BRAINWARE UNIVERSITY

Term End Examination 2021 - 22

Programme – Bachelor of Technology in Computer Science & Engineering

Course Name – Data Analytics

Course Code - PEC-702A

(Semester VII)

Time : 1 Hr.25 Min. **Full Marks : 70**

[The figure in the margin indicates full marks.]

Group-A

(Multiple Choice Type Question) 1 x 70=70

Choose the correct alternative from the following :

(1) Which of the following is the most important language for data science?

a) Java b) Ruby

c) R d) none of the mentioned

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

a) correlation b) cofounding

c) causation d) none of the mentioned

(3) Which of the following approach should be used if you can't fix the variable?

a) randomize it b) non stratify it

c) generalize it d) none of the mentioned

(4) Which of the following data mining technique is used to uncover patterns in data?

a) data bagging b) data booting

c) data merging d) data dredging

(5) The goal of _____ is to focus on summarizing and explaining a specific set of d ata

a) inferential statistics b) descriptive statistics

c) none of these d) all of these

(6) 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

(7) Which of the following mentioned standard probability density functions is applicable t o discrete random variables?

a) gaussian distribution b) poisson distribution

c) rayleigh distribution
d) exponential distribution

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

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

(9) The expected value of a discrete random variable 'x' is given by ______

a) P(x)
b) $\sum$ P(x)
c) $\sum$ x P(x)
d) 1

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

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

(11) 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

(12) A fair six-sided die is rolled 6 times. What is the probability of getting all outcomes as unique?

a) 0.01543
b) 0.01993
c) 0.23148
d) 0.03333

(13) Some test scores follow a normal distribution with a mean of 18 and a standard deviati on 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

(14) What is the number of restrictions in the calculation of the F-statistics in question 22 ab ove?

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

(15) The process of constructing a mathematical model or function that can be used to predi ct or determine one variable by another variable is called

a) regression
b) correlation
c) residual
d) outlier plot

(16) Which of the following is an assumption of one-way ANOVA comparing samples from three or more experimental treatments?

a) the samples associated with each population are randomly selected and are independent fr om all other samples
b) the response variable within each of the k po pulations have equal variances
c) all the response variables within the k popula tions follow a normal distributions
d) all of these

(17) Assume that there is no overlap between the box and whisker plots for three drug treat ments where each drug was administered to 35 individuals. The box plots for these dat a:

a) represent evidence against the null hypothesi s 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

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

a) describe those groups that have reliable diffe
b) set the critical value for the F test (or chi-squ

rences between group means

c) determine if any statistically significant grou p differences have occurred

are)

d) none of these

(19) 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

(20) 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

(21) The probability of type 1 error is referred as?

a) 1-α

b) β

c) A

d) 1-β

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

a) that there are a greater number of explained vs. unexplained observations

b) that the statistical model fits the data well

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

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

(23) Logistic regression assumes a

a) linear relationship between continuous predi ctor variables and the outcome variable

b) linear relationship between continuous predi ctor variables and the logic of the outcome v ariable

c) linear relationship between continuous predi ctor variables

d) linear relationship between observations

(24) In supervised learning, class labels of the training samples are

a) known

b) unknown

c) does not matter

d) partially known

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

a) EM

b) PCA

c) ML

d) any of these

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

a) gamboost

b) gbm

c) ada

d) All of these

(27) Which of the following is the advantage/s of decision trees?

a) possible scenarios can be added

b) use a white box model, if given result is prov ided by a model

c) use a white box model, if given result is prov ided by a model

d) all of the mentioned

(28) How can you prevent a clustering algorithm from getting stuck in bad local optima?

a) set the same seed value for each run

b) use multiple random initialization

c) both set the same seed value for each run an d use multiple random initialization

d) none of these

(29) 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

(30) Statement 1: It is possible to train a network well by initializing all the weights as 0 Sta tement 2: It is possible to train a network well by initializing biases as 0 Which of the s tatements given above is true?

a) statement 1 is true while statement 2 is false
b) statement 2 is true while statement 1 is false
c) both statement 1 is true while statement 2 is false & statement 2 is true while statement 1 is false are true
d) both statement 1 is true while statement 2 is false & statement 2 is true while statement 1 is false are false

(31) Which of the following would have a constant input in each epoch of training a deep le arning 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

(32) What is stability plasticity dilemma?

a) system can neither be stable nor plastic
b) static inputs and categorization can't be hand led
c) dynamic inputs and categorization can't be h andled
d) none of the mentioned

