Introduction to Big Data Science

Homework 3 Reference Answer

2.

1. non-linear regression
2. B

3.

1. In multivariate linear problem, input $\mathbf{X}$, then the corresponding output is

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}.$$

Then

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

By differential $RSS(\mathbf{w})$ with respect to $\mathbf{w}$ and let it be zero, we get

$$\frac{\partial RSS(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = \mathbf{0}.$$

It gives

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

Thus, the linear regression predictor is

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

2. Since we have

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

then if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{P}$, we must have

$$\lambda\mathbf{x} = \mathbf{P}\mathbf{x} = \mathbf{P}^2\mathbf{x} = \lambda^2\mathbf{x}.$$

Eigenvectors are by definition nonzero, thus, $\lambda = \lambda^2$ must hold, which gives λ can only be 0 or 1, i.e., $\mathbf{P}$ has only 0 and 1 eigenvalues.

3. By definition, we get

$$\mathbf{E}(\hat{\mathbf{w}}) = \mathbf{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] = \mathbf{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w}+\epsilon)] = \mathbf{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w}] = \mathbf{E}[\mathbf{w}] = \mathbf{w},$$

and

$$\begin{aligned}
\mathrm{Var}(\hat{\mathbf{w}}) &= \mathbf{E}[(\hat{\mathbf{w}} - \mathbf{E}(\hat{\mathbf{w}}))(\hat{\mathbf{w}} - \mathbf{E}(\hat{\mathbf{w}}))^T] = \mathbf{E}[(\hat{\mathbf{w}} - \mathbf{w})(\hat{\mathbf{w}} - \mathbf{w})^T] \\
&= \mathbf{E}\left[\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\right)\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\right)^T\right] \\
&= \mathbf{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] \\
&= \mathbf{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right]\mathbf{E}[\epsilon\epsilon^T] \\
&= \mathbf{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]\mathrm{Var}(\epsilon) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2.
\end{aligned}$$

4. *Proof.* From (1.), we know that $\mathbf{X}^T(\mathbf{y}-\hat{\mathbf{y}}) = \mathbf{X}^T(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}}) = \mathbf{0}$. Since the first column of $\mathbf{X}$ is just $\mathbf{1}$, we know that $\mathbf{1}^T(\mathbf{y}-\hat{\mathbf{y}}) = 0$. Therefore,

$$\begin{aligned}
SS_{tot} &= \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) \\
&= SS_{res} + SS_{reg} + 2(\mathbf{y}-\hat{\mathbf{y}})^T(\mathbf{X}\hat{\mathbf{w}} - \overline{y}\mathbf{1}) \\
&= SS_{res} + SS_{reg} + 2\underbrace{(\mathbf{y}-\hat{\mathbf{y}})^T\mathbf{X}}_{=0}\hat{\mathbf{w}} - 2\overline{y}\underbrace{(\mathbf{y}-\hat{\mathbf{y}})^T\mathbf{1}}_{=0} \\
&= SS_{res} + SS_{reg}.
\end{aligned}$$

□

4.

1. Let

$$f^{[k]}(\mathbf{w}) = \sum_{i=1, i\neq k}^{n} (y_i - \mathbf{x}_i^T\mathbf{w})^2 + \lambda\|\mathbf{w}\|_2^2 = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 - (y_k - \mathbf{x}_k^T\mathbf{w})^2 + \lambda\|\mathbf{w}\|_2^2,$$

then by differential $f^{[k]}(\mathbf{w})$ with respect to $\mathbf{w}$ and let it be zero, we get

$$\frac{\partial f^{[k]}(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + 2\mathbf{x}_k(y_k - \mathbf{x}_k^T\mathbf{w}) + 2\lambda\mathbf{w} = \mathbf{0}.$$

It follows

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} - \mathbf{x}_k\mathbf{x}_k^T)\mathbf{w} = \mathbf{X}^T\mathbf{y} - \mathbf{x}_k y_k,$$

which gives

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

2. Denote $\mathbf{A} = \mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$, which is clearly nonsingular and $-\mathbf{x}_k^T\mathbf{A}^{-1}\mathbf{x}_k \neq -1$ (by choosing proper λ), applying the Sherman-Morrison formula, we get

$$\begin{aligned}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} - \mathbf{x}_k\mathbf{x}_k^T)^{-1} = (\mathbf{A} + (-\mathbf{x}_k)\mathbf{x}_k^T)^{-1} &= \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}(-\mathbf{x}_k)\mathbf{x}_k^T\mathbf{A}^{-1}}{1 + \mathbf{x}_k^T\mathbf{A}^{-1}(-\mathbf{x}_k)} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} + \frac{(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_k\mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}}{1 - \mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_k}.\end{aligned}$$

Notice that

$$\mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_k = p_{kk} \quad \text{and} \quad \hat{y}_k = \mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

then we have

$$\begin{aligned}&\mathbf{x}_k^T\hat{\mathbf{w}}^{[k]} - y_k \\ &= \mathbf{x}_k^T\left[(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} + \frac{(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_k\mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}}{1 - \mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_k}\right](\mathbf{X}^T\mathbf{y} - \mathbf{x}_k y_k) - y_k \\ &= \mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} - \mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_k y_k \\ &\quad + \frac{\mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_k\mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}}{1 - \mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_k} \\ &\quad - \frac{\mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_k\mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_k y_k}{1 - \mathbf{x}_k^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x}_k} - y_k \\ &= \hat{y}_k - p_{kk}y_k + \frac{p_{kk}\hat{y}_k}{1 - p_{kk}} - \frac{p_{kk}p_{kk}y_k}{1 - p_{kk}} - y_k = \frac{\hat{y}_k - y_k}{1 - p_{kk}}.\end{aligned}$$

