BRAINWARE UNIVERSITY

Term End Examination 2020 - 21

Programme – Master of Computer Applications

Course Name – Machine Learning

Course Code - MCA501B

Semester / Year - Semester V

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) The SVM's are less effective when:

a) The data is linearly separable

b) The data is clean and ready to use

c) The data is noisy and contains overlapping points

d) None of these

(ii) 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

(iii) The cost parameter in the SVM means:

a) The number of cross-validations to be made

b) The kernel to be used

c) The tradeoff between misclassification and simplicity of the model

d) None of these

(iv) 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

(v) 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

(vi) Which of the following option would you more likely to consider iterating SVM next time?

a) You want to increase your data points

b) You want to decrease your data points

c) You will try to calculate more variables

d) You will try to reduce the features

(vii) 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

(viii) In above question suppose you want to change one of it's(SVM) hyperparameter so that effect would be same as previous questions i.e model will not under fit?

a) We will increase the parameter C

b) We will decrease the parameter C

c) Changing in C don't affect

d) None of these

(ix) 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

(x) 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. How many times we need to train our SVM model in such case?

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

(xi) 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

(xii) Suppose you are dealing with 2 class classification problem and you want to train a SVM model on the data for that you are using One-vs-all method. What would you think how many times we need to train SVM in such case?

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

(xiii) 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

(xiv) 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

(xv) Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods? 1. Both methods can be used for classification task 2. Random Forest is use for classification whereas Gradient Boosting is use for regression task 3. Random Forest is use for regression whereas Gradient Boosting is use for Classification task 4. Both methods can be used for regression task

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

(xvi) 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

(xvii) 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

(xviii) 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

(xix) 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

(xx) The average squared difference between classifier predicted output and actual output.

a) mean squared error

b) root mean squared error

c) mean absolute error

d) mean relative error

(xxi) 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

(xxii) 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 initializations

c) Both Set the same seed value for each runand Use multiple random initializations

d) None of these

(xxiii) To calculate the Median

a) Middle value of samples

b) Arrange the samples in ascending order

c) Calculate middle position

d) All of these

(xxiv) The measures of dispersion include

a) Range
b) Standard Deviation
c) Variance
d) None of these

(xxv) The range is

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

(xxvi) Variance is

a) Sample mean of the squared deviations from the arithmetic mean
b) Arithmetic mean of the squared deviations from the sample mean
c) Sample mean of the squared deviations from the sample mean
d) None of these

(xxvii) Standard deviation is the

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

(xxviii) Coefficient of the correlation ranges from

a) -1 to +1
b) 0 to +1
c) -1 to 0
d) None of these

(xxix) Chebysheff's theorem deals with

a) Range
b) Variance
c) Standard deviation
d) None of these

(xxx) Conditional probability is related with

a) Naïve Bayes theorem
b) Chebysheff's Theorem
c) Pythagorian theorem
d) None of these

(xxxi) 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

(xxxii) Which of the following options is/are true for K-fold cross-validation? 1. Increase in K will result in higher time required to cross validate the result. 2. Higher values of K will result in higher confidence on the cross-validation result as compared to lower value of K. 3. If K=N, then it is called Leave one out cross validation, where N is the number of observations.

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

(xxxiii) 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

(xxxiv) What would you do in PCA to get the same projection as SVD?

a) Transform data to zero mean
b) Transform data to zero median
c) Not possible
d) None of these

(xxxv) 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

(xxxvi) For which of the following hyperparameters, higher value is better for decision tree algorithm? 1. Number of samples used for split 2. Depth of tree 3. Samples for leaf

a) 1 and 2
b) 2 and 3
c) 1 and 3
d) Can't say

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

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

(xxxviii) One of the most common uses of Machine Learning today is in the domain of Robotics. Robotic tasks include a multitude of ML methods tailored towards navigation, robotic control and a number of other tasks. Robotic control includes controlling the actuators available to the robotic system. An example of this is control of a painting arm in automotive industries. The robotic arm must be able to paint every corner in the automotive parts while minimizing the quantity of paint wasted in the process. Which of the following learning paradigms would you select for training such a robotic arm?

