91

ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

Student's Name: Dev Kewlani

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, Dev Kewlani, 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

Dev Kewlani

Student Signature *Date*

Exam must be turned in by: 3:14 pm Dev Kewlani

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 tree models the responses in rectangles in a plane, where each rectangle represents a region of a split we built the tree on. We can have many such rectangles & regions based on the splits we do.

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!

we use a greedy algorithm, so basically for full range of values for each X_1 & X_2 we decide on a value and consider that value as split point, so for eg: for X_1: split at c, for X_2: split at d. then we have regions of the form - $X_1 < c$ $X_1 \geq c$ or $X_2 < d$ & $X_2 \geq d$ we send all values to whichever split they belong to and take the mean of all

values in each split, that's our prediction for that split.

So we have $\hat{y}(x_1 < c)$, $\hat{y}(x_1 \geq c)$

or

$\hat{y}(x_2 < d)$, $\hat{y}(x_2 \geq d)$

Then we find residuals comparing each y_i with $\hat{y}$, and sum up across both $(y_i - \hat{y}_1(x_1 < c))^2 + (y_i - \hat{y}(x_1 \geq c))^2$

or

$y_i - \hat{y}_1(x_2 < d)^2 + (y_i - \hat{y}(x_2 \geq d))^2$

whichever from ~~#~~ X_1 or X_2 yields a lower RSS is selected as the split and we do this for a range of values for both X_1 and X_2. That is how we decide on our first split.

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 the data between train and test (80%) (20%)

2) Since we are using bootstrap:
we will draw 10 bootstrap samples of training data [which is sampling training data with replacement and sample size same as training data], so some observation will naturally be left out. That is our OOB samples for validation.

3) we define a grid of both K & λ values.

4) for each K and each λ,
we train KNN & ridge regression respectively on the bootstrap sample and test the performance on OOB samples, we keep a track of RSS across all 10 OOB samples.

5) We repeat this for all K and all λ.

6) So let's say we have a grid of 10 K and 10 λ values. we get a total of 20 RSS values averaged across 10

10 bootstrapped OOB observations.

7) we select the best λ and best K whichever has the lowest RSS among all ridge and all KNN

8) we then train the best K KNN model and best λ ridge model on the full ~~data~~ training data (80%) and evaluate its performance on test set (20% remaining)

9) whichever yields better results we select that model.

10) We train that model on the full Dataset. [This is the best model].

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

1) Either we stop when our terminal nodes exceed a threshold.
2) We can stop when number of observations in leaf node is less than a certain threshold.

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

1) ReLU
2) Tanh

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

- & Time Series tasks [Next token prediction, Stock market prediction]
- when the output we want to model is sequential.

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
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. — False
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. — False.
i. KNN provides a discriminant for classifying our observations — False
j. The Naive Bayes provides a discriminant for classifying our observations — True

-9

-9

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

$$h_1(X) = I(c_1 \le X < c_2), \ldots, h_{M-1} = I(c_{M-1} \le X < c_M), h_M(X) = I(X > c_M)$$

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?

$$\beta_0 = \frac{\sum I(X_i < C_1)}{n}$$

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

For $\beta_1 \Rightarrow \beta_2 \cdots \beta_M = 0 \Rightarrow C_1 \le X_i < C_2$

$$\beta_1 = \frac{\sum I(C_1 \le X_i < C_2)}{n} - \frac{\sum I(X_i < C_1)}{n}$$

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

Tree Depth (d): how much should the tree be extended vertically based on how many splits. [more depth more overfit]

Learning Rate (λ): how fast or slow we change the movement of our current output based on our current prediction.

No. of trees to fit: more trees, more overfitting

subsampling ratio : how many rows or columns to select for building the current tree [Idea of Stochastic gradient Boosting]

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

By considering only a subset of columns at each split (usually $m \approx \sqrt{p}$) it decorelates individual trees, which lead to more improved stable estimates when we take mean across all trees.

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

1) start with $\hat{f}^0(x) = 0$ & r_i = Residual = y_i

2) for tree $t = 1 \dots T$
fit a tree [usually a small tree as here learner called a weak learner]
get prediction $\hat{f}^t(x)$;
update $\hat{f}(x) = \hat{f}^0(x) + \lambda \hat{f}^t(x)$
update $r_i = r_{i-1} - \lambda \hat{f}^t(x)$
continue this process for all iterations [Trees T]

3) for a new sample, our prediction is,

$$\hat{f}(x) = \sum_{t=1}^{T} \lambda \hat{f}^t(x)$$

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

Support vectors are those data points which lie inside the margin or on the wrong side of hyperplane. Our margin is influenced only by these data points. Based on our cost, we can have a wider or

a narrower margin. More wider margin leads to more points violating ⇒ more support vectors, less wide ⇒ narrow margin ⇒ less support vectors.

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.

Let's say we have 3 classes, {1, 2, 3}.
then we train 3 models, 1 vs (2 & 3)
where (x & y) means that 2 vs (1 & 3)
x & y are assigned same class 3 vs (1 & 2)
label ⊛. doing this we get probability for each class based on our model fit. whichever has the highest - we select that class.

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

Because K-means is heavily influenced by the initial configuration that we start with when we assign each data point in one of the K clusters, hence it is important to run multiple times to select the one with minimum loss to get confidence in our algo.

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

- we find all pairwise dissimilarities between observations from cluster A and cluster B.
- we then select the minimum of all these pairwise dissimilarities to get our dissimilarity

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

A biplot is a powerful tool to visualize the (direction) loadings of principal components and data points of original features that existed in our original space. It is an effective

visualization tool that helps us understand and see which principal component have how much coefficients or to put it losely, how much content in which direction and from the original feature space.

eg: PC1 has $0.2x - 0.2y + 0.7z$ then we can see the directions of these original features and PC1 w.r.t data points through the biplot.