the F test (or chi-square)

c) determine if any statistically significant group differences have occurred

d) none of these

(xviii) Big data is used to uncover

a) hidden patterns and unknown correlations

b) market trends and customer preferences

c) other useful information

d) all of these

(xix) Which of the following is defined as the rule or formula to test a null hypothesis?

a) test statistic

b) population statistic

c) variance statistic

d) null statistic

(xx) Consider a hypothesis H0 where ?0 = 5 against H1 where ?1 > 5. The test is?

a) right tailed

b) left tailed

c) center tailed

d) cross tailed

(xxi) Logistic regression is used when you want to

a) predict a dichotomous variable from continuous or dichotomous variables

b) predict a continuous variable from dichotomous variables

c) predict any categorical variable from several other categorical variables

d) predict a continuous variable from dichotomous or continuous variable

(xxii) Large values of the log-likelihood statistic indicate

a) that there are a greater number of

b) that the statistical model fits the data

explained vs. unexplained observations

well

c) that as the predictor variable increases, the likelihood of the outcome occurring decreases

d) that the statistical model is a poor fit of the data

(xxiii) Logistic regression assumes a

a) linear relationship between continuous predictor variables and the outcome variable

b) linear relationship between continuous predictor variables and the logic of the outcome variable

c) linear relationship between continuous predictor variables

d) linear relationship between observations

(xxiv) If Sw is singular and N < D, its rank is at most (N is total number of samples, D dimension of data, C is number of classes)

a) N+C

b) N

c) C

d) N-C

(xxv) If Sw is singular and N < D the alternative solution is to use (N is total number of samples, D dimension of data)

a) EM

b) PCA

c) ML

d) any of these

(xxvi) Which of the following is statistical boosting based on additive logistic regression?

a) gamboost

b) gbm

c) ada

d) All of these

(xxvii) What is the purpose of performing cross-validation?

a) to assess the predictive performance of the models

b) to judge how the trained model performs outside the sample on test data

c) both to assess the predictive performance of the models and to judge how the trained

d) none of these

model performs outside the sample on test data

(xxviii) You run gradient descent for 15 iterations with a=0.3 and compute J (theta) after each iteration. You find that the value of J (Theta) decreases quickly and then levels off. Based on this, which of the following conclusions seems most plausible?

a) rather than using the current value of a, use a larger value of a (say a=1.0)

b) rather than using the current value of a, use a smaller value of a (say a=0.1)

c) a=0.3 is an effective choice of learning rate

d) none of these

(xxix) Suppose you have trained a logistic regression classifier and it outputs a new example x with a prediction ho(x) = 0.2. This means

a) our estimate for P(y=1 | x)

b) our estimate for P(y=0 | x)

c) All of these

d) None

(xxx) Which of the following would have a constant input in each epoch of training a deep learning model?

a) weight between input and hidden layer

b) weight between hidden and output layer

c) biases of all hidden layer neurons

d) activation function of output layer

(xxxi) What is true regarding back propagation rule?

a) it is a feedback neural network

b) actual output is determined by computing the outputs of units for each hidden layer

c) hidden layer's output is not all important, they are only meant for supporting input and output layers

d) none of the mentioned

(xxxii) How are input layer units connected to second layer in competitive learning networks?

a) feed forward manner

b) feedback manner

c) feed forward and feedback
d) feed forward or feedback

(xxxiii) What is the name of the model in figure below?

a) rosenblatt perceptron model
b) mcculloch-pitts model
c) widrow's adaline model
d) none of the mentioned

(xxxiv) In random forest you can generate hundreds of trees (say T1, T2 ....Tn) and then aggregate the results of these tree. Which of the following is true about individual (Tk) tree in random forest? 1.Individual tree is built on a subset of the features 2.Individual tree is built on all the features 3.Individual tree is built on a subset of observations Individual tree is built on full set of observations

a) 1 and 3
b) 1 and 4
c) 2 and 3
d) 2 and 4

(xxxv) In random forest or gradient boosting algorithms, features can be of any type. For example, it can be a continuous feature or a categorical feature. Which of the following option is true when you consider these types of features?

a) only random forest algorithm handles real valued attributes by discretizing them
b) only gradient boosting algorithm handles real valued attributes by discretizing them
c) both algorithms can handle real valued attributes by discretizing them
d) none of these

(xxxvi) The cell body of neuron can be analogous to what mathematical operation?

a) summing
b) differentiator
c) integrator
d) none of the mentioned

(xxxvii) What consist of a basic counter propagation network?

a) a feed forward network only
b) a feed forward network with hidden layer
c) two feed forward network with hidden
d) none of the mentioned

layer

(xxxviii) How do you handle missing or corrupted data in a dataset?

a) drop missing rows or columns

b) replace missing values with mean/median/mode

c) assign a unique category to missing values

d) all of these

(xxxix) Which of the following scenario prefers failover cluster instance over standalone instance in SQL server?

a) high confidentiality

b) high availability

c) high integrity

d) none of the mentioned

(xl) A windows failover cluster can support up to _____ nodes

a) 12

b) 14

c) 16

d) 18

(xli) Which of the following is a windows failover cluster quorum mode?