(33) What is true regarding back propagation rule?

a) it is a feedback neural network
b) actual output is determined by computing th e outputs of units for each hidden layer
c) hidden layer's output is not all important, the y are only meant for supporting input and ou tput layers
d) none of the mentioned

(34) Correlation learning law can be represented by equation?

a) $\Delta wij = \mu(si)\, aj$
b) $\Delta wij = \mu(bi - si)\, aj$
c) $\Delta wij = \mu(bi - si)\, aj\, Á(xi)$, where Á(xi) is deri vative of xi
d) $\Delta wij = \mu\, bi\, aj$

(35) 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

(36) 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

(37) In random forest you can generate hundreds of trees (say T1, T2 ....Tn) and then aggreg ate 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 t ree is built on full set of observations

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

(38) Which of the following algorithm would you take into the consideration in your final m odel building on the basis of performance? Suppose you have given the following grap h which shows the ROC curve for two different classification algorithms such as rando m forest (Red) and logistic regression (Blue)

a) random forest
b) logistic regression
c) both random forest & logistic regression
d) none of these

(39) In random forest or gradient boosting algorithms, features can be of any type. For exam ple, it can be a continuous feature or a categorical feature. Which of the following opti on 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

(40) The cell body of neuron can be analogous to what mathematical operation?
a) summing
b) differentiator
c) integrator
d) none of the mentioned

(41) 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 layer
d) none of the mentioned

(42) 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

(43) 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

(44) A windows failover cluster can support up to _____ nodes
a) 12
b) 14
c) 16
d) 18

(45) 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

(46) Point out the wrong statement
a) the system configuration checker will verify the system state of your computer before set up 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

(47) Which of the following model model include a backwards elimination feature selection routine?
a) MCV
b) MARS
c) MCRS
d) all of the mentioned

(48) Point out the correct statement
a) all z nodes are ephemeral, which means they are describing a "temporary" state
b) hbase/replication/state contains the list of Region Servers in the main cluster
c) offline snapshots are coordinated by the Master using zoo keeper to communicate with the Region servers using a two-phase-commit-like transaction
d) none of the mentioned

(49) _____has a design policy of using zoo keeper only for transient data
a) hive
b) imphala
c) hbase
d) oozie

(50) According to analysts, for what can traditional IT systems provide a foundation when t hey'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

(51) What are the five V's of big data?

a) volume
b) velocity
c) variety
d) all of these

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

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

(53) Facebook tackles big data with _____ based on Hadoop

a) projectprism
b) prism
c) projectdata
d) projectbid

(54) As companies move past the experimental phase with Hadoop, many cite the need for a dditional capabilities, including _____

a) improved data storage and information retrie val
b) improved extract, transform and load feature s for data integration
c) improved data warehousing functionality
d) improved security, workload management an d SQL support

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

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

(56) 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

(57) 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

(58) The important 3vs big data are

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

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

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

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

a) so that attribution of authorship can be decid ed
b) to better coordinate the selection of appropri ate sampling methods and data collection ins truments
c) so that conflicts of interest could be identifie d
d) how data will be archived can be planned

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

a) casual
b) predictive
c) mechanistic
d) none of these

(62) Which of the following focuses on the discovery of (previous) unknown properties of t he data?

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

(63) The model which consists of management philosophy, behavioral tools and statistical methods as key steps towards improvement is considered as

a) serial improvement process model
b) behavioral improvement process model
c) quality improvement process model
d) statistics improvement process model

(64) The analysis based on study of price fluctuations, production of commodities and depo sits in banks is classified as

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

(65) What is one of the benefits of active learning?

a) students learn by listening to the teacher
b) students are interested and motivated to be p assive listeners
c) students are interested and motivated to parti cipate in the learning process
d) students pay attention to the teacher and foll ow instructions

(66) 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

(67) Which of the following involves predicting a categorical response?

a) regression
b) summarization
c) clustering
d) classification

(68) _____ is a JavaScript charting library and feature-rich API set that lets you build int eractive flash or HTML5 charts.

a) instantatlas
b) alterian
c) zingchart
d) paleots

(69) Analysis of variance in short form is

a) ANOV
b) AVA
c) ANVA
d) ANOVA

(70) Which of the following metrics can be used for evaluating regression models? i) R Squ ared ii) Adjusted R Squared iii) F Statistics iv) RMSE / MSE / MAE

a) ii and iv
b) i and ii
c) ii, iii and iv
d) i, ii, iii and iv