HW3b

`Tianyunyang Wang` `N14855366` `tw1719`

1.

For **approach a**, the problem could be that when we choose the whole data as the training set, there is little space for us to testify whether this linear regressor is indeed optimal or not. We could not easily replace it when it shows a bad representation.
For **approach b**, because we choose the half of N samples randomly, it is possible that the accuracy of prediction is related to the grouping of the original data. Sometimes the number of the samples in the training set is too small to represent the attribution of the maternal sample.
For **approach c** The K-fold cross validation could efficiently avoid the effect of under-fitting and over-fitting with relatively convincible regression results. However, it is hard when we decide what is the value of K. A too small K will do less favor to our regression model when a too large K will critically increase the effort we will pay.
Therefore, when N is relatively small, I will use the hold-out method in order to avoid heavy work, and use K-fold cross validation when N is very large.
When using K-fold cross validation to predict a new sample, I will calculate the mean of K sets of parameters and combine them into a new regressor.

2.

In method 1:
$y_1 = \beta^1 X$
$y_2 = \beta^2 X$
.
.
.
$y_k = \beta^k X$
$\bar{y} = \Sigma(y_1 + y_2 + ... + y_k)/K$
$\bar{y} = \Sigma(\beta^1 X + \beta^2 X + ... + \beta^k X)/K$
$\bar{y} = \Sigma(\beta^1 + \beta^2 + ... + \beta^k)X/K$
$\bar{y} = \bar{\beta} X$
which means two methods are equivalent.

3.

(1) When N is very large, we use K-fold cross validation to establish the prediction model.

- Create a K-fold cross validation object and LASSO model
- Set a test value α
- Compute the LASSO path for the split
- Compute the mean and standard deviation over different folds
- Find the optimal α by the following steps:
  - Find the α with the minimum test MSE
  - Set mse_tgt = minimum MSE + 1 std dev MSE
  - Find the least complex model (highest alpha) such that MSE < mse_tgt

- Finally, we recompute the coefficients using all the training data at the correct alpha. The optimal subset consist of Non-zero coefficients
- With the optimal subset we got, we would be able to use the linear regression method directly with cross validation to evaluate the test error and to determine the mean regression coefficient.

(2)When N is relatively small, we use the hold-out method to establish the prediction model
All things about LASSO would be the same, however, when it comes to the linear regression, we only set two subsets among N samples. We use one of them as the training set to train our model and the other as the testing set to test the accuracy of our model.

4.

Instead of using LASSO, we may also use Ridge Regression, Filtering method or Wrapper method.

5.

(a)

$$\bar{x}_{i,j} = \frac{\Sigma x_{i,j}}{j}$$

$$\bar{x}_{i,j} = \Sigma(x^{\gamma}_{i,j} - j \cdot \bar{x}_j)/(\sigma_j \cdot j)$$

$$\because j \cdot \bar{x}_j = \Sigma x^{\gamma}_{i,j}$$

$$\therefore \bar{x}_{i,j} = 0$$

And it’s the same with $\bar{y}_{i,j}$

The variance of x $D(x_{i,j})$ could be represented as below:

$$D(x_{i,j}) = (\Sigma(x_{i,j} - \bar{x}_{i,j}))^2/j$$

$$D(x_{i,j}) = (\Sigma(x^{\gamma}_{i,j} - \bar{x}_j)/\sigma_j)^2/j$$

$$= j/j = 1$$

And it’s the same with $D(y_i)$

(b)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij}$$

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{i1} + \beta_2 \bar{x}_{i2} + \ldots + \beta_j \bar{x}_{ij}$$

$$\because \bar{y}_i = 0, \bar{x}_{11} = \bar{x}_{22} = \bar{x}_{33} = \ldots = \bar{x}_{ij} = 0$$

$$\therefore \beta_0 = 0$$

(c)

$$y^{\gamma} = \beta^{\gamma}_0 + \beta^{\gamma}_1 x^{\gamma}_1 + \ldots + \beta^{\gamma}_j x^{\gamma}_j$$

$$y^{\gamma} = \hat{y} \cdot \sigma_y + \bar{y}$$

$$y^{\gamma} = \beta[(x^{\gamma} - \bar{x})/\sigma_x]\sigma_y + \bar{y}$$

$$y^{\gamma} = \sigma_y/\sigma_x \cdot (x^{\gamma} - \bar{x}) \cdot \beta + \bar{y}$$

$$\beta^{\gamma}_0 = \bar{y} - \sigma_y/\sigma_x \cdot \bar{x} \cdot \beta_0 = \bar{y}$$

$$\therefore \beta^{\gamma} = [\bar{y}, \sigma_y/\sigma_x \cdot \beta_1, \sigma_y/\sigma_x \cdot \beta_2, \ldots, \sigma_y/\sigma_x \cdot \beta_j]$$

6.

Sometimes some data will be relatively huge compared to other data in the same data set. We have to normalize them in order to let their influence factor stay same, otherwise, when the data range changes, it is hard to tell how much influence it puts on the model.

What’s more, with mean removal, the β_0 intercept is 0 and the regularization term would be a simply L1 or L2 norm of coefficient vector.

7.

The loss function of Ridge Regression can be expressed as :

$J(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{d} |\beta_j|^2$

It tends to lead to many ‘small’ coefficients instead of eliminate them. It’s easy to solve multicolinearity with L2 penalty, but it cannot shrink parameters to zero, so it’s hard to do the variable selection.

