69

ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

Student's Name: Jaswinder Kaur

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, _____ 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

Jaswinder Kaur — 30th Apr 2025

STUDENT SIGNATURE *DATE*

Exam must be turned in by: _____

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

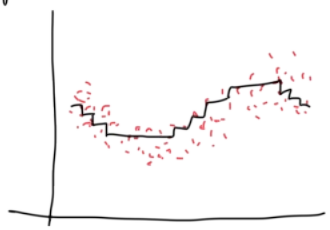
Please write your answers below each question. You should not have access nor use any materials during this exam.

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?

It model the surface by explaining the maximum variability present in the model

eg 

−5

Top up: greedy method

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 criteria used to determine first split depend on the decrease in SSE it will provide if the split is made over there. The difference in residuals & observed is predicted & squared & then both are added for both split eg. $S_1 \geq 5 \rightarrow$ (residuals − observed)^2 $S_2 < 5 \rightarrow$ (residuals − observ)^2] added

Whenever we find the maximum decrease in loss function, that split is decided upon. & it is done for both multiple predictor as well as the values within those predictor & Based on decrease in loss function, & the first split is made

−5

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.

Split the data set → Can be 70% train & 30% test
or 80% train & 20% test

Use the train data set :-
This train data set can further be splited into 10% validation & rest internal training.

Using the internal training data set :- we tune the hyperparameters

For KNN → we determine the best value of k that can be used it can be anywhere from $k = 1, \ldots, K$
Best value will provide us with least error.

For Ridge regression → Best value for λ will be determine by tuning on internal dataset & picking the one with least error rate.

The validation test will be used for further confirming these value. After multiple repeatition, the Best value would be selected & using the tuned parameter,

The models will be trained on entire training dataset (Internal Training dataset + validation)

→ Once fitted on training, the models will be applied on testing dataset & the one with lesser RMSE and better performance can be choosed & then applied over entire dataset.

bootstrap

-3

-3

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

① Either we can put threshold on number of splits for eg:- 5, so that models stops after 5 splits

② Or we can put threshold on min. no. of observations present with predictor to be used for splitting.

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

→ sigmoid

→ Relu

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

for predicting on text or word relating problems

ex:- predicting on reviews of a movies based on reading specific words.

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. 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, can be done based on similarity

h. In a standard multilayer neural network, all inputs are 'connected to' all first level activations. True

layer ?

i. KNN provides a discriminant for classifying our observations True

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

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 intercept for the model unlike every regression model. It is the value of y when all other predictors are not significant or held constant.

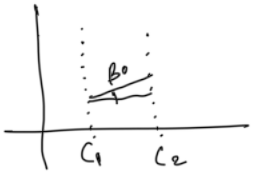
−1

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

β_1 is the unit change when value of X is more than Knot C_1 & while moving towards C_2 But on the edge of both knots, given when all other are parameters are kept '0'

$C_1 \leq X < C_2$

−2



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

① No. of trees to be used ✓

② No. of splits / Branches ✓

③ Presence of minimum observations −1

−4

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

In random forest, like bagged tree model, we train trees on train data & predict on natural test (out of Bag) data. In random tree however, specific weights are given to residuals performing poorly with higher values. Hence on aggregation the variance is considerably reduced. which is not be the case in basic bagged model.

-4

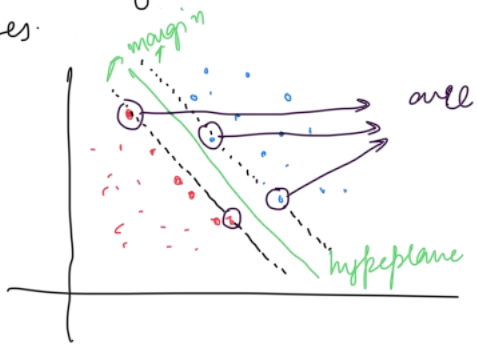
11. Describe the algorithm for fitting a basic boosted regression tree model.

In boosted regression tree model:-

Initially most basic tree is generated & the residuals of this tree are estimated. Now moving forward, these residuals are used as starting value & using them, new tree is fit. Again the residuals of next tree is taken & used as input for generating next tree. This way multiples trees are generated, targetting areas where residuals are high & model is struggling to perform well there. This way variance of multiple Trees once reduced by aggregation & finally better tree is choosen with reduced loss function & better predictions.

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

Support vectors are those observation that lie on edge of our margin & guide us to the hyperplane seperating two classes.



once support vectors.

-4

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.

In one versus on approach
lets say we have 3 class of M = 1, 2, 3

then according to this approach
1 v/s 2
1 v/s 3
2 v/s 3
Now using these classifiers.

we will fit the model, independent of each other
& whichever class came up majority of votes or majority of times will be decided as the predicted class.

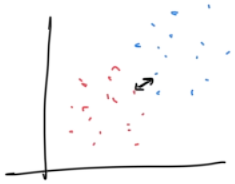
14. Why do we often run the kmeans clustering algorithm multiple times?

Because its a unsupervised learning, hence we don't have anything to compare & check with. Running multiple times provides the surity of seeing if things have stopped changing & started clustering in a specific way over multiple iteration.

−1

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

Its the minimum distance among the two closest points of two different clusters. This distance can be even smaller than distance among points within a cluster hences creating dissimilarity with cluster.



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

Totally guessing :-

Is it the plot function in 'R' that help us plot the tree?

−5