Fall 2018 MATH895 HW2

Qinghong(Jackie) Xu

October 2, 2018

1 Ex3.5

The ridge regression problem is

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\}$$

After reparametrization using centered inputs: each x_{ij} gets replaced by $x_{ij} - \bar{x}_j$, we can rewrite this problem as

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} \bar{x}_j\beta_j - \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\}$$

Denote $\beta_0^c = \beta_0 + \sum_{j=1}^{p} \bar{x}_j\beta_j$ and $\beta_j^c = \beta_j$ for $j \geq 1$, we have

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} \bar{x}_j\beta_j - \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \sum_{i=1}^{N}(y_i - \beta_0^c - \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\beta_j^c)^2 + \lambda \sum_{j=1}^{p}(\beta_j^c)^2$$

So we can say the ridge regression problem is equivalent to

$$\hat{\beta}^{c} = \underset{\beta^c}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i - \beta_0^c - \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\beta_j^c)^2 + \lambda \sum_{j=1}^{p}(\beta_j^c)^2\}$$

The solution to this problem is

$$\hat{\beta}^c = ((\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}}) + \lambda\mathbf{I})^{-1}(\mathbf{X} - \bar{\mathbf{X}})^T\mathbf{y}$$

where $\bar{\mathbf{X}}$ is a matrix

$$\begin{bmatrix} \vdots & \vdots & \vdots & & \vdots \\ 0 & \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & \vdots & & \vdots \end{bmatrix}$$

The lasso problem is

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

Using the same notation above, it is also equivalent to

$$\hat{\beta}^c = \underset{\beta^c}{\operatorname{argmin}}\{\sum_{i=1}^{N}(y_i - \beta_0^c - \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)\beta_j^c)^2 + \lambda \sum_{j=1}^{p} |\beta_j^c|\}$$

We can see the ridge regression and the lasso problem are similar, the L^2 ridge penalty $\sum_{j=1}^{p} |\beta_j|^2$ is replaced by the L^1 lasso penalty $\sum_{j=1}^{p} |\beta_j|$. This constraint makes the solution nonlinear in the y_i, and there is no closed form expression for the lasso problem.

2 Ex3.9

Suppose we have the QR decomposition for the $N \times q$ matrix $\mathbf{X}_1$ in a multiple regression problem with response $\mathbf{y}$. Suppose $[\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_q]$ are the columns of $\mathbf{Q}$, then we know that our current residual $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \sum_{j=1}^{q}(\mathbf{z}_j^T \mathbf{y})\mathbf{z}_j$ is orthogonal to $\mathbf{Q}$. If we include one variable from $\mathbf{X}_2$, and suppose the variable has been orthonormalized, then the new residual will be

$$\mathbf{r}_{new} = \mathbf{y} - \hat{\mathbf{y}}_{new} = \mathbf{r} - (\mathbf{z}_{q+1}^T \mathbf{y})\mathbf{z}_{q+1}$$

The residual sum of squares is computed by

$$\begin{aligned}
\mathbf{r}_{new}^T \mathbf{r}_{new} &= (\mathbf{r} - (\mathbf{z}_{q+1}^T \mathbf{y})\mathbf{z}_{q+1})^T (\mathbf{r} - (\mathbf{z}_{q+1}^T \mathbf{y})\mathbf{z}_{q+1}) \\
&= (\mathbf{r}^T - (\mathbf{z}_{q+1}^T \mathbf{y})\mathbf{z}_{q+1}^T)(\mathbf{r} - (\mathbf{z}_{q+1}^T \mathbf{y})\mathbf{z}_{q+1}) \\
&= \mathbf{r}^T \mathbf{r} - (\mathbf{z}_{q+1}^T \mathbf{y})\mathbf{z}_{q+1}^T \mathbf{r} - \mathbf{r}^T (\mathbf{z}_{q+1}^T \mathbf{y})\mathbf{z}_{q+1} + (\mathbf{z}_{q+1}^T \mathbf{y})^2 \mathbf{z}_{q+1}^T \mathbf{z}_{q+1}
\end{aligned}$$

Since $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ and $\mathbf{z}_{q+1}^T \mathbf{z}_{q+1} = 1$, we have

$$\mathbf{r}_{new}^T \mathbf{r}_{new} = \mathbf{r}^T \mathbf{r} - 2(\mathbf{z}_{q+1}^T \mathbf{y})\mathbf{z}_{q+1}^T (\mathbf{y} - \hat{\mathbf{y}}) + (\mathbf{z}_{q+1}^T \mathbf{y})^2 = \mathbf{r}^T \mathbf{r} - (\mathbf{z}_{q+1}^T \mathbf{y})^2$$

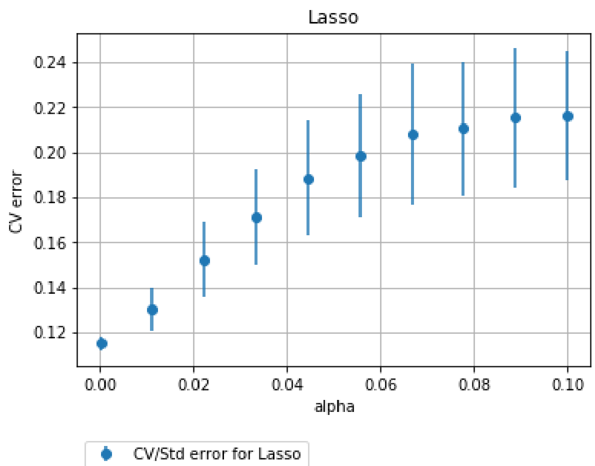
So including one additional variable the residual will be reduced by $(\mathbf{z}_{q+1}^T \mathbf{y})^2$. Our goal for this algorithm is to determine which one of these additional variables will reduce the residual-sum-of squares the most.

