PS132B Problem Set 4

Due 12:29PM Thursday March 2, 2023

Please submit this assignment by uploading your R Markdown code file (`.Rmd`) AND either an html (`.html`) or pdf (`.pdf`) output onto bCourses before the due time, with easy-to-recognize file names (e.g., `pset4_KirkBansak.Rmd`). Your homework will be graded based on completeness, accuracy, and readability of both code and written answers.

Unless a package is explicitly mentioned in the problem set, you should not use any `R` packages beyond the following packages that are automatically loaded by default when you open up `RStudio`: `base`, `datasets`, `grDevices`, `graphics`, `methods`, `stats`, `utils`. In addition, you may also use `ggplot2`.

The point allocation in this problem set is given by:

Q1.1	Q1.2	Q1.3	Q1.4	Q1.5	Q1.6	Q1.7	Q1.8	Q1.9	Q2.1	Q2.2	Total
5	15	10	5	5	5	5	5	15	5	25	100

Q1: Training vs. Test Error

Our goal in this exercise is distinguish between in-sample and out-of-sample model fit. To do so, we have taken a real dataset and added several fake predictors, which are generated specifically to be completely unrelated to the response variable (i.e. noise variables that have no systematic relationship with the response variable). We will explore how their inclusion impacts our fitted regression, in terms of how well the regression fits training data vs. held-out test data.

We will be using the data set `vote92plus.csv`, which is a modified version of survey data containing self-reports on political attitudes and choices during the 1992 U.S. Presidential election, from the 1992 American National Election Studies. We will use these data to predict a respondent's ideological difference from the Democratic candidate, Bill Clinton.

The data set contains two sets of variables:

- Real variables:

  `clintondist`: a numeric ideological difference score, which measures the difference between respondent's self-placement on a scale measure of political ideology and the respondent's placement of the Democratic candidate, Bill Clinton. **We will treat this as our response variable.**

  `vote`: a categorical variable denoting who the respondent voted for.

  `dem`: an indicator variable, 1 if the respondent reports identifying with the Democratic party, 0 otherwise.

  `rep`: an indicator variable, 1 if the respondent reports identifying with the Republican party, 0 otherwise.

  `female`: an indicator variable, 1 if the respondent is female, 0 otherwise.

  `persfinance`: a numeric variable, -1 if the respondent reports that their personal financial situation has gotten worse over the last 12 months, 0 for no change, 1 if better.

  `natlecon`: a numeric variable, -1 if the respondent reports that national economic conditions have gotten worse over the last 12 months, 0 for no change, 1 if better.

- Fake variables:

  `fake1`

  `fake2`

  `fake3`

  `fake4`

  `fake5`

1) Begin by running the following code. Explain what each line is doing and its purpose.

```
dat <- read.csv("vote92plus.csv")
set.seed(5000)
k <- sample(1:nrow(dat), round(nrow(dat)*2/3), replace = FALSE)
train.dat <- dat[k,]
test.dat <- dat[-k,]
```

2) Using a multiple linear regression, and only the training data, regress the response variable (`clintondis`) on all of the **real** predictor variables. Call this Model A, and do the following:

   (a) Compute the mean squared error of the fitted regression on the training data set.

   (b) Use the fitted regression to make predictions on the test (held-out) data, and compute the mean squared error for the test set.

3) Now perform the same process as in (2) but add the variables `fake1`, `fake2`, `fake3`, `fake4`, and `fake5` to the regression in addition to all of the real predictors. Call this Model B. Again, compute the mean squared error for both the training and the test sets.

4) Do the same thing, adding in $\texttt{fake1}^2$, $\texttt{fake2}^2$, $\texttt{fake3}^2$, $\texttt{fake4}^2$, and $\texttt{fake5}^2$ (i.e. the squared versions of the fake variables) in addition to everything included in Model B. Call this Model C.

5) Once more, do the same thing, now adding in $\texttt{fake1}^3$, $\texttt{fake2}^3$, $\texttt{fake3}^3$, $\texttt{fake4}^3$, and $\texttt{fake5}^3$ (i.e. the cubed versions of the fake variables) in addition to everything included in Model C. Call this Model D.

Note) You will now have fit four regressions, and for each regression, you will have computed the mean squared error for the training set and for the test set.

6) Produce a table that reports the training set mean squared error for Models A, B, C, and D.

7) Produce a table that reports the test set mean squared error for Models A, B, C, and D.

8) Explain your results. What accounts for the trends you see in the mean squared error for the training vs. test sets?

9) Now employ a different random split of the data into training and test sets (allocating the same proportion of data to each set as previously) by setting the seed to 202—i.e. `set.seed(202)`—and reproduce new versions of the tables in (6) and (7) using this new split. How do the results compare to the results from your original split? Are they the same or different? Why?

Q2: Conceptual Questions

1) Which of the following statements is correct (justify your answer)?
The lasso, relative to ordinary least squares regression, is:

   - More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
   - More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.
   - Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
   - Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

2) Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

for a particular value of λ. For parts (a) through (e), indicate which one of i. through v. is correct. Justify your answer.

   a) As we increase λ from 0, the training mean squared error will:

      i. Increase initially, and then eventually start decreasing in an inverted U shape.

      ii. Decrease initially, and then eventually start increasing in a U shape.

      iii. Steadily increase.

      iv. Steadily decrease.

      v. Remain constant.

   b) Repeat (a) for test mean squared error.

   c) Repeat (a) for variance.

   d) Repeat (a) for (squared) bias.

   e) Repeat (a) for the irreducible error.