DS 303: Final Exam

12:00 - 2:15 pm, Dec. 13, 2021

General Instructions:

1. On your solution sheet legibly write your:
   - Name
   - Student ID
   - **Signature** certifying that all the work in this exam is your own original work and you have upheld Iowa State University's policy on academic dishonesty. You agree not to disclose the content of this exam to students not in this class.
2. You have between 12:00 pm to 2:15 pm to finish the exam and upload it to Canvas.
3. This midterm is open note/open book/ open internet. However, you cannot consult with anyone else on the exam or solicit any help for the exam - **all work must be your own**. Any violations of this course's academic policy will result in an automatic 0.
4. Upload your solutions and `R` script to Canvas **before** 2:15 pm. **Make sure you upload all your solutions**. Any work not uploaded will not be counted. **Any exams not submitted by the deadline will receive a 0. No exceptions.**
5. The exam is made up of 2 parts: a conceptual review and a coding component. The exam is **3 pages**.
6. If you get stuck on one problem, move on and come back to that problem at the end.
7. I will be available via Zoom from **12 - 2 pm** for any questions that may come up.

Part 1: Concept Review (18 points)

State whether the following statements are True or False and **briefly justify.**

a. For a given data set, we can directly calculate the bias and variance of a regularized regression model to see whether or not the decrease in variance is enough to offset the increase in bias. Based on this, we can choose an optimal λ.

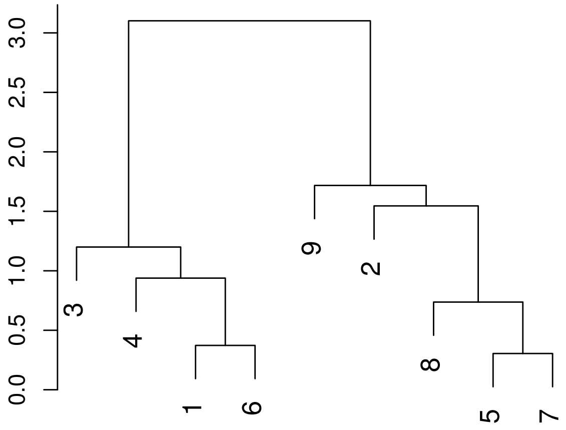
b. When choosing the optimal degree for a polynomial regression model, we can implement cross-validation on the entire dataset and choose the degree d that obtains the smallest cross-validated MSE.

c. We split a data set into a training and test set and train a classifier on the training set. If the true decision boundary is linear, we would expect QDA to perform better than LDA on the training set.

d. The amount of variance reduction we can achieve from using bagging or random forest depends on the correlation between the bagged trees.

e. Since ridge regression always returns the full model (with all p predictors), its test MSE will always be smaller than that of Lasso.

f. In the dendrogram below, we can say that observations 3 and 4 are quite similar to each other:



g. (Bonus) In k-means clustering, the total within-cluster variation is guaranteed to decrease (or stay the same) for each step of the k-means algorithm.

Part 2a: Classification (22 points)

We will use a heart disease data set for this problem. To see information on the data and how to load it into R, check the `heart.R` file. You can directly download the data from Canvas or read the data from the website directly. To save time, you may directly copy/paste raw `R` output. Split the dataset so that the first 100 observations are the test set and the remaining observations are the training set. The goal is to carry out classification of `chd` using all other variables as predictors.

a. Train LDA on the training set with `chd` as the response. Use all other variables as predictors. What is the threshold that will minimize the overall misclassification rate? Report this threshold. Using this threshold, report the confusion matrix and misclassification rate on the test set.

b. Let's implement random forest on the training set with `chd` as the response and all other variables as predictors. Set the number of trees to bootstrap to be 500. Select the optimal m (number of predictors to be considered at each split) using 5-fold cross-validation. Before you create your 5-folds, `set.seed(1)` so that we all get the same answers. The possible m values to consider are $m = 2, 3, 4, 5$. Report the 5-fold CV error for each m and report the m you decide to use.

c. Implement random forest on the training set with `chd` as the response and all other variables as predictors. Set the number of trees to bootstrap to be 500 and use the optimal m found in part (b). Report the misclassification rate on the test set. How does this compare to the out-of-bag error estimation from the training set?

d. Based on our random forest from part (c), what is the $\hat{P}(Y = 1|X)$ for the 2nd observation in our test set? Report that probability here. (Hint: use `type = ‘prob’`).

e. Bootstrap the standard error of $\hat{P}(Y = 1|X)$ for the 2nd observation in our test set. You can bootstrap from the entire dataset. To save time, you can reduce the number of trees to be 25. Report the standard error here.

Part 2b: Clustering (10 points)

We will continue working with the heart disease dataset. We want to carry out clustering so we will ignore the response `chd`.

a. We want to cluster the dataset into two groups: those with a risk for CHD and those with a low risk for CHD. What type of clustering algorithm would be appropriate here: k-means or hierarchical clustering? Explain.

b. Should we scale our features for this application? Explain your reasoning.

c. Based on your answers from (a) and (b), implement clustering using all of the variables except `famhist` (and of course `chd`). Report a table that compares your cluster labels to the actual labels for `chd` from our dataset. Comment on what you observe.

End of Exam.