SOLUTION MANUAL FOR PATTERN RECOGNITION AND MACHINE LEARNING

EDITED BY

ZHENGQI GAO

the State Key Lab. of ASIC and System
School of Microelectronics
Fudan University

NOV.2017

Hence, we go back to deal with the Gaussian terms:

$$\text{Gaussian terms} = (\frac{\beta}{2\pi})^{N/2}\frac{|\mathbf{S_N}|^{1/2}}{|\mathbf{S_0}|^{1/2}}exp\{-(b_N-b_0)\beta\}$$

If we substitute the expressions above into $p(\mathbf{t})$, we will obtain (3.118) immediately.

0.4 Linear Models Classification

Problem 4.1 Solution

If the convex hull of $\{\mathbf{x_n}\}$ and $\{\mathbf{y_n}\}$ intersects, we know that there will be a point $\mathbf{z}$ which can be written as $\mathbf{z} = \sum_n \alpha_n \mathbf{x_n}$ and also $\mathbf{z} = \sum_n \beta_n \mathbf{y_n}$. Hence we can obtain:

$$\begin{aligned}
\widehat{\mathbf{w}}^T\mathbf{z}+w_0 &= \widehat{\mathbf{w}}^T(\sum_n \alpha_n\mathbf{x_n}) + w_0 \\
&= (\sum_n \alpha_n \widehat{\mathbf{w}}^T\mathbf{x_n}) + (\sum_n \alpha_n)w_0 \\
&= \sum_n \alpha_n(\widehat{\mathbf{w}}^T\mathbf{x_n} + w_0) \quad (*)
\end{aligned}$$

Where we have used $\sum_n \alpha_n = 1$. And if $\{\mathbf{x_n}\}$ and $\{\mathbf{y_n}\}$ are linearly separable, we have $\widehat{\mathbf{w}}^T\mathbf{x_n}+w_0 > 0$ and $\widehat{\mathbf{w}}^T\mathbf{y_n}+w_0 < 0$, for $\forall \mathbf{x_n}$, $\mathbf{y_n}$. Together with $\alpha_n \geq 0$ and $(*)$, we know that $\widehat{\mathbf{w}}^T\mathbf{z}+w_0 > 0$. And if we calculate $\widehat{\mathbf{w}}^T\mathbf{z}+w_0$ from the perspective of $\{\mathbf{y_n}\}$ following the same procedure, we can obtain $\widehat{\mathbf{w}}^T\mathbf{z}+w_0 < 0$. Hence contradictory occurs. In other words, they are not linearly separable if their convex hulls intersect.

We have already proved the first statement, i.e., "convex hulls intersect" gives "not linearly separable", and what the second part wants us to prove is that "linearly separable" gives "convex hulls do not intersect". This can be done simply by contrapositive.

The true converse of the first statement should be if their convex hulls do not intersect, the data sets should be linearly separable. This is exactly what Hyperplane Separation Theorem shows us.

Problem 4.2 Solution

Let's make the dependency of $E_D(\widetilde{\mathbf{W}})$ on w_0 explicitly:

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\{(\mathbf{XW}+\mathbf{1w_0}^T-\mathbf{T})^T(\mathbf{XW}+\mathbf{1w_0}^T-\mathbf{T})\}$$

Then we calculate the derivative of $E_D(\widetilde{\mathbf{W}})$ with respect to $\mathbf{w_0}$:

$$\frac{\partial E_D(\widetilde{\mathbf{W}})}{\partial \mathbf{w_0}} = 2N\mathbf{w_0} + 2(\mathbf{XW}-\mathbf{T})^T\mathbf{1}$$

Where we have used the property:

$$\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}\left[(\mathbf{AXB}+\mathbf{C})(\mathbf{AXB}+\mathbf{C})^T\right] = 2\mathbf{A}^T(\mathbf{AXB}+\mathbf{C})\mathbf{B}^T$$

We set the derivative equals to 0, which gives:

$$\mathbf{w_0} = -\frac{1}{N}(\mathbf{XW}-\mathbf{T})^T\mathbf{1} = \bar{\mathbf{t}} - \mathbf{W}^T\bar{\mathbf{x}}$$

Where we have denoted:

$$\bar{\mathbf{t}} = \frac{1}{N}\mathbf{T}^T\mathbf{1}, \quad \text{and} \quad \bar{\mathbf{x}} = \frac{1}{N}\mathbf{X}^T\mathbf{1}$$

If we substitute the equations above into $E_D(\widetilde{\mathbf{W}})$, we can obtain:

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{XW}+\bar{\mathbf{T}}-\bar{\mathbf{X}}\mathbf{W}-\mathbf{T})^T(\mathbf{XW}+\bar{\mathbf{T}}-\bar{\mathbf{X}}\mathbf{W}-\mathbf{T})\right\}$$

Where we further denote

$$\bar{\mathbf{T}} = \mathbf{1}\bar{\mathbf{t}}^T, \quad \text{and} \quad \bar{\mathbf{X}} = \mathbf{1}\bar{\mathbf{x}}^T$$

Then we set the derivative of $E_D(\widetilde{\mathbf{W}})$ with regard to $\mathbf{W}$ to 0, which gives:

$$\mathbf{W} = \widehat{\mathbf{X}}^{\dagger}\widehat{\mathbf{T}}$$

Where we have defined:

$$\widehat{\mathbf{X}} = \mathbf{X}-\bar{\mathbf{X}}, \quad \text{and} \quad \widehat{\mathbf{T}} = \mathbf{T}-\bar{\mathbf{T}}$$

Now consider the prediction for a new given $\mathbf{x}$, we have:

$$\begin{aligned}
\mathbf{y}(\mathbf{x}) &= \mathbf{W}^T\mathbf{x}+\mathbf{w_0} \\
&= \mathbf{W}^T\mathbf{x}+\bar{\mathbf{t}}-\mathbf{W}^T\bar{\mathbf{x}} \\
&= \bar{\mathbf{t}} + \mathbf{W}^T(\mathbf{x}-\bar{\mathbf{x}})
\end{aligned}$$

If we know that $\mathbf{a}^T\mathbf{t_n}+b = 0$ holds for some $\mathbf{a}$ and b, we can obtain:

$$\mathbf{a}^T\bar{\mathbf{t}} = \frac{1}{N}\mathbf{a}^T\mathbf{T}^T\mathbf{1} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{a^T}\mathbf{t_n} = -b$$

Therefore,

$$\begin{aligned}
\mathbf{a}^T\mathbf{y}(\mathbf{x}) &= \mathbf{a}^T\left[\bar{\mathbf{t}} + \mathbf{W}^T(\mathbf{x}-\bar{\mathbf{x}})\right] \\
&= \mathbf{a}^T\bar{\mathbf{t}} + \mathbf{a}^T\mathbf{W}^T(\mathbf{x}-\bar{\mathbf{x}}) \\
&= -b + \mathbf{a}^T\widehat{\mathbf{T}}^T(\widehat{\mathbf{X}}^{\dagger})^T(\mathbf{x}-\bar{\mathbf{x}}) \\
&= -b
\end{aligned}$$

Where we have used:

$$\begin{aligned}\mathbf{a}^T\widehat{\mathbf{T}}^T &= \mathbf{a}^T(\mathbf{T}-\bar{\mathbf{T}})^T = \mathbf{a}^T(\mathbf{T}-\frac{1}{N}\mathbf{1}\mathbf{1}^\mathbf{T}\mathbf{T})^T \\ &= \mathbf{a}^T\mathbf{T}^T - \frac{1}{N}\mathbf{a}^T\mathbf{T}^T\mathbf{1}\mathbf{1}^\mathbf{T} = -b\mathbf{1}^T + b\mathbf{1}^T \\ &= \mathbf{0}^T\end{aligned}$$

