Machine Learning &Artificial Intelligence

End Semester Exam

Time: 3 hours

Max marks: 100

1. Answer all 5 questions. The question paper has 4 pages.

2. Please be precise and brief in your answers.

3. 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.

Answer all Questions (12\*1=12 Marks)

1.To apply bagging to regression trees which of the following is/are true in such case?

1. We build the N regression with N bootstrap sample
2. We take the average the of N regression tree
3. Each tree has a high variance with low bias

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

2.How to select best hyper parameters in tree based models?

A) Measure performance over training data  
B) Measure performance over validation data  
C) Both of these  
D) None of these

3.**Suppose you have given the following scenario for training and validation error for Gradient Boosting. Which of the following hyper parameter would you choose in such case?**

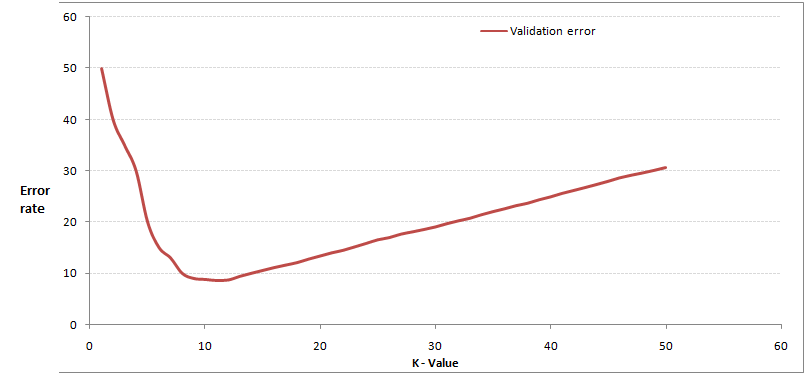
|  |  |  |  |
| --- | --- | --- | --- |
| **Scenario** | **Depth** | **Training Error** | **Validation Error** |
| 1 | 2 | 100 | 110 |
| 2 | 4 | 90 | 105 |
| 3 | 6 | 50 | 100 |
| 4 | 8 | 45 | 105 |
| 5 | 10 | 30 | 150 |

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

4. **[**True or False] k-NN algorithm does more computation on test time rather than train time.

A) TRUE  
B) FALSE

16. In the image below, which would be the best value for k assuming that the algorithm you are using is k-Nearest Neighbor.

[](https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/08/01195427/Pic21.jpg)

A) 3  
B) 10  
C) 20  
D 50

5.Which of the following distance metric cannot be used in k-NN?

A) Manhattan  
B) Minkowski  
C) Tanimoto  
D) Jaccard  
E) Mahalanobis  
F) All can be used

6.Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?

A) K-NN  
B) Linear Regression  
C) Logistic Regression

7.Which of the following is true about Manhattan distance?

A) It can be used for continuous variables  
B) It can be used for categorical variables  
C) It can be used for categorical as well as continuous  
D) None of these

8.Which of the following distance measure do we use in case of categorical variables in k-NN?

1. Hamming Distance
2. Euclidean Distance
3. Manhattan Distance

9.Which of the following will be Euclidean Distance between the two data point A (1,3) and B (2,3)?

A) 1  
B) 2  
C) 4  
D) 8

10.Which of the following will be true about k in k-NN in terms of variance?

A) When you increase the k the variance will increases  
B) When you decrease the k the variance will increases  
C) Can’t say  
D) None of these

11.A company has built a kNN classifier that gets 100% accuracy on training data. When they deployed this model on client side it has been found that the model is not at all accurate. Which of the following thing might gone wrong?

Note: Model has successfully deployed and no technical issues are found at client side except the model performance  
A) It is probably a over fitted model  
B) It is probably a under fitted model  
C) Can’t say  
D) None of these

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

**1) Which of the following statement is true in following case?**

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) Both of these

Answer any Eight (6 Marks Each ) 8\*6=48 Marks

1. **Can you explain the difference between a Test Set and a Validation Set?**
2. **List any three Machine learning algorithms**
3. **Assume you need to generate a predictive model using multiple regression. Explain how you intend to validate this model**
4. **What is a decision tree? Explain the steps in building a decision tree.**
5. **Explain about steps to manage Large datasets using R**
6. **Write a pseudo code for K-means Cluster analysis**
7. **Explain about various hyper parameters in a xg boost model**
8. **Explain difference between hierarchical and K-means Clustering**
9. **Explain difference between Linear and Logistic Regression**
10. **What is difference between R square and Adjusted R square**

Answer any two (2\*20=40 Marks)

1. What are the different types of recommender systems and the key steps in implementing a recommendation system?
2. Explain about the following Evaluation Metrics
   1. ROC Curve
   2. Precision
   3. Recall
   4. Sigmoid Function
   5. Lasso and Ridge Regression
3. Imagine you have been provided with the sales data of an Online Retail. The key task is to reduce the return rate.
   1. Identify the key data points required for this exercise
   2. Highlight the data validation and cleansing steps to be followed
   3. Which model you would use
   4. Highlight model validation and implementation methodology

Return Rate : Total Products Returned Back/Total Products Purchased