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?

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!

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.

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

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

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

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
    b. kNN models generally require you to standardize your predictors
    c. The number of trees we use in a random forest model is important because we can overfit with too many trees.
    d. When using BART we need to remove the first few prediction models.
    e. SVM models can only be used in classification tasks.
    f. KMeans clustering does not necessarily create the same clusters in each run of the algorithm.
    g. Hierarchical clustering requires you to know the 'true' underlying groupings to use it effectively.
    h. In a standard multilayer neural network, all inputs are 'connected to' all first level activations.
    i. KNN provides a discriminant for classifying our observations
    j. The Naive Bayes provides a discriminant for classifying our observations

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?

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

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

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

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

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

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

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

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