Statistical Machine Learning

Home Work Three YI CHEN YC3356

1) $n=2$, $p=2$. $x_{11}=x_{12}=x_1$, $x_{21}=x_{22}=x_2$

(a) $\sum_{i=1}^{2}(y_i-\hat{\beta}_1x_1-\hat{\beta}_2x_2)^2+\lambda(\hat{\beta}_1^2+\hat{\beta}_2^2)$.

$\underset{(\hat{\beta}_1,\hat{\beta}_2)}{\operatorname{argmin}}$

(b) $\frac{\partial L}{\partial \hat{\beta}_1}=\hat{\beta}_1(x_1^2+x_2^2+\lambda)+\hat{\beta}_2(x_1^2+x_2^2)-y_1x_1-y_2x_2=0$

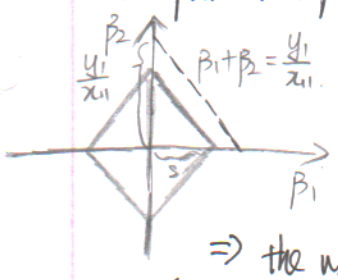
$\frac{\partial L}{\partial \hat{\beta}_2}=\hat{\beta}_1(x_1^2+x_2^2)+\hat{\beta}_2(x_1^2+x_2^2+\lambda)-y_1x_1-y_2x_2=0$

Then we get: $\hat{\beta}_1=\hat{\beta}_2$

(c) $L_1=\underset{(\hat{\beta}_1,\hat{\beta}_2)}{\operatorname{argmin}}(y_1-\hat{\beta}_1x_1-\hat{\beta}_2x_2)^2+(y_2-\hat{\beta}_1x_1-\hat{\beta}_2x_2)^2+\lambda(|\hat{\beta}_1|+|\hat{\beta}_2|)$

(d) $\begin{cases}\min (y_1-\hat{\beta}_1x_1-\hat{\beta}_2x_2)^2+(y_2-\hat{\beta}_1x_1-\hat{\beta}_2x_2)^2 \Rightarrow \min 2(y_1-(\hat{\beta}_1+\hat{\beta}_2)x_{11})^2 \\ \text{s.t. } |\hat{\beta}_1|+|\hat{\beta}_2|\le S\end{cases}$

This optimization problem has a simple solution $\hat{\beta}_1+\hat{\beta}_2=\frac{y_1}{x_{11}}$



This is a line that parallel to the edge of LASSO diamond $|\hat{\beta}_1|+|\hat{\beta}_2|=S$. Now, as we know, the solution must be the contours of function $(y_1-(\hat{\beta}_1+\hat{\beta}_2)x_{11})^2$ that touch the Lasso diamond $|\hat{\beta}_1|+|\hat{\beta}_2|=S$.

$\Rightarrow$ the whole line of $\hat{\beta}_1+\hat{\beta}_2=S$ or $\hat{\beta}_1+\hat{\beta}_2=-S$ would be the solution.

$\Rightarrow \begin{cases}\hat{\beta}_1\ge 0,\ \hat{\beta}_2\ge 0: & \hat{\beta}_1+\hat{\beta}_2=S \Rightarrow \text{it's not necessary to let } \hat{\beta}_1=\hat{\beta}_2 \\ \hat{\beta}_1\le 0,\ \hat{\beta}_2\le 0: & \hat{\beta}_1+\hat{\beta}_2=-S\end{cases}$

2) $\hat{g}_1=\underset{g}{\operatorname{argmin}}\left(\sum_{i=1}^{n}(y_i-g(x_i))^2+\lambda\int[g^{(3)}(x)]^2dx\right)$

$\hat{g}_2=\underset{g}{\operatorname{argmin}}\left(\sum_{i=1}^{n}(y_i-g(x_i))^2+\lambda\int[g^{(4)}(x)]^2dx\right)$

(1) As $\lambda\to\infty$, $\hat{g}_2$ will have the smaller training RSS.

Reason: the penalty term for $\hat{g}_2$ is higher in order. Thus, we also need the polynomial of $g(x_i)$ to have a higher order to ensure that the deviation exist.

In that way, the model is more flexible in $\hat{g}_2$ and thus has lower training RSS.

(2) As $\lambda\to\infty$, we already know that the $\hat{g}_2$ is more flexible and have lower RSS. Thus the data has higher probability to overfit. and has higher test RSS. But, it will also depend on how flexible the data is.

In summary: $\hat{g}_1$ has smaller test RSS.