MA234 · Introduction to Big Data Science

Assignment 3

刘乐奇 / 12011327

March 30, 2022

1 Exercise 1

It was done on the website.

2 Exercise 2

1. Non-linear regression problem.
2. B

3 Exercise 3

1. For linear model $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$, we need to minimize the error ϵ, i.e., we need to minimize

$$RSS(\mathbf{w}) = ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 \tag{1}$$

Since

$$\frac{\partial RSS(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) \tag{2}$$

let the expression above equal to $\mathbf{0}$, we have

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3}$$

So

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{4}$$

2. When let $\mathbf{P}\mathbf{x} = \lambda\mathbf{x}$, we have $\mathbf{P}^2\mathbf{x} = \mathbf{P}\lambda\mathbf{x} = \lambda\mathbf{P}\mathbf{x} = \lambda^2\mathbf{x}$. Since

$$\mathbf{P}^2 = (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{P} \tag{5}$$

we have $\mathbf{P}^2\mathbf{x} = \mathbf{P}\mathbf{x}$, i.e., $\lambda^2\mathbf{x} = \lambda\mathbf{x} \implies \lambda(\lambda - 1)\mathbf{x} = 0$. For $\mathbf{x} \neq 0$, we can get the eigenvalues are 0 and 1.

3. Since $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, we have

$$\begin{aligned}
\mathbf{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{E}(\epsilon) = 0 \\
Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon) &= ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T Var(\epsilon) \\
&= ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1})\sigma^2\mathbf{I} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbf{E}(\hat{\mathbf{w}}) &= \mathbf{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] \\
&= \mathbf{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w} + \epsilon)] \\
&= \mathbf{E}[\mathbf{w} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon] \\
&= \mathbf{E}(\mathbf{w}) + \mathbf{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon] \\
&= \mathbf{w}
\end{aligned} \tag{6}$$

Similarly,

$$\begin{aligned}
Var(\hat{\mathbf{w}}) &= \mathbf{E}[(\hat{\mathbf{w}} - \mathbf{E}(\hat{\mathbf{w}}))(\hat{\mathbf{w}} - \mathbf{E}(\hat{\mathbf{w}}))^T] \\
&= \mathbf{E}[(\mathbf{w} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon - \mathbf{w})(\mathbf{w} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon - \mathbf{w})^T] \\
&= \mathbf{E}[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon)((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon)^T] \\
&= \mathbf{E}[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon - \mathbf{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon))((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon - \mathbf{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon))^T] \\
&= Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2
\end{aligned} \tag{7}$$

4. Since

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y}) + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) \tag{8}$$

we need to prove that $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 0$.
By least squares for linear regression, we want to minimize $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ where $\hat{y}_i = w_0 + w_1x_{i1} + w_2x_{i2} + ... + w_px_{ip}$. It must satisfy that

$$\frac{\partial RSS}{\partial w_0} = -2\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0 \tag{9}$$

and

$$\begin{aligned}
\frac{\partial RSS}{\partial w_1} &= -2\sum_{i=1}^{n}x_{i1}(y_i - \hat{y}_i) = 0 \\
&\vdots \\
\frac{\partial RSS}{\partial w_p} &= -2\sum_{i=1}^{n}x_{ip}(y_i - \hat{y}_i) = 0
\end{aligned} \tag{10}$$

That is,

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0 \tag{11}$$

and

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i &= \sum_{i=1}^{n}(y_i - \hat{y}_i)(w_0 + w_1x_{i1} + w_2x_{i2} + ... + w_px_{ip}) \\
&= w_0\sum_{i=1}^{n}(y_i - \hat{y}_i) + w_1\sum_{i=1}^{n}(y_i - \hat{y}_i)x_{i1} + w_2\sum_{i=1}^{n}(y_i - \hat{y}_i)x_{i2} + ... + w_p\sum_{i=1}^{n}(y_i - \hat{y}_i)x_{ip} \\
&= 0
\end{aligned} \tag{12}$$

Thus $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i - \overline{y}\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$ has been proved. Therefore, $SS_{tot} = SS_{reg} + SS_{res}$.

