BRAINWARE UNIVERSITY

Term End Examination 2021 - 22
Programme – Master of Computer Applications
Course Name – Machine Learning
Course Code - MCA401A
(Semester IV)

Time allotted : 1 Hrs.15 Min. **Full Marks : 60**

[The figure in the margin indicates full marks.]

Group-A

(Multiple Choice Type Question) 1 x 60=60

Choose the correct alternative from the following :

(1) Support Vector machine is a

a) Clustering algorithm
b) Feature Selection Algorithm
c) Classification algorithm
d) None of these

(2) Y=mx+c . Here m is

a) Y intercept
b) X intercept
c) Slope of line
d) None of these

(3) Which data is used to build a data mining model?

a) Validation data
b) Testing Data
c) Training data
d) None of these

(4) What is Machine learning?

a) The autonomous acquisition of knowledge through the use of computer programs
b) The autonomous acquisition of knowledge through the use of manual programs
c) The selective acquisition of knowledge through the use of computer programs
d) The selective acquisition of knowledge through the use of manual programs

(5) Let's say, you are using activation function X in hidden layers of neural network. At a particular neuron for any given input, you get the output as "-0.0001". Which of the following activation function could X represent?

a) ReLU
b) tanh
c) SIGMOID
d) None of these

(6) Suppose you want to project high dimensional data into lower dimensions. The two most famous dimensionality reduction algorithms used here are PCA and t-SNE. Let's say you have applied both algorithms respectively on data "X" and you got the datasets "X_projected_PCA" , "X_projected_tSNE". Which of the following statements is true for "X_projected_PCA" & "X_projected_tSNE"?

a) X_projected_PCA will have interpretation in the nearest neighbour space.
b) X_projected_tSNE will have interpretation in the nearest neighbour space.
c) Both will have interpretation in the nearest neighbour space
d) None of them will have interpretation in the nearest neighbour space.

(7) Imagine, you are solving a classification problem with highly imbalanced class. The majority class is observed 99% of times in the training data. Your model has 99% accuracy after taking the

predictions on test data. Which of the following is true in such a case?1. Accuracy metric is not a good idea for imbalanced class problems.2. Accuracy metric is a good idea for imbalanced class problems.3. Precision and recall metrics are good for imbalanced class problems.4. Precision and recall metrics aren't good for imbalanced class problems.

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

(8) Which of the following is a good test dataset characteristic?

a) Large enough to yield meaningful results
b) Is representative of the dataset as a whole
c) Both Large enough to yield meaningful resultsand Is representative of the dataset as a whole
d) None of these

(9) A multiple regression model has

a) Only one independent variable
b) More than one dependent variable
c) More than one independent variable
d) None of these

(10) A nearest neighborhood approach is best used

a) With large size data set
b) When irrelevant attributes are removed from data
c) When a generalized model of data is desirable
d) When an explanation of what has been found is of primary importance

(11) To find the minimum or the maximum of a function, we set the gradient to zero because:

a) The value of the gradient at extrema of a function is always zero
b) Depends on the type of problem
c) Both The value of the gradient at extrema of a function is always zero and Depends on the type of problem
d) None of these

(12) Which of the following is a disadvantage of decision trees?

a) Factor analysis
b) Decision trees are robust to outliers
c) Decision trees are prone to be overfit
d) None of these

(13) Logistics Regression is a

a) Ternary Classifier
b) Binary Classifier
c) MultiValued Classified
d) None of these

(14) Which of the following can be used to impute data sets based only on information in the training set?

a) postProcess
b) preProcess
c) process
d) All of the Mentioned

(15) Point out the correct statement:

a) Asymptotics are used for inference usually
b) caret includes several functions to pre-process the predictor data
c) The function dummyVars can be used to generate a complete set of dummy variables from one or more factors
d) All of the Mentioned

(16) A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

a) Feature F1 is an example of nominal variable.
b) Feature F1 is an example of ordinal variable.
c) It doesn't belong to any of the above category.
d) All of these

