

# Machine Learning Exercise Sheet 1

## Math Refresher

### Group\_369

Fan XUE – fan98.xue@tum.de

Xing ZHOU – xing.zhou@tum.de

Jianzhe LIU – jianzhe.liu@tum.de

November 17, 2021

#### Problem 6

#### Problem 7

We have

$$p(\mathbf{w}, \beta \mid D) \propto p(D \mid \mathbf{w}, \beta) \cdot p(\mathbf{w}, \beta)$$

So

$$\begin{aligned} \log p(\mathbf{w}, \beta \mid D) &= \log \prod_{i=1}^N \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(\Phi \mathbf{w} - \mathbf{y})^T (\Phi \mathbf{w} - \mathbf{y})} \\ &\quad + \log \left( \frac{1}{\sqrt{(2\pi)^M |\beta^{-1} \mathbf{S}_0|}} e^{-\frac{\beta}{2}(\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0)} \cdot \frac{b_0^{a_0} \beta^{a_0-1} e^{-b_0 \beta}}{\Gamma(a_0)} \right) \\ &= \frac{M}{2} \log \beta - \frac{\beta}{2} \mathbf{w}^T (\mathbf{S}_0^{-1} + \Phi^T \Phi) \mathbf{w} + \beta (\mathbf{m}_0^T \mathbf{S}_0^{-1} + \mathbf{y}^T \Phi) \mathbf{S}_N \mathbf{S}_N^{-1} \\ &\quad + \left( \frac{N}{2} + a_0 - 1 \right) \log \beta - \left( b_0 + \frac{1}{2} \mathbf{y}^T \mathbf{y} + \frac{1}{2} \mathbf{m}_0^T \mathbf{S}^{-1} \mathbf{m}_0 \right) \beta + \text{const.} \end{aligned}$$

We can expand the  $p(\mathbf{w}, \beta \mid D)$

$$\begin{aligned} \log p(\mathbf{w}, \beta \mid D) &= \frac{M}{2} \log \beta - \frac{\beta}{2} \mathbf{w}^T \mathbf{S}_N^{-1} \mathbf{w} + \beta \mathbf{m}_N^T \mathbf{S}_N^{-1} \mathbf{w} + (a_N - 1) \log \beta \\ &\quad - \beta \left( \frac{1}{2} \mathbf{m}_N^T \mathbf{S}_N^{-1} \mathbf{m}_N - b_N \right) \end{aligned}$$

Comparing the two expressions, we can find that

$$\begin{aligned}\mathbf{m}_N &= ((\mathbf{m}_0^T \mathbf{S}_0^{-1} + \mathbf{y}^T \Phi) \mathbf{S}_N)^T \\ \mathbf{S}_N &= (\mathbf{S}_0^{-1} + \Phi^T \Phi)^{-1} \\ a_N &= \frac{N}{2} + a_0 \\ b_N &= b_0 + \frac{1}{2} (\mathbf{m}_0^T \mathbf{S}_0^{-1} \mathbf{m}_0 - \mathbf{m}_N^T \mathbf{S}_N^{-1} \mathbf{m}_N + \mathbf{y}^T \mathbf{y})\end{aligned}$$

### Problem 8

$$\begin{aligned}E_{ridge}(\mathbf{w}) &= \frac{1}{2} \sum_{i=1}^N (\mathbf{w}^T \phi \mathbf{x}_i - y_i)^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} \\ &= \frac{1}{2} (\Phi \mathbf{w} - \mathbf{y})^T (\Phi \mathbf{w} - \mathbf{y}) + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}\end{aligned}$$

The gradient of  $E_{ridge}(\mathbf{w})$  is

$$\begin{aligned}\nabla_{\mathbf{w}} E_{ridge}(\mathbf{w}) &= \Phi^T \Phi \mathbf{w} - \Phi^T \mathbf{y} + \lambda \mathbf{w} \\ &= (\Phi^T \Phi + \lambda \mathbf{I}) \mathbf{w} - \Phi^T \mathbf{y}\end{aligned}$$

Let the gradient be zero, we get

$$(\Phi^T \Phi + \lambda \mathbf{I}) \mathbf{w} - \Phi^T \mathbf{y} = 0$$

Therefore

$$\mathbf{w}^* = (\Phi^T \Phi + \lambda \mathbf{I})^{-1} \Phi^T \mathbf{y}$$

If  $N < M$ , the matrix  $\Phi^T \Phi \in \mathbb{R}^{M \times M}$  is not invertible. The equation  $(\Phi^T \Phi) \mathbf{w} - \Phi^T \mathbf{y} = 0$  does not have only solution.

With the regulation, the normal equation is changed to  $(\Phi^T \Phi + \lambda \mathbf{I}) \mathbf{w} - \Phi^T \mathbf{y} = 0$ , and the problem is fixed.

### Problem 9

a) We want to find the same prediction, which means

$$\hat{y}_i = \mathbf{w}^{*T} \mathbf{x}_i = \mathbf{w}_{new}^T \mathbf{x}_{i,new} = a \mathbf{w}_{new}^T \mathbf{x}_i$$

Hence,

$$\mathbf{w}_{new} = \frac{1}{a} \mathbf{w}^*$$

b) According to the result of Problem 8, the solution for  $\mathbf{w}^*$  on the original dataset  $\mathbf{X}$  is

$$\mathbf{w}^* = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

We want to find the new regulation factor  $\lambda_{new}$  for  $\mathbf{X}_{new}$ , it is to find the  $\lambda_{new}$ , such that  $\mathbf{w}_{new}^* = \frac{1}{a}\mathbf{w}^*$ .

We have

$$\begin{aligned}\mathbf{w}_{new} &= (\mathbf{X}_{new}^T \mathbf{X}_{new} + \lambda_{new} \mathbf{I})^{-1} \mathbf{X}_{new}^T \mathbf{y} \\ &= (a^2 \mathbf{X}^T \mathbf{X} + \lambda_{new} \mathbf{I})^{-1} a \mathbf{X}^T \mathbf{y}\end{aligned}$$

We can observe that if we let  $\lambda_{new} = a^2 \lambda$ , we will have

$$\begin{aligned}\mathbf{w}_{new} &= a (a^2 \mathbf{X}^T \mathbf{X} + a^2 \lambda \mathbf{I})^{-1} a \mathbf{X}^T \mathbf{y} \\ &= \frac{1}{a} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \frac{1}{a} \mathbf{w}^*\end{aligned}$$

which satisfies the condition.