a) Supervised learning
b) Unsupervised learning
c) Combination of supervised and unsupervised learning
d) Reinforcement learning

(xxxix) Let u be a n × 1 vector, such that u T u = 1. Let I be the n × n identity matrix. The n × n matrix A is given by (I ? kuuT), 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

(xl) Consider the following 4 training examples • x = ?1, y = 0.0319 • x = 0, y = 0.8692 • x = 1, y = 1.9566 • x = 2, y = 3.0343 We want to learn a function f(x) = ax+b which is parametrized by (a, b). Using squared error as the loss function, which of the following parameters would you use to model this function.

a) (1, 1)
b) (1, 2)
c) (2, 1)
d) (2, 2)

(xli) Which of the following are recommended applications of PCA?

a) To get more features to feed into a
b) Data compression: Reduce the

learning algorithm.

dimension of your data, so that it takes up less memory / disk space.

c) Preventing overfitting: Reduce the number of features (in a supervised learning problem), so that there are fewer parameters to learn.

d) As a replacement for (or alternative to) linear regression: For most learning applications, PCA and linear regression give substantially similar results.

(xlii) 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).

(xliii) Suppose you have trained an anomaly detection system for fraud detection, and your system that flags anomalies when p(x) is less than ?, and you find on the cross-validation set that it is missing many fraudulent transactions (i.e., failing to flag them as anomalies). What should you do?

a) Increase ?

b) Decrease ?

c) Both Increase ? and Decrease ?

d) None of these

(xliv) 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

(xlv) Which of the following is a categorical outcome?

a) RMSE
b) RSquared
c) Accuracy
d) All of these

(xlvi) 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

(xlvii) K-fold cross-validation is

a) linear in K
b) quadratic in K
c) cubic in K
d) exponential in K

(xlviii) 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

(xlix) Suppose you are using SVM with linear kernel of polynomial degree 2. think that you increase the complexity (or degree of polynomial of this kernel). What would you think will happen?

a) Increasing the complexity will overfit the data
b) Increasing the complexity will underfit the data
c) Nothing will happen since your model was already 100% accurate
d) None of these

(l) K-Means clustering algorithm is example of which model?

a) Connectivity models
b) Centroid models
c) Distribution models
d) Density Models

(li) Expectation-maximization algorithm is example of which model?

a) Connectivity models
b) Centroid models
c) Distribution models
d) Density Models

(lii) DBSCAN and OPTICS are example of which model?

a) Connectivity models
b) Centroid models
c) Distribution models
d) Density Models

(liii) In K Means Clustering algorithm, K denotes

a) Number of associations
b) Number of regressions
c) Number of clusters
d) None of these

(liv) Hierarchical Clustering algorithm terminates when _____ .

a) there is only a single cluster left.
b) two nearest clusters are merged into the same cluster.
c) all the data points assigned to a cluster of their own
d) None of these.

(lv) K means is

a) Clustering algorithm
b) Classification algorithm.
c) Association algorithm
d) None of these.

(lvi) The time complexity of K Means is

a) linear
b) quadratic
c) cubic
d) none of these

(lvii) The time complexity of hierarchical clustering is

a) linear
b) quadratic
c) cubic
d) none of these

(lviii) Which metrics are used for deciding the closeness of two clusters?

a) Euclidean distance
b) Manhattan distance

c) Maximum distance
d) All of these

(lix) Components of machine learning are

a) Representation
b) Evaluation
c) Optimization
d) All of these

(lx) The sample of data used to fit the model is called _____.

a) Training dataset
b) Validation dataset
c) Test dataset
d) None of these

(lxi) Suppose that a psychologist wants to evaluate the effectiveness of a new learning strategy. She randomly assigns students to two groups and assigns each student the same passage on a particular topic to study for half an hour. Subsequently each student participates in an individual assessment on the topic, where students of the one group use the new learning strategy, and students of the other group use any strategy they prefer. Which among the following is an extraneous variable in the above experiment?

a) The choice of using two groups
b) The amount of time given to study the passage
c) Existing knowledge about the passage among the students
d) The amount of time given to complete the assessment

(lxii) I have trained a classifier, and to evaluate its performance I perform a 10-fold validation. I have obtained the following accuracies on the validation set in each of the runs - 0.90, 0.98, 0.95, 0.98, 0.97, 0.96, 0.94, 0.99, 0.96, 0.96. What is the sample standard deviation for the accuracies?

a) 0.0243
b) 0.0256
c) 0.0006544399999999999
d) 0.000589

(lxiii) A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather.

