STAT 6340 (Statistical and Machine Learning, Spring 2020)

Mini Project 5

Instructions:

- Due date: Apr 13, 2020.
- Total points = 30.
- Submit a typed report.
- It is OK to discuss the project with others and to search on the internet, but each student must write their own code and answers. If your submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job.
- You must use the following template for your report:
  Mini Project #
  Name
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.
- Section 1 of the report must be limited to **five** pages. Also, only those output should be provided in this section that are referred to in the report.

---

1. Consider the `state.x77` dataset available from `datasets` package in `R`. The 50×8 data matrix gives statistics on 8 features of the 50 states in the US from 1977.

   (a) Perform an exploratory analysis of the data.

   (b) Do you think standardizing the variables before performing the analysis would be a good idea?

   (c) Regardless of your answer in (b), standardize the variables, and perform a principal components analysis (PCA) of the data. Summarize the results using appropriate tables and graphs. How many PCs would you recommend?

   (d) Focus on the first two PCs obtained in (c). Prepare a table showing the correlations of the standardized variables with the components and the cumulative percentage of the total variability explained by the two components. Also, display the scores on the two components and the loadings on them using a biplot. Interpret the results. Can you identify, for example, a "southern" component?

2. First, read carefully about *latent semantic analysis*—an application of PCA—from a PDF of the same name posted on eLearning. This is an excerpt from the book *Advanced Data Analysis from an Elementary Point of View* by C. R. Shalizi. The excerpt presents an analysis of a dataset consisting of 102 news stories from the *New York Times*. We will divide the dataset into a training and test dataset consisting of 80 and 22 news stories, respectively. These are available as `nyt.training.csv` and `nyt.test.csv`.

(a) Perform the analysis described in the excerpt but using only the training data. Your results won't match exactly because you are working with a subset of the data. However, your conclusions must be reported along the same lines and must include an analog of Figure 15.6 for the training data. Additionally, comment on whether the total number of PCs is consistent with what you expect.

(b) What does your analysis in (a) say about using only the first two PCs of a story to predict whether the story is about art or music?

(c) Use the training data to fit a logistic regression model that predicts the class of a story using its scores on the first two PCs from (a) as predictors. Report estimate of the training error rate and superimpose the decision boundary in the figure from (a). Comment on the result.

(d) Compute scores on the first two PCs of the stories in the test data and use them to get their class predicted by the model in (c). Compute test error rate and class-specific error rates. Comment on the results. [*Note*: You may use `predict` function to compute the PC scores for test data. If you compute them directly, be sure to take into account of the centering done in (a).]

(e) The above analysis used only two PCs but we can work in general with M PCs. Right? How would you choose M? How does this approach compare with PCR?

3. Consider `College` data from the ISLR package. These are described on page 54 of the textbook. We would like to predict the number of applications received using the other variables. All data will be taken as training data.

(a) Fit a linear model using least squares and report the LOOCV estimate of the test error.

(b) Fit a ridge regression model with λ chosen by LOOCV. Report the estimated test error.

(c) Fit a lasso model with λ chosen by LOOCV. Report the estimated test error.

(d) Fit a PCR model with M chosen by LOOCV. Report the estimated test error.

(e) Fit a PLS model with M chosen by LOOCV. Report the estimated test error.

(f) Compare the results from the five models. Which model(s) would you recommend? Justify your conclusion.