96

ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

Student's Name: Kevin Kronk

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, Kevin Kronk 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 4/30/2025 *Date*

Exam must be turned in by: 5:48pm KK

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

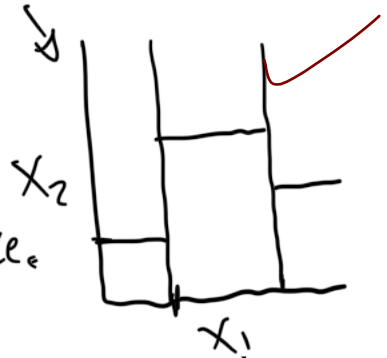
Please write your answers below each question. You should not have access nor use any materials during this exam.

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?

Its modeled as a decision tree with splits at certain points of each feature. For instance ↘



Then the mean response in each region is calculated. So it has different levels in a 3D plot where y shows the response surface.

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!

Its a greedy algorithm so it chooses the split that minimizes the sum of squared error based on the mean response in each region and summing over all regions. Then it goes to the next split from either X_1 or X_2 that once again minimizes SSE. It does this by calculating the SSE for many points where the tree splits across each predictor and then choosing the one that gave the lowest SSE.

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.

KNN
RR
bootstrap
tuning params

(1) Start by splitting the data into a training set (say 80% of total data) and test set (20%)

(2) Set up a tuning grid for both models. In KNN we use K - the number of nearest neighbors. In ridge regression α - the regularization parameter. The grid will have a range of potential values for each.

(3) On the training set we randomly sample with replacement until we get a sample size equal to the train dataset. (the values that aren't used are set aside as out of bag values for testing). This gives the bootstrap sample from which we train each model with one given hyperparam from the tuning grid. This is done many times for each hyperparam

(4) Each model is tested on the OOB values and the bootstrap error (like MSE) is averaged over all the bootstrap samples. (for a model with one hyperparam value)

(5) this is done for every hyperparam value and the one with the lowest avg error is chosen for both kNN and Ridge Regression

(6) These best models are then tested on the test set from the beginning and the one with the lowest test error (MSE) is the best model.

(7) This final overall best model is then fit on the full dataset.

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

- At a certain tree depth (like only allowing 3 splits in a row)
- At a minimum number of observations in a leaf node or below

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

- Sigmoid - gives output between 0 and 1
- Relu - outputs 0 below some cutoff and a linear function above it.

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

Text Analysis like classifying sentiment

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

F a. Random forest and bagged tree models generally require you to standardize your predictors

T b. kNN models generally require you to standardize your predictors

~~T~~ c. The number of trees we use in a random forest model is important because we can overfit with too many trees.

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

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

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

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

T h. In a standard multilayer neural network, all inputs are 'connected to' all first level activations.

F i. KNN provides a discriminant for classifying our observations

F 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, ..., 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?

Its the mean response for values in the region specified by $X < C_1$ ✓

$\hat{\beta}_0 = E[Y_i]$

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

Its the contribution to the mean response for values in the region specified by $C_1 \leq X < C_2$ ✓

$\hat{\beta}_1 = E[Y_i] - \beta_0 \qquad E[Y_i] = \beta_0 + (1)\beta_1$

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

- The learning rate ✓
- The number of iterations of learning ✓
- The number of predictors it can choose from when making a modification

-1

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

Random Forests only choose from a subset of the total predictors for splits on each tree. This helps ensure that no single predictor has too much weight in every tree. Thus it decorrelates the trees leading to better predictions.

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

It starts with a base learner tree (like random forest). This is then fit on the data. The original tree is then modified in some way - changing the predictions, pruning the tree, or adding branches - that reduces MSE. This is moderated by a learning rate that only allows it to change by a certain amount. This new tree is then trained on the residuals of the predictions from the previous one. The process continues for a specified amount (that needs to be tuned so the model doesn't under or overfit).

ok

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

They are individual data values that fall within the margins of the SVM, and are therefore important for determining the hyperplane and margins.

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.

For a model with K classes we compare all the classifications for each combination. So classifying a datapoint as K=1 or K=2. We then take the majority vote of which class won the most in these match ups to be that data point's classification.

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

There are many possible local minimums when reducing the intercluster distance summed over all clusters. Starting in different locations allows us to choose the fit that minimizes this the most.

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

It calculates the distance between all points in two clusters and finds the two that are closest to serve as the dissimilarity measure.

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

For PCA this shows the contribution of each feature to two (usually the first two) principle components. This gives a sense for what each PC represents.