Homework 3

Questions: https://github.com/SUSTech-ML-Course/Machine-Learning-Course/blob/05b4e0bd4063b9d486f34ce557720b519947cbe3/lesson-homework/hw3/HW3.md

Question 1

Compute the derivative with respect to $\boldsymbol{w}$

$$\nabla E_D(\boldsymbol{w}) = \sum_{n=1}^{N} r_n[t_n - \boldsymbol{w}^T\phi(\boldsymbol{x}_n)]\phi(\boldsymbol{x}_n)^T$$

Set the derivative to 0, we have

$$\sum_{n=1}^{N} r_n t_n \phi(\boldsymbol{x}_n)^T = \boldsymbol{w}^T \sum_{n=1}^{N} r_n \phi(\boldsymbol{x}_n)\phi(\boldsymbol{x}_n)^T$$

If we denote $\phi'(\boldsymbol{x_n}) = \sqrt{r_n}\phi(\boldsymbol{x_n})$ and $t'_n = \sqrt{r_n}t_n$, we can get the same equation (3.14) in the PRML textbook

$$\sum_{n=1}^{N} t'_n \phi'(\boldsymbol{x}_n)^T = \boldsymbol{w}^T \sum_{n=1}^{N} \phi'(\boldsymbol{x}_n)\phi'(\boldsymbol{x}_n)^T$$

Then we can get the solution

$$\boldsymbol{w}^* = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{t}$$

where

- $\boldsymbol{t} = [\sqrt{r_1}t_1, \sqrt{r_2}t_2, \ldots, \sqrt{r_n}t_n]^T$
- Φ is an $N \times M$ **design matrix**, whose elements are given by $\Phi_{ij} = \sqrt{r_i}\phi_j(\boldsymbol{x_i})$

$$\Phi = \begin{bmatrix} \sqrt{r_1}\phi_0(\boldsymbol{x_1}) & \sqrt{r_1}\phi_1(\boldsymbol{x_1}) & \cdots & \sqrt{r_1}\phi_{M-1}(\boldsymbol{x_1}) \\ \sqrt{r_2}\phi_0(\boldsymbol{x_2}) & \sqrt{r_2}\phi_1(\boldsymbol{x_2}) & \cdots & \sqrt{r_2}\phi_{M-1}(\boldsymbol{x_2}) \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{r_N}\phi_0(\boldsymbol{x_N}) & \sqrt{r_N}\phi_1(\boldsymbol{x_N}) & \cdots & \sqrt{r_N}\phi_{M-1}(\boldsymbol{x_N}) \end{bmatrix}$$

The two alternative interpretation:

(i) If we substitute β^{-1} by $r_n\beta^{-1}$ in the summation term, the equation in the question stem will become the expression above

(ii) r_n can also be viewed as the effective number of observation of $(\boldsymbol{x_n}, t_n)$

Question 2

$$
\begin{aligned}
p(\boldsymbol{w}, \boldsymbol{\beta}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{\beta}^{-1}\mathbf{S}_0)\,\mathrm{Gam}(\boldsymbol{\beta}|\boldsymbol{a}_0, \boldsymbol{b}_0) \\
&\propto (\frac{\beta}{|\mathbf{S_0}|})^2 exp(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m_0})^T\beta\mathbf{S_0^{-1}}(\boldsymbol{w}-\boldsymbol{m_0}))b_0^{a_0}\beta^{a_0-1}exp(-b_0\beta) \qquad (i)
\end{aligned}
$$

$$
\begin{aligned}
p(\mathbf{t}|\mathbf{X}, \boldsymbol{w}, \boldsymbol{\beta}) &= \prod_{n=1}^{N} \mathcal{N}(t_n|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n), \boldsymbol{\beta}^{-1}) \\
&\propto \prod_{n=1}^{N} \beta^{1/2} exp\big[-\frac{\beta}{2}(t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2\big] \qquad (ii)
\end{aligned}
$$