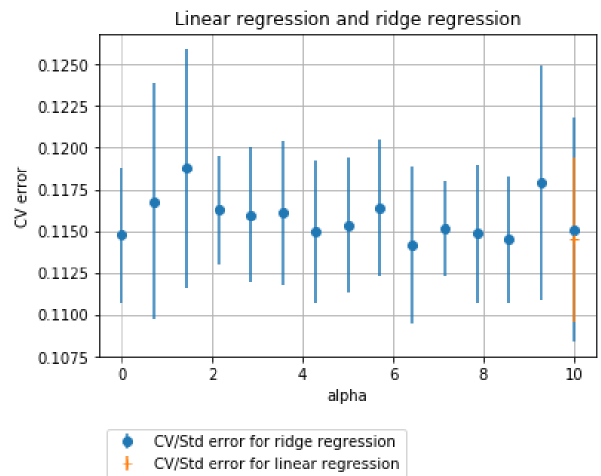
Algorithm 1 Forward stepwise regression

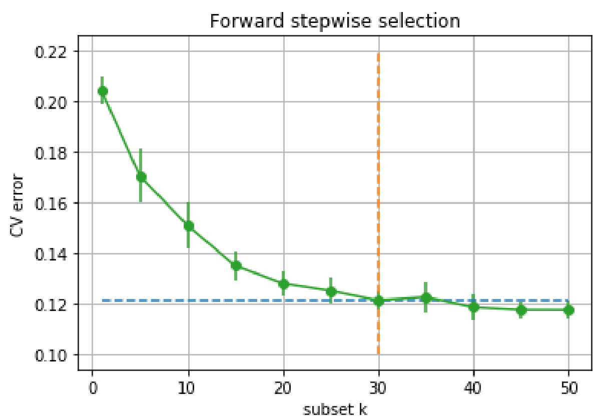
Suppose $[\mathbf{x}_{q+1}, \mathbf{x}_{q+2}, ..., \mathbf{x}_p]$ are the columns of $\mathbf{X}_2$.
for $j = q+1, ..., p$ **do**
Regress $\mathbf{x}_j$ on $\mathbf{z}_1, ..., \mathbf{z}_q$ to produce coefficients $\hat{\gamma}_{lj} = \mathbf{z}_l^T \mathbf{x}_j$, $l = 1, ..., q$ and the residual vector $\mathbf{z}_j = \mathbf{x}_j - \sum_{l=1}^{q} \hat{\gamma}_{lj} \mathbf{z}_l$.
Orthonormalize the vector $\mathbf{v}_j = \frac{\mathbf{z}_j}{||\mathbf{z}_j||}$ and compute $|\mathbf{v}_j^T \mathbf{y}|$.
end for
Choose the best variable included in $\mathbf{X}_1$ to be the k-th variable where $k = \underset{q+1 \leq j \leq p}{\operatorname{argmax}} |\mathbf{v}_j^T \mathbf{y}|$ and set $\mathbf{z}_{q+1} = \mathbf{v}_k$.

3 Ex3.17

We perform the plain linear regression, ridge regression, lasso and the forward stepwise regression on the email/spam data. The prediction error estimates and their standard errors were obtained







by tenfold cross-validation. From the figure we can see the CV/Std errors for the ridge regression stay steady when the α range from 1 to 10, while the lasso method is more sensitive to the choice of α.The CV errors for the ridge regression floated around the CV errors for the linear regression, same as the lasso when $\alpha < 0.01$. When $\alpha > 0.01$, the lasso produced almost twice times lager CV errors. For the forward stepwise selection, the CV errors are decreasing as we increase the subset size, and the largest size $k = 57$ is just the plain linear regression. Using the cross validation strategy we can conclude that $k = 30$ is the optimal value. We also provide a table recording the errors for these four models:

Error	Linear regression	Ridge regression($\alpha = 5.2$)	Lasso($\alpha = 0.001$)	Forward stepwise($k = 30$)
CV	0.1151	0.1141	0.1150	0.1212
Std	0.003	0.003	0.005	0.003

4 Ex3.23(a)

Show that

$$\frac{1}{N}|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{u}(\alpha)\rangle| = (1-\alpha)\lambda,\ j = 1, ..., p$$

and hence the correlations of each $\mathbf{x}_j$ with the residuals remain equal in magnitude as we progress toward $\mathbf{u}$.

Proof.

$$\begin{aligned}\frac{1}{N}|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{u}(\alpha)\rangle| &= \frac{1}{N}|\langle \mathbf{x}_j, \mathbf{y} - \alpha \mathbf{X}\hat{\beta}\rangle| = \frac{1}{N}|\langle \mathbf{x}_j, \mathbf{y}\rangle - \alpha\langle \mathbf{x}_j, \mathbf{X}\hat{\beta}\rangle| \\ &= \frac{1}{N}(1-\alpha)|\langle \mathbf{x}_j, \mathbf{y}\rangle| = \lambda(1-\alpha)\end{aligned}$$

□