API 222 Problem Set 2

Machine Learning and Big Data Analytics

Due Oct. 13 by Midnight EST

Conceptual Questions

1. The table below contains a training dataset of 6 observations, 3 predictors and 1 qualitative outcome variable. Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 1$ using K-nearest neighbors. **(2 pts)**

Obs.	X_1	X_2	X_3	Y
1	0	2	2	Orange
2	2	1	2	White
3	0	0	0	White
4	0	1	1	Black
5	0	1	2	White
6	1	1	3	Black

   (a) Compute the Euclidean distance between observation 3 and the test point, $X_1 = X_2 = X_3 = 1$

   (b) Using Euclidean distance, what is our prediction for an observation with $X_1 = X_2 = X_3 = 1$ with K = 1? Why?

   (c) Using Euclidean distance, what is our prediction with K = 3? Why?

   (d) Suppose the table looked like this instead (i.e. a regression problem). What would the prediction be for K=3 if the test point is still, $X_1 = X_2 = X_3 = 1$?

Obs.	X_1	X_2	X_3	Y
1	0	2	2	0
2	2	1	2	3
3	0	0	0	1.5
4	0	1	1	0.5
5	0	1	2	2.5
6	1	1	3	1

2. Let's say I have a model with 30 potential covariates. How many potential variants on the models can I have? What does this imply about the tradeoff between forward/backward selection over best subset selection. **(0.5 pts)**

3. If the underlying data is highly linear, we would expect QDA to outperform LDA. True or False? **(0.5 pts)**

4. We have a dataset of genetics sequencing outcome, with 30 observations and 4000 variables. You are trying to determine the best method for regression analysis. Colleague A is advocating for KNN, Colleague B is advocating for linear regression, Colleague C thinks Colleague A and B are both wrong. Who should you side with? **(1 pt)**

5. We have a dataset of genetics sequencing outcome, with 3,000 observations and 40,000 variables. You are trying to determine the best method for classification analysis. Colleague A is advocating for QDA but Colleague C is worried. Why might she be concerned? **(1 pt)**

6. You have a dataset that is all dummy variables (i.e. 0/1 categorical variables). If you want to use a linear decision boundary, would you expect LDA or a logistic regression to perform better? **(0.5 pts)**

7. What is the shrinkage penalty for ridge regression? **(0.5pts)**

8. What is the shrinkage penalty for lasso regression? **(0.5pts)**

9. How do the different shrinkage penalties influence variable selection for lasso vs ridge? **(0.5pts)**

Data Questions

For the next part, please use the `Fatalities` data set from the `AER` package. Consider the prediction problem where you want to predict number of single vehicle fatalities given all other variables available in the data set.

1. How many observations have missing values for at least one feature? Drop those observations for now. **(1 pt)**

2. Which variables are categorical variables? How many classes do each of these categorical variables have? **(1 pt)**

Now, consider the prediction problem where you want to predict number of single vehicle fatalities. (`Fatalities`) given all other variables available in the data set.

3. Convert the categorical variables to indicator variables (also called "dummy" variables) and run a linear regression. What is the adjusted R^2? **(1 pt)**

4. Run lasso regression with cross-validation using the canned function `cv.glmnet` from the package `glmnet`. You can use the λ sequence generated by grid function we used in section notes 4. In order to receive credit for this question, make the line immediately preceding this command say `set.seed(222)` and run the two lines together. Please report all numbers by rounding to three decimal places. **(2 pts)**

    - Which λ had the lowest mean cross-validation error for 5 fold cross validation?
    - What was the cross-validation error?
    - What was the standard error of the mean cross-validation error for this value of λ?
    - What was the largest value of λ whose mean cross validation error was within one standard deviation of the lowest cross-validation error?

5. Using the same data, implement your own 5-fold cross-validation routine for KNN for $k = 1, ..., 20$ (e.g. write the cross-validation routine yourself rather than using a canned package). In the 90s, a popular policy response to high rates of alcohol related fatalities was to increase taxes on alcohol. Consider the prediction problem of predicting beer tax (tax on cases of beer) using all of the other variables. Include the snippet of code you wrote here. It should not exceed 20 lines. Which k is best according to CV? **(2 pts)**

6. Plot mean cross-validation MSE as a function of k. Label the y-axis "Mean CV MSE" and the x-axis "k". **(1 pt)**