In this setting, what is E?

a) The weather prediction task.

b) The probability of it correctly predicting a future date's weather

c) The process of the algorithm examining a large amount of historical weather data

d) None of these.

(lxiv) Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? (Select all that apply.) In each case, assume some appropriate dataset is available for your algorithm to learn from.

a) Given 50 articles written by male authors, and 50 articles written by female authors, learn to predict the gender of a new manuscript's author (when the identity of this author is unknown).

b) Examine a large collection of emails that are known to be spam email, to discover if there are sub-types of spam mail

c) Given data on how 1000 medical patients respond to an experimental drug (such as effectiveness of the treatment, side effects, etc.), discover whether there are different categories or "types" of patients in terms of how they respond to the drug, and if so what these categories are

d) All of these

(lxv) Suppose that for some linear regression problem (say, predicting housing prices as in the lecture), we have some training set, and for our training set we managed to find some ?0, ?1 such that J(?0,?1)=0. Which of the statements below must then be true?

a) For this to be true, we must have ?0=0 and ?1=0 so that h?(x)=0

b) Our training set can be fit perfectly by a straight line, i.e., all of our training examples lie perfectly on some straight line

c) For this to be true, we must have y(i)=0 for every value of i=1,2,…,m.

d) We can perfectly predict the value of y even for new examples that we have not yet seen.

(lxvi) You run gradient descent for 15 iterations with ?=0.3 and compute J(?) after each iteration. You find that the value of J(?) decreases slowly and is still decreasing after 15 iterations. Based on this, which of the following conclusions seems most plausible?

a) Rather than use the current value of ?, it'd be more promising to try a larger value of ? (say ?=1.0).

b) Rather than use the current value of ?, it'd be more promising to try a smaller value of ? (say ?=0.1).

c) ?=0.3 is an effective choice of learning rate.

d) None of these

(lxvii) Which of the following are reasons for using feature scaling?

a) It speeds up gradient descent by making it require fewer iterations to get to a good solution

b) It is necessary to prevent gradient descent from getting stuck in local optima

c) It speeds up gradient descent by making each iteration of gradient descent less expensive to compute.

d) It prevents the matrix XTX (used in the normal equation) from being non-invertable (singular/degenerate)

(lxviii) Suppose Theta1 is a 5x3 matrix, and Theta2 is a 4x6 matrix. You set thetaVec = [Theta1(:), Theta2(:)]. Which of the following correctly recovers?

a) reshape(thetaVec(16 : 39), 4, 6)

b) reshape(thetaVec(15 : 38), 4, 6)

c) reshape(thetaVec(16 : 24), 4, 6)

d) reshape(thetaVec(15 : 39), 4, 6)

(lxix) Suppose you have a dataset with n = 10 features and m = 5000 examples. After training your logistic regression classifier with gradient descent, you find that it has underfit the training set and does not achieve the desired performance on the training or cross validation sets. Which of the following might be promising steps to take?

a) Increase the regularization parameter ?.

b) Use an SVM with a Gaussian Kernel.

c) Use an SVM with a linear kernel, without introducing new features.

d) Reduce the number of examples in the training set

(lxx) Suppose you have implemented regularized logistic regression to classify

what object is in an image (i.e., to do object recognition). However, when you test your hypothesis on a new set of images, you find that it makes unacceptably large errors with its predictions on the new images. However, your hypothesis performs well (has low error) on the training set. Which of the following are promising steps to take?

a) Try adding polynomial features.

b) Use fewer training examples.

c) Try using a smaller set of features.

d) Try evaluating the hypothesis on a cross validation set rather than the test set.