81

ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

Student's Name: Sanith Rao

Date of Exam: Monday, April 28, 2025 - Wednesday, April 30, 2025
Time Limit: 90 minutes
Allowed Materials: None (closed book & closed notes)

Student – NC State University Pack Pledge

I, Sanith Rao have neither given nor received unauthorized aid on this exam or assignment. I have read the instructions and acknowledge that this is the correct exam.

Student's Printed Name

Sanith Rao

Student Signature

04/30/2025

Date

Exam must be turned in by: _____

Exam End Time *Student's Initial Agreement*

NOTE: Failure to turn in exam on time may result in penalties at the instructor's discretion.

Final Exam

Please write your answers below each question. You should not have access nor use any materials during this exam.

A reminder that, by taking this exam, you are required to uphold the NC State honor pledge:

"I have neither given nor received unauthorized aid on this test or assignment."

1. We know that a multiple linear regression model fits a (hyper) plane as the response surface (or a curved hyperplane with higher order polynomial or interaction terms). How does a standard regression tree model the response surface?

So once the tree is designed (using the method I described in the second question), for a test observation, we basically make our way through the entire tree until we find the node with no children. That is our destination node and we take the average of all values in the node. ok

2. For a standard regression tree that uses recursive binary splitting, suppose we have two predictors X1 and X2. What criterion is used to determine the first split? Describe how this first split is decided upon. Be specific on both of these!

When we are doing a regression task, the criterion we look at is minimizing the RSS. This is how the final feature and threshold and split are chosen. To decide the split, we consider all features and all possible thresholds and analyze which combination gives us the lowest RSS. That features is then chosen.

0

3. Suppose we have a large data set where we want to perform a regression task. We want to determine the best overall model between a kNN model and a ridge regression model. We want to use a train test split and compare the best kNN and ridge regression model on the test set. We wish to determine the appropriate tuning parameters on the training set only using the bootstrap. Fully outline the process for splitting the data, tuning, comparing, and fitting a final overall best model.

The first to do is to split the data into a testing and training set (20-80 or 30-70 etc). Once we do this, we need to tune the hyperparameter (λ for Ridge & k for KNN). In bootstrapping, we essentially randomly pick data points with replacement until we have the same number of data points as our training. Now obviously some data points are not picked and this becomes part of our test set. Now we create a grid of hyperparameter values (for λ & k). For each hyperparameter value, we run a certain number of bootstraps where each time, we train the model, and test it on the OOB test set (this is different for each bootstrap). We aggregate the performance metric (like RMSE) for each value of hyperparameter and choose the one with the lowest error rate. Now train both models with the chosen hyper-parameter with the original training set. Calculate error rate for both models based on the TEST set. Choose the one with lowest error. fit to entire data -1

4. We discussed two ways to do 'early stopping' in a regression or classification tree. What are those two methods?

we can do early stopping when max tree depth is reached or when we don't have enough data points available.

5. In a standard multilayer feed-forward neural network, what are two common activation functions?

Two activation functions are sigmoid, and log loss. -2

6. What task is a Recurrent neural network well-suited for?

Recurrent Neural Networks are well suited when we are dealing with sequential or time series data.

7. True or False questions (write True or false next to each letter):

   a. Random forest and bagged tree models generally require you to standardize your predictors False

   b. kNN models generally require you to standardize your predictors False

   c. The number of trees we use in a random forest model is important because we can overfit with too many trees. False

   d. When using BART we need to remove the first few prediction models. True

   e. SVM models can only be used in classification tasks. False

   f. KMeans clustering does not necessarily create the same clusters in each run of the algorithm. True

   g. Hierarchical clustering requires you to know the 'true' underlying groupings to use it effectively. False

   h. In a standard multilayer neural network, all inputs are 'connected to' all first level activations. True

   i. KNN provides a discriminant for classifying our observations false

   j. The Naive Bayes provides a discriminant for classifying our observations False

-5

8. Consider the piecewise polynomial regression model. Here we define our knots to be c_1, ..., c_M and use the indicator functions

$h_1(X) = I(c_1 \leq X < c_2), \ldots, h_{M-1} = I(c_{M-1} \leq X < c_M), h_M(X) = I(X > c_M)$

in our regression equation given by

$$Y_i = \beta_0 + h_1(X_i)\beta_1 + \cdots + h_M(X_i)\beta_M + \epsilon_i$$

Suppose we have n observations and we fit the model.

a. What is the estimate of β_0 in this model?

β_0 essentially gives us the value of Y_i when all other indicator functions are 0. This means when we are looking at the region $< c_1$. We take the mean of all observations in this region.

b. What is the estimate of β_1 in the model?

β_1 essentially gives us the slope when we consider a value between $c_1 \leq X < c_2$. Here we aggregate and take the mean of all the data points in $c_1 \leq X < c_2$. ok

9. What are the three most common tuning parameters associated with a boosted tree model?

The most common tuning parameters are the learning rate, and the parameter values for L1 and L2 regularization. -2

-2

10. Why do random forests for a regression task generally improve prediction over the basic bagged tree model?

Random Forest builds on normal bagged tree model by randomly choosing a subset of features (P/3 for example) to consider at every single node of a tree. This reduces the ~~variance~~ and is able to give us better results. −1

11. Describe the algorithm for fitting a basic boosted regression tree model.

A boosted tree is different from bagging in that it tries to minimize the error of the previous tree. We start with a very weak condition initially. Split the data and find the RSS. In every subsequent tree, we look minimize the RSS. This is a sequential tree building process and stops when have the desired number of trees. −1

12. When fitting a support vector machine model for classification, what are support vectors?

Support vectors are essentially used when there is no clear split between classes. Because of this, we can't draw a proper plane. Here support vectors are used to find the best possible planes that will minimize the number of misclassified observations. −1

−3

13. When we wish to apply the SVM model to a classification task with more than two levels, we discussed the one-versus-one approach. Describe how this SVM model works.

Here, we basically compare every pair of classes. And once we have the hyperplane for all combinations of classes, we look at the metric that shows how confident each class was in classifying the test observations. We then aggregate this individually for all classes and it is assigned to the class that had the highest confidence (furthest on average from from hyperplane). -3

14. Why do we often run the kmeans clustering algorithm multiple times?

The way k means clustering works is by randomly choosing k centroids first. And so the results that we get are dependent on the centroids we choose. This is one reason to run it multiple. We also do this to ensure we consider different k values to find the optimal one.

15. When doing hierarchical clustering, how does the 'single' linkage create a dissimilarity measure?

When we do a single linkage, we try to find the two closest points between any two clusters and the clusters that are closest to each other are grouped together. The potential issue from this is chaining.

16. What is a biplot and how can it be useful?

A biplot is basically useful for SVMs where it measures the level of confidence each class had in correctly classifying an observation. It compares two classes and shows which class was more confident -5

-8