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

82

DELTA Testing Services

Student Name: Alex Devoid Date: 4-30-25

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

Course: ST 563 601 Exam #: FINAL EXAM

Start Time: 10:02 am End Time: 12:17 pm

Proctor's Name (Print): Jessica Snow

Proctor's Signature: Jessica Snow

Institution: Southwestern Community 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: Alex Devoid

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, Alex Devoid (STUDENT'S PRINTED NAME) 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.

Alex Devoid (STUDENT SIGNATURE) 4/30/25 (DATE)

Exam must be turned in by: 12:17 pm (EXAM END TIME) AD (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?

It splits the prediction space into regions.

-1

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!

For the first split recursive binary splitting finds the split that optimises the reduction in the loss fuction (RSS) for the two resulting regions.

RBS is a Greedy algorythm.

-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.

- Split data for trang/testing, for example (80%/20%)
- Create grids of possible hyperparameters for tuning for the knn model and the ridge regression model
- for each model, Iterate over the values in the hyperparameter grids
  - train the given model using each hyperparameter
    - each of these models will be trained on the bootstrapped data.
      - from the trang data, a sample of the data is taken of the same size. with replacement.
    - each model is tested on the out of bag observations
- Combine errs
- We choose the knn model and the ridge regression model the minimized the loss function when tested on the out of bag observations.
- using the selected hyperparameters, we now train a knn and ridge regression model on the trang data.
- Then we test the fitted models on the test data.
- We select the final model (knn or Ridge regression) based on the one that preforms the best with the best model metic score, minimizing the loss fuction.

- fit best to all data

-1

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

- we can use a tree depth HyPerParameter and test it with CV.
- we can use a metric threshold like a cost complexity limit.

-2

One benifit of XGboost is that it allows for early stopping

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

ReLu
Soft max

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

NLP - An Rnn is well suited for natural language processing because it is good for sequencial 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

-3 b. kNN models generally require you to standardize your predictors ~~False~~

-3 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 False

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

-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?

The Average response for when ~~X~~ $< C_1$

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

The Average response for when ~~$C_1 \leq X$~~ $< C_2$

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

- Number of Iterations or models ✓
- learning rate ✓
- tree depth ✓

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

Random forests will randomly select a subset of predictors in each node when training the bootstraped models. This will help control for any predictors that are very important and may dominate the nodes. This process results in more variability across the bootstraped models. But this variability is smoothed out when the models are aglegated.

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

Boosted regression ~~tree~~ models are slow learners.

The algorithm for fitting a basic boosted regression tree model trains sequencial trees. The target variable for these trees are the loss fuction values. Each sequencial model learns from the previous model's loss fuction. Predictors with higher erros are given more weight on each iteration to correct errors resulting from those predictors. The learning rate specifies how much each model will learn from the previous model. A lower rate will result slower learning. A High rate will learn more from each previous model resulting in faster learning.

ok

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

Support vectors are the points that are used to seporate classes. They are the point that a kernal fuction uses to seporate the classes into a higher ~~D~~demention without fully transforming the dataset.

0

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.

m=3

1_m vs 2_m

1_m v 3_m

2_m v 3_m

we apply the one vs. one svm model to the observations for all the combinations of classes and we take the majority vote response from all one vs. one models

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

It often does not create the same clusters each time the algorithm is run. -1

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

Single linkage uses the min distance between all the pairwise points from two clusters.

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

Elbow Plot - we look for the initial lowest point on the graph, using the reduction in variance.

-4

-5