According to Bayesian Inference, we have $p(\boldsymbol{w}, \beta|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{X}, \boldsymbol{w}, \beta)p(\boldsymbol{w}, \beta)$

- The **quadratic term** in the exponent is

$$
\begin{aligned}
\text{quadratic term} &= -\frac{\beta}{2}\boldsymbol{w}^T\mathbf{S_0}^{-1}\boldsymbol{w} + \sum_{n=1}^{N} -\frac{\beta}{2}\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n})\boldsymbol{\phi}(\boldsymbol{x_n})^T\boldsymbol{w} \\
&= -\frac{\beta}{2}\boldsymbol{w}^T[\boldsymbol{S_0}^{-1} + \sum_{n=1}^{N}\phi(\boldsymbol{x_n})\phi(\boldsymbol{x_n})^T]\boldsymbol{w}
\end{aligned}
$$

  where the first term is from (i) and the second from (ii), then we can get

$$
\boldsymbol{S_N} = \left[\boldsymbol{S_0}^{-1} + \sum_{n=1}^{N}\phi(\boldsymbol{x_n})\phi(\boldsymbol{x_n})^T\right]^{-1} \qquad (1)
$$

- The **linear term** in the exponent is

$$
\begin{aligned}
\text{linear term} &= \beta\boldsymbol{m_0}^T\boldsymbol{S_0}^{-1}\boldsymbol{w} + \sum_{n=1}^{N}\beta t_n\phi(\boldsymbol{x_n})^T\boldsymbol{w} \\
&= \beta\left[\boldsymbol{m_0}^T\mathbf{S_0}^{-1} + \sum_{n=1}^{N} t_n\phi(\boldsymbol{x_n})^T\right]\boldsymbol{w}
\end{aligned}
$$

  Similarly, the first term is from (i) and the second from (ii), then we can also get

$$
\begin{aligned}
\boldsymbol{m}_N{}^T\boldsymbol{S}_N{}^{-1} &= \boldsymbol{m}_0{}^T\boldsymbol{S}_0{}^{-1} + \sum_{n=1}^{N} t_n\phi(\boldsymbol{x}_n)^T \\
\Rightarrow \qquad \boldsymbol{m_N} &= \boldsymbol{S_N}\left[\boldsymbol{S_0}^{-1}\boldsymbol{m_0} + \sum_{n=1}^{N} t_n\phi(\boldsymbol{x_n})\right] \qquad (2)
\end{aligned}
$$

- The **constant term** in the exponent is

$$
\begin{aligned}
\text{constant term} &= (-\frac{\beta}{2}\boldsymbol{m_0}^T\boldsymbol{S_0}^{-1}\boldsymbol{m_0} - b_0\beta) - \frac{\beta}{2}\sum_{n=1}^{N} t_n^2 \\
&= -\beta\left[\frac{1}{2}\boldsymbol{m_0}^T\boldsymbol{S_0}^{-1}\boldsymbol{m_0} + b_0 + \frac{1}{2}\sum_{n=1}^{N} t_n^2\right]
\end{aligned}
$$

  From which we can obtain

$$\frac{1}{2}\boldsymbol{m}_N{}^T\boldsymbol{S}_N{}^{-1}\boldsymbol{m}_N + b_N = \frac{1}{2}\boldsymbol{m}_0{}^T\boldsymbol{S}_0{}^{-1}\boldsymbol{m}_0 + b_0 + \frac{1}{2}\sum_{n=1}^{N} t_n^2$$

$$\Rightarrow \quad b_N = \frac{1}{2}\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + b_0 + \frac{1}{2}\sum_{n=1}^{N} t_n^2 - \frac{1}{2}\boldsymbol{m}_N^T\boldsymbol{S}_N^{-1}\boldsymbol{m}_N \qquad (3)$$

- The **exponential term** based on β is

$$\begin{aligned}\text{exponent term} &= (2 + a_0 - 1) + \frac{N}{2} \\ &= 2 + a_N - 1 \\ \Rightarrow \quad a_N &= a_0 + \frac{N}{2} \qquad (4)\end{aligned}$$