Problem 4.3 Solution

Suppose there are Q constraints in total. We can write $\mathbf{a_q}^T\mathbf{t_n}+b_q = 0$, $q = 1,2,...,Q$ for all the target vector $\mathbf{t_n}$, $n = 1,2...,N$. Or alternatively, we can group them together:

$$\mathbf{A}^T\mathbf{t_n} + \mathbf{b} = \mathbf{0}$$

Where $\mathbf{A}$ is a $Q \times Q$ matrix, and the qth column of $\mathbf{A}$ is $\mathbf{a_q}$, and meanwhile $\mathbf{b}$ is a $Q \times 1$ column vector, and the qth element is $\mathbf{b_q}$. for every pair of $\{\mathbf{a_q}, b_q\}$ we can follow the same procedure in the previous problem to show that $\mathbf{a_q}\mathbf{y}(\mathbf{x}) + b_q = 0$. In other words, the proofs will not affect each other. Therefore, it is obvious :

$$\mathbf{A}^T\mathbf{y}(\mathbf{x}) + \mathbf{b} = \mathbf{0}$$

Problem 4.4 Solution

We use Lagrange multiplier to enforce the constraint $\mathbf{w}^T\mathbf{w} = 1$. We now need to maximize :

$$L(\lambda, \mathbf{w}) = \mathbf{w}^T(\mathbf{m_2} - \mathbf{m_1}) + \lambda(\mathbf{w}^T\mathbf{w} - 1)$$

We calculate the derivatives:

$$\frac{\partial L(\lambda, \mathbf{w})}{\partial \lambda} = \mathbf{w}^T\mathbf{w} - 1$$

And

$$\frac{\partial L(\lambda, \mathbf{w})}{\partial \mathbf{w}} = \mathbf{m_2} - \mathbf{m_1} + 2\lambda\mathbf{w}$$

We set the derivatives above equals to 0, which gives:

$$\mathbf{w} = -\frac{1}{2\lambda}(\mathbf{m_2} - \mathbf{m_1}) \propto (\mathbf{m_2} - \mathbf{m_1})$$

Problem 4.5 Solution

We expand (4.25) using (4.22), (4.23) and (4.24).

$$\begin{aligned}J(\mathbf{w}) &= \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \\ &= \frac{||\mathbf{w}^T(\mathbf{m_2} - \mathbf{m_1})||^2}{\sum_{n \in C_1}(\mathbf{w}^T\mathbf{x_n} - m_1)^2 + \sum_{n \in C_2}(\mathbf{w}^T\mathbf{x_n} - m_2)^2}\end{aligned}$$

The numerator can be further written as:

$$\text{numerator} = \left[\mathbf{w}^T(\mathbf{m_2}-\mathbf{m_1})\right]\left[\mathbf{w}^T(\mathbf{m_2}-\mathbf{m_1})\right]^T = \mathbf{w}^T\mathbf{S_B}\mathbf{w}$$

Where we have defined:

$$\mathbf{S_B} = (\mathbf{m_2}-\mathbf{m_1})(\mathbf{m_2}-\mathbf{m_1})^T$$

And ti is the same for the denominator:

$$\begin{aligned}\text{denominator} &= \sum_{n\in C_1}[\mathbf{w}^T(\mathbf{x_n}-\mathbf{m_1})]^2 + \sum_{n\in C_2}[\mathbf{w}^T(\mathbf{x_n}-\mathbf{m_2})]^2 \\ &= \mathbf{w}^T\mathbf{S_{w1}}\mathbf{w} + \mathbf{w}^T\mathbf{S_{w2}}\mathbf{w} \\ &= \mathbf{w}^T\mathbf{S_w}\mathbf{w}\end{aligned}$$

Where we have defined:

$$\mathbf{S_w} = \sum_{n\in C_1}(\mathbf{x_n}-\mathbf{m_1})(\mathbf{x_n}-\mathbf{m_1})^T + \sum_{n\in C_2}(\mathbf{x_n}-\mathbf{m_2})(\mathbf{x_n}-\mathbf{m_2})^T$$

Just as required.

Problem 4.6 Solution

Let's follow the hint, beginning by expanding (4.33).

$$\begin{aligned}(4.33) &= \sum_{n=1}^{N}\mathbf{w}^T\mathbf{x_n}\mathbf{x_n} + w_0\sum_{n=1}^{N}\mathbf{x_n} - \sum_{n=1}^{N}t_n\mathbf{x_n} \\ &= \sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T\mathbf{w} - \mathbf{w}^T\mathbf{m}\sum_{n=1}^{N}\mathbf{x_n} - (\sum_{n\in C_1}t_n\mathbf{x_n} + \sum_{n\in C_2}t_n\mathbf{x_n}) \\ &= \sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T\mathbf{w} - \mathbf{w}^T\mathbf{m}\cdot(N\mathbf{m}) - (\sum_{n\in C_1}\frac{N}{N_1}\mathbf{x_n} + \sum_{n\in C_2}\frac{-N}{N_2}\mathbf{x_n}) \\ &= \sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T\mathbf{w} - N\mathbf{w}^T\mathbf{m}\mathbf{m} - N(\sum_{n\in C_1}\frac{1}{N_1}\mathbf{x_n} - \sum_{n\in C_2}\frac{1}{N_2}\mathbf{x_n}) \\ &= \sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T\mathbf{w} - N\mathbf{m}\mathbf{m}^T\mathbf{w} - N(\mathbf{m_1}-\mathbf{m_2}) \\ &= [\sum_{n=1}^{N}(\mathbf{x_n}\mathbf{x_n}^T) - N\mathbf{m}\mathbf{m}^T]\mathbf{w} - N(\mathbf{m_1}-\mathbf{m_2})\end{aligned}$$

If we let the derivative equal to 0, we will see that:

$$[\sum_{n=1}^{N}(\mathbf{x_n}\mathbf{x_n}^T) - N\mathbf{m}\mathbf{m}^T]\mathbf{w} = N(\mathbf{m_1}-\mathbf{m_2})$$

Therefore, now we need to prove:

$$\sum_{n=1}^{N}(\mathbf{x_n}\mathbf{x_n}^T) - N\mathbf{m}\mathbf{m}^T = \mathbf{S_w} + \frac{N_1N_2}{N}\mathbf{S_B}$$

Let's expand the left side of the equation above:

$$\begin{aligned}
\text{left} &= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T - N(\frac{N_1}{N}\mathbf{m_1} + \frac{N_2}{N}\mathbf{m_2})^2 \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T - N(\frac{N_1^2}{N^2}||\mathbf{m_1}||^2 + \frac{N_2^2}{N^2}||\mathbf{m_2}||^2 + 2\frac{N_1N_2}{N^2}\mathbf{m_1}\mathbf{m_2}^T) \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T - \frac{N_1^2}{N}||\mathbf{m_1}||^2 - \frac{N_2^2}{N}||\mathbf{m_2}||^2 - 2\frac{N_1N_2}{N}\mathbf{m_1}\mathbf{m_2}^T \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T + (N_1 + \frac{N_1N_2}{N} - 2N_1)||\mathbf{m_1}||^2 + (N_2 + \frac{N_1N_2}{N} - 2N_2)||\mathbf{m_2}||^2 - 2\frac{N_1N_2}{N}\mathbf{m_1}\mathbf{m_2}^T \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T + (N_1 - 2N_1)||\mathbf{m_1}||^2 + (N_2 - 2N_2)||\mathbf{m_2}||^2 + \frac{N_1N_2}{N}||\mathbf{m_1} - \mathbf{m_2}||^2 \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T + N_1||\mathbf{m_1}||^2 - 2\mathbf{m_1}\cdot(N_1\mathbf{m_1}^T) + N_2||\mathbf{m_2}||^2 - 2\mathbf{m_2}\cdot(N_2\mathbf{m_2}^T) + \frac{N_1N_2}{N}\mathbf{S_B} \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T + N_1||\mathbf{m_1}||^2 - 2\mathbf{m_1}\sum_{n\in C_1} x_n^T + N_2||\mathbf{m_2}||^2 - 2\mathbf{m_2}\sum_{n\in C_2} x_n^T + \frac{N_1N_2}{N}\mathbf{S_B} \\
&= \sum_{n\in C_1} \mathbf{x_n}\mathbf{x_n}^T + N_1||\mathbf{m_1}||^2 - 2\mathbf{m_1}\sum_{n\in C_1} x_n^T \\
&\quad + \sum_{n\in C_2} \mathbf{x_n}\mathbf{x_n}^T + N_2||\mathbf{m_2}||^2 - 2\mathbf{m_2}\sum_{n\in C_2} x_n^T + \frac{N_1N_2}{N}\mathbf{S_B} \\
&= \sum_{n\in C_1} (\mathbf{x_n}\mathbf{x_n}^T + ||\mathbf{m_1}||^2 - 2\mathbf{m_1}x_n^T) + \sum_{n\in C_2} (\mathbf{x_n}\mathbf{x_n}^T + ||\mathbf{m_2}||^2 - 2\mathbf{m_2}\mathbf{x_n}^T) + \frac{N_1N_2}{N}\mathbf{S_B} \\
&= \sum_{n\in C_1} ||\mathbf{x_n} - \mathbf{m_1}||^2 + \sum_{n\in C_2} ||\mathbf{x_n} - \mathbf{m_2}||^2 + \frac{N_1N_2}{N}\mathbf{S_B} \\
&= \mathbf{S_w} + \frac{N_1N_2}{N}\mathbf{S_B}
\end{aligned}$$

Just as required.