Hence

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{x}_k^T\hat{\mathbf{w}}^{[k]} - y_k\right)^2 = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{\hat{y}_k - y_k}{1 - p_{kk}}\right)^2.$$

3. We can write $V(\lambda)$ as

$$\begin{aligned}
V(\lambda) &= \frac{1}{n}\sum_{k=1}^{n} w_k \left(\mathbf{x}_k^T \hat{\mathbf{w}}^{[k]} - y_k\right)^2 = \frac{1}{n}\sum_{k=1}^{n} \left(\frac{1-p_{kk}}{\frac{1}{n}\mathrm{tr}(\mathbf{I}-\mathbf{P})}\right)^2 \left(\frac{\hat{y}_k - y_k}{1-p_{kk}}\right)^2 \\
&= \frac{1}{n}\sum_{k=1}^{n} \left(\frac{1-p_{kk}}{\frac{1}{n}\mathrm{tr}(\mathbf{I}-\mathbf{P})} \cdot \frac{\hat{y}_k - y_k}{1-p_{kk}}\right)^2 = \frac{1}{n}\sum_{k=1}^{n} \left(\frac{\hat{y}_k - y_k}{\frac{1}{n}\mathrm{tr}(\mathbf{I}-\mathbf{P})}\right)^2 \\
&= \frac{1}{n}\left(\frac{1}{\frac{1}{n}\mathrm{tr}(\mathbf{I}-\mathbf{P})}\right)^2 \sum_{k=1}^{n} (\hat{y}_k - y_k)^2 = \frac{1}{n}\left(\frac{1}{\frac{1}{n}(\mathrm{tr}(\mathbf{I}) - \mathrm{tr}(\mathbf{P}))}\right)^2 \|\hat{\mathbf{y}} - \mathbf{y}\|^2 \\
&= \frac{1}{n}\left(\frac{1}{\frac{1}{n}(n - \mathrm{tr}(\mathbf{P}))}\right)^2 \|\mathbf{P}\mathbf{y} - \mathbf{y}\|^2 = \frac{\frac{1}{n}\|(\mathbf{I}-\mathbf{P})\mathbf{y}\|^2}{\left[1 - \mathrm{tr}(\mathbf{P})/n\right]^2}.
\end{aligned}$$

5.

Rewrite the LASSO problem as

$$\min_{\mathbf{w},\mathbf{z}}[\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\lambda\|\mathbf{z}\|_1],\quad \text{subject to } \mathbf{w}-\mathbf{z}=0.$$

The augmented Lagrange function is defined as

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

Then the ADMM steps are given by

1. $\mathbf{w}^{(k+1)}=\arg\min_{\mathbf{w}} L(\mathbf{w},\mathbf{z}^{(\mathbf{k})},\mathbf{u}^{(\mathbf{k})})$.

From that

$$\nabla_{\mathbf{w}}L(\mathbf{w},\mathbf{z}^{(\mathbf{k})},\mathbf{u}^{(\mathbf{k})})=-2\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w})+\mathbf{u}^{(\mathbf{k})}+(\mathbf{w}-\mathbf{z}^{(\mathbf{k})}),$$

solving $\nabla_{\mathbf{w}}L(\mathbf{w},\mathbf{z}^{(\mathbf{k})},\mathbf{u}^{(\mathbf{k})})=0$ gives that

$$\mathbf{w}^{(k+1)}=(\mathbf{I}+\mathbf{2X^TX})^{-1}(2\mathbf{X^Ty}-\mathbf{u}^{(\mathbf{k})}+\mathbf{z}^{(\mathbf{k})}).$$

2. $\mathbf{z}^{(k+1)}=\arg\min_{\mathbf{z}} L(\mathbf{w}^{(\mathbf{k+1})},\mathbf{z},\mathbf{u}^{(\mathbf{k})})$.

From that

$$\partial_{\mathbf{z}}L(\mathbf{w}^{(\mathbf{k+1})},\mathbf{z},\mathbf{u}^{(\mathbf{k})})=\lambda\partial\|\mathbf{z}\|_1-\mathbf{u}^{(\mathbf{k})}-(\mathbf{w}^{(\mathbf{k+1})}-\mathbf{z}),$$

where $\mathbf{z}=(z_1,\cdots,z_m)^T$, $\partial_{z_j}\|\mathbf{z}\|_1=\begin{cases}1, & z_j>0,\\ [-1,1], & z_j=0,\\ -1, & z_j<0,\end{cases}$ is the subgradient of $\|\mathbf{z}\|_1$ in z_j, then the optimal condition $0\in\partial_{z_j}L(\mathbf{w}^{(\mathbf{k+1})},\mathbf{z},\mathbf{u}^{(\mathbf{k})})=0$ gives that

$$z_j^{(k+1)}=\begin{cases}u_j^{(k)}+w_j^{(k+1)}-\lambda, & u_j^{(k)}+w_j^{(k+1)}>\lambda,\\ 0, & u_j^{(k)}+w_j^{(k+1)}\in[-\lambda,\lambda],\\ u_j^{(k)}+w_j^{(k+1)}+\lambda, & u_j^{(k)}+w_j^{(k+1)}<\lambda,\end{cases}\quad j=1,\cdots,n.$$

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