81

ST 563 601 – SPRING 2025 – POST
Final Exam Tablet

Student's Name: Henry van Eijk

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, Henry van Eijk 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

Henry van Eijk 4/28

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?

A standard regression tree won't be a smooth surface. It will resemble a step function in multiple dimensions. In 2D, imagine lines are drawn for each split which defines the regions

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!

A split in the regression tree is based upon a "greedy" approach, i.e., what step will be the best for this iteration. For the first split, it will find the best threshold across predictors X1 and X2 that minimize some loss function the most. Eg, $X_1 < 10$

-1

-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.

First, we split data into train/test set, possibly 80/20 or 90/10. Using only the training data, we take B samples of size n from the training data with replacement (ie nonparametric bootstrap). We use the sampled values for the fitting and predict on the out-of-bag observations, ie observations that weren't sampled. After doing this procedure B times, we choose the best for some default grid, eg $\lambda = \{0.1, 1, 10\}$ tuning parameters for KNN and best λ for Ridge. We then refit across all training data with the optimal tuning parameters. We evaluate our models on the test set, then choose the model with lower loss. We refit across all data with the best model and use that for any real-life predictions.

-1

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

Setting a max depth where each tree can only have a # of levels, or pruning where we grow a large tree then "trim" off nodes

-2

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

Sigmoid and ReLu

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

Text generation

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. True
   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

-5

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 $X < c_1$ with all other parameters fixed

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

β_1 is the slope estimate when X is between c_1 and c_2. Ie, when X is between c_1 and c_2, it causes a β_1 increase/decrease to Y, (with all other parameters fixed).

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

1) Number of trees: too many trees can cause overfitting.

2) Max-depth: how deep each tree should be.

3) Parameter to control the # of random features to consider for a single split

-2

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

The trees in a RF are less correlated w/ eachother since we introduce a parameter to consider splitting on a random ~~x~~ features instead of all p features. This makes the prediction better by generating trees that are different from eachother instead of many very similar trees, i.e decrease variance.

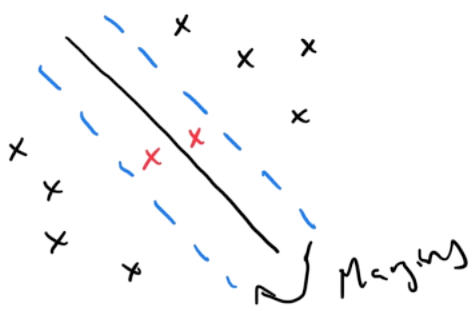
11. Describe the algorithm for fitting a basic boosted regression tree model.

Boosting starts out with a naive prediction such as mean of Y. I.e $\hat{y} = \bar{Y}$. Then it calculates the residuals $r_i = y_i - \hat{x}$. Then the residuals $r_1, \ldots, r_n$ are passed to the tree along with features $x_1, \ldots, x_p$. The goal is to predict the residuals. We repeat this process many times by producing a tree with residuals, generating new residuals with those predictions.

combine for pred −1

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

Support vectors are the points inside the margins. They are the points that dictate where the margins are.



✓

−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.

We model each class against all others. Eg classes 1, 2, 3: we model 1 vs. 2/3, 2 vs. 1/3, 3 vs. 1/2. So 3 SVM models. To make a prediction, we pass the observation to each and assign it to the class with largest absolute value.

-3

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

kmeans doesn't find the global minimum, it only finds a local minimum. Thus we run it multiple times since the local minimums will change from iteration to iteration. We also randomly assign each observation to a cluster which will change from iteration to iteration.

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

-2

Single looks at the ~~largest~~ pairwise distance between observations from 2 classes. It computes all pairwise then finds the largest one. Black line represents the dissimilarity between red cluster and blue.

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

Biplot is useful when examining the principal components in PCA. ~~How much of the variation was captured by each component.~~

-4

-9