Final Exam Practice Questions

The final exam is cumulative. You should go back through the exam 1 and 2 practice and actual exam questions.

Your actual exam will have questions similar to those below with space below each question to give answers. I’ve tried to emulate that here but I haven’t taken the time to print the questions out and hand write the answers, which means there may be too much or too little space in this document. I’ll make sure an appropriate amount of space is given on the actual exam (which can also help you gauge how detailed of an answer to give!)

* Please note that the example questions below are not exhaustive!
* You may notice there are no programming questions below (that is, no R or python syntax at all)
* There are some pseudo code questions. Here you are writing out the logic of the process and how you would go about doing it within a programming language without worrying about the syntax or the process.
* There is very little calculation required for these questions. For most answers that involve output and reading/using it, you do not need to simplify calculations.
* If you have any other questions about the content or structure of the exam, please post to the discussion forum!

1. What is the standard way for a classification tree to model the response over a particular region? That is, given a particular leaf of a classification tree, how would we make a classification over that region?  
     
   Usually we use the majority rule as the classification. That is, we use the most prevalent class as the classification.
2. What issue can occur if we grow a very large regression or classification tree? If we’ve grown a large tree, how can we try to deal with this issue?  
     
   We may overfit to the data! The tree wouldn’t generalize well. We can use pruning to drop off leaves that are deemed not as useful by some criterion (such as cost-complexity).
3. We discussed the idea of variable importance in (single) CART models and ensemble CART models such as the random forest model. Usually, how do we determine the variable importance?  
     
   We can keep track of the overall reduction in our RSS attributed to each variable. Each time the variable is used in the model we can look at the reduction in the RSS and sum those up to get a total measure of feature importance.
4. When fitting a classification tree, we discussed three major node impurity measures. Name these three! (One turned out not to be great for actually determining the splits.)  
     
   Misclassification rate, Gini index, and negative log loss (or cross-entropy or deviance).
5. True or False questions:
   1. CART models generally require you to standardize your predictors (F)
   2. Boosted tree models generally require you to standardize your predictors (F)
   3. Regularized MLR and logistic regression models generally require you to standardize your predictors (T)
   4. The number of trees we use in a boosted model is important because we can overfit with too many trees. (T)
   5. In deep learning models we still need to be worried about overfitting. (T)
6. Describe what (non-parametric) bootstrapping is and how it is used to fit a random forest model for a regression task. Then describe how out-of-bag observations can be used as a prediction set.   
     
   A non-parametric bootstrap is where we resample from the data with replacement. We take a sample the same size as the original sample size. As we sample from the data with replacement, we then have some observations not included in our bootstrap sample. These are called out-of-bag observations. We repeatedly create bootstrap resamples and, for each resample, we fit a tree to the data. At each branch of the tree we randomly select a certain number of predictors to consider for that split (this is a tuning parameter for the model). Once each tree is fit, we then use it to predict on the out-of-bag observations. The predictions are then averaged to create an overall prediction.
7. What difference does stochastic gradient boosted trees do compared to a standard boosted tree model?  
     
   We randomly sample a subset of the data with replacement at each iteration of the boosting algorithm.
8. Name three advantages the XGBoost models provide over basic boosted models.  
     
   Optimized for distributed computing, regularization, early stopping, different loss functions, saving the existing model, different base learners.
9. We discussed using Bayesian Additive Regression Trees. What are the three common ‘perturbations’ that are used in these models?  
     
   Using the same shape of the tree but using different predictions. Using a pruned tree. Using a tree with additional branches.
10. Suppose we are doing a classification problem (two classes) and have two continuous predictors. What is meant by the term separable data here?  
      
    If we have separable data then we can fit a ‘straight line’ the perfectly separates the classes.
11. When using the maximal margin classifier, how do we try to determine the ‘optimal’ separating hyperplane? What is a disadvantage of selecting the hyperplane with this method?  
      
    We try to maximize the ‘margin’ which is the minimum distance between the hyperplane and the data points. The issue with his is, since we rely on a few points to define where the margin is, small changes in the data can lead to very different hyperplanes.
12. When we don’t have separable data in a classification problem, we said we could use the support vector classifier. In this setting we described ‘slack variables’ associated with each observation. These variables took on 0, a value between 0 and 1, or a value greater than 1. Describe the situation where each of these values occur.  
      
    The slack variables take on 0 if the ith observation is located on the correct side of the margin. The value is between 0 and 1 if the observations falls in the margin but is still on the correct side of the hyperplane. The value is greater than 1 if it falls on the wrong side of the hyperplane.
13. When we wish to apply the SVM model to a classification task with more than two levels, we discussed the one-versus-all approach. Describe how this SVM model works.  
      
    We compare a given class against all the remaining classes pooled together. We classify a new point by finding the model that gives the largest value (in absolute sense) and use that as the classification.
14. In SVM models, what is meant by a kernel function?  
      
    Kernel functions are the functions that we use to obtain (possibly) non-linear transformations of the data.
15. When doing unsupervised learning we discussed the KMeans algorithm for clustering. Describe the objective function this algorithm attempts to minimize for a given number of clusters, k. You don’t have to write out the math here if you prefer to use words.  
      
    We first assign each observation to one of the k clusters. We then look at the distance between each of our observations in the cluster, sum those distances up, and divide by the number of observations in the cluster. We add these within cluster variations together and then attempt to minimize that quantity.
16. When doing hierarchical clustering, what is meant by the term agglomeration?  
      
    Doing clustering from the bottom up.
17. When doing hierarchical clustering, how does the ‘complete’ linkage create a dissimilarity measure?  
      
    Complete linkage uses the largest pairwise dissimilarity between all points in two clusters to determine the dissimilarity.
18. When doing hierarchical clustering, how does the ‘average’ linkage create a dissimilarity measure?  
      
    Average linkage finds all pairwise dissimilarities between each point and the other cluster and averages across all of those to create the dissimilarity.
19. What is our major goal when using the Principal Components Analysis technique and how does PCA accomplish this goal?  
      
    We attempt to reduce the dimension of our data. That is, we take our p predictors (or variables) and try to find linear combinations of the predictors that account for as much variability in the data as we can. Each subsequent linear combination must be orthogonal to the previous PCs. We hope that m < p PCs can account for the majority of the variability in the original data.
20. What task is a convolutional neural network well-suited for?  
      
    Image recognition
21. In a convolutional neural network, what is a filter layer and what is the idea of a pooling layer?  
      
    A filter layer attempts to take a part of an image and extract information from it. For instance, we might consider the differences in values across a 3x3 window of an image. A pooling layer attempts to summarize a block, such as using the maximum value over that block. This reduces down the size of our data.
22. In a Recurrent neural network, what is meant by a bidirectional RNN?  
      
    A bidirectional RNN looks at the sequence of text (or data) in both a ‘forward’ and ‘backward’ sense. Although we might not be able to pick up patterns in this way, the model may.