DS 5230 Unsupervised Machine Learning and Data Mining
Homework 2 By Yanchi Li

Exercise 1

a. We directly derive the target equation from the closed form solution of the weights $\boldsymbol{w}_*$ that minimize the loss

$$\begin{aligned}
\boldsymbol{w}_* &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T y \\
(\boldsymbol{X}^T\boldsymbol{X} + \lambda I)\boldsymbol{w}_* &= \boldsymbol{X}^T y \\
\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}_* + \lambda\boldsymbol{w}_* &= \boldsymbol{X}^T y \\
\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}_* + \lambda\boldsymbol{w}_* - \boldsymbol{X}^T y &= 0 \\
\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w}_* - y) + \lambda\boldsymbol{w}_* &= 0 \\
2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w}_* - y) + 2\lambda\boldsymbol{w}_* &= 0
\end{aligned} \tag{1}$$

This is exactly the same equation with the target equation below

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}^*} = 0 \tag{2}$$

b. By definition,

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2} \cdot \frac{(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^2}{\sigma^2}) \tag{3}$$

and

$$p(\boldsymbol{w}) = \frac{1}{\sqrt{2\pi s^2}} \exp(-\frac{1}{2} \cdot \frac{{\boldsymbol{w}_n}^2}{s^2}) \tag{4}$$

Then we break $p(\boldsymbol{w}|\boldsymbol{y})$ into $p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})$ and use logarithm to calculate the $\arg\max$

value.

$$
\begin{aligned}
\boldsymbol{w}_* &= \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y}) \\
&= \arg\max_{\boldsymbol{w}} \log(p(\boldsymbol{w}|\boldsymbol{y})) \\
&= \arg\max_{\boldsymbol{w}} \log[p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})] \\
&= \arg\max_{\boldsymbol{w}} (\log p(\boldsymbol{y}|\boldsymbol{w}) + \log p(\boldsymbol{w})) \\
&= \arg\max_{\boldsymbol{w}} [\log \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2} \cdot \frac{(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^2}{\sigma^2}) + \log \frac{1}{\sqrt{2\pi s^2}} \exp(-\frac{1}{2} \cdot \frac{\boldsymbol{w}^2}{s^2})] \\
&= \arg\max_{\boldsymbol{w}} [-\frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^2 - \frac{1}{2s^2}\boldsymbol{w}^2 + \log \frac{1}{\sqrt{2\pi\sigma^2}} + \log \frac{1}{\sqrt{2\pi s^2}}] \\
&= \arg\max_{\boldsymbol{w}} [-\frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^2 - \frac{1}{2s^2}\boldsymbol{w}^2] \\
&= \arg\min_{\boldsymbol{w}} [-\frac{1}{2\sigma^2}E(\boldsymbol{w}) - \frac{1}{2s^2}|\boldsymbol{w}|^2] \\
&= \arg\min_{\boldsymbol{w}} [-\frac{1}{2\sigma^2}(E(\boldsymbol{w}) + \frac{\sigma^2}{s^2}|\boldsymbol{w}|^2)] \\
&= \arg\min_{\boldsymbol{w}} [-\frac{1}{2\sigma^2}(E(\boldsymbol{w}) + \lambda|\boldsymbol{w}|^2)]
\end{aligned}
\tag{5}
$$

Where

$$\lambda = \frac{\sigma^2}{s^2} \tag{6}$$

c. When applying linear regression, we have a equation of $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\epsilon}$, where $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{w}$. Then we want to apply the function of Expectation and Covariance to $\boldsymbol{y}$.

$$
\begin{aligned}
\mathbb{E}(\boldsymbol{y}) &= \mathbb{E}(\boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}) \\
&= \mathbb{E}(\boldsymbol{X}\boldsymbol{w}) + \mathbb{E}(\boldsymbol{\epsilon}) \\
&= \boldsymbol{X}\mathbb{E}(\boldsymbol{w}) \\
&= \boldsymbol{X}m_0
\end{aligned}
\tag{7}
$$

and

$$
\begin{aligned}
Cov(\boldsymbol{y}) &= Cov(\boldsymbol{f}) + Cov(\boldsymbol{\epsilon}) \\
&= Cov(\boldsymbol{X}\boldsymbol{w}) + Cov(\boldsymbol{\epsilon}) \\
&= \boldsymbol{X}Cov(\boldsymbol{w})\boldsymbol{X}^T + Cov(\boldsymbol{\epsilon}) \\
&= \boldsymbol{X}S_0\boldsymbol{X}^T + \sigma^2 I
\end{aligned}
\tag{8}
$$