Problem 4.7 Solution

This problem is quite simple. We can solve it by definition. We know that logistic sigmoid function has the form:

$$\sigma(a) = \frac{1}{1 + exp(-a)}$$

Therefore, we can obtain:

$$\begin{aligned}
\sigma(a) + \sigma(-a) &= \frac{1}{1 + exp(-a)} + \frac{1}{1 + exp(a)} \\
&= \frac{2 + exp(a) + exp(-a)}{[1 + exp(-a)][1 + exp(a)]} \\
&= \frac{2 + exp(a) + exp(-a)}{2 + exp(a) + exp(-a)} = 1
\end{aligned}$$

Next we exchange the dependent and independent variables to obtain its inverse.

$$a = \frac{1}{1+exp(-y)}$$

We first rearrange the equation above, which gives:

$$exp(-y) = \frac{1-a}{a}$$

Then we calculate the logarithm for both sides, which gives:

$$y = \ln(\frac{a}{1-a})$$

Just as required.

Problem 4.8 Solution

According to (4.58) and (4.64), we can write:

$$\begin{aligned} a &= \ln\frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)} \\ &= \ln p(\mathbf{x}|C_1) - \ln p(\mathbf{x}|C_2) + \ln\frac{p(C_1)}{p(C_2)} \\ &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_1})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_1}) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_2})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_2}) + \ln\frac{p(C_1)}{p(C_2)} \\ &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})\mathbf{x} - \frac{1}{2}\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} + \frac{1}{2}\boldsymbol{\mu_2}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_2} + \ln\frac{p(C_1)}{p(C_2)} \\ &= \mathbf{w}^T\mathbf{x} + w_0 \end{aligned}$$

Where in the last second step, we rearrange the term according to **x**, i.e., its quadratic, linear, constant term. We have also defined :

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})$$

And

$$w_0 = -\frac{1}{2}\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} + \frac{1}{2}\boldsymbol{\mu_2}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_2} + \ln\frac{p(C_1)}{p(C_2)}$$

Finally, since $p(C_1|\mathbf{x}) = \sigma(a)$ as stated in (4.57), we have $p(C_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}+w_0)$ just as required.

Problem 4.9 Solution

We begin by writing down the likelihood function.

$$\begin{aligned} p(\{\phi_{\mathbf{n}}, t_n\}|\pi_1, \pi_2, ..., \pi_K) &= \prod_{n=1}^{N}\prod_{k=1}^{K}[p(\boldsymbol{\phi_n}|C_k)\,p(C_k)]^{t_{nk}} \\ &= \prod_{n=1}^{N}\prod_{k=1}^{K}[\pi_k\,p(\boldsymbol{\phi_n}|C_k)]^{t_{nk}} \end{aligned}$$

Hence we can obtain the expression for the logarithm likelihood:

$$\ln p = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left[\ln\pi_k + \ln p(\boldsymbol{\phi_n}|C_k)\right] \propto \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ \ln\pi_k$$

Since there is a constraint on π_k, so we need to add a Lagrange Multiplier to the expression, which becomes:

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ \ln\pi_k + \lambda(\sum_{k=1}^{K}\pi_k - 1)$$

We calculate the derivative of the expression above with regard to π_k:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N}\frac{t_{nk}}{\pi_k} + \lambda$$

And if we set the derivative equal to 0, we can obtain:

$$\pi_k = -(\sum_{n=1}^{N} t_{nk})/\lambda = -\frac{N_k}{\lambda} \qquad (*)$$

And if we preform summation on both sides with regard to k, we can see that:

$$1 = -(\sum_{k=1}^{K} N_k)/\lambda = -\frac{N}{\lambda}$$

Which gives $\lambda = -N$, and substitute it into $(*)$, we can obtain $\pi_k = N_k/N$.

Problem 4.10 Solution

This time, we focus on the term which dependent on $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma}$ in the logarithm likelihood.

$$\ln p = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left[\ln\pi_k + \ln p(\boldsymbol{\phi_n}|C_k)\right] \propto \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ \ln p(\boldsymbol{\phi_n}|C_k)$$

Provided $p(\phi|C_k) = \mathcal{N}(\phi|\boldsymbol{\mu_k},\boldsymbol{\Sigma})$, we can further derive:

$$\ln p \propto \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left[-\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})^T\right]$$

We first calculate the derivative of the expression above with regard to $\boldsymbol{\mu_k}$:

$$\frac{\partial \ln p}{\partial \boldsymbol{\mu_k}} = \sum_{n=1}^{N} t_{nk}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})$$

We set the derivative equals to 0, which gives:

$$\sum_{n=1}^{N} t_{nk}\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi_n} = \sum_{n=1}^{N} t_{nk}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_k} = N_k\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_k}$$

Therefore, if we multiply both sides by $\boldsymbol{\Sigma}/N_k$, we will obtain (4.161). Now let's calculate the derivative of $\ln p$ with regard to $\boldsymbol{\Sigma}$, which gives:

$$\begin{aligned}
\frac{\partial \ln p}{\partial \boldsymbol{\Sigma}} &= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\,(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}) - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})^T \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} -\frac{t_{nk}}{2}\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}\sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})^T \\
&= \sum_{n=1}^{N} -\frac{1}{2}\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}}\sum_{k=1}^{K} N_k \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S_k}) \\
&= -\frac{N}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\sum_{k=1}^{K} N_k \boldsymbol{\Sigma}^{-1}\mathbf{S_k}\boldsymbol{\Sigma}^{-1}
\end{aligned}$$

Where we have denoted

$$\mathbf{S_k} = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})^T$$

Now we set the derivative equals to 0, and rearrange the equation, which gives:

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K}\frac{N_k}{N}\mathbf{S_k}$$

Problem 4.11 Solution

Based on definition, we can write down

$$p(\boldsymbol{\phi}|C_k) = \prod_{m=1}^{M}\prod_{l=1}^{L}\mu_{kml}^{\phi_{ml}}$$

Note that here only one of the value among ϕ_{m1}, ϕ_{m2}, ... ϕ_{mL} is 1, and the others are all 0 because we have used a $1-\mathrm{of}-L$ binary coding scheme, and also we have taken advantage of the assumption that the M components of $\boldsymbol{\phi}$ are independent conditioned on the class C_k. We substitute the expression above into (4.63), which gives:

$$a_k = \sum_{m=1}^{M}\sum_{l=1}^{L}\phi_{ml}\cdot\ln\mu_{kml} + \ln p(C_k)$$

Hence it is obvious that a_k is a linear function of the components of $\boldsymbol{\phi}$.

Problem 4.12 Solution

Based on definition, i.e., (4.59), we know that logistic sigmoid has the form:

$$\sigma(a) = \frac{1}{1+exp(-a)}$$

Now, we calculate its derivative with regard to a.

$$\frac{d\sigma(a)}{da} = \frac{exp(a)}{[\,1+exp(-a)\,]^2} = \frac{exp(a)}{1+exp(-a)} \cdot \frac{1}{1+exp(-a)} = [\,1-\sigma(a)\,]\cdot\,\sigma(a)$$

Just as required.

