

80

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

Student's Name: Nick Zehnle

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, Nick Zehnle 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

Nick Zehnle  
STUDENT SIGNATURE

4/30/25  
DATE

**Exam must be turned in by:** 12:58

EXAM END TIME

NZ

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?

Average of the responses within the split by the leaf node. Allows for <sup>highly</sup> non-linear surfaces.

OK

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!

RSS; Computes RSS for splittable values of  $X_1$  and  $X_2$  on the entire training set, chooses split that produces the smallest RSS.

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) Randomly split data into a training and test set (e.g. 80-20, 70-30)
- (2) Tuning with bootstrapping: For each value of the hyperparameter in a tuning grid ( $k$  for KNN and  $\lambda$  for ridge regression) take  $B$  bootstrap samples on the training data. These are random samples with replacement on  $N_{\text{train}}$ . The data points that are not selected in the training set (since we are sampling with replacement and can draw on the same data point multiple times) are used as a test set. We then measure the MSE, for example, on this test set, repeat  $B$  times for each hyperparameter value, and average across  $B$  to get an estimated MSE for each hyperparameter value. Choose the value that produces the lowest MSE.
- (3) Fit tuned models to entire training set
- (4) Predict on the test set and choose based on test MSE (w.l.g. can use RMSE, MAE, etc.), i.e. whichever model's is smaller
- (5) Fit selected model to entire dataset

7

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

- based on whether the additional splits reduce error effectively or model remains roughly the same
- set a maximum depth
- look at whether there are ineffective splits (e.g. splitting hairs where there aren't many data points left to split or if a split puts <sup>essentially</sup> all the remaining points into one branch)

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

- Rectified linear unit (ReLU)
- Sigmoid
- Linear unit

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

Sequential data (e.g. time-series)

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 F
- b. kNN models generally require you to standardize your predictors X
- c. The number of trees we use in a random forest model is important because we can overfit with too many trees. F (bagging with random feature selection at splits reduces this worry)
- d. When using BART we need to remove the first few prediction models. T
- e. SVM models can only be used in classification tasks. F
- f. KMeans clustering does not necessarily create the same clusters in each run of the algorithm. X (unless you change K)
- g. Hierarchical clustering requires you to know the 'true' underlying groupings to use it effectively. F
- h. In a standard multilayer neural network, all inputs are 'connected to' all first level activations. T
- i. KNN provides a discriminant for classifying our observations F
- j. The Naive Bayes provides a discriminant for classifying our observations F

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?

$\hat{\beta}_0$  is the average of the response below  
 $c_1$ .  $\beta_0$  is the intercept. ✓

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

$\hat{\beta}_1$  is the average of the response  
between  $c_1$  and  $c_2$ .  $\beta_1$  is the slope  
between  $c_1$  and  $c_2$ . ok

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

- learning rate ✓
- early stopping ✓
- error / loss function ✓
- max depth ✓
- initial weak learner ✓
- number of trees ✓

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

Basic bagged tends to overfit since it does not take a random sample of features to be chosen from at each split like random forests does. Thus, in a basic bagged tree model you have  $n$  trees where every feature is considered for each node. By taking a random subset at each split in random forests, trees are decorrelated, introduces more bias, less variability  $\Rightarrow$  better generalization.

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

- Fit a weak learner, i.e. a tree with depth 1 or 2
- measure residuals
- fit a corrective tree onto the residuals
- add the first tree estimates to the prediction of residuals multiplied by a learning rate to get estimate of the second model
- repeat

$$x_{i+1} = x_i + \eta R_{i+1} \leftarrow \begin{matrix} \text{learning rate} \\ \text{residual predictors} \\ \text{by subsequent tree} \end{matrix}$$

(In stochastic gradient boosting, the difference is you take a subset on the residuals to fit on in the subsequent tree)

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

- represent the margin
- with separable data, parallel plane to the hyperplane through the nearest correctly classified point on each side to the hyperplane
- can have soft margins for non-separable data where points are allowed to be in the margin

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.

- Fit a SVM for each class against all other classes
- for a new given point, each SVM will produce a signed distance from its hyperplane  
(negative means classified as another class) -3
- choose the class where the SVM produces the greatest of these values (deepest within that class)

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

- to check that the clusters do not vary between runs
- to observe whether data can be clustered better into more/less clusters - |

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

Single linkage is the <sup>distance between</sup> closest pairwse points of ~~between~~ two clusters. If small, boundary of clusters are nearby (similar).

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

Allows you to see dissimilarity between clusters.

- 5

- 9