Final Exam Key

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 regression tree breaks the predictor space up into rectangular regions and then models the response as a constant over those 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!  
     
   We try to minimize the sum of squared errors over the two possible regions. That is, we look at the sum of squares of all response values in one region against that mean in that region and add that to the similar term for the second region. To determine the best split, we look at all possible split points for X1 and for X2 and choose the one that minimizes this (sum of) sums of squares.
3. We discussed two ways to do ‘early stopping’ in a regression or classification tree. What are those two methods?  
     
   Restrict the tree depth (the total number of splits down a side of the tree) or restrict the number of observations allowed in any terminal node.
4. True or False questions:
   1. Random forest and bagged tree models generally require you to standardize your predictors (F)
   2. kNN models generally require you to standardize your predictors (T)
   3. The number of trees we use in a random forest model is important because we can overfit with too many trees. (F)
   4. When using BART we need to remove the first few prediction models. (T)
   5. SVM models can only be used in classification tasks. (F)
   6. KMeans clustering does not necessarily create the same clusters in each run of the algorithm. (T)
   7. Hierarchical clustering requires you to know the ‘true’ underlying groupings to use it effectively. (F)
   8. In a standard multilayer neural network, all inputs are ‘connected to’ all first level activations. (T)
   9. KNN provides a discriminant for classifying our observations (FALSE)
   10. The Naive Bayes provides a discriminant for classifying our observations (FALSE)
5. Why do random forests for a regression task generally improve prediction over the basic bagged tree model?  
     
   If a really strong predictor exists then the first split of each tree tends to be similar. Generally, predictions from the trees are more correlated when this is the case. By using a random subset of the predictors at each split, we get trees that are less similar. When averaging the tree predictions to get our final prediction we get a larger decrease in the variance – leading to a smaller MSE.
6. What are the three most common tuning parameters associated with a boosted tree model?  
     
   The number of splits or terminal nodes to use, the learning rate, and the number of trees to fit.
7. Describe the algorithm for fitting a basic boosted regression tree model.  
     
   First we set our predictions to 0 and the residuals become the response values. We then fit a tree with a given number of splits. We update our prediction by moving in the direction of the current model’s predictions (we include a learning rate parameter here). We then update the residuals. This process is repeated B times!
8. When fitting a support vector machine model for classification, what are support vectors?  
     
   The support vectors are the points from the data (observations) that fall on the margin or violate the margin.
9. 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 compute all pairwise SVM classifiers. We then create the prediction using each classifier and use a majority vote to classify an observation.
10. Why do we often run the kmeans clustering algorithm multiple times?  
      
    We don’t know if we are going to get a ‘global’ minimum when we run the algorithm. By running it multiple times and taking the best run, we are attempting to get the best run we can.
11. When doing hierarchical clustering, how does the ‘single’ linkage create a dissimilarity measure?  
      
    Single linkage uses the smallest pairwise dissimilarity between all points in two clusters to determine the dissimilarity.
12. What is a biplot and how can it be useful?  
      
    A biplot is a scatterplot of the observations across two PCs. We can map the original variables onto the biplot and attempt to understand their relationship to each PC. This may lead to a reasonable interpretation of each PC.
13. In a standard multilayer feed-forward neural network, what are two common activation functions?  
      
    ReLU and sigmoid (softmax, etc.)
14. What task is a Recurrent neural network well-suited for?  
      
    text and time-series data
15. Consider the piecewise polynomial regression model. Here we define our knots to be , …, and use the indicator functions

* in our regression equation given by

Suppose we have observations and we fit the model.

1. What is the estimate of in this model?  
     
    corresponds to the case when all the indicator functions are 0. This implies that our observation must fall in the region less than . is then the mean of all responses corresponding to these observations.
2. What is the estimate of in the model?  
     
    corresponds to the slope on the indicator that is between and . That means our observation must fall in this region. is then the difference in the mean of all responses corresponding to these observations and the intercept.
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.  
     
   Here we would first take the full data set and split it into a training and test set (likely 80/20 or 70/30).  
     
   Using the training set only we’ll select our tuning parameters for the two models.  
     
   For ridge regression, we create a tuning grid the penalty term (). We take a (non-parametric) bootstrap sample from the training data. Let be the number of observations in the training data. We randomly select observations from the training data with replacement. On this bootstrap sample we fit our model with one value of our tuning parameter. As we sampled with replacement, some observations were not used in the fitting of the model. These are called out-of-bag observations and we can use these observations to obtain predictions on unseen data. We find our metric on those prediction. This process is repeated B times for that tuning parameter value. We combine the test errors into one training metric for this tuning parameter value.  
     
   We now repeat this process for the next value of the tuning parameter and so on. We choose the tuning parameter for the ridge regression model that has the lowest test MSE found through this process.  
     
   For the KNN model, the process is similar. We create a grid of values for k. We then create a bootstrap sample and fit the model to that for a specific value of k. We test on the out-of-bag observations. We repeat many times and combine the test errors into one training metric for this tuning parameter value.  
     
   We repeat the process for every value of k. We choose the k that minimizes this test MSE found in this way.  
     
   We now fit our RR and KNN models to the training data with the specified tuning parameter values. These models are then tested on the test set and we pick an overall best model.  
     
   This best model is then fit to the entire data set!