86

ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

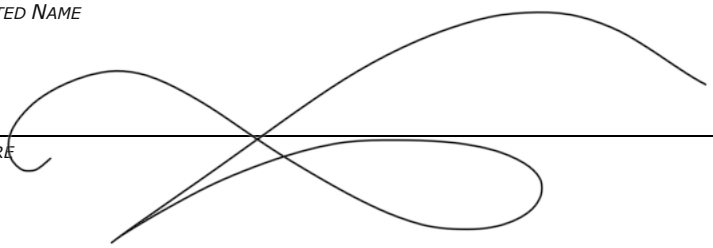
Student's Name: Julia Fish

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, Julia Fish 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



4/30/25

Student Signature *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.

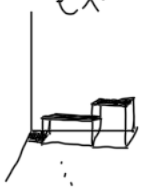
A reminder that, by taking this exam, you are required to uphold the NC State honor pledge:

"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?

A standard regression tree models the response as a stepwise plane. For example, if we have 2 predictors (X_1 & X_2), the response will be a surface covering that 2D plane as well as it being at the "height" that corresponds to its response. This can be generalized (but not visualized, really) to 4 space, 5 space, etc.

ex:



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!

The criterion used to determine this first split is the total squared deviations from the average value on each side of the potential cut. The cut is made where the average values differ enough as well as the squared deviations from that value are (hopefully) as small as they can get.

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.

First, we will split the data into a training set and a test set. Typically, 80-20 splits are used (or 70-30).

We will first train a KNN model. We create a grid of K values to train/tune the model. Then, we train the KNN model using that tuning grid as well as a many bootstrap resample of the training data. The out of bag observations become the "test set" for each KNN fit. The K value with the desired metric value (min RMSE, etc.) is chosen and refit to the entire training set.

Next, we will fit our ridge regression model. We create a grid of penalty values for our β coefficients (to train the model). Then, we train a ridge regression model using bootstrap resamples as well as different values for the penalty term (one per each model fit). The out of bag observations become a "test set" for each model fit. Then, the penalty value that produces the desired metric value (min RMSE, etc.) is chosen. This model is refit to the entire training set using only that penalty value.

Lastly, we predict the values for our test set on the two overall best fit models (the final KNN and final ridge regression models). The model with the "better performance" on the test set (the one with the lower RMSE, etc.) is the overall best fit model. Fit that model with the predetermined tuning parameter value to the full data set.

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

Defining a minimum value of observations allowed in terminal nodes as well as defining a certain amount of splits allowed in the model.

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

Two common activation functions for standard multilayer feed-forward neural networks are convolution and pooling.

-2

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

An RNN model is well suited for classification tasks of text string data (ex: good or bad movie review).

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 F
   b. kNN models generally require you to standardize your predictors T
   c. The number of trees we use in a random forest model is important because we can overfit with too many trees. T
   d. When using BART we need to remove the first few prediction models. T
   e. SVM models can only be used in classification tasks. F
   f. KMeans clustering does not necessarily create the same clusters in each run of the algorithm. T
   g. Hierarchical clustering requires you to know the 'true' underlying groupings to use it effectively. F
   h. In a standard multilayer neural network, all inputs are 'connected to' all first level activations. T
   i. KNN provides a discriminant for classifying our observations F
   j. The Naive Bayes provides a discriminant for classifying our observations F

-3

-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?

Our estimate of β_0 in this model is the constant value that would best fit all values before knot c_1 (since the basis functions are all only indicator values as well as all values besides those before knot c_1 are in a basis function).

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

Our estimate of β_1 in this model is the constant value that would best model the values between c_1 (inclusive) and c_2 (exclusive). That is because the basis function is only an indicator function (no transformation used).

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

B (number of trees fit)

The number of predictors that can be randomly sampled at each cut

λ (learning rate)

-1

-1

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

Random forests generally improve prediction over basic bagged trees because a random sample of predictors is chosen to be considered for each ~~tree~~ split. This allows us to have trees where significant predictors are considered as well as some where they are not (can analyze other predictors). Also, an "average" is taken over all of these trees, resulting in a better fit than one single fit model.

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

A basic boosted regression tree model is fit by a slow learning process. At each split of this tree model, only a random ~~sample~~ of the predictors can be considered (this can be tuned and/or chosen). The process essentially makes a cut, learns from that decision, and proceeds forward.

-4

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

Support vectors are the few observations that hold a lot of weight when fitting an SVM model. These are the observations closest to the margins/classifier or past the classifier. When these values change, the overall fit can change drastically.

-4

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.

The one versus ~~one~~ all approach compares a certain class to all of the remaining classes (as if they were one class). This allows us to act as though we have a binary classification (correct/incorrect) for the model fit. -3

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

We often run the kmeans clustering algorithm multiple times in order to try to ensure we have a good fit (choose the best fit of them all). This is because kmeans fits can be drastically different from one another depending on choice of number of ~~classes~~ for the model fit. -1

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

Single linkage creates a dissimilarity measure by finding the distance between the closest 2 points of different classes.

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

A biplot plots the first two principal components alongside one another (one per axis). This can be useful in order to understand and interpret the principal components (in terms of what variables are being represented and how).

-4