Problem 4.13 Solution

Let's follow the hint.

$$\begin{aligned}
\nabla E(\boldsymbol{w}) &= -\nabla \sum_{n=1}^{N} \{\, t_n \ln y_n + (1-t_n)\ln(1-y_n)\,\} \\
&= -\sum_{n=1}^{N} \nabla\{\, t_n \ln y_n + (1-t_n)\ln(1-y_n)\,\} \\
&= -\sum_{n=1}^{N} \frac{d\{\, t_n \ln y_n + (1-t_n)\ln(1-y_n)\,\}}{dy_n} \frac{dy_n}{da_n} \frac{da_n}{d\mathbf{w}} \\
&= -\sum_{n=1}^{N} \left(\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}\right)\cdot y_n\,(1-y_n)\cdot \boldsymbol{\phi_n} \\
&= -\sum_{n=1}^{N} \frac{t_n - y_n}{y_n(1-y_n)} \cdot y_n\,(1-y_n)\cdot \boldsymbol{\phi_n} \\
&= -\sum_{n=1}^{N} (t_n - y_n)\boldsymbol{\phi_n} \\
&= \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi_n}
\end{aligned}$$

Where we have used $y_n = \sigma(a_n)$, $a_n = \mathbf{w}^T \boldsymbol{\phi_n}$, the chain rules and (4.88).

Problem 4.14 Solution

According to definition, we know that if a dataset is linearly separable, we can find $\mathbf{w}$, for some points $\mathbf{x_n}$, we have $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_n}) > 0$, and the others $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_m}) < 0$. Then the boundary is given by $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) = 0$. Note that for any point $\mathbf{x_0}$ in the dataset, the value of $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_0})$ should either be positive or negative, but it can not equal to 0.

Therefore, the maximum likelihood solution for logistic regression is trivial. We suppose for those points $\mathbf{x_n}$ belonging to class C_1, we have $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_n}) > 0$ and $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_m}) < 0$ for those belonging to class C_2. According to (4.87), if $|\mathbf{w}| \to \infty$, we have

$$p(C_1|\boldsymbol{\phi}(\mathbf{x_n})) = \sigma(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_n})) \to 1$$

Where we have used $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_n}) \to +\infty$. And since $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_m}) \to -\infty$, we can also obtain:

$$p(C_2|\boldsymbol{\phi}(\mathbf{x_m})) = 1 - p(C_1|\boldsymbol{\phi}(\mathbf{x_m})) = 1 - \sigma(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_m})) \to 1$$

In other words, for the likelihood function, i.e.,(4.89), if we have $|\mathbf{w}| \to \infty$, and also we label all the points lying on one side of the boundary as class C_1, and those on the other side as class C_2, the every term in (4.89) can achieve its maximum value, i.e., 1, finally leading to the maximum of the likelihood.

Hence, for a linearly separable dataset, the learning process may prefer to make $|\mathbf{w}| \to \infty$ and use the linear boundary to label the datasets, which can cause severe over-fitting problem.

Problem 4.15 Solution(Waiting for update)

Since y_n is the output of the logistic sigmoid function, we know that $0 < y_n < 1$ and hence $y_n(1-y_n) > 0$. Then we use (4.97), for an arbitrary non-zero real vector $\mathbf{a} \neq \mathbf{0}$, we have:

$$\begin{aligned}\mathbf{a}^T\mathbf{H}\mathbf{a} &= \mathbf{a}^T\Big[\sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\Big]\mathbf{a} \\ &= \sum_{n=1}^{N} y_n(1-y_n)(\boldsymbol{\phi}_n^T\mathbf{a})^T(\boldsymbol{\phi}_n^T\mathbf{a}) \\ &= \sum_{n=1}^{N} y_n(1-y_n)\,b_n^2\end{aligned}$$

Where we have denoted $b_n = \boldsymbol{\phi}_n^T\mathbf{a}$. What's more, there should be at least one of $\{b_1, b_2, ..., b_N\}$ not equal to zero and then we can see that the expression above is larger than 0 and hence $\mathbf{H}$ is positive definite.

Otherwise, if all the $b_n = 0$, $\mathbf{a} = [a_1, a_2, ..., a_M]^T$ will locate in the null space of matrix $\boldsymbol{\Phi}_{N\times M}$. However, with regard to the *rank-nullity theorem*, we know that Rank($\boldsymbol{\Phi}$) + Nullity($\boldsymbol{\Phi}$) =M, and we have already assumed that those M features are independent, i.e., Rank($\boldsymbol{\Phi}$) = M, which means there is only $\mathbf{0}$ in its null space. Therefore contradictory occurs.

Problem 4.16 Solution

We still denote $y_n = p(t = 1|\boldsymbol{\phi_n})$, and then we can write down the log likelihood by replacing t_n with π_n in (4.89) and (4.90).

$$\ln p(\mathbf{t}|\mathbf{w}) = \sum_{n=1}^{N}\{\pi_n \ln y_n + (1-\pi_n)\ln(1-y_n)\}$$

Problem 4.17 Solution

We should discuss in two situations separately, namely $j = k$ and $j \neq k$. When $j \neq k$, we have:

$$\frac{\partial y_k}{\partial a_j} = \frac{-exp(a_k)\cdot exp(a_j)}{[\sum_j exp(a_j)]^2} = -y_k \cdot y_j$$

And when $j = k$, we have:

$$\frac{\partial y_k}{\partial a_k} = \frac{exp(a_k)\sum_j exp(a_j) - exp(a_k)exp(a_k)}{[\sum_j exp(a_j)]^2} = y_k - y_k^2 = y_k(1-y_k)$$

Therefore, we can obtain:

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$$

Where I_{kj} is the elements of the indentity matrix.

Problem 4.18 Solution

We derive every term $t_{nk}\ln y_{nk}$ with regard to a_j.

$$\begin{aligned}\frac{\partial t_{nk}\ln y_{nk}}{\partial \mathbf{w_j}} &= \frac{\partial t_{nk}\ln y_{nk}}{\partial y_{nk}}\frac{\partial y_{nk}}{\partial a_j}\frac{\partial a_j}{\partial \mathbf{w_j}} \\ &= t_{nk}\frac{1}{y_{nk}}\cdot y_{nk}(I_{kj} - y_{nj})\cdot \boldsymbol{\phi_n} \\ &= t_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi_n}\end{aligned}$$

Where we have used (4.105) and (4.106). Next we perform summation over n and k.

$$\begin{aligned}\nabla_{\mathbf{w_j}}E &= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi_n} \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\,y_{nj}\,\boldsymbol{\phi_n} - \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\,I_{kj}\,\boldsymbol{\phi_n} \\ &= \sum_{n=1}^{N}\big[(\sum_{k=1}^{K} t_{nk})y_{nj}\,\boldsymbol{\phi_n}\big] - \sum_{n=1}^{N} t_{nj}\boldsymbol{\phi_n} \\ &= \sum_{n=1}^{N} y_{nj}\,\boldsymbol{\phi_n} - \sum_{n=1}^{N} t_{nj}\boldsymbol{\phi_n} \\ &= \sum_{n=1}^{N}(y_{nj} - t_{nj})\boldsymbol{\phi_n}\end{aligned}$$

Where we have used the fact that for arbitrary n, we have $\sum_{k=1}^{K} t_{nk} = 1$.

Problem 4.19 Solution

We write down the log likelihood.

$$\ln p(\mathbf{t}|\mathbf{w}) = \sum_{n=1}^{N}\big\{t_n\ln y_n + (1 - t_n)\ln(1 - y_n)\big\}$$

Therefore, we can obtain:

$$\begin{aligned}\nabla_{\mathbf{w}}\ln p &= \frac{\partial \ln p}{\partial y_n}\cdot\frac{\partial y_n}{\partial a_n}\cdot\frac{\partial a_n}{\partial \mathbf{w}} \\ &= \sum_{n=1}^{N}(\frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n})\Phi'(a_n)\boldsymbol{\phi_n} \\ &= \sum_{n=1}^{N}\frac{y_n - t_n}{y_n(1 - y_n)}\Phi'(a_n)\boldsymbol{\phi_n}\end{aligned}$$

Where we have used $y = p(t=1|a) = \Phi(a)$ and $a_n = \mathbf{w}^T \boldsymbol{\phi_n}$. According to (4.114), we can obtain:

$$\Phi'(a) = \mathcal{N}(\theta|0,1)\big|_{\theta=a} = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}a^2)$$

Hence, we can obtain:

$$\nabla_{\mathbf{w}} \ln p = \sum_{n=1}^{N} \frac{y_n - t_n}{y_n(1-y_n)} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} \boldsymbol{\phi_n}$$

To calculate the Hessian Matrix, we need to first evaluate several derivatives.

