DELTA Testing Services

go.ncsu.edu/testing

Campus Box 7555
1730 Varsity Dr.
Venture IV, Suite 236
Raleigh, NC 27695-7113

919.515.1560 phone
919.515.7180 fax

delta-testing@ncsu.edu

NC STATE UNIVERSITY

DELTA Testing Services

Student Name: David Grant **Date:** 4/28/25

Student's NCSU Email Address: dgrant@ncsu.edu

Course: ST 563 601 **Exam #:** Final Exam

Start Time: 10:09 am **End Time:** 11:09

Proctor's Name (Print): Giana De Rosa

Proctor's Signature: Giana De Rosa

Institution: Ramapo College

PLEASE SIGN & DATE THIS SHEET AND RETURN ALONG WITH THE EXAM

Proctoring Guidelines

If you are unable to comply with the following, please destroy the exam and have the student submit the name of another proctor for approval.

1. Please ask student for their photo ID.
2. **Have the student put their name on the exam and exam answer sheet.**
3. The test should be conducted in an atmosphere conducive to good concentration (quiet, good lighting, etc.).
4. The student must take the exam without outside help. Have the students leave all materials (except blank paper, pen or pencil, or calculator, as needed) outside the testing room. This includes notes, books, calculators, phones, etc. (excluding materials required for the exam).
5. Close and constant supervision must be provided.
6. Please scan and email the proctoring form, completed exam, and any formula sheets permitted for the assessment to delta-testing@ncsu.edu or fax to 919-515-7180.
7. Not including exams that permit all notes or textbooks, students should not be permitted to leave the testing room with formula sheets or scrap paper unless explicitly stated.
8. DO NOT GIVE THE EXAM TO THE STUDENT TO MAIL BACK

If you have any questions, please contact DELTA Testing Services at our main Venture IV location via phone: (919)-515-1560 or e-mail: delta-testing@ncsu.edu.

Thank you for assisting our students.
DELTA Testing Services
NC State University

Updated March 2022

ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

Student's Name: David Grant

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, David Grant 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

David Grant 4/28/25

Student Signature *Date*

Exam must be turned in by: 11:40 DG

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 by using a variable to split the tree, repeating that process, until it eventually has its terminal leaves. The response gets assigned the value of whatever leaf it falls into based on the path it followed down the tree.

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 that determines the first split is the RSS. This first split is decided upon based on whichever optimal value of each predictor produces the lowest RSS at the split. This is considered a greedy algorithm because it only looks at the next split and not further down the tree.

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 split the data into a train/test set. Either 70/30 or 80/20 is usually optimal.

For the kNN algorithm, we have to tune the value of k (# of nearest neighbors), which in this case we do via bootstrap resampling on the training set. We take a sample using replacement on the training set, and test on the observations that weren't included in the sample (out-of-bag observations).

For ridge regression, we do the same thing - except the tuning parameter in this case is the penalty term, and we find that optimal value.

We compare the 2 optimally tuned models by finding the RSS of each, produced by the test set.

Whichever model produced the lower test RSS is our best model, and so we'll re-fit that model using the entire data set.

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

The two methods of early stopping is controlling the tree height, and the number of observations in the terminal leaves.

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

Two common activation functions is the one that considers a small subset (like a 3x3 block), and the regularization function.

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

The task that a recurrent neural network is well-suited for is image recognition.

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

False 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.

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

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

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

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

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

False i. KNN provides a discriminant for classifying our observations

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

8. Consider the piecewise polynomial regression model. Here we define our knots to be $c_1, \ldots, c_M$ 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 is the estimate of the intercept when X does not fall into any range specified by the indicator functions.

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

β_1 is the estimate of the change of the response for a unit change in the predictor, when the predictor's value is within the specified range.

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

The 3 most common tuning parameters associated with a boosted tree model are the number of observations in the terminal leaves, the number of trees, and the height of each tree.

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

Random forests for a regression task generally improve prediction over the basic bagged tree model due to the fact that each individual tree uses a specific subset of predictors. In a basic bagged tree model, not all predictors may be used.

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

To fit a basic boosted regression tree model, we do bootstrap resampling and use the out-of-bag observations to consider the splits of each tree. We create different trees based on changing the values of the tuning parameters, and find the tree that is optimally tuned, using our out-of-bag observations as the test data to tune the trees.

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

The support vectors are the hyperplanes that separate the classification groups.

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 1-v-1 approach considers all pairwise comparisons between each group. i.e. if we had 3 levels, then we comare 1 vs. 2, 1 vs. 3, and 2 vs. 3. Whichever class comes up the most (majority rule) is the one that's predicted.

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

We run kmeans clustering multiple times because each run could produce different groups in the end. So finding the optimal set of groups across multiple runs will produce better results.

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

Single linkage creates a dissimilarity measure by looking at the pairwise dissimilarity between the closest points of each cluster.

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

A biplot is the plot used to view a hierarchical clustering. It can be useful because you can see how many clusters there are from a given height on this plot.