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 20, 2021

Problem 6

$$\left|\begin{bmatrix} 1 & 2 & 3 \\ 4 & ui & \boldsymbol{w} \end{bmatrix}\right| \int_a^b \sin x \cos ab_{\mathrm{fd}} ab_{\mathrm{fdsf}} ab_{fdsf} \vec{v}$$

We have,

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

So,

$$\log p(\boldsymbol{w}, \beta \mid D) = \log \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}}$$

Problem 7

We have

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

So

$$\begin{aligned}
\log p(\boldsymbol{w}, \beta, \mid D) = \log \prod_{i=1}^{N} \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(\boldsymbol{\Phi w}-\boldsymbol{y})^T(\boldsymbol{\Phi w}-\boldsymbol{y})} & \\
+ \log \left(\frac{1}{\sqrt{(2\pi)^M \left|\beta^{-1}\boldsymbol{S_0}\right|}} e^{-\frac{\beta}{2}(\boldsymbol{w}-\boldsymbol{m_0})^T \boldsymbol{S_0}^{-1}(\boldsymbol{w}-\boldsymbol{m_0})} \cdot \frac{b_0^{a_0}\beta^{a_0-1}e^{-b_0\beta}}{\Gamma(a_0)} \right) & \\
= \frac{M}{2}\log\beta - \frac{\beta}{2}\boldsymbol{w}^T\left(\boldsymbol{S_0^{-1}} + \boldsymbol{\Phi^T\Phi}\right)\boldsymbol{w} + \beta\left(\boldsymbol{m_0}^T\boldsymbol{S_0}^{-1} + \boldsymbol{y}^T\boldsymbol{\Phi}\right)\boldsymbol{S_N S_N^{-1}} & \\
+ \left(\frac{N}{2} + a_0 - 1\right)\log\beta - \left(b_0 + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} + \frac{1}{2}\boldsymbol{m_0}^T\boldsymbol{S}^{-1}\boldsymbol{m_0}\right)\beta + \text{const.} &
\end{aligned}$$

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

$$\begin{aligned}\log p(\boldsymbol{w}, \beta \mid D) = \frac{M}{2}\log\beta - \frac{\beta}{2}\boldsymbol{w}^{-1}\boldsymbol{S_N}^{-1}\boldsymbol{w} + \beta\boldsymbol{m_N}^T\boldsymbol{S_N}^{-1}\boldsymbol{w} + (a_N - 1)\log\beta \\ - \beta\left(\frac{1}{2}\boldsymbol{m_N}^T\boldsymbol{S_N}^{-1}\boldsymbol{m_N} - b_N\right)\end{aligned}$$

Comparing the two expressions, we can find that

$$\begin{aligned}\boldsymbol{m_N} &= \left(\left(\boldsymbol{m_0}^T\boldsymbol{S_0}^{-1} + \boldsymbol{y}^T\boldsymbol{\Phi}\right)\boldsymbol{S_N}\right)^T \\ \boldsymbol{S_N} &= \left(\boldsymbol{S_0}^{-1} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1} \\ a_N &= \frac{N}{2} + a_0 \qquad\qquad \to\to \\ b_N &= b_0 + \frac{1}{2}\left(\boldsymbol{m_0}^T\boldsymbol{S_0}^{-1}\boldsymbol{m_0} - \boldsymbol{m_N}^T\boldsymbol{S_N}^{-1}\boldsymbol{m_N} + \boldsymbol{y}^T\boldsymbol{y}\right)\end{aligned}$$

Problem 8

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

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

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

Let the gradient be zero, we get

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

Therefore

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

If $N < M$, the matrix $\boldsymbol{\Phi}^T\boldsymbol{\Phi} \in \mathbb{R}^{M\times M}$ is not invertible. The equation $\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w} - \boldsymbol{\Phi}^T\boldsymbol{y} = 0$ does not have only solution.
With the regulation, the normal equation is changed to $\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I}\right)\boldsymbol{w} - \boldsymbol{\Phi}^T\boldsymbol{y} = 0$, and the problem is fixed.

Problem 9

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

$$\hat{y}_i = \boldsymbol{w}^{*T}\boldsymbol{x_i} = \boldsymbol{w}_{new}^T x_{i,new} = a\boldsymbol{w}_{new}^T\boldsymbol{x_i}$$

Hence,

$$\boldsymbol{w_{new}} = \frac{1}{a}\boldsymbol{w}^*$$

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

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

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

We have

$$\begin{aligned}\boldsymbol{w_{new}} &= \left(\boldsymbol{X_{new}}^T\boldsymbol{X_{new}} + \lambda_{new}\boldsymbol{I}\right)^{-1}\boldsymbol{X_{new}}^T\boldsymbol{y}\\ &= \left(a^2\boldsymbol{X}^T\boldsymbol{X} + \lambda_{new}\boldsymbol{I}\right)^{-1}a\boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

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

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

which satisfies the condition.