Assignment 4

CS 750/850 Machine Learning

- **Due**: February 24th at 11:59PM
- **Submisssion**: Turn in both a **PDF** and the **source code** on MyCourses
- **Questions**: Piazza and Office hours: *Marek*: Wed 1:30-3:00pm, Soheil: Mon 2-4pm, Xihong: Thu 1:30-3:30pm
- **Extra credit**: Especially good questions or helpful answers on Piazza regarding the assignment earn up to 5 points extra credit towards the assignment's grade.

Problem 1 [25%]

In this exercise, we will predict the number of applications received using the other variables in the College (`ISLR::College`) data set.

1. Fit a linear model using least squares on the training set, and report the **test** error obtained.
2. Use best subset selection with cross-validation. Report the test error obtained.
3. Fit a ridge regression model on the training set, with λ chosen by cross-validation.
4. Fit a lasso model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.
5. Briefly comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these approaches?

Problem 2 [25%]

We will try to predict per capita crime rate in the `Boston` dataset.

1. Try out best subset selection, the lasso, ridge regression, and PCR on this problem. Present and discuss results for the approaches that you consider.
2. Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, cross-validation, or some other reasonable alternative, as opposed to using training error.

Problem 3 [25%]

Suppose we have a linear regression problem with P features. We estimate the coefficients in the linear regression model by minimizing the RSS for the first p features:

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

where $p \leq P$. For parts (1) through (5), indicate which of i. through v. is correct. Briefly **justify** your answer.

1. As we increase p from 1 to P, the training RSS will *typically*:

i. Remain constant.
ii. Steadily increase.
iii. Steadily decrease.
iv. Increase initially, and then eventually start decreasing in an inverted U shape.
v. Decrease initially, and then eventually start increasing in a U shape.

2. Repeat (1) for test MSE.
3. Repeat (1) for squared bias.
4. Repeat (1) for variance.
5. Repeat (1) for the irreducible error (Bayes error).

Problem 4 [25%]

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

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j| \leq s$$

for a particular value of s. For parts (1) through (5), indicate which of i. through v. is correct. **Justify** your answer.

1. As we increase s from 0, the training RSS will *typically*:

i. Remain constant.
ii. Steadily increase.
iii. Steadily decrease.
iv. Increase initially, and then eventually start decreasing in an inverted U shape.
v. Decrease initially, and then eventually start increasing in a U shape.

2. Repeat (1) for test RSS.
3. Repeat (1) for (squared) bias.
4. Repeat (1) for variance.
5. Repeat (1) for the irreducible error (Bayes error).

Problem O4 [30%]

This problem can be substituted for Problem 4 above, for up to 5 points extra credit. The better score from problems 4 and O4 will be considered.

Solve Exercise *3.6* in [Bishop, C. M. (2006). Pattern Recognition and Machine Learning].