STAT4001 Data Mining and Statistical Learning

Homework 2

Due on Nov.13 (Friday)

Instructions:

- Please write down **detailed** calculations and derivations to receive credits.
- Please include **all the R codes and outputs** (including numerical values, plots and so on) in your homework as **pdf form**. You may use R markdown to summarize, or you can simply print screen for ALL R codes and outputs.
- Please submit **only one single pdf file** through blackboard by 5pm on the date the assignment is due. We will only grade **ONE** pdf file, so please compress ALL your answers (derivations, R codes, outputs and so on) into ONE pdf file.
- File formats other than pdf form (both for the written part and R code, output part) will NOT be graded. Re-submission may be treated as late submission with partial marks deducted.

1. (15 marks) Ridge regression v.s. Least squares
   Given data $(y_i, x_i)_{i=1,\ldots,n}$, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, where\ \epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$, $(x_i)_{i=1,\ldots,n}$ is known and fixed.
   Least square estimate $(\hat{\beta}_0^{LS}, \hat{\beta}_1^{LS}) = \underset{\beta_0,\beta_1}{argmin} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$.
   Ridge regression $(\hat{\beta}_0^{Ridge}, \hat{\beta}_1^{Ridge}) = \underset{\beta_0,\beta_1}{argmin} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + \lambda\beta_1^2$.

   (a) Show that the least square estimate is unbiased by showing $\mathbb{E}(\hat{\beta}_0^{LS}) = \beta_0$ and $\mathbb{E}(\hat{\beta}_1^{LS}) = \beta_1$.

   (b) Show that the ridge regression estimate is biased by calculating $\mathbb{E}(\hat{\beta}_0^{Ridge})$ and $\mathbb{E}(\hat{\beta}_1^{Ridge})$.

   Hint: You may directly use some derivations in the lecture.

2. (20 marks) Invariant of linear regression to scaling, but not ridge regression without standardization

   (a) Consider the following data set: $y = (2.2, 3.3, 3.8), x = (1, 2, 3)$. Fit $y = \beta_0 + \beta_1 x$.

i. Calculate least square parameter estimates $\hat{\beta}_0^{LS}$, $\hat{\beta}_1^{LS}$ and ridge regression parameter estimates $\hat{\beta}_0^{Ridge}$, $\hat{\beta}_1^{Ridge}$ with $\lambda = 1$.

ii. Calculate $\hat{y}^{LS}$ and $\hat{y}^{Ridge}$ for x.

(b) Consider the data set in (a) with $x' = 10x$, i.e. $x' = (10, 20, 30)$.

i. Calculate least square parameter estimates $\hat{\beta}_0^{L}$, $\hat{\beta}_1^{L}$ and ridge regression parameter estimates $\hat{\beta}_0^{R}$, $\hat{\beta}_0^{R}$ with $\lambda = 1$.

ii. Compare $\hat{\beta}_0^{R}$ and $\hat{\beta}_1^{R}$ in 2b(i) with $\hat{\beta}_0^{Ridge}$ and $\frac{\hat{\beta}_1^{Ridge}}{10}$ in 2a(i) which are without scaling, also compare $\hat{\beta}_0^{L}$ and $\hat{\beta}_1^{L}$ 2b(i) with $\hat{\beta}_0^{LS}$ and $\frac{\hat{\beta}_1^{LS}}{10}$ in 2a(i) for least square.

iii. Calculate $\hat{y}^{L}$ and $\hat{y}^{R}$ for x', and compare with 2a(ii).

(Note: You will see that scaling x will have an effect on $\hat{y}$ for ridge regression, but not in least square)

Hint: You may directly use some derivations in the lecture.

3. (15 marks) Cyclic coordinate descent for LASSO
Given $f(\beta_j) = a\beta_j^2 - 2b\beta_j + \lambda|\beta_j|$, where $a > 0, \lambda > 0$. Show that when $b < -\frac{\lambda}{2} < 0, \hat{\beta}_j = \frac{2b+\lambda}{2a}$ minimizes $f(\beta_j)$.

4. (35 marks) Variance and bias for Linear regression v.s. Ridge regression
Fit the data with model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, where\ \epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$.
Least square parameter estimtates: $\hat{\beta}_0^{LS} = \bar{y} - \hat{\beta}_1^{LS}\bar{x}$, and $\hat{\beta}_1^{LS} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$

Ridge regression parameter estimtates: $\hat{\beta}_0^{Ridge} = \bar{y} - \hat{\beta}_1^{Ridge}\bar{x}$, and $\hat{\beta}_1^{Ridge} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2 + \lambda}$

For a new point x_0, calculate bias and variance for

(a) Linear regression

(b) Ridge regression

Where $bias^2 = [\beta_0 + \beta_1 x_0 - \mathbb{E}(\hat{y}_0)]^2$ and $variance = \mathbb{E}[\hat{y}_0 - \mathbb{E}(\hat{y}_0)]^2$.

(Note: You will see that compared with linear regression, the $bias^2$ for ridge is larger, but the variance is smaller. At high dimensional setting, ridge regression (and lasso) will be better because of the smaller variance.)

Hint: You may directly use some derivations in the lecture.
$Var(A + B) = Var(A) + Var(B) + 2cov(A, B)$
$Var(\alpha A) = \alpha^2 Var(A)$, *where* α *is a scalar.*

5. (15 marks) R code exercise

(a) Use the rnorm() function to generate a predictor X of length $n = 100$, $\mu_X = 0$, $\sigma_X = 1$, as well as a noise vector ϵ of length $n = 100$, $\mu_\epsilon = 0$, $\sigma_\epsilon = 0.1$.

(b) Generate a response vector Y of length $n = 100$ according to the model $Y = 1 + X + X^2 + X^3 + \epsilon$.

(c) Fit a lasso model to the simulated data, using $X, X^2, ..., X^{10}$ as predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error (i.e. Mean-Square error v.s. $log(\lambda)$) as a function of λ. Report the resulting coefficient estimates.

(d) Now re-generate a response vector Y according to the new model $Y = 1 + X^7 + \epsilon$. Again, re-fit a lasso model using $X, X^2, ..., X^{10}$ as predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error (i.e. Mean-Square error v.s. $log(\lambda)$) as a function of λ. Report the resulting coefficient estimates.

(Note: You will see that when the true data-generating model is sparser, cross-validation tends to select a sparser model.)

Hint: You may refer to the tutorial notes 'Tutorial05'.

- End -