4 Exercise 4

1. Since $\sum_{i=1,i\neq k}^{n}(y_i - \mathbf{x}_i^T\mathbf{w})^2 = ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 - (y_k - \mathbf{x}_k^T\mathbf{w})^2$, we can let $EST(\mathbf{w}) = ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 - (y_k - \mathbf{x}_k^T\mathbf{w})^2 + \lambda||\mathbf{w}||_2^2$ and minimize it.
We have

$$\frac{\partial EST(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + 2\mathbf{x}_k(y_k - \mathbf{x}_k^T\mathbf{w}) + 2\lambda\mathbf{w} = -2(\mathbf{X}^T\mathbf{y} - \mathbf{x}_k y_k) + 2(\mathbf{X}^T\mathbf{X} - \mathbf{x}_k\mathbf{x}_k^T + \lambda\mathbf{I})\mathbf{w} \tag{13}$$

Let it equal to $\mathbf{0}$, we have

$$\hat{\mathbf{w}}^{[k]} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} - \mathbf{x}_k\mathbf{x}_k^T)^{-1}(\mathbf{X}^T\mathbf{y} - \mathbf{x}_k y_k) \tag{14}$$

2.

3.

5 Exercise 5

For solving LASSO problem

$$\min_{\mathbf{w}}[||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2+\lambda||\mathbf{w}||_1] \tag{15}$$

turn it to

$$\min_{\mathbf{w},\mathbf{z}}[||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2+\lambda||\mathbf{z}||_1],\ s.t.\ \mathbf{w}=\mathbf{z} \tag{16}$$

The augmented Lagrange function is

$$L(\mathbf{w},\mathbf{z},\mathbf{u})=||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2+\lambda||\mathbf{z}||_1+\mathbf{u}^T(\mathbf{w}-\mathbf{z})+\frac{1}{2}||\mathbf{w}-\mathbf{z}||_2^2 \tag{17}$$

Let

$$\frac{\partial L(\mathbf{w},\mathbf{z},\mathbf{u})}{\partial \mathbf{w}}=-2\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w})+\mathbf{u}+(\mathbf{w}-\mathbf{z}) \tag{18}$$

equal to $\mathbf{0}$, we can get

$$\mathbf{w}=(2\mathbf{X}^T\mathbf{X}+I)^{-1}(2\mathbf{X}^T\mathbf{y}+\mathbf{u}+\mathbf{z}) \tag{19}$$

That is, $\mathbf{w}^{(k+1)}=(2\mathbf{X}^T\mathbf{X}+I)^{-1}(2\mathbf{X}^T\mathbf{y}+\mathbf{u}^{(k)}+\mathbf{z}^{(k)})$.
To find $\mathbf{z}^{(k+1)}$, we can start from

$$\begin{aligned}
\mathbf{z}&=\arg\min_{\mathbf{z}} L(\mathbf{w},\mathbf{z},\mathbf{u})\\
&=\arg\min_{\mathbf{z}}\{\lambda||\mathbf{z}||_1+\mathbf{u}^T(\mathbf{w}-\mathbf{z})+\frac{1}{2}||\mathbf{w}-\mathbf{z}||_2^2\}\\
&=\arg\min_{\mathbf{z}}\{2\lambda||\mathbf{z}||_1+2\mathbf{u}^T(\mathbf{w}-\mathbf{z})+||\mathbf{w}-\mathbf{z}||_2^2\}\\
&=\arg\min_{\mathbf{z}}\{2\lambda||\mathbf{z}||_1+||\mathbf{u}+(\mathbf{w}-\mathbf{z})||_2^2-||\mathbf{u}||_2^2\}\\
&=\arg\min_{\mathbf{z}}\{2\lambda||\mathbf{z}||_1+||\mathbf{z}-(\mathbf{u}+\mathbf{w})||_2^2\}\\
&=S_\lambda(\mathbf{u}+\mathbf{w})
\end{aligned} \tag{20}$$

where S_λ is a soft thresholding operator for

$$z_i=[S_\lambda(\mathbf{u}+\mathbf{w})]_i=\begin{cases} u_i+w_i-\lambda, & u_i+w_i>\lambda\\ 0, & |u_i+w_i|\le\lambda\\ u_i+w_i+\lambda, & u_i+w_i<-\lambda\end{cases}$$

So, $\mathbf{z}^{(k+1)}=S_\lambda(\mathbf{u}^{(k)}+\mathbf{w}^{(k+1)})$.
Therefore, we have

$$\mathbf{w}^{(k+1)}=(2\mathbf{X}^T\mathbf{X}+I)^{-1}(2\mathbf{X}^T\mathbf{y}+\mathbf{u}^{(k)}+\mathbf{z}^{(k)}) \tag{21}$$

$$\mathbf{z}^{(k+1)}=S_\lambda(\mathbf{u}^{(k)}+\mathbf{w}^{(k+1)}) \tag{22}$$

$$\mathbf{u}^{(k+1)}=\mathbf{u}^{(k)}+\mathbf{w}^{(k+1)}-\mathbf{z}^{(k+1)} \tag{23}$$