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)}\} &= \frac{\partial}{\partial y_n}\{\frac{y_n - t_n}{y_n(1-y_n)}\} \cdot \frac{\partial y_n}{\partial a_n} \cdot \frac{\partial a_n}{\partial \mathbf{w}} \\
&= \frac{y_n(1-y_n) - (y_n - t_n)(1-2y_n)}{[y_n(1-y_n)]^2} \Phi'(a_n)\boldsymbol{\phi_n} \\
&= \frac{y_n^2 + t_n - 2y_n t_n}{y_n^2(1-y_n)^2} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} \boldsymbol{\phi_n}
\end{aligned}$$

And

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\{\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} &= \frac{\partial}{\partial a_n}\{\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} \frac{\partial a_n}{\partial \mathbf{w}} \\
&= -\frac{a_n}{\sqrt{2\pi}} exp(-\frac{a_n^2}{2}) \boldsymbol{\phi_n}
\end{aligned}$$

Therefore, using the chain rule, we can obtain:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} &= \frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)}\} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} + \frac{y_n - t_n}{y_n(1-y_n)} \frac{\partial}{\partial \mathbf{w}}\{\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} \\
&= [\frac{y_n^2 + t_n - 2y_n t_n}{y_n(1-y_n)} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} - a_n(y_n - t_n)] \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi} y_n(1-y_n)} \boldsymbol{\phi_n}
\end{aligned}$$

Finally if we perform summation over n, we can obtain the Hessian Matrix:

$$\begin{aligned}
\mathbf{H} &= \nabla\nabla_{\mathbf{w}} \ln p \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} \cdot \boldsymbol{\phi_n} \\
&= \sum_{n=1}^{N} [\frac{y_n^2 + t_n - 2y_n t_n}{y_n(1-y_n)} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} - a_n(y_n - t_n)] \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi} y_n(1-y_n)} \boldsymbol{\phi_n} \boldsymbol{\phi_n}^T
\end{aligned}$$

Problem 4.20 Solution(waiting for update)

We know that the Hessian Matrix is of size $MK \times MK$, and the (j,k)th block with size $M \times M$ is given by (4.110), where $j,k = 1,2,...,K$. Therefore, we can obtain:

$$\mathbf{u^T H u} = \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbf{u_j^T H_{j,k} u_k} \qquad (*)$$

Where we use $\mathbf{u_k}$ to denote the kth block vector of $\mathbf{u}$ with size $M \times 1$, and $\mathbf{H_{j,k}}$ to denote the (j,k)th block matrix of $\mathbf{H}$ with size $M \times M$. Then based on (4.110), we further expand (4.110):

$$\begin{aligned}
(*) &= \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbf{u_j^T}\{-\sum_{n=1}^{N} y_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}\mathbf{u_k} \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{n=1}^{N} \mathbf{u_j^T}\{-y_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}\mathbf{u_k} \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{n=1}^{N} \mathbf{u_j^T}\{-y_{nk}I_{kj}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}\mathbf{u_k} + \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{n=1}^{N} \mathbf{u_j^T}\{y_{nk}y_{nj}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}\mathbf{u_k} \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbf{u_k^T}\{-y_{nk}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}\mathbf{u_k} + \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{n=1}^{N} y_{nj}\mathbf{u_j^T}\{\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}y_{nk}\mathbf{u_k}
\end{aligned}$$

Problem 4.21 Solution

It is quite obvious.

$$\begin{aligned}
\Phi(a) &= \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)d\theta \\
&= \frac{1}{2} + \int_{0}^{a} \mathcal{N}(\theta|0,1)d\theta \\
&= \frac{1}{2} + \int_{0}^{a} \mathcal{N}(\theta|0,1)d\theta \\
&= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_{0}^{a} exp(-\theta^2/2)d\theta \\
&= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\frac{\sqrt{\pi}}{2}\int_{0}^{a} \frac{2}{\sqrt{\pi}}exp(-\theta^2/2)d\theta \\
&= \frac{1}{2}(1 + \frac{1}{\sqrt{2}}\int_{0}^{a} \frac{2}{\sqrt{\pi}}exp(-\theta^2/2)d\theta) \\
&= \frac{1}{2}\{1 + \frac{1}{\sqrt{2}}erf(a)\}
\end{aligned}$$

Where we have used

$$\int_{-\infty}^{0} \mathcal{N}(\theta|0,1)d\theta = \frac{1}{2}$$

Problem 4.22 Solution

If we denote $f(\boldsymbol{\theta}) = p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$, we can write:

$$\begin{aligned} p(D) &= \int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta} = \int f(\boldsymbol{\theta})\,d\boldsymbol{\theta} \\ &= f(\boldsymbol{\theta}_{MAP})\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \\ &= p(D|\boldsymbol{\theta}_{MAP})p(\boldsymbol{\theta}_{MAP})\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \end{aligned}$$

Where $\boldsymbol{\theta}_{MAP}$ is the value of $\boldsymbol{\theta}$ at the mode of $f(\boldsymbol{\theta})$, $\mathbf{A}$ is the Hessian Matrix of $-\ln f(\boldsymbol{\theta})$ and we have also used (4.135). Therefore,

$$\ln p(D) = \ln p(D|\boldsymbol{\theta}_{MAP}) + \ln p(\boldsymbol{\theta}_{MAP}) + \frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{A}|$$

Just as required.

Problem 4.23 Solution

According to (4.137), we can write:

$$\begin{aligned} \ln p(D) &= \ln p(D|\boldsymbol{\theta}_{MAP}) + \ln p(\boldsymbol{\theta}_{MAP}) + \frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{A}| \\ &= \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{V_0}| - \frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m}) \\ &\quad + \frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{A}| \\ &= \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}\ln|\mathbf{V_0}| - \frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{A}| \end{aligned}$$

Where we have used the definition of the multivariate Gaussian Distribution. Then, from (4.138), we can write:

$$\begin{aligned} \mathbf{A} &= -\nabla\nabla\ln p(D|\boldsymbol{\theta_{MAP}})p(\boldsymbol{\theta_{MAP}}) \\ &= -\nabla\nabla\ln p(D|\boldsymbol{\theta_{MAP}}) - \nabla\nabla\ln p(\boldsymbol{\theta_{MAP}}) \\ &= \mathbf{H} - \nabla\nabla\Big\{-\frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m})\Big\} \\ &= \mathbf{H} + \nabla\{\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m})\} \\ &= \mathbf{H} + \mathbf{V_0}^{-1} \end{aligned}$$

Where we have denoted $\mathbf{H} = -\nabla\nabla\ln p(D|\boldsymbol{\theta_{MAP}})$. Therefore, the equation above becomes:

$$\begin{aligned}\ln p(D) &= \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}}-\mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}}-\mathbf{m}) - \frac{1}{2}\ln\{|\mathbf{V_0}|\cdot|\mathbf{H}+\mathbf{V_0^{-1}}|\} \\ &= \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}}-\mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}}-\mathbf{m}) - \frac{1}{2}\ln\{|\mathbf{V_0}\mathbf{H}+\mathbf{I}|\} \\ &\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}}-\mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}}-\mathbf{m}) - \frac{1}{2}\ln|\mathbf{V_0}| - \frac{1}{2}\ln|\mathbf{H}| \\ &\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}}-\mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}}-\mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \text{const}\end{aligned}$$

Where we have used the property of determinant: $|\mathbf{A}|\cdot|\mathbf{B}| = |\mathbf{AB}|$, and the fact that the prior is board, i.e. $\mathbf{I}$ can be neglected with regard to $\mathbf{V_0H}$. What's more, since the prior is pre-given, we can view $\mathbf{V_0}$ as constant. And if the data is large, we can write:

$$\mathbf{H} = \sum_{n=1}^{N}\mathbf{H_n} = N\widehat{\mathbf{H}}$$

Where $\widehat{\mathbf{H}} = 1/N\sum_{n=1}^{N}\mathbf{H_n}$, and then

$$\begin{aligned}\ln p(D) &\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}}-\mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}}-\mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \text{const} \\ &\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}}-\mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}}-\mathbf{m}) - \frac{1}{2}\ln|N\widehat{\mathbf{H}}| + \text{const} \\ &\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}}-\mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}}-\mathbf{m}) - \frac{M}{2}\ln N - \frac{1}{2}\ln|\widehat{\mathbf{H}}| + \text{const} \\ &\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{M}{2}\ln N\end{aligned}$$

This is because when $N >> 1$, other terms can be neglected.

