86

ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

Student's Name: Nirmal Tirimanna.

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, Nirmal Tirimanna 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

NT 4/30/2025

Student Signature *Date*

Exam must be turned in by: 90 mins NT

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

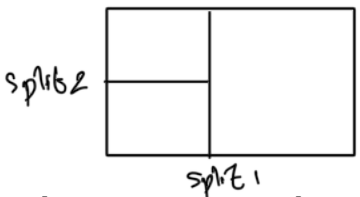
Please write your answers below each question. You should not have access nor use any materials during this exam.

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 response surface by dividing the predictor space into regions and makes predictions. The predictions for that region are taken by using the mean/average of the space for regression tasks. And for classification it uses majority vote.

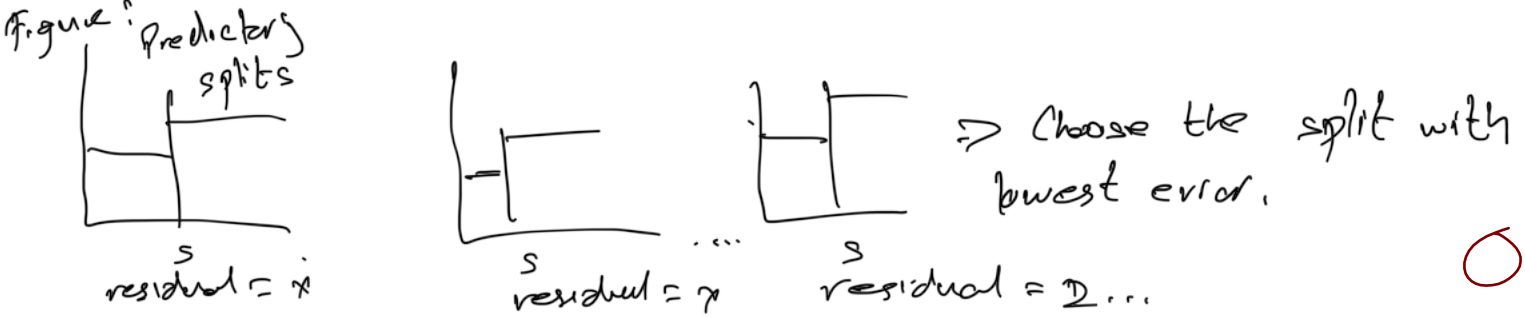


& response surface.

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!

→ The standard regression tree uses greedy down approach to make the split. It calculates the error or loss for all possible first splits and uses the one that has lowest error or loss for first split.

For the first split it uses all the possible 's' values & uses the combination (j, s) that has lowest error. and uses it for split. It does not care for the further 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.

a) kNN

Step 1: split the data into train set & test set (70:30 or 80:20)

Step 2: Hyperparameter tuning
a) Take the training set & create bootstrap samples B.
b) Define k; k is a pre-specified grid for no. of neighbours to use for the model
c) For each bootstrap sample fit the model with a k value & use out-of-bag samples (OOB) to evaluate the fit using model metrics such as. MAE, RMSE, R^2, & average and for that k.
d) Repeat 'c' for all values of k & choose the one with lowest evaluation metric (eg. lowest value of MAE)

Step 3: Fit the tuning parameter value on full training set & evaluate on test set.

b) For Ridge Regression:

Step 1 -> Same ;

Step 2: -> a -> same
b -> define λ; λ is the penalty term in ridge regression which shrinks coefficients.
c) -> same but using λ values
d) -> same & choose the one with lowest metric.

Step 3: Same

c) Step 4: Compare the mode metrics from step 3 of two models & Choose the one with lowest (eg: lowest MAE).

d) Step 5: Use the model with lowest metric & fit on the full data for final use.

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

→ The two ways to do early stopping are:
- → define the no. of split.
- → define the final number of observation on terminal node. (or minimum)

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

→ The two common activations are: -2
- transformation (log, sqrt)
- Radial basis transformation
- Polynomials (dth degree)
- Neural network transformation.

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

→ RNNs are well suited for sequential data where order & sequence matters; eg: text, time series data. RNNs can remember the prior input values, so it is well suited for tasks where order & sequence matters.

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. True, OR Support Vector regression for regression tasks.

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. → All observation have own cluster & are grouped as we go based on similarity with dendogram. and get to 1 cluster.

