66

ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

Student's Name: Zach Ginder

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, Zach Ginder 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

Student Signature 04/30/2025 *Date*

Exam must be turned in by: 3:00pm *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.

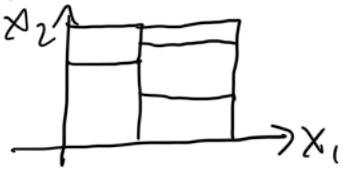
A reminder that, by taking this exam, you are required to uphold the NC State honor pledge:

"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 surface as regions. These regions are rectangles or in a multi predictor instance are 3D surfaces

graphs for visualization:



-1 Plot

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 to determine the first split is the split at which the sum of squared errors is minimized. The issue with this method is that it is the best split at that node however it may not be the best split overall because this method is not forward looking. To determine the best split it is evaluated at each predictor (X_1, X_2) at each point and comparing SSE. ✓

-1

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 determine the percentage of observations in our training and test set (say 80% training and 20% test). We split our data into these two datasets. The tuning parameters for KNN is k (the number of neighbors considered) and for ridge regression the tuning parameter is λ (the shrinkage tuning parameter). For bootstrapping we sample with replacement from our training data set a sample equivalent to the sample size of our training data. We construct a tuning grid of our tuning parameters k and λ. Using the bootstrap sample we construct KNN and ridge regression models for each value of k and λ respectively in the tuning grid. We test each model on the "out of bag" observations, those observations in our training dataset that were not included in the bootstrap sample. The tuning parameters k and λ that result in the models with the smallest test error using the out of bag observations are then chosen as the optimal KNN and ridge regression models respectively. Both these models are then retrained to the entire training dataset. To determine an overall best model each retrained KNN and ridge regression is evaluated using the test set (which has remained untouched up until this point). The model with the lowest test error (normally lowest RMSE) is considered to be the overall best model. and is refit to the entire dataset

done repeatedly

−1

−2

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

Setting a minimum number of observations allowed in each terminal node or setting the number of splits allowed in a given tree.

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

Forward activation

bidirectional activation

-2

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

Image identification/classification

-3

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 TRUE

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

   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 ~~TRUE~~

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

-9

-14

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?

B_0 is the value of the model for all values below the value of C_1. This corresponds to a straight line where $Y_i = B_0$ and is also therefore the intercept

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

B_1 is the additional value added to B_0 when considering values $C_1 \leq X < C_2$ so within knots C_1 and C_2 (not including C_2). So for values $C_1 \leq X < C_2$ the equation is $Y_i = B_0 + B_1$

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

- pruning parameter
- # of predictors available at each node to determine the optimal split
- minimum # of observations allowed in each terminal node

-2

-2

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

Random forests aggregate multiple trees together to attempt to account for the inefficient method that occurs with regression trees where the optimal split at each node is not necessarily the optimal split for the whole model. We then see the subsequent importance of each predictor based on how often it was utilized among all of the trees which attempts to help reduce overfitting, which pruning also helps account for.

-4

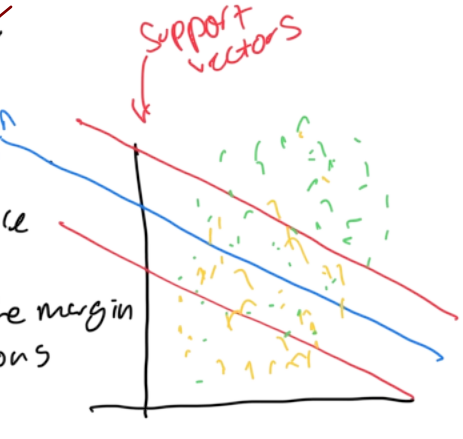
11. Describe the algorithm for fitting a basic boosted regression tree model.

At each node of the tree a subset of predictors are randomly available for splitting (the # of predictors for this is a tuning parameter) with the best split occurring with the optimal SSE.

-5

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

Support vectors are additional boundaries added to the margin classifier is SVM models. Many times they are utilized when creating slack variables. A slack variable of 0 means an obs is correctly classified and on the correct side of the margin. A slack variable of between 0-1 indicates a obs is within the support vector but on the correct side of the margin and a value of 1 means the obs is on the wrong side of the margin.



-9

#12 (continued)

Since margins rely on a small number of observations they can be highly variable and so the support vectors provide an additional boundary to the margin's classifications especially when combined with slack variables.

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.

An SVM model attempts to construct a hyperplane to seperate the data into classes based on p predictors. In a one-versus-one approach, each classifyer is compared to a subsequent classifier for each observation (unlike one-versus-all when the other classifiers are pooled). The most probabale classifier for each observation is then that observations class.

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

In K means clustering the algorithm does not necessarily produce the same clusters in each run of the algorithm therefore no single run is necessarily the "optimal" run.

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

Single linkage involves comparing the dissimilarity between the largest dissimilar values in a cluster. This dissimilarity value is calculated for each cluster and we attempt to minimize this value.

-2

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

A biplot is an output used during unsupervised learning that helps show the RSS values associated with each additional principal component when using PCA. It helps show the "elbow effect" in which additional principal components are no longer significantly useful in contributing to the model.

-4

-6