(17) 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

(18) 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) Our estimate for P(y=1 | x)
d) Our estimate for P(y=0 | x)

(19) Which of the following sentence is FALSE regarding regression?

a) It relates inputs to outputs.
b) It is used for prediction.
c) It may be used for interpretation.
d) It discovers causal relationships

(20) What is the dimensionality of the null space of the following matrix? 1 1 1 A= 1 1 1 1 1 1

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

(21) The advantage of Grid search is (are),

a) It can be applied to non-di_erentiable functions.
b) It can be applied to non-continuous functions.
c) It is easy to implement.
d) All of these

(22) Consider a linear-regression model with N = 3 and D = 1 with input-ouput pairs as follows: y1 = 22, x1 = 1, y2 = 3, x2 = 1, y3 = 3, x3 = 2. What is the gradient of mean-square error (MSE) with respect to _1 when _0 = 0 and _1 = 1?

a) 2.99
b) 1.92
c) 1.66
d) None of these

(23) Computational complexity of Gradient descent is,

a) linear in D
b) linear in N
c) polynomial in D
d) dependent on the number of iterations

(24) What do you mean by generalization error in terms of the SVM?

a) How far the hyperplane is from the support vectors
b) How accurately the SVM can predict outcomes for unseen data
c) The threshold amount of error in an SVM
d) None of these

(25) What do you mean by a hard margin?

a) The SVM allows very low error in classification
b) The SVM allows high amount of error in classification
c) All of these
d) None of these

(26) The effectiveness of an SVM depends upon:

a) Selection of Kernel
b) Kernel Parameters
c) Soft Margin Parameter C
d) All of these

(27) Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?

a) The model would consider even far away points from hyperplane for modeling
b) The model would consider only the points close to the hyperplane for modeling
c) The model would not be affected by distance of points from hyperplane for modeling
d) None of these

(28) Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of it's hyper parameter. Based upon that give the answer for following question. What would happen when you use very large value of C(C->infinity)? Note: For small C was also classifying all data points correctly

a) We can still classify data correctly for given setting of hyper parameter C
b) We can't classify data correctly for given setting of hyper parameter C
c) Can't Say
d) None of these

(29) Which of the following are real world applications of the SVM?

a) Text and Hypertext Categorization
b) Image Classification
c) Clustering of News Articles
d) All of these

(30) Suppose you gave the correct answer in previous question. What do you think that is actually happening? 1. We are lowering the bias 2. We are lowering the variance 3. We are increasing the bias 4. We are increasing the variance

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

(31) We usually use feature normalization before using the Gaussian kernel in SVM. What is true about

feature normalization? 1. We do feature normalization so that new feature will dominate other 2. Sometimes, feature normalization is not feasible in case of categorical variables 3. Feature normalization always helps when we use Gaussian kernel in SVM

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

(32) Suppose you are dealing with 4 class classification problem and you want to train a SVM model on the data for that you are using One-vs-all method. Suppose you have same distribution of classes in the data. Now, say for training 1 time in one vs all setting the SVM is taking 10 second. How many seconds would it require to train one-vs-all method end to end?

a) 20
b) 40
c) 60
d) 80

(33) What is/are true about kernel in SVM? 1. Kernel function map low dimensional data to high dimensional space 2. It's a similarity function

a) 1
b) 2
c) 1 and 2
d) None of these

(34) Which of the following is/are true about boosting trees? 1. In boosting trees, individual weak learners are independent of each other 2. It is the method for improving the performance by aggregating the results of weak learners

a) 1
b) 2
c) 1 and 2
d) None of these

(35) 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 4. 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

(36) Which of the following algorithm doesn't uses learning Rate as of one of its hyperparameter? 1. Gradient Boosting 2. Extra Trees 3. AdaBoost 4. Random Forest

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

(37) Which of the following is true about the Gradient Boosting trees? 1. In each stage, introduce a new regression tree to compensate the shortcomings of existing model 2. We can use gradient decent method for minimize the loss function

a) 1
b) 2
c) 1 and 2
d) None of these

(38) A feed-forward neural network is said to be fully connected when

a) all nodes are connected to each other
b) all nodes at the same layer are connected to each other
c) all nodes at one layer are connected to all nodes in the next higher layer
d) all hidden layer nodes are connected to all output layer nodes

