Homework 2 - Machine Learning

Felix Nguyen

February 2019

1 Exercise 1

a. Without considering interaction terms, the number of number of possible regression model we can fit is the number of subsets of a set of p elements. This is equal to:

$$\sum_{k=0}^{p} \binom{p}{k}$$

Using binomial theorem where $x = y = 1$, we can derive that:

$$\sum_{k=0}^{p} \binom{p}{k} = 2^p$$

Thus, the number of possible regression models we can fit is 2^p.
b. If we consider all possible two-way interaction terms in our models, then we can see that for each possible model in 1.a. there are now n possible models, with n is the number of possible combinations from all two-way interactions.

We can also see that the number of two-way interaction terms is $\binom{p}{2} = \frac{p(p-1)}{2}$.
Therefore, similar to 1.a., we can derive $n = 2^{\frac{p(p-1)}{2}}$.
So, the number of possible regression models if we consider two-way interaction terms is:

$$2^p.2^{\frac{p(p-1)}{2}} = 2^{\frac{p(p+1)}{2}}.$$

2 Exercise 2

a. From OLS formulae, we have:

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Plugging the data given, we can calculate $\hat{\beta_1} = 2.1875$.
Similarly, we have $\hat{\beta_0} = \bar{y} - \bar{x}\hat{\beta_1} = 40 - 0 * 2.1875 = 40$

b. From the Ridge regression formulae, we have:

$$\hat{\beta}_{ridge} = (X^T X + I\lambda)^{-1} X^T y$$

From our data, we can calculate $diag(X^T X) = 16$, hence we have $diag(X^T X + I\lambda) = 16 + \lambda$. Therefore,

$$\begin{aligned}(X^T X + I\lambda)^{-1} X^T &= \frac{1}{(16+\lambda)} \begin{pmatrix} -2 & -1 & \cdots & 2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{-2}{16+\lambda} & \frac{-1}{16+\lambda} & \cdots & \frac{2}{16+\lambda} \end{pmatrix} \\ (X^T X + I\lambda)^{-1} X^T y &= \begin{pmatrix} \frac{-2}{16+\lambda} & \frac{-1}{16+\lambda} & \cdots & \frac{2}{16+\lambda} \end{pmatrix} \begin{pmatrix} 35 \\ 40 \\ \vdots \\ 43 \end{pmatrix} = \frac{35}{16+\lambda}\end{aligned}$$

Therefore, the ridge regression fit is $y = 40 + \frac{35}{16+\lambda} x + \varepsilon$
c. With $\lambda = 0.5$, we have the ridge regression fit:

$$\begin{aligned} y &= 40 + \frac{35}{16+0.5} x + \varepsilon \\ y &= 40 + \frac{35}{16.5} x + \varepsilon \\ y &= 40 + \frac{70}{33} x + \varepsilon \end{aligned}$$

d. We have:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Using partial derivation of β_0 and β_1 for minimization (each equals 0), we have $\frac{\delta W}{\delta \beta_0} = 0 \Leftrightarrow \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$. For $\frac{\delta W}{\delta \beta_1} = 0$, we have:

$$-2x_i \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1) + \lambda \frac{|\beta_1|}{\beta_1} = 0$$

4 Plug in β_0 and after some basic transformations, we can derive:

$$\hat{\beta}_1 = \frac{\pm\frac{\lambda}{2} + \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

With $+\frac{\lambda}{2}$ if $\beta_1 < 0$ and $-\frac{\lambda}{2}$ if $\beta_1 > 0$. Plug in the data given, we have:

$$\hat{\beta}_1 = \begin{cases} \frac{35+0.5\lambda}{16} & (\beta_1 < 0) \\ \frac{35-0.5\lambda}{16} & (\beta_1 > 0) \end{cases} \quad ; \ \hat{\beta}_0 = 40$$

e. With $\lambda = 14$, we have $\hat{\beta}_1 = \frac{35-7}{16} = 1.75$, so LASSO fit is $y = 40 + 1.75x$

3 Exercise 3

From Bayes theorem, we have

$$Pr(\beta|D) \propto Pr(D|\beta) * Pr(\beta) = N(y - X\beta, \sigma^2 I)N(0, \tau^2 I)$$

From this, using Maximum likelihood estimation, we then have

$$\begin{aligned} log(Pr(\beta|D)) &= log(Pr(D|\beta)) + log(Pr(\beta)) \\ &= \frac{-1}{2}\frac{(y - X\beta)^T(y - X\beta)}{\sigma^2} - \frac{\beta\beta^T}{2\tau^2} + C \end{aligned}$$

Since C is constant, if we multiply the above with $-2\sigma^2$, we can have

$$\hat{\beta} = \arg\min_{\beta}(y - X\beta)^T(y - X\beta) + \frac{\sigma^2}{\tau^2}\beta\beta^T$$

If we call $\frac{\sigma^2}{\tau^2} = \lambda$, the above is exactly $\hat{\beta}_{ridge}$.