Thus, we may conclude that $\boldsymbol{\mu} = \boldsymbol{X}m_0$ and $\boldsymbol{\Sigma} = \boldsymbol{X}S_0\boldsymbol{X}^T + \sigma^2 I$. And

$$p(\boldsymbol{y}) \sim Normal(\boldsymbol{X}m_0, \boldsymbol{X}S_0\boldsymbol{X}^T + \sigma^2 I) \tag{9}$$

d. According to the standard identity on Gaussians, we have

$$\boldsymbol{\alpha}|\boldsymbol{\beta} \sim Normal(\boldsymbol{\alpha} + CB^{-1}(\boldsymbol{\beta} - \boldsymbol{b}), A - CB^{-1}C^T) \tag{10}$$

when

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \sim Normal(\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}) \tag{11}$$

Here we want to substitute $\boldsymbol{\alpha}$ with $\boldsymbol{y}_*$ and $\boldsymbol{\beta}$ with $\boldsymbol{y}$. Since we already have the mean value and covariance matrix of $\boldsymbol{y}$ in part d, we only need to determine the covariance matrix C between $\boldsymbol{y}$ and $\boldsymbol{y}_*$ and the mean value $\boldsymbol{\mu}_*$ of $\boldsymbol{y}_*$. Here we generalize the relationship we derived in the previous part to get

$$\mathbb{E}(\boldsymbol{y}_*) = \boldsymbol{x}_* m_0 \tag{12}$$

$$Cov(\boldsymbol{y}_*) = \boldsymbol{x}_* S_0 \boldsymbol{x}_*^T + \sigma^2 I \tag{13}$$

$$Cov(\boldsymbol{y}_*, \boldsymbol{y}) = Cov(\boldsymbol{y}, \boldsymbol{y}_*)^T = \boldsymbol{x}_* S_0 \boldsymbol{X}^T \tag{14}$$

Hence

$$\boldsymbol{f}_* = \boldsymbol{x}_* m_0 + \boldsymbol{x}_* S_0 \boldsymbol{X}^T (\boldsymbol{X} S_0 \boldsymbol{X}^T + \sigma^2 I)^{-1} (\boldsymbol{y} - \boldsymbol{X} m_0) \tag{15}$$

$$\boldsymbol{\sigma}_* = \boldsymbol{x}_* S_0 \boldsymbol{x}_*^T + \sigma^2 I - \boldsymbol{x}_* S_0 \boldsymbol{X}^T (\boldsymbol{X} S_0 \boldsymbol{X}^T + \sigma^2 I)^{-1} (\boldsymbol{x}_* S_0 \boldsymbol{X}^T)^T \tag{16}$$

e. When we substitute m_0 by 0 and S_0 by s^2 in the equation of f_*, we have

$$\begin{aligned} \boldsymbol{f}_* &= \boldsymbol{x}_* s^2 \boldsymbol{X}^T (\boldsymbol{X} s^2 \boldsymbol{X}^T + \sigma^2 I)^{-1} \boldsymbol{y} \\ &= \boldsymbol{x}_* \boldsymbol{X}^T (\boldsymbol{X}\boldsymbol{X}^T + \lambda I)^{-1} \boldsymbol{y} \end{aligned} \tag{17}$$

Then we apply a formula from matrix cookbook, which is

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T (BPB^T + R)^{-1} \tag{18}$$

In our context here, P is just the identity I, B is the matrix $\boldsymbol{X}$ and R is λI. We substitute everything into equation(18) to get

$$\begin{aligned} \boldsymbol{f}_* &= \boldsymbol{x}_* \boldsymbol{X}^T (\boldsymbol{X}\boldsymbol{X}^T + \lambda I)^{-1} \boldsymbol{y} \\ &= \boldsymbol{x}_* (I + \boldsymbol{X}^T \frac{1}{\lambda} \boldsymbol{X})^{-1} \boldsymbol{X}^T \frac{1}{\lambda} I \boldsymbol{y} \\ &= \boldsymbol{x}_* (\lambda I + \boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \end{aligned} \tag{19}$$

Where $(\lambda I + \boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{w}_*$ which is the closed form solution of ridge regression and the original form in line three of equation(19) is just kernel ridge regression. Here I really don't know how to solve the equation given in item e,

but my intuition tells me that the result should have some relationship with ridge regression and kernel ridge regression. When I derive through equation and just find it similar to the solution of ridge regression.
Then we determine $\mathbb{E}[f(\boldsymbol{x}_*)]$

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{x}_*)] &= \mathbb{E}[\boldsymbol{y}_* - \boldsymbol{\epsilon}] \\ &= \mathbb{E}[\boldsymbol{y}_*] \\ &= \boldsymbol{X} m_0 \\ &= 0\end{aligned} \tag{20}$$