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

77

Student Name: Matthew Bray **Date:** 30Apr25

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

Course: ST563-601 **Exam #:** Final / Exam 3

Start Time: 1 35 **End Time:** 2 50

Proctor's Name (Print): Anthony Winters

Proctor's Signature:

Institution: Bridgewater State University Testing

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: Matthew Bray

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, Matt Bray 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 30 Apr 25 *Date*

Exam must be turned in by: 3 00 *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

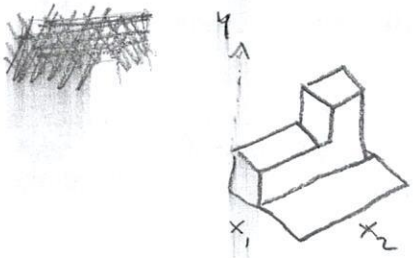
Please write your answers below each question. You should not have access nor use any materials during this exam.

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 series of "steps", in the step regions need to be next to each other



✓

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!

All possible splits are compared. The split that reduces the variance the most, or increases the homogeneity of the splits the most, is selected as the first split.

−1

$\frac{1}{n_1}$ or $\frac{1}{n_1-1}$ would change things

−1

① → continued... set aside the non-sampled obs from each bootstrap to use as internal test dataset for tuning.

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.

1) Split data using SRS stratified at least on the response variable, typically 70%/30% or 80%/20% train/test.

2) Create grid of k (for kNN) and λ (for ridge).

3) Randomly sample the training dataset using SRS w/ replacement to create bootstrap sample training datasets. These should contain the same number of observations as the original training dataset. Create, say, 10 of these "bootstrap" training samples. ①

4) Fit each model for each value of k or λ on all bootstrap samples and collect appropriate training metrics (AIC, adjusted R^2, MSE). Take the average performance for each tuning parameter. Fit each trained model for each value of k or λ on the out of bag samples. Collect the appropriate test metrics and compare each tested model for each value of k (kNN) vs. all others and select model with best OOB test metric. Compare each tested model for each value of λ with all other ridge models and select best test metric.

5) Train each tuned model on full training dataset.

6) Fit each new fully tuned and fully trained model to test set.

7) Model with best test metric (kNN vs Ridge) is the winner

8) Fit overall best model on full dataset.

on what?

−1

−1

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

1) Define minimum number of obs in a leaves.

2) Define maximum tree depth

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

reLU

-2

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

Language processing

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

-3 ~~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.

-3 ~~True~~ i. KNN provides a discriminant for classifying our observations

-3 ~~True~~ j. The Naive Bayes provides a discriminant for classifying our observations

-11

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 value of Y_i when [illegible] $X_i = 0$

-2

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

β_1 is the slope between [illegible] c_1 and c_2

-1

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

Pruning ~~[illegible]~~ depth ✓

Min # obs in leaves

trees

Learning speed ✓

-1

-4

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

The predictors are randomly sampled at each split. This allows for more exploration of the feature space in the cases where the earlier split variables may be most important to the model, so that the first few splits would be the same in multiple or all trees. Random forest can help to reduce that "sameness" of the first few splits.

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

1) Regression tree is fit, residuals collected.
2) Perturbations are introduced based on residuals.
3) Regression tree refit and 1 & 2 repeated until # of iterations reached.

Perturbations can be pruning leaves and moving terminal nodes up a level. or adjusting minimum obs in a leaf or adjusting split rules

-3

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

These are the vectors that define the shortest distance from the observations to the hyper plane that is the decision boundary.

-1

-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.

A hyperplane is created between each possible comparison of 2 classes.

then... -2

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

We don't always reach the global minima. We can run multiple times ✓ and use the best fit.

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

Each observation is compared to every other observation and the two observations that have the smallest distance are "agglomerated" into a larger cluster.

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

It can show a comparison of two principal components and allow for a visualization of which predictors are important for a particular PC.

-2