Problem 4.24 Solution(Waiting for updating)

Problem 4.25 Solution

We first need to obtain the expression for the first derivative of probit function $\Phi(\lambda a)$ with regard to a. According to (4.114), we can write down:

$$\begin{aligned}\frac{d}{da}\Phi(\lambda a) &= \frac{d\Phi(\lambda a)}{d(\lambda a)}\cdot\frac{d\lambda a}{da} \\ &= \frac{\lambda}{\sqrt{2\pi}}exp\{-\frac{1}{2}(\lambda a)^2\}\end{aligned}$$

Which further gives:

$$\frac{d}{da}\Phi(\lambda a)\Big|_{a=0} = \frac{\lambda}{\sqrt{2\pi}}$$

And for logistic sigmoid function, according to (4.88), we have

$$\frac{d\sigma}{da} = \sigma(1-\sigma) = 0.5\times0.5 = \frac{1}{4}$$

Where we have used $\sigma(0) = 0.5$. Let their derivatives at origin equals, we have:

$$\frac{\lambda}{\sqrt{2\pi}} = \frac{1}{4}$$

i.e., $\lambda = \sqrt{2\pi}/4$. And hence $\lambda^2 = \pi/8$ is obvious.

Problem 4.26 Solution

We will prove (4.152) in a more simple and intuitive way. But firstly, we need to prove a trivial yet useful statement: Suppose we have a random variable satisfied normal distribution denoted as $X \sim \mathcal{N}(X|\mu,\sigma^2)$, the probability of $X \leq x$ is $P(X \leq x) = \Phi(\frac{x-\mu}{\sigma})$, and here x is a given real number. We can see this by writing down the integral:

$$\begin{aligned} P(X \leq x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}(X-\mu)^2\right] dX \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2}\gamma^2)\sigma\, d\gamma \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}\gamma^2) d\gamma \\ &= \Phi(\frac{x-\mu}{\sigma}) \end{aligned}$$

Where we have changed the variable $X = \mu + \sigma\gamma$. Now consider two random variables $X \sim \mathcal{N}(0,\lambda^{-2})$ and $Y \sim \mathcal{N}(\mu,\sigma^2)$. We first calculate the conditional probability $P(X \leq Y\,|\,Y = a)$:

$$P(X \leq Y\,|\,Y = a) = P(X \leq a) = \Phi(\frac{a-0}{\lambda^{-1}}) = \Phi(\lambda a)$$

Together with Bayesian Formula, we can obtain:

$$\begin{aligned} P(X \leq Y) &= \int_{-\infty}^{+\infty} P(X \leq Y\,|\,Y = a) pdf(Y = a) dY \\ &= \int_{-\infty}^{+\infty} \Phi(\lambda a)\mathcal{N}(a|\mu,\sigma^2) da \end{aligned}$$

Where $pdf(\cdot)$ denotes the probability density function and we have also used $pdf(Y) = \mathcal{N}(\mu,\sigma^2)$. What's more, we know that $X - Y$ should also satisfy normal distribution, with:

$$E[X-Y] = E[X] - E[Y] = 0 - \mu = -\mu$$

And

$$var[X-Y] = var[X] + var[Y] = \lambda^{-2} + \sigma^2$$

Therefore, $X - Y \sim \mathcal{N}(-\mu, \lambda^{-2} + \sigma^2)$ and it follows that:

$$P(X - Y \leq 0) = \Phi(\frac{0-(-\mu)}{\sqrt{\lambda^{-2}+\sigma^2}}) = \Phi(\frac{\mu}{\sqrt{\lambda^{-2}+\sigma^2}})$$

Since $P(X \leq Y) = P(X - Y \leq 0)$, we obtain what have been required.

0.5 Neural Networks

Problem 5.1 Solution

Based on definition of $tanh(\cdot)$, we can obtain:

$$\begin{aligned} tanh(a) &= \frac{e^a - e^{-a}}{e^a + e^{-a}} \\ &= -1 + \frac{2e^a}{e^a + e^{-a}} \\ &= -1 + 2\frac{1}{1 + e^{-2a}} \\ &= 2\sigma(2a) - 1 \end{aligned}$$

If we have parameters $w_{ji}^{(1s)}, w_{j0}^{(1s)}$ and $w_{kj}^{(2s)}, w_{k0}^{(2s)}$ for a network whose hidden units use logistic sigmoid function as activation and $w_{ji}^{(1t)}, w_{j0}^{(1t)}$ and $w_{kj}^{(2t)}, w_{k0}^{(2t)}$ for another one using $tanh(\cdot)$, for the network using $tanh(\cdot)$ as activation, we can write down the following expression by using (5.4):

$$\begin{aligned} a_k^{(t)} &= \sum_{j=1}^{M} w_{kj}^{(2t)} tanh(a_j^{(t)}) + w_{k0}^{(2t)} \\ &= \sum_{j=1}^{M} w_{kj}^{(2t)} [2\sigma(2a_j^{(t)}) - 1] + w_{k0}^{(2t)} \\ &= \sum_{j=1}^{M} 2w_{kj}^{(2t)} \sigma(2a_j^{(t)}) + [-\sum_{j=1}^{M} w_{kj}^{(2t)} + w_{k0}^{(2t)}] \end{aligned}$$

What's more, we also have :

$$a_k^{(s)} = \sum_{j=1}^{M} w_{kj}^{(2s)} \sigma(a_j^{(s)}) + w_{k0}^{(2s)}$$

To make the two networks equivalent, i.e., $a_k^{(s)} = a_k^{(t)}$, we should make sure:

$$\begin{cases} a_j^{(s)} = 2a_j^{(t)} \\ w_{kj}^{(2s)} = 2w_{kj}^{(2t)} \\ w_{k0}^{(2s)} = -\sum_{j=1}^{M} w_{kj}^{(2t)} + w_{k0}^{(2t)} \end{cases}$$

Note that the first condition can be achieved by simply enforcing:

$$w_{ji}^{(1s)} = 2w_{ji}^{(1t)}, \quad \text{and} \quad w_{j0}^{(1s)} = 2w_{j0}^{(1t)}$$

Therefore, these two networks are equivalent under a linear transformation.

Problem 5.2 Solution

It is obvious. We write down the likelihood.

$$p(\mathbf{T}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{t_n}|\mathbf{y}(\mathbf{x_n},\mathbf{w}),\beta^{-1}\mathbf{I})$$

Taking the negative logarithm, we can obtain:

$$E(\mathbf{w},\beta) = -\ln p(\mathbf{T}|\mathbf{X},\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\left[(\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n})^T(\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n})\right]-\frac{NK}{2}\ln\beta+\text{const}$$

Here we have used const to denote the term independent of both $\mathbf{w}$ and β. Note that here we have used the definition of the multivariate Gaussian Distribution. What's more, we see that the covariance matrix $\beta^{-1}\mathbf{I}$ and the weight parameter $\mathbf{w}$ have decoupled, which is distinct from the next problem. We can first solve $\mathbf{w_{ML}}$ by minimizing the first term on the right of the equation above or equivalently (5.11), i.e., imaging β is fixed. Then according to the derivative of $E(\mathbf{w},\beta)$ with regard to β, we can obtain (5.17) and hence β_{ML}.

