ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

Student's Name: VIDYUL JAIN

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, VIDYUL JAIN 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

Vidyul 04/28

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?

By dividing the entire space into different regions (usually rectangular regions). So, in essence it divides the space into specific regions using the predictors. then uses a constant -1

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 RSS to determine the first split. This is done to see that for which predictor and value does the split return the lowest RSS.

The feature/predictor and its corresponding value that gives the lowest value to the metric is then used to do a split.

-1

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. The data is first split into training and test set. In practice, we generally use a 70:30 or 80:20 split wherein 70% or 80% of data is used for training & 20% or 30% is used for testing respectively.

2. The training set is further used for bootstrapping for tuning the parameters, under which we randomly select datapoints with replacement, essentially giving us a natural validation set (Out of bag samples) to tune our parameters on.

3. Using this, we tune the parameters for the KNN & ridge regression model & test it on the validation set.

4. We use the best parameters for both the models and make prediction on the test set.

5. We select the best model with its best parameters based on the performance of the test set and fit that particular on the entire dataset.

—

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

1. Controlling depth of the tree.

2. Minimum samples needed for a split, controlling number of samples in the terminal node.

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

ReLu and Sigmoid.

$z\,\mathbb{1}(z>0)$ $\qquad \frac{1}{1+e^{-z}}$

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

It is best suited for tasks such as text processing & time series data.

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. 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

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

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

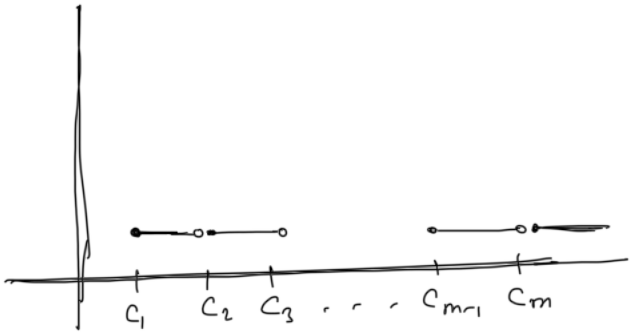
in our regression equation given by

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

Suppose we have n observations and we fit the model.

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

$\hat{\beta}_0 = \bar{Y}_i$



-2

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

$\hat{\beta}_1 = \dfrac{\bar{Y}_i - \beta_0}{1}$

-2

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

same

1. Depth of the trees.
2. Learning rate
3. Number of splits.

-1

-5

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

Random forests are bagged trees with an additional step of selecting a random number of predictors at each a node is split. This generally means at each node we are selecting a random sample of features instead of splitting on the entire set of predictors. This leads to reduced overfitting and better generalization of the model's performance. −1

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

Under the boosted regression tree model, trees are fit sequentially and subsequent trees are weighted according the performance of the subsequent trees. ie.

1. A base learner is fit, which predicts 0 for all points. This means, the residuals $r_i = y_i$
2. After this a tree is fit on the residuals and $\hat{f}^b(x)$ is updated from 0.
3. This process is repeated, essentially by giving higher weights to incorrectly classified trees in hope for next tree to accurately predict its result.

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

Support vectors are all those points that either lie on the margins, or beyond the margin but still on the correct side of the hyperplane & points that are on the wrong side of the hyperplane. eg.

All the points marked by arrows are the support vectors.

−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.

Under the One Vs One approach, the following is done.

Eg. $K = 1, 2, 3$.

We run for: $K=1$ Vs $K=2$

$K=1$ Vs $K=3$

$K=2$ Vs $K=3$.

After this, for the classification task we take the majority vote for each run of the above mentioned classes & assign the class with the majority of the votes.

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

Since the result we get may be sensitive to the initialization of clusters to the datapoints & affects how centroids are formed and clusters are made. For this reason, we generally run the algorithm multiple times to get the best cluster assignment.

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

By taking the distance between the two closest points with the least distance

Eg.  → Distance between the two points in red will be computed to get the dissimilarity measure.

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

-5

-5