(39) To calculate the Median

a) Middle value of samples
b) Arrange the samples in ascending order
c) Calculate middle position
d) All of these

(40) The range is

a) Highest value-Lowest Value
b) Lowest Value- Highest value
c) Mean Value- Highest value
d) None of these

(41) Standard deviation is the

a) Square of the variance
b) Cube of the variance
c) Square root of the variance
d) None of these

(42) Chebysheff's theorem deals with

a) Range
b) Variance

c) Standard deviation
d) None of these

(43) Adding a non-important feature to a linear regression model may result in. 1. Increase in R-square 2. Decrease in R-square

a) Only 1 is correct
b) Only 2 is correct
c) Either 1 or 2
d) None of these

(44) Which of the following option is true for overall execution time for 5-fold cross validation with 10 different values of "max_depth"?

a) Less than 100 seconds
b) 100 – 300 seconds
c) 300 – 600 seconds
d) More than or equal to 600 seconds

(45) Which of the following options can be used to get global minima in k-Means Algorithm? 1. Try to run algorithm for different centroid initialization 2. Adjust number of iterations 3. Find out the optimal number of clusters

a) 2 and 3
b) 1 and 3
c) 1 and 2
d) All of these

(46) How does generalization performance change with increasing size of training set?

a) Improves
b) Deteriorates
c) No Change
d) None

(47) Let u be a $n \times 1$ vector, such that $u^T u = 1$. Let I be the $n \times n$ identity matrix. The $n \times n$ matrix A is given by $(I - kuu^T)$, where k is a real constant. u itself is an eigenvector of A, with eigenvalue -1. What is the value of k?

a) -2
b) -1
c) 2
d) 0

(48) Which of the following constitute Type I errors?

a) the null hypothesis is rejected when it is true.
b) the null hypothesis is accepted when it is false.
c) the null hypothesis is accepted when it is true.
d) the alternate hypothesis is accepted when it is true.

(49) For which of the following problems would anomaly detection be a suitable algorithm?

a) From a large set of primary care patient records, identify individuals who might have unusual health conditions.
b) Given data from credit card transactions, classify each transaction according to type of purchase (for example: food, transportation, clothing).
c) Given an image of a face, determine whether or not it is the face of a particular famous individual.
d) From a large set of hospital patient records, predict which patients have a particular disease (say, the flu).

(50) 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 model performs outside the sample on test data
d) None of these

(51) Which of the following statements about regularization is not correct?

a) Using too large a value of lambda can cause your hypothesis to underfit the data.
b) Using too large a value of lambda can cause your hypothesis to overfit the data.
c) Using a very large value of lambda cannot hurt the performance of your hypothesis.
d) None of these

(52) If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?

a) Underfitting
b) Nothing, the model is perfect
c) Overfitting
d) None of these

(53) Confusion matrix is used for

a) Predicting values
b) Assesses a model
c) Summarizing of prediction results on a classification problem.
d) None of these

(54) Accuracy formula for Confusion Matrix

a) (TN)/(TP+FN)
b) (TP+TN)/(TP+TN+FP+FN)
c) (TP+TN)/(TP+FN)
d) None of these

(55) High recall, low precision in Confusion Matrix defines

a) That we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive
b) That we miss a lot of negative examples (high FN) but those we predict as positive are indeed negative
c) Most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.
d) None of these

(56) Which is method of cross validation?

a) K Fold
b) Precision
c) Recall
d) None of these

(57) LOOCV is

a) Leave out one cross-validation
b) Leave out one cross-validation
c) Leave one out cross-validation
d) None of these

(58) Bootstrap Method is

a) method of cross validation
b) method of validation
c) classifier performance measure
d) None of these

(59) 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

(60) Which of the following is an example of a deterministic algorithm?

a) PCA
b) PDF
c) DCA
d) DCA