Problem 5.3 Solution

Following the process in the previous question, we first write down the negative logarithm of the likelihood function.

$$E(\mathbf{w},\mathbf{\Sigma}) = \frac{1}{2}\sum_{n=1}^{N}\left\{[\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n}]^T\mathbf{\Sigma}^{-1}[\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n}]\right\} + \frac{N}{2}\ln|\mathbf{\Sigma}| + \text{const} \quad (*)$$

Note here we have assumed $\mathbf{\Sigma}$ is unknown and const denotes the term independent of both $\mathbf{w}$ and $\mathbf{\Sigma}$. In the first situation, if $\mathbf{\Sigma}$ is fixed and known, the equation above will reduce to:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{[\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n}]^T\mathbf{\Sigma}^{-1}[\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n}]\right\} + \text{const}$$

We can simply solve $\mathbf{w_{ML}}$ by minimizing it. If $\mathbf{\Sigma}$ is unknown, since Σ is in the first term on the right of $(*)$, solving $\mathbf{w_{ML}}$ will involve $\mathbf{\Sigma}$. Note that in the previous problem, the main reason that they can decouple is due to the independent assumption, i.e., $\mathbf{\Sigma}$ reduces to $\beta^{-1}\mathbf{I}$, so that we can bring β to the front and view it as a fixed multiplying factor when solving $\mathbf{w_{ML}}$.

Problem 5.4 Solution

Based on (5.20), the current conditional distribution of targets, considering mislabel, given input $\mathbf{x}$ and weight $\mathbf{w}$ is:

$$p(t=1|\mathbf{x},\mathbf{w}) = (1-\epsilon)\cdot p(t_r=1|\mathbf{x},\mathbf{w}) + \epsilon\cdot p(t_r=0|\mathbf{x},\mathbf{w})$$

Note that here we use t to denote the observed target label, t_r to denote its real label, and that our network is aimed to predict the real label t_r not t, i.e., $p(t_r = 1|\mathbf{x}, \mathbf{w}) = y(\mathbf{x}, \mathbf{w})$, hence we see that:

$$p(t = 1|\mathbf{x}, \mathbf{w}) = (1-\epsilon)\cdot y(\mathbf{x}, \mathbf{w}) + \epsilon \cdot \left[1 - y(\mathbf{x}, \mathbf{w})\right] \qquad (*)$$

Also, it is the same for $p(t = 0|\mathbf{x}, \mathbf{w})$:

$$p(t = 0|\mathbf{x}, \mathbf{w}) = (1-\epsilon)\cdot \left[1 - y(\mathbf{x}, \mathbf{w})\right] + \epsilon \cdot y(\mathbf{x}, \mathbf{w}) \qquad (**)$$

Combing $(*)$ and $(**)$, we can obtain:

$$p(t|\mathbf{x}, \mathbf{w}) = (1-\epsilon)\cdot y^t (1-y)^{1-t} + \epsilon \cdot (1-y)^t y^{1-t}$$

Where y is short for $y(\mathbf{x}, \mathbf{w})$. Therefore, taking the negative logarithm, we can obtain the error function:

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \ln\left\{(1-\epsilon)\cdot y_n^{t_n}(1-y_n)^{1-t_n} + \epsilon\cdot(1-y_n)^{t_n} y_n^{1-t_n}\right\}$$

When $\epsilon = 0$, it is obvious that the equation above will reduce to (5.21).

Problem 5.5 Solution

It is obvious by using (5.22).

$$\begin{aligned}
E(\mathbf{w}) &= -\ln \prod_{n=1}^{N} p(\mathbf{t}|\mathbf{x_n}, \mathbf{w}) \\
&= -\ln \prod_{n=1}^{N} \prod_{k=1}^{K} y_k(\mathbf{x_n}, \mathbf{w})^{t_{nk}} \left[1 - y_k(\mathbf{x_n}, \mathbf{w})\right]^{1-t_{nk}} \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} \ln\left\{y_k(\mathbf{x_n}, \mathbf{w})^{t_{nk}} \left[1 - y_k(\mathbf{x_n}, \mathbf{w})\right]^{1-t_{nk}}\right\} \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} \ln\left[y_{nk}^{t_{nk}} (1 - y_{nk})^{1-t_{nk}}\right] \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} \left\{t_{nk} \ln y_{nk} + (1 - t_{nk}) \ln(1 - y_{nk})\right\}
\end{aligned}$$

Where we have denoted

$$y_{nk} = y_k(\mathbf{x_n}, \mathbf{w})$$

Problem 5.6 Solution

We know that $y_k = \sigma(a_k)$, where $\sigma(\cdot)$ represents the logistic sigmoid function. Moreover,

$$\frac{d\sigma}{da} = \sigma(1-\sigma)$$

$$\begin{aligned}
\frac{dE(\mathbf{w})}{da_k} &= -t_k \frac{1}{y_k}[y_k(1-y_k)] + (1-t_k)\frac{1}{1-y_k}[y_k(1-y_k)] \\
&= [y_k(1-y_k)]\,[\frac{1-t_k}{1-y_k} - \frac{t_k}{y_k}] \\
&= (1-t_k)y_k - t_k(1-y_k) \\
&= y_k - t_k
\end{aligned}$$

Just as required.

Problem 5.7 Solution

It is similar to the previous problem. First we denote $y_{kn} = y_k(\mathbf{x_n}, \mathbf{w})$. If we use softmax function as activation for the output unit, according to (4.106), we have:

$$\frac{dy_{kn}}{da_j} = y_{kn}(I_{kj} - y_{jn})$$

Therefore,

$$\begin{aligned}
\frac{dE(\mathbf{w})}{da_j} &= \frac{d}{da_k}\{-\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn}\ln y_k(\mathbf{x_n}, \mathbf{w})\} \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{d}{da_j}\{t_{kn}\ln y_{kn}\} \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn}\frac{1}{y_{kn}}[y_{kn}(I_{kj} - y_{jn})] \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K}(t_{kn}I_{kj} - t_{kn}y_{jn}) \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn}I_{kj} + \sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn}y_{jn} \\
&= -\sum_{n=1}^{N} t_{jn} + \sum_{n=1}^{N} y_{jn} \\
&= \sum_{n=1}^{N}(y_{jn} - t_{jn})
\end{aligned}$$

Where we have used the fact that only when $k = j$, $I_{kj} = 1 \neq 0$ and that $\sum_{k=1}^{K} t_{kn} = 1$.

Problem 5.8 Solution

It is obvious based on definition of 'tanh', i.e., (5.59).

$$\begin{aligned}
\frac{d}{da}tanh(a) &= \frac{(e^a + e^{-a})(e^a + e^{-a}) - (e^a - e^{-a})(e^a - e^{-a})}{(e^a + e^{-a})^2} \\
&= 1 - \frac{(e^a - e^{-a})^2}{(e^a + e^{-a})^2} \\
&= 1 - tanh(a)^2
\end{aligned}$$

Problem 5.9 Solution

We know that the logistic sigmoid function $\sigma(a) \in [0,1]$, therefore if we perform a linear transformation $h(a) = 2\sigma(a) - 1$, we can find a mapping function $h(a)$ from $(-\infty, +\infty)$ to $[-1,1]$. In this case, the conditional distribution of targets given inputs can be similarly written as:

$$p(t|\mathbf{x},\mathbf{w}) = [\frac{1+y(\mathbf{x},\mathbf{w})}{2}]^{(1+t)/2}[\frac{1-y(\mathbf{x},\mathbf{w})}{2}]^{(1-t)/2}$$

Where $[1+y(\mathbf{x},\mathbf{w})]/2$ represents the conditional probability $p(C_1|x)$. Since now $y(\mathbf{x},\mathbf{w}) \in [-1,1]$, we also need to perform the linear transformation to make it satisfy the constraint for probability.Then we can further obtain:

$$\begin{aligned} E(\mathbf{w}) &= -\sum_{n=1}^{N}\{\frac{1+t_n}{2}\ln\frac{1+y_n}{2} + \frac{1-t_n}{2}\ln\frac{1-y_n}{2}\} \\ &= -\frac{1}{2}\sum_{n=1}^{N}\{(1+t_n)\ln(1+y_n) + (1-t_n)\ln(1-y_n)\} + N\ln 2 \end{aligned}$$

Problem 5.10 Solution

It is obvious. Suppose $\mathbf{H}$ is positive definite, i.e., (5.37) holds. We set $\mathbf{v}$ equals to the eigenvector of $\mathbf{H}$, i.e., $\mathbf{v} = \mathbf{u_i}$ which gives:

$$\mathbf{v}^T\mathbf{H}\mathbf{v} = \mathbf{v}^T(\mathbf{H}\mathbf{v}) = \mathbf{u_i}^T\lambda_i\mathbf{u_i} = \lambda_i||\mathbf{u_i}||^2$$

Therefore, every λ_i should be positive. On the other hand, If all the eigenvalues λ_i are positive, from (5.38) and (5.39), we see that $\mathbf{H}$ is positive definite.

