MATH 373: Introduction to Machine Learning
Homework 1

Computational Problems

For each of these problems, feel free to use R and/or Python.

1. Write a function *kfold.cv.lm()* which performs the following. This function is to be made from scratch, i.e. you cannot simply use a function for cross validation in any package.

   Input Arguments:

   - *k*: integer number of disjoint sets
   - *seed*: numeric value to set random number generator seed for reproducability
   - *X*: $n \times p$ design matrix
   - *y*: $n \times 1$ numeric response
   - *which.betas* : $p \times 1$ logical specifying which predictors to be included in a regression

   Output: *Avg.MSPE*, *Avg.MSE*

   Description: Function performs k-fold cross-validation on the linear regression model of y on X for predictors *which.betas*. Returns both the average MSE of the training data and the average MSPE of the test data.

2. Download the *College* data set from the following link:

   `http://www-bcf.usc.edu/~gareth/ISL/College.csv`

   This data describes several interesting summary characteristics of American colleges and universities in 2013, including the University of San Francisco!

   Suppose that we are curious about what factors at a university play an important role in the room and board each semester (column *Room.Board*). Answer the following questions.

   (a) Based on some research into the area, you believe that the five most important predictors for the room and board amount are
   - the number of students who accepted admission *Accept*
   - the number of students who are currently enrolled *Enroll*
   - the out of state tuition for a semester *Outstate*
   - the average cost of books per year *Books*
   - the graduation rate of the students *Grad.Rate*

   Plot a pairwise scatterplot of these variables along with the room and board cost, and comment on any trends.

(b) Use your *kfold.cv.lm()* function from the first question to run 10 - fold cross-validation on each of the $2^5 = 32$ possible regression models of *Room.Board* on the every subset of the above 5 predictors. For each model, run cross validation 100 times to get a distribution of the average MSPE. Which model would you choose? What are the estimates and standard errors of your parameter estimates? Plot a histogram of the average MSE and MSPE from the 100 runs of 10-fold cross validation for your chosen model.

(c) Run best subset selection on the linear regression model of the room and board cost on the above 5 predictors using the adjusted R^2 as your model assessment criterion. State your selected model and comment on any differences between the two chosen models. *Note: it is OK to use pre-built functions to do this.*

(d) Do the Normal assumptions hold for your chosen model in (b)? Perform formal hypothesis tests here.

(e) Perform ridge regression across a grid of λ values for a regression of room and board cost on the remaining 17 predictors in the *College* data set. Here, only set the *Private* covariate to be a factor and keep the remaining continuous variables. Plot the MSPE assoicated with each value of λ. Refine the grid of λ to identify the point of minimum MSPE. What is your final model? What is the final average MSPE? Comment on any differences between the parameter estimates estimated by ridge with those estimated by standard linear regression. *Note: it is OK to use pre-built functions to do this.*

(f) Repeat (e) using the Lasso. *Note: it is OK to use pre-built functions to do this.*

(g) Run the Elastic Net for the regression of room and board on the remaining 17 predictors with $\alpha = 0.5$ using 10-fold cross validation. What λ minimizes the cross-validation MSPE? What is your final model? What is the final average MSPE? Comment on the similarities and differences between the Lasso, Ridge and Elastic net. *Note: it is OK to use pre-built functions to do this.*

(h) Of the models selected from ridge regression, the Lasso, and Elastic Net, which model do you prefer? Discuss this in terms of variance, interpretability, inference, and prediction.

Conceptual Problems

1. In this problem we will prove the relationship between bias, variance, and MSPE in a regression problem. The proof just relies on properties of expectation. Let Y and Z be two independent random variables with means μ_Y and μ_Z, respectively. Answer the following questions.

   (a) By expansion, show that

   $$\mathbb{E}[(Z - \mu_Y)^2] = \text{Var}(Z) + (\mathbb{E}[Y - Z])^2$$

Hint: think carefully about what is needed to obtain $\text{Var}(Z)$ on the right hand side of the equality above.