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

-3

-3

-6

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

$$Y_i = \beta_0 + h_1(X_i)\beta_1 + \cdots + h_M(X_i)\beta_M + \epsilon_i$$

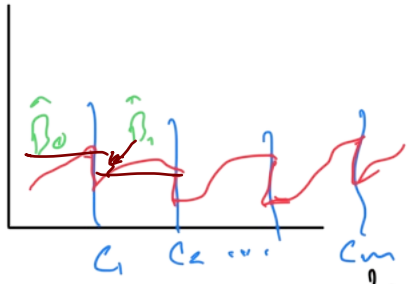
Suppose we have n observations and we fit the model.

a. What is the estimate of β_0 in this model?

⇒ β_0 is for the section when all indicators are '0'. All the observations falls at $I(x \le c_1)$ and has the estimate of $\hat{\beta}_0$.

$\hat{Y}_i = \beta_0 + \epsilon_i$

Mean −1



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

⇒ β_1 is for the section for between c_1 & c_2. All the observations for this falls between $I(c_1 \le x \le c_2)$ and it uses $\hat{\beta}_1$.

$$\hat{Y}_i = \beta_0 + I(c_1 \le x < c_2)\beta_1 + \epsilon_i$$

−1

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

⇒ The three most common tuning parameters associated with a boosted tree model are:

a) B = no. of trees; we need to be carful not to fit too many or to less trees (note, B doesn't mean bootstrap samples here) to avoid over or underfitting

b) γ = learning rate (eg: 0.001, 0.01)

c) d = no. of splits in the tree

−2

Also, few strong predictors dominate the prediction, which is improved by random forest by using random subset of predictor for each split.

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

→ The basic bagged tree model grows tree independently & splits the data using all available predictors. This causes trees to be highly correlated & the model does not generalize well. Random forest, on other hand, is a extension of bagging with one modification. It uses random subset of predictors at each step, which reduces correlation bet^n trees & the model predics better.

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

→ Step 1: Split the data into training & test set (80:20 or 70:30%)

Step 2: Tune the hyperparameters (B, d, λ) using training set only
- a) Use cross validation (5 or 10) on the training set to create internal training set & validation set.
- b) For each hyperparmeters train on training set & evaluate on the validation set. Repeate the set & average the model metrics (MAE, RMSE) for that hyperparameters.
- c) Choose the hyperparameters with lowest metric.

Step 3: Fit a weak learners (such as for mean or 0); (around 6 to 8 splits)

Step 4: Calculate the residual

Step 5: Fit a new tree on residuals. It tries to improve on the signals left by the previous tree & improves model.

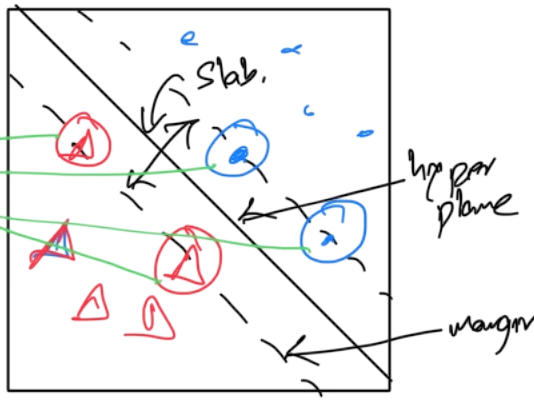
Step 6: Repeat 4 & 5. for B times.

Step 7: The prediction is obtained using weighted average from all trees.

learning rate

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

Support vectors are the observation or data point that determines margins & hyperplane. They are the minimum distance points from the the hyperplane. for each dataset.



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.

→ One Vs one approach compares all the pairs one by one to obtain the classifier. The classification is done by using majority vote. Example,

Examples. 🔴 Vs 🔵
🔴 Vs 🟡
🟡 Vs 🔵

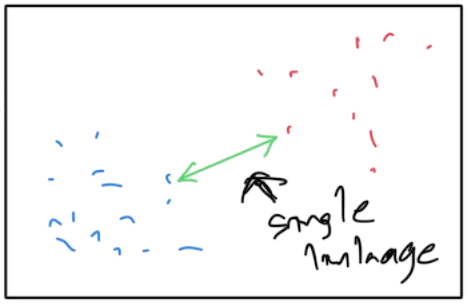
If there are 3, there will be 3 classifiers & uses majority vote to classify the new observation.

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

→ In unsupervised methods, eg: kmeans, there is no response variable or target variable. So, there is no standard thing that we shoot for. Thus, we run the kmeans algorithm multiple times to get the model evaluation metrics and we choose the best one. Running more than one time same kmean algorithm also gives different results

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

Single linkage is a intercluster dis-similarity measure which measures the distance between nearest point of two clusters.



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

→ Biplot shows the variance and principal components used. It helps to map the variance each principal component accounts & the cumulative biplot helps in determining the no. of PC to be used.

Eg: We can use 90% cumulative variance to know how many PC's to use.