Question 3

According to the standard form of a multivariate normal distribution, we know that

$$\int \frac{1}{(2\pi)^{M/2}}\frac{1}{|\mathbf{A}|^{1/2}} exp\left[-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)^T\mathbf{A}(\boldsymbol{w}-\boldsymbol{m}_N)\right]d\boldsymbol{w} = 1$$

$$\Rightarrow \int exp\left[-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)^T\mathbf{A}(\boldsymbol{w}-\boldsymbol{m}_N)\right]d\boldsymbol{w} = (2\pi)^{M/2}|\mathbf{A}|^{1/2}$$

Since $E(\boldsymbol{m}_N)$ doesn't depend on $\boldsymbol{w}$, we can get the equation

$$\int \exp[-E(\mathbf{w})]\mathrm{d}\mathbf{w} = \exp[-E(\mathbf{m}_N)](2\pi)^{M/2}|\mathbf{A}|^{-1/2}$$

Then we substitute it into the equation (3.78) in PRML textbook

$$p(\mathbf{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\left[-E(\mathbf{w})\right]\mathrm{d}\mathbf{w} \qquad (3.78)$$

Finally we can get

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi)$$

Question 4

$$\begin{aligned}F(a) &= \frac{1}{2}\sum_i (Y_i - aX_i)^2 \\ \frac{\partial F(a)}{\partial a} &= \sum_i \hat{a}X_i^2 - X_iY_i = 0 \\ \hat{a} &= \frac{\sum_i X_iY_i}{\sum_i X_i^2}\end{aligned}$$

Question 5

$$
\begin{aligned}
\text{Likelihood} \quad L(\theta|y_1, y_2, \ldots, y_n) &= P(X = x_1|\theta)P(X = x_2|\theta)\cdots P(X = x_n|\theta) \\
&= e^{-\theta}\frac{\theta^{y_1}}{y_1!}\cdots e^{-\theta}\frac{\theta^{y_n}}{y_n!} \\
&= e^{-n\theta}\frac{\theta^{y_1+y_2+\ldots+y_n}}{y_1!y_2!\cdots y_n!} \\
&= e^{-n\theta}\frac{\theta^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!}
\end{aligned}
$$

$$
\text{Log-likelihood} \quad lnL(\theta|y_1, y_2, \ldots, y_n) = -n\theta + \left(\sum_{i=1}^n y_i\right)\ln\theta - \ln(\prod_{i=1}^n y_i!)
$$

Then we can get the MLE of θ

$$
\begin{aligned}
\frac{\partial lnL(\theta|y_1, y_2, \ldots, y_n)}{\partial\theta} &= -n + \frac{\sum_{i=1}^n y_i}{\hat{\theta}} = 0 \\
\hat{\theta} &= \frac{\sum_{i=1}^n y_i}{n}
\end{aligned}
$$

Question 6

$$
\begin{aligned}
\text{Likelihood} \quad L(\boldsymbol{X}|\alpha, \lambda) &= \prod_{i=1}^n f(X_i|\alpha, \lambda) \\
&= \left(\frac{\lambda^\alpha}{\Gamma(\alpha)}\right)^n \prod_{i=1}^n X_i^{\alpha-1}e^{-\lambda X_i}
\end{aligned}
$$

$$
\text{Log-likelihood} \quad lnL(\boldsymbol{X}|\alpha, \lambda) = (\alpha - 1)\sum_{i=1}^n lnX_i - \lambda\sum_{i=1}^n X_i + n\alpha ln\lambda - nln\Gamma(\alpha)
$$

$$
\frac{\partial lnL(\boldsymbol{X}|\alpha, \lambda)}{\partial\lambda} = -\sum_{i=1}^n X_i + \frac{n\alpha}{\hat{\lambda}} = 0
$$

$$
\Rightarrow \quad \hat{\lambda} = \frac{n\alpha}{\sum_{i=1}^n X_i} = \frac{\alpha}{\bar{X}}
$$