Stat 427/627, Statistical Machine Learning

Homework 4

Due: Friday, June 14, 2024

Contents

- This assignment covers VIF, ridge regression, lasso, PCA and PLS.
- Although B-spline and smoothing methods are not included this assignment, be sure to review the smoothing methods in the notes and the lab.
- Finish part (a) - (d) after Tuesday's class, and the rest after Thursday's class.
- 32 Points

Question	a	b	c	d	e	f	g	h	i	Total
427	4	4	4	4	4	4	4	4		32
627	4	4	4	4	4	4	3	3	2	32

1 College Applications (p.286, 44pts)

Predict the number of applications received based on the other variables in the `College` data set in {`ISLR2`}.

Use the following code to split the data set randomly. Use the subset `college.data` for the most of the analysis. Set aside `holdout.data` data until the last 2 sub-questions.

```
library(ISLR2)
data("College")

my.college <- College[-484, ] # remove an extreme case.
#my.college <- College[College$Apps <=16000, ] # remove several extreme case.
train.pct <- 0.78
set.seed(2024)
Z <- sample(nrow(my.college), floor(train.pct*nrow(my.college)))
college.data <- my.college[Z, ]
holdout.data <- my.college[-Z, ]
```

Recall that we fit the data in homework 1.

- we fit the **full** model with all 17 predictors. That is:

  ```
  college.lmF <- lm(Apps ~ ., data = college.data)
  college.lmF
  ```

- Using stepwise selection, AIC is the smallest with 12 predictors:

Apps ~ Private + Accept + Enroll + Top10perc + Top25perc + F.Undergrad + Outstate + Room.Board + PhD + perc.alumni + Expend + Grad.Rate

- Using the best-subset algorithm. BIC is the smallest with 7 predictors:

  Apps ~ Private + Accept + Top10perc + Outstate + Room.Board + PhD + Expend

Continue working on this data set and the above models.

(a) (4 pt) Compute the variance inflation factor and comment on the severity of the collinearity of the data. Why is "collinearity" a concern, even if the model is correct?

(b) (4 pts) Evaluate prediction accuracy of your selected models based on AIC and BIC. Estimate the prediction mean squared error by 10-fold cross-validation. Recall that `glm( ..., family=gaussian)` fits linear regression and its outcome can be used in `cv.glm()` (`boot` package) for cross-validation.

(c) (4 pts) Consider the **full** model with all 17 predictors (reminder: use data frame `college.data` you recreated at the beginning). Use functions in package `glmnet` to fit a ridge regression.

- Select λ chosen by (default 10-fold) cross-validation.
- Plot the results of the cross-validation.
- Report the esimated MSE of the model based on your selected λ.

(d) (4 pts) Consider the **full** model with all 17 predictors (reminder: use data frame `college.data` you recreated at the beginning). Use functions in package `glmnet` to fit a LASSO regression.

- Select λ chosen by (default 10-fold) cross-validation.
- Plot the results of the cross-validation.
- Report the esimated MSE of the model based on your selected λ.

(e) (4 pts) Fit a PCR model on `college.data`, with M (the number of principal components) chosen by cross validation. Prepare a validation plot. Report the estimated test error (MSE), along with the value of M selected by cross-validation.

(f) (4 pts) Fit a PLS (partial least squares) model on `college.data`, with M (the number of principal components) chosen by cross validation. Prepare a validation plot. Report the estimated test error (MSE), along with the value of M selected by cross-validation.

(g) (3 pts) Summarize and comment on the results obtained from the following models. Recommend a model, and justify your choice.

Method	Number of predictors	Estimated prediction MSE
Least Squares 1: model with the smallest AIC		
Least Squares 2: model with the smallest BIC		
Ridge Regression (lambda.min)		
Lasso (lambda.min)		
Lasso (lambda.1se)		
PCR		
PLS		

(h) (3 pts) Apply the above models to the hold-out data `holdout.data` that we created at the beginning. Which model wins this contest in terms of prediction accuracy? (This should be the first time you use observations in `holdout.data` data frame.)

(i) (2 pts) **Stat-627** Compare your estimated prediction MSE from the training data `college.data` (part i) and the resulting MSE from the `holdout.data` (part j). Is there anything "surprising" that worth investigation? If yes, what are the possible causes? (Note. It is not surprising to see a tuned "best" model not to perform the best on the testing data.)

—— This is the end of HW 4. ——