Machine Learning &Artificial Intelligence

End Semester Exam

Time: 3 hours

Max marks: 100

1. Please be precise and brief in your answers.

2. Please answer all parts of a question together do not scatter the parts of a question all over the answer script.

4. You can consult only your own handwritten notes. Photocopies, printouts and any electronic gadgets are not allowed.

Section 1: Objective Type (10\*1=10 Marks)

1.Decision Trees can be used for Classification Tasks.  
a) True  
b) False

2. Choose from the following that are Decision Tree nodes  
a) Decision Nodes  
b) End Nodes  
c) Chance Nodes  
d) All of the mentioned

2.Which of the following is/are true about bagging trees?

1. In bagging trees, individual trees are independent of each other
2. Bagging 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

3.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

4.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

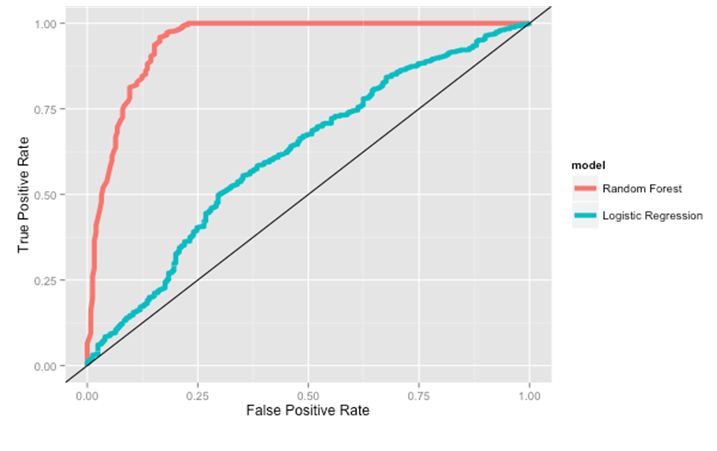
5. Which of the following is true about “max\_depth” hyperparameter in Gradient Boosting?

1. Lower is better parameter in case of same validation accuracy
2. Higher is better parameter in case of same validation accuracy
3. Increase the value of max\_depth may overfit the data
4. Increase the value of max\_depth may underfit the data

A) 1 and 3  
B) 1 and 4  
C) 2 and 3  
D) 2 and 4

6.Which of the following algorithm would you take into the consideration in your final model building on the basis of performance?

Suppose you have given the following graph which shows the ROC curve for two different classification algorithms such as Random Forest(Red) and Logistic Regression(Blue)

[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/07/12181501/7_image_dt.jpg)A) Random Forest  
B) Logistic Regression  
C) Both of the above  
D) None of these

7.In random forest or Logistic Regression 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 Logistic Regression algorithm handles real valued attributes by discretizing them  
C) Both algorithms can handle real valued attributes by discretizing them  
D) None of these

8.True-False: The bagging is suitable for high variance low bias models?

A) TRUE  
B) FALSE

9.Consider the hyperparameter “number of trees” and arrange the options in terms of time taken by each hyperparameter for building the Random Forest Model?

Note: remaining hyperparameters are same

1. Number of trees = 100
2. Number of trees = 500
3. Number of trees = 1000

A) 1~2~3  
B) 1<2<3

C) 1>2>3  
D) None of these

10. [True or False] Cross validation can be used to select the number of iterations in boosting; this procedure may help reduce overfitting.

A) TRUE  
B) FALSE

Answer any Ten (5 Marks Each) 10\*5=50 Marks

1. How is KNN different from k-means clustering?
2. What is deep learning, and how does it contrast with other machine learning algorithms?
3. Which is more important – model accuracy, or model performance? Explain
4. When should you use classification over regression?
5. How do you ensure you’re not overfitting with a model?
6. How would you evaluate a logistic regression model?
7. **What are Recommender Systems?**
8. **What do you understand by the term Normal Distribution?**
9. **What are various steps involved in an analytics project?**
10. **Can you cite some examples where a false negative important than a false positive?**
11. **Can you explain the difference between a Test Set and a Validation Set?**
12. **List any three Machine learning algorithms**

Answer any two (2\*10 =20 Marks)

1. Imagine you have a million tweets from which you need to categorize topics. What are the steps you would perform to extract the top 10 topics from the tweets?
2. Imagine you work as part of a service desk –Call center and the key target is to increase the Closure rate of tickets? What methods would you suggest? You would be provided with 3 years of service desk data
3. A) Can Structured and Unstructured data be used within the same model? If so provide an example and steps to use both the variables within the same model

B) Imagine you have built a model for identifying the probability of a customer to purchase an Insurance Policy. How would you evaluate the model? Highlight the key benefits to business and cost savings by implementing the same.

Answer any one (1\*20=20 Marks)

1. A company has a large number of documents that need to be sorted into one of three categories: Research & Development, Finance or Marketing. They have been able to identify a number of phrases which are commonly used in these documents and may help disambiguate them but there are a large number (thousands) of these phrases and each one only appears in a small number of documents. Thankfully someone has labeled a few hundred documents for you but you need to build a method to automatically label the rest. Highlight the List of Steps to be performed in this case to solve this problem
2. Risk varies from customer to customer and understanding the risk and mitigating it is a main issue across the insurance industry. Imagine that you are working as a data scientist for Insurance major. Many factors contribute to the frequency and severity of car accidents including how, where and under what conditions people drive, as well as what they are driving. Bodily Injury Liability Insurance covers other people’s bodily injury or death for which the insured is responsible. Considering this situation how would you proceed with collecting the required data and suggest a solution using Machine Learning algorithm. Explain in detail about data cleansing, validation and modelling steps.