a) node majority

b) no majority: read only

c) file read majority

d) none of the mentioned

(xlii) Point out the wrong statement

a) the system configuration checker will verify the system state of your computer before setup continues

b) micro soft lync server 2010 supports clustering for micro soft SQL server 2005 only

c) the system configuration checker will run one more set of rules to validate your computer configuration with the SQL server features you have specified

d) none of the mentioned

(xliii) Which of the following model model include a backwards elimination feature selection routine?

a) MCV
b) MARS
c) MCRS
d) all of the mentioned

(xliv) Which of the following argument is used to set importance values?

a) scale
b) set
c) value
d) all of the mentioned

(xlv) To register a watch on a z node data, you need to use the _____ commands to access the current content or metadata.

a) stat
b) put
c) receive
d) gets

(xlvi) According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop?

a) big data management and data mining
b) data warehousing and business intelligence
c) management of Hadoop clusters
d) collecting and storing unstructured data

(xlvii) What are the different features of big data analytics?

a) open source
b) scalability
c) data recovery
d) all of these

(xlviii) What is a unit of data that flows through a flume agent?

a) Record
b) event
c) row
d) log

(xlix) As companies move past the experimental phase with Hadoop, many cite the need for additional capabilities, including _____

a) improved data storage and information retrieval
b) improved extract, transform and load features for data integration
c) improved data warehousing functionality
d) improved security, workload management and SQL support

(l) _____the jobs are optimized for scalability but not latency

a) map reduce
b) drill
c) hive
d) oozie

(li) The data node and name node are, respectively, which of the following?

a) master and worker nodes
b) worker and master nodes
c) both worker nodes
d) both master nodes

(lii) What is the process of examining large and varied data sets?

a) big data analytics
b) Small data analytics
c) machine learning
d) none of these

(liii) The important 3vs big data are

a) volume, vulnerability, variety
b) volume, variety, velocity
c) variety, vulnerability, volume
d) velocity, vulnerability, variety

(liv) Active learning, creating data for analytics through reinforcement learning

a) performance element
b) changing element
c) learning element
d) none of these

(lv) Statistical analysis advice should be obtained at the stage of initial planning in a study:

a) so that attribution of authorship can be decided
b) to better coordinate the selection of appropriate sampling methods and data collection instruments
c) so that conflicts of interest could be identified
d) how data will be archived can be planned

(lvi) Which of the following is characteristic of best machine learning method?

a) casual
b) predictive

c) mechanistic
d) none of these

(lvii) Which of the following focuses on the discovery of (previous) unknown properties of the data?

a) velocity
b) variety
c) volume
d) none of these

(lviii) The analysis based on study of price fluctuations, production of commodities and deposits in banks is classified as

a) sample series analysis
b) time series analysis
c) numerical analysis
d) experimental analysis

(lix) What is one of the benefits of small group discussions?

a) it encourages smaller classrooms
b) teachers can teach to a smaller group
c) it allows students to contribute and discuss their ideas in a less intimidating environment than the full classroom
d) it allows teachers to identify the ideal lesson plans

(lx) The IBM _____ analytics appliances combine high-capacity storage for big data with a massively-parallel processing platform for high-performance computing.

a) watson
b) netezza
c) infosight
d) lityxeq

(lxi) Which of the following contains pre-built predictive tools?

a) alteryx
b) fossil
c) paleots
d) ssas

(lxii) What is predicting y for a value of x that is within the interval of points that we saw in the original data called?

a) regression
b) extrapolation
c) intra polation
d) polation

(lxiii) In a simple linear regression model (one independent variable), if we change the input variable by 1 unit. How much output variable will change?

a) by 1
b) no change
c) by intercept
d) by its slope

(lxiv) What is the role of exploratory graphs in data analysis?

a) they are made for formal presentations
b) they are typically made very quickly
c) axes, legends, and other details are clean and exactly detailed
d) they are used in place of formal modeling

(lxv) After SVM learning, each lagrange multiplier ?i takes either zero or non-zero value. What does it indicate in each situation?

a) a non-zero ?i indicates the data point i is a support vector, meaning it touches the margin boundary
b) a non-zero ?i indicates that the learning has not yet converged to a global mini-mum
c) a zero ?i indicates that the data point i has become a support vector data point, on the margin
d) a zero ?i indicates that the learning process has identified support for vector i

(lxvi) What term is applied to the random reappearance of a behavior after extinction?

a) operant conditioning
b) spontaneous recovery
c) random acquisition
d) reconditioning

(lxvii) What are the two forms of ratio schedules?

a) fixed and variable ration schedules
b) operant and classical
c) interval and punishment
d) reward and punishment

(lxviii) An appropriate learning algorithm for the SVM is

a) quadratic programming of soft margins
b) quadratic programming via gradient descent
c) gradient descent with lagrange multiplier
d) quadratic programming via sequential

constraints

minimal optimization

(lxix) What are the three essential components of a learning system?

a) model, gradient descent, learning algorithm

b) error function, model, learning algorithm

c) accuracy, sensitivity, specificity

d) model, error function, cost function

(lxx) Which of the following model include a backwards elimination feature selection routine?

a) MCV

b) MARS

c) MCRS

d) all of these