(b) Use part (a) and expansion to show that

$$\mathbb{E}[(Y - Z)^2] = \text{Var}(Y) + \text{Var}(Z) + (\mathbb{E}[Y - Z])^2$$

(c) Consider setting $Y = f(X) + \epsilon$, and $Z = \hat{f}(X)$. Now use parts (a) and (b) to conclude that

$$\mathbb{E}[\text{MSPE}(\hat{f}(X))] = \text{Var}(\hat{f}(X)) + \text{Var}(\epsilon) + \text{Bias}(\hat{f}(X))^2$$

2. Suppose that we peform best subset, forward stepwise, and backward stepwise selection for linear regression on a single data set. For each approach, we obtain $p+1$ models, containing $0, 1, \ldots, p$ predictors. Explain your answers for each of the following:

(a) Which of the three methods with k predictors has the smallest *training* MSE?

(b) Which of the three methods with k predictors has the smallest *test* MSPE?

(c) (TRUE or FALSE): The predictors in the k-variable model identified by forward stepwise selection are a subset of the predictors in the $(k+1)$-variable model of forward stepwise selection.

(d) (TRUE or FALSE): The predictors in the k-variable model identified by best subset selection are a subset of the predictors in the $(k+1)$-variable model of best subset selection.

3. Suppose that we observe data (x, y) and that we fit polynomials of increasing degree M to the data. Namely, we fit models like

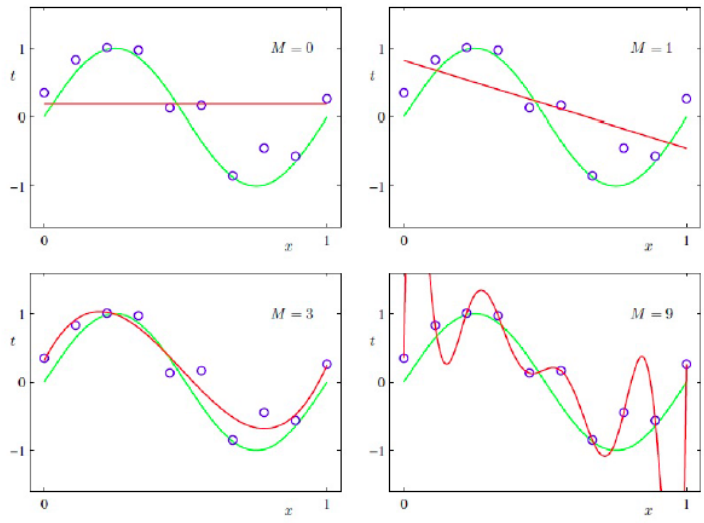
$$f(X) = \sum_{j=0}^{M} a_j x^j$$

In the plots below, we show the data (blue points), the fitted model (red line) and the true model (green line). Let $\hat{f}_M$ be the model fitted in each plot.

(a) Suppose that all of the data observed in the plots are used as training data. Suppose that we observed a new data point (X_o, y_o) *outside* the original data set. Rank the models in terms of $\text{MSPE}(\hat{f}_M)$. Explain your answer.

(b) Suppose that all data observed in the plots are used as training data. Rank the models in terms of $\text{MSE}(\hat{f}_M)$. Explain your answer.

(c) Rank the models in terms of $\text{Var}(\hat{f}_M)$. Explain your answer.



4. Consider fitting a model $y = X\beta + \epsilon$ where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

   (a) Using the normal equations for $\hat{\beta}_{OLS}$, show that

   - $\mathbb{E}[\hat{\beta}_{OLS}] = \beta$ and
   - $\text{Var}(\hat{\beta}_{OLS}) = \sigma^2(X^TX)^{-1}$

   (b) Let $Z = (X^TX)$. Using the regularization definition of Ridge Regression with tuning parameter λ, show that

   $$\mathbb{E}[\hat{\beta}_{Ridge}] = (I + \lambda Z^{-1})^{-1}\beta$$

   Comment on the bias of $\hat{\beta}_{Ridge}$ and $\hat{\beta}_{OLS}$.

   (c) Show that

   $$\text{Var}(\hat{\beta}_{Ridge}) = \sigma^2(I + \lambda Z^{-1})^{-1}Z^{-1}(I + \lambda Z^{-1})^{-1}$$

   (d) Consider an example where we have a design matrix:

   $$X = \begin{pmatrix} 1 & 0.7 \\ 1 & 0.69 \end{pmatrix}$$

   Using the formulae above, calculate $\text{Var}(\hat{\beta}_{OLS})$ and $\text{Var}(\hat{\beta}_{Ridge})$ for $\lambda = 2$ when $\text{Var}(\epsilon) = 2$. Comment on the differences between these two values, and why they might be so extreme.