

58

# ST 563 601 – SPRING 2025 – POST Final Exam Tablet

Student's Name:

Lilian Ngonadi

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, Lilian Ngonadi, 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

29/04/2025

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?

The standard regression model the response surface by splitting the surface into a straight line that divides the surface into 2 parts say  $X_1$  and  $X_2$  by a straight line - 1

2. For a standard regression tree that uses recursive binary splitting, suppose we have two predictors  $X_1$  and  $X_2$ . What criterion is used to determine the first split? Describe how this first split is decided upon. Be specific on both of these!

The criterium use to determine the first split is the sign if we have  $X_1$  and  $X_2$  our first split could be  $X_1 > C_1$  and  $X_2 < C_2$  where  $C_1$  and  $C_2$  are constant values - 4

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.

We first start with the entire dataset where we split the data into training (70%) and test set (30%). Then we further split our training set using the bootstrap method to obtain -2 the optimal  $k$  by fitting a kNN model, then we do the same for the lasso regression where we further split our training set (70%) using the bootstrap method to obtain the optimal  $\lambda$ . After getting the optimal  $k$  and optimal  $\lambda$  for the kNN and Lasso regression respectively, we will go ahead to fit a kNN model and Lasso regression model using the entire training set and our optimal  $k$  and  $\lambda$  respectively. We then use our test set for prediction. We then select the model using our metrics for regression task. (using MSE,  $R^2$ , AIC, BIC). Then the one that have the least MSE, AIC, BIC will be selected as the overall best model. fit to entire data set. -1

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

1) Set the number of ~~trees~~ to be used  
2) Set the ~~stopping~~ rate - 4

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

The ~~two~~ common activation functions are  
RELU and -2

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

Recurrent neural network is suited for  
processing text and time series 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~~ ~~True~~

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. ~~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~~

-12

-18

8. Consider the piecewise polynomial regression model. Here we define our knots to be  $c_1, \dots, c_M$  and use the indicator functions

$h_1(X) = I(c_1 \leq X < c_2), \dots, h_{M-1} = I(c_{M-1} \leq X < c_M), h_M(X) = I(X > c_M)$   
in our regression equation given by

$$Y_i = \beta_0 + h_1(X_i)\beta_1 + \dots + h_M(X_i)\beta_M + \epsilon_i$$

Suppose we have  $n$  observations and we fit the model.

- a. What is the estimate of  $\beta_0$  in this model?

The  $\beta_0$  is the intercept in the model

-2

- b. What is the estimate of  $\beta_1$  in the model?

$\beta_1$  is the slope in the model

-2

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

B - number of trees

$\lambda$  - stopping rate

-1

-5

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

Random forest improve prediction because it does variable selection and can be interpretable unlike the basic bagged tree that we struggle with interpretability - 4

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

$$\frac{1}{B} \sum_{l=1}^B \hat{f}(x)$$

first we get B. number of trees - 3

$$B_1, B_2, \dots, B_n$$

Then we fit a model  $f(x)$  and then find the predictions  $\hat{f}(x)$  and then sum the predictions for the number of trees and then take the average by the number of trees

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

The support vectors are hyperplane when it cannot be splitted by a straight line. - 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.

When we have more than two levels, we don't work with all the levels all at once, for instance if we have 3 levels, we first have  ${}^3 C_2 = \frac{3!}{(3-2)!2!} = \frac{3!}{1!2!} = \frac{3 \times 2!}{2!} = 3$  and we then use the SVM model for ~~1, 2~~, ~~1, 3~~ and ~~2, 3~~. So we apply the model for each of these and then take the one that has the highest proportion.

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

We run the kmeans clustering multiple times in order to obtain the optimal  $k$  OK

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

The single linkage measures pairwise dissimilarities between observations in cluster A and observations in cluster B. Then observations that have minimal dissimilarities are then used.

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

Biplot is use in finding the relationship between two principal component like PC1 and PC2. It is useful for finding relationship and it is use in visualization J