The loss function of LASSO can be expressed as :

$J(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{d} |\beta_j|$

Although L1 penalty is hard to calculate but it shrinks some parameter to zero. With variable selection valid, LASSO is not consistent and could at most choose n variables when n is quite small. And it cannot deal with group selection as well.

8.

$J(\beta) = ||A\beta - y||^2 + \alpha||\beta||^2$

$\nabla J(\beta) = 2A^T(A\beta - y) + 2\alpha\beta$

$2(A^TA + \alpha I)\beta = 2A^Ty$

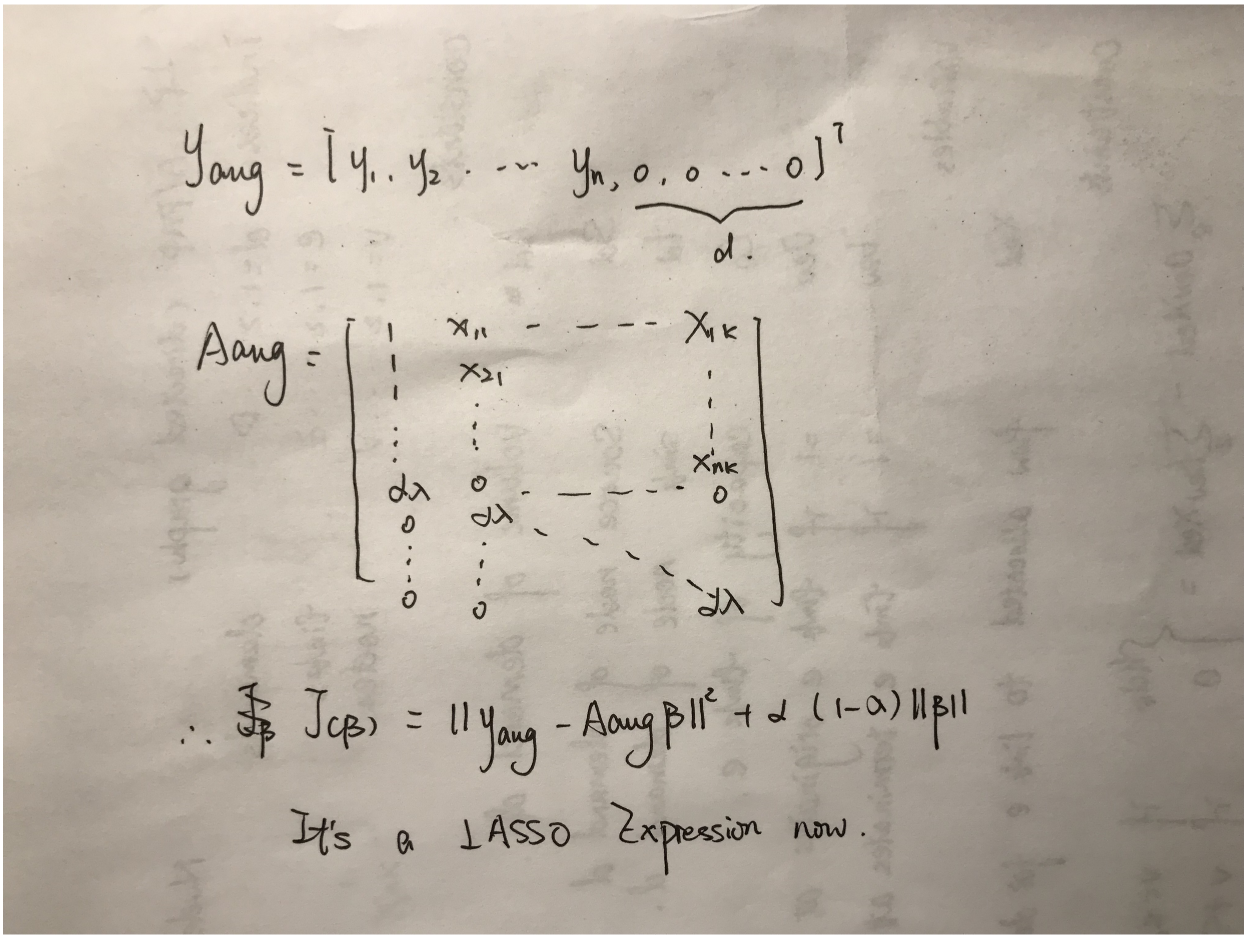
$\beta_{opt} = (A^TA + \alpha I)^{-1}A^Ty$

9.

$J(\beta) = ||A\beta - y||^2 + \alpha(\lambda||\beta||^2 + (1-\lambda)||\beta||_1$

$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha[\sum_{j=1}^{d} \lambda|\beta_j|^2 + \sum_{j=1}^{d}(1-\lambda)|\beta_j|$

Augmented version of A and y:



$$Y_{aug} = [y_1 . \, y_2 \cdots y_n, \underbrace{0, 0 \cdots 0}_{d.}]^T$$

$$A_{aug} = \begin{bmatrix} 1 & x_{11} & - - - & x_{1k} \\ 1 & x_{21} & & \vdots \\ 1 & \vdots & & \vdots \\ \vdots & \vdots & & x_{nk} \\ d\lambda & 0 & - - - & 0 \\ 0 & d\lambda & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & & d\lambda \end{bmatrix}$$

$$\therefore \cancel{J}_{\beta} \; J(\beta) = \| y_{aug} - A_{aug}\beta \|^2 + \alpha \, (1-\alpha) \|\beta\|$$

It's a LASSO Expression now.