81 A-11

ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

Student's Name: Naman Pujani

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, Naman Pujani 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 4/26/25 *Date*

Exam must be turned in by: 2:26pm NP

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?

When we create a regression tree, we model the response by analyzing which tree/splits give us the lowest amount of variance

For SVMs, this is decided by maximizing the margin (minimum distance from each point to the margin)

-5

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!

The first split is decided using a greedy approach. We pick the split that would be best at that point, not for the overall fit. We do this by looking at the pair that has the least amount of variability.

-2

-7

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 split our dataset into a training and test set (70|30). We then create a grid of possible k values and use non-parametric bootstrap to sample from those values. We sample with replacement and we end up with observations that don't get sampled (out of bag observations), which we can use to make predictions. From these values, we find the k that gives us the lowest Euclidean distance from our data (tuning). We then create a grid of λ values (penalty term) for ridge regression. We repeat the same bootstrapping process for the different values of λ for our data and find the best tuned λ. We then use them to make predictions on the out of bag observations.

repeat bootstrap

sample for train set, not test

We then combine the different euclidean distances and average them and find the average of the penalty rates as well. We fit both to our train set and see which model gives us the lowest MSE (RR) or higher accuracy (KNN).

We then fit the best model out of the two to our test set and then fit it to our entire data.

-3

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

1) Deciding on the number of observations in a cluster

2) Deciding on the # of splits

ok

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

1) ReLU

2) Sigmoid

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

A recurrent neural network is suited for text analysis task (analyzing movie reviews, for example).

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 True

   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

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

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?

β_0 is the estimate when X is not between c_1 & c_2 (indicator function is 0). It's the intercept so it's the estimate when our $h_M(X)$ functions are all 0. i.e. $X < c_1$ -1

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

β_1 is the estimate when X is between c_1 and c_2 and the indicator function is 1. It's the estimated slope associated to the indicator function when X is between c_1 and c_2. -1

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

B = # of trees to fit ✓

λ = the learning rate (.0001, .001, etc.) ✓

d = # of splits ✓

-2

10. Why do random forests for a regression task generally improve prediction over the basic bagged tree model? They improve prediction because if there is a strong feature, then that feature is likely the first split in a lot of the trees in our bagged model. Random forests randomly select a subset of trees in order to avoid fitting trees with similar splits. This helps decrease variance & any correlation that existed across trees and provides us with better predictions.

11. Describe the algorithm for fitting a basic boosted regression tree model.
We start with $f_k(x)=0$. From there, we randomly assigns observations and predictors to a split and measure our prediction error. Every subsequent function will have the learning rate applied to it (λ) and will try to learn & fix the mistake from the function before (normally, it tries to fix the largest prediction error.) These functions are uncorrelated/orthogonal to each other. We will then iterate this over different values of B (# of trees to fit) and d (# of splits), which are tuning params. We will stop when we can't decrease our prediction error anymore and the learning rate slows down to find the optimal number of trees and splits without overfitting.

-1

close

12. When fitting a support vector machine model for classification, what are support vectors?
Support vectors are the points that are on or close to the classifier or the line that separates our classes. These are the points that decide where the line/classifier splits our classes.

-1

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.

The one vs one approach looks at the pairs of classes. For example, if we have 3 classes, we'll look at: 1 vs 2, 1 vs 3, & 2 vs 3. We will look at $\binom{M}{2}$ classes and choose the most prevalent class (majority vote).

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

The kmeans algorithm tries to minimize an objective function that assigns each obs to a random cluster (K total). We sum the euclidean distances and divide by the total # of obs. and average the within cluster variation. Doing multiple runs allows us to find the optimal # of clusters and find the best within cluster variation that minimizes that objective function.

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

The single linkage looks at the pairwise differences from cluster A & cluster B and measures the dissimilarity based on the 2 closest points between the 2 clusters.

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

A biplot shows us the different weights associated with our different principal components (PC1 vs PC2 for example). It can be useful b/c it helps us understand the weights of the different features for each principal component. This can help compare the linear transformations across the different components.