ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

Student's Name: Dezhong Xu

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, Dezhong Xu 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

Dezhong Xu. April 30th

Student Signature *Date*

Exam must be turned in by: DX

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 trees model the response surface with piecewise lines or multiple discrete planes depending on the dimension of the input variables. Regression tree would not make continous prediction, so the associated hyperplane is discrete.

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 found based on whether any split in X_1 or X_2 would make the biggest reduction in RSS.

(1) We calculate the RSS of not making any split, in this case the RSS would representing the error between observations and average of all values.

(2) Then we could test if making a split on any value from X_1 or X_2 that could make the biggest reduction in RSS.
→ for example, make split on X_1. Calculate average value of the observation on two regions and RSS is the sum of the RSS for each region.

(3) Make the split at that value with the largest RSS Reduction.

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, randomly split the data by a seperation of 70% to 30% (70% training / 30% test).

② Tunning by bootstrap.

KNN

- define a tuning grid for K

△ for each K value:
- bootstrap some portion of data from training dataset to form bootstramp data, repeat B times
- the training data not selected are treated as out of bag data

△ for each bootstrap sample in the given K value:
- train the KNN model based on the bootstrap data, and validate its performance use oob data.
- the overall performance for the K value model is based on the average of the model performance on oob data
- the optimal K is selected based on the metric values.

Ridge Regression.

- define a tuning grid for λ.

△ for each λ value:
- bootstrap some portion of data from training dataset to form bootstramp data, repeat B times.
- the training data not selected are treated as out of bag data

△ for each bootstrap sample in the given λ:
- train the ridge regression model based on bootstramp data, and validat its performance use oob data
- the overall performance for the λ value model is based on the average of the model performance on oob data
- the optimal λ is selected based on the metric values.

-1 ③ fit the optimal K KNN & optimal λ ridge regresion to the 30% test dataset

-1 fit to whole ④ Compare the MSE of two models, the one with fewer MSE is better. -2

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

(1) define the # of splits

(2) define the minimum observations in nodes.

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

(1) signomoid

(2) rectified linear unit (ReLU.)

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

To analyze the sequential data such as documentation

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

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

F 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. burn-in

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 not directly though

-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

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

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?

$\hat{\beta}_0$ is the expected value for any X that is not fall into the the criterial of C_1 to infinity.

ok

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

$\hat{\beta}_1$ is the expected increase to ~~the~~ expected value of response variable when X falls into anyvalue between C_1 & C_2.

OK

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

B : the number of trees.

λ : the learning rate (usually 0.01 or 0.001).

d : the number of split when growing the tree, usually 1 ↙ stumpf tree.

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

The random forests tree further reduce the variance of the tree and by randomly select a subset of predictors (not all predictors) when deciding a split. It reduces the correlation btw the trees, so it reduces the variance and prevent overfitting.

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

(1) Initialize a group of trees with few splits. (# of B).

(2) for each initialized tree, grow the tree successively:

2.1 find the residual in the previous tree

2.2 grow a stumpt tree from residual.

2.3 Combine the previous tree result and part of the stumpt tree result

2.4 Repeat above steps to grow the tree to minimize the RSS, until the tree reached to certain split numbers threshold (d).

(3) Tally the result from all trees and form the embedded trees for prediction, the overall performance would be the average of all trees.

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

The support vectors are the observations that has the minimal and equal distance to the hyperplane. They sit on the margin and support defining the hyperplane. Hence they are called 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.

1 vs 1 approach is

① Assume K classes, build a total of $\binom{K}{2}$ SVMs for each pair of the response classes.

② Use each of SVMs to make classification for every observation.

③ The final classification for the observation is decided based on majority vote. The most frequently classified class is assigned to that observation.

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

Since K-means clustering is an unsupervised learning technique, we never know if the prediction is right or not. we ran multiple times to make sure the result is relatively robust and reliable.

-1

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

The single linkage refers to the shortest distance exists between the observations of the two clusters. By understanding the single linkage, we could understand how far the clusters seperate to each other to understand the dissimilarity. If single linkage is large, means clusters are more dissimilar. but also, it neglects that the observations may be disperse.

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

The biplot is the plot contains two key informations in same plot:

① the data distribution of the first two most prominent variables with largest variability

② the rest of the prominent variables relationship projected to the 2 dimensional plot.

It helps to understand the basic relationship among the most prominent variables with single plot.