Problem 5.11 Solution

It is obvious. We follow (5.35) and then write the error function in the form of (5.36). To obtain the contour, we enforce $E(\mathbf{w})$ to equal to a constant C.

$$E(\mathbf{w}) = E(\mathbf{w}^*) + \frac{1}{2}\sum_i \lambda_i\alpha_i^2 = C$$

We rearrange the equation above, and then obtain:

$$\sum_i \lambda_i\alpha_i^2 = B$$

Where $B = 2C - 2E(\mathbf{w}^*)$ is a constant. Therefore, the contours of constant error are ellipses whose axes are aligned with the eigenvector $\mathbf{u_i}$ of the Hessian Matrix $\mathbf{H}$. The length for the jth axis is given by setting all $\alpha_i = 0, s.t. i \neq j$:

$$\alpha_j = \sqrt{\frac{B}{\lambda_j}}$$

In other words, the length is inversely proportional to the square root of the corresponding eigenvalue λ_j.

Problem 5.12 Solution

If **H** is positive definite, we know the second term on the right side of (5.32) will be positive for arbitrary **w**. Therefore, $E(\mathbf{w}^*)$ is a local minimum. On the other hand, if $\mathbf{w}^*$ is a local minimum, we have

$$E(\mathbf{w}^*) - E(\mathbf{w}) = -\frac{1}{2}(\mathbf{w}-\mathbf{w}^*)^T\mathbf{H}(\mathbf{w}-\mathbf{w}^*) < 0$$

In other words, for arbitrary **w**, $(\mathbf{w}-\mathbf{w}^*)^T\mathbf{H}(\mathbf{w}-\mathbf{w}^*) > 0$, according to the previous problem, we know that this means **H** is positive definite.

Problem 5.13 Solution

It is obvious. Suppose that there are W adaptive parameters in the network. Therefore, **b** has W independent parameters. Since **H** is symmetric, there should be $W(W+1)/2$ independent parameters in it. Therefore, there are $W + W(W+1)/2 = W(W+3)/2$ parameters in total.

Problem 5.14 Solution

It is obvious. Since we have

$$E_n(w_{ji}+\epsilon) = E_n(w_{ji}) + \epsilon E_n'(w_{ji}) + \frac{\epsilon^2}{2}E_n''(w_{ji}) + O(\epsilon^3)$$

And

$$E_n(w_{ji}-\epsilon) = E_n(w_{ji}) - \epsilon E_n'(w_{ji}) + \frac{\epsilon^2}{2}E_n''(w_{ji}) + O(\epsilon^3)$$

We combine those two equations, which gives,

$$E_n(w_{ji}+\epsilon) - E_n(w_{ji}-\epsilon) = 2\epsilon E_n'(w_{ji}) + O(\epsilon^3)$$

Rearrange the equation above, we obtain what has been required.

Problem 5.15 Solution

It is obvious. The back propagation formalism starts from performing summation near the input, as shown in (5.73). By symmetry, the forward propagation formalism should start near the output.

$$J_{ki} = \frac{\partial y_k}{\partial x_i} = \frac{\partial h(a_k)}{\partial x_i} = h'(a_k)\frac{\partial a_k}{\partial x_i} \qquad (*)$$

Where $h(\cdot)$ is the activation function at the output node a_k. Considering all the units j, which have links to unit k:

$$\frac{\partial a_k}{\partial x_i} = \sum_j \frac{\partial a_k}{\partial a_j}\frac{\partial a_j}{\partial x_i} = \sum_j w_{kj}h'(a_j)\frac{\partial a_j}{\partial x_i} \qquad (**)$$

Where we have used:

$$a_k = \sum_j w_{kj} z_j, \quad z_j = h(a_j)$$

It is similar for $\partial a_j / \partial x_i$. In this way we have obtained a recursive formula starting from the input node:

$$\frac{\partial a_l}{\partial x_i} = \begin{cases} w_{li}, \text{if there is a link from input unit } i \text{ to } l \\ 0, \text{if there isn't a link from input unit } i \text{ to } l \end{cases}$$

Using recursive formula $(**)$ and then $(*)$, we can obtain the Jacobian Matrix.

Problem 5.16 Solution

It is obvious. We begin by writing down the error function.

$$E = \frac{1}{2}\sum_{n=1}^{N} ||\mathbf{y_n} - \mathbf{t_n}||^2 = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}(y_{n,m} - t_{n,m})^2$$

Where the subscript m denotes the $m$the element of the vector. Then we can write down the Hessian Matrix as before.

$$\mathbf{H} = \nabla\nabla E = \sum_{n=1}^{N}\sum_{m=1}^{M} \nabla \mathbf{y_{n,m}} \nabla \mathbf{y_{n,m}} + \sum_{n=1}^{N}\sum_{m=1}^{M}(y_{n,m} - t_{n,m})\nabla\nabla \mathbf{y_{n,m}}$$

Similarly, we now know that the Hessian Matrix can be approximated as:

$$\mathbf{H} \simeq \sum_{n=1}^{N}\sum_{m=1}^{M} \mathbf{b}_{n,m}\mathbf{b}_{n,m}^T$$

Where we have defined:

$$\mathbf{b}_{n,m} = \nabla y_{n,m}$$

Problem 5.17 Solution

It is obvious.

$$\begin{aligned} \frac{\partial^2 E}{\partial w_r \partial w_s} &= \frac{\partial}{\partial w_r}\frac{1}{2}\int\int 2(y-t)\frac{\partial y}{\partial w_s} p(\mathbf{x},t) d\mathbf{x} dt \\ &= \int\int [(y-t)\frac{\partial y^2}{\partial w_r \partial w_s} + \frac{\partial y}{\partial w_s}\frac{\partial y}{\partial w_r}] p(\mathbf{x},t) d\mathbf{x} dt \end{aligned}$$

Since we know that

$$\begin{aligned} \int\int (y-t)\frac{\partial y^2}{\partial w_r \partial w_s} p(\mathbf{x},t) d\mathbf{x} dt &= \int\int (y-t)\frac{\partial y^2}{\partial w_r \partial w_s} p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt \\ &= \int \frac{\partial y^2}{\partial w_r \partial w_s}\{\int (y-t) p(t|\mathbf{x}) dt\} p(\mathbf{x}) d\mathbf{x} \\ &= 0 \end{aligned}$$

Note that in the last step, we have used $y = \int tp(t|\mathbf{x})dt$. Then we substitute it into the second derivative, which gives,

$$\begin{aligned}\frac{\partial^2 E}{\partial w_r \partial w_s} &= \int\int \frac{\partial y}{\partial w_s}\frac{\partial y}{\partial w_r} p(\mathbf{x},t)d\mathbf{x}dt \\ &= \int \frac{\partial y}{\partial w_s}\frac{\partial y}{\partial w_r} p(\mathbf{x})d\mathbf{x}\end{aligned}$$

Problem 5.18 Solution

By analogy with section 5.3.2, we denote w_{ki}^{skip} as those parameters corresponding to skip-layer connections, i.e., it connects the input unit i with the output unit k. Note that the discussion in section 5.3.2 is still correct and now we only need to obtain the derivative of the error function with respect to the additional parameters w_{ki}^{skip}.

$$\frac{\partial E_n}{\partial w_{ki}^{\text{skip}}} = \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial w_{ki}^{\text{skip}}} = \delta_k x_i$$

Where we have used $a_k = y_k$ due to linear activation at the output unit and:

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} z_j + \sum_i w_{ki}^{\text{skip}} x_i$$

Where the first term on the right side corresponds to those information conveying from the hidden unit to the output and the second term corresponds to the information conveying directly from the input to output.

Problem 5.19 Solution

The error function is given by (5.21). Therefore, we can obtain:

$$\begin{aligned}\nabla E(\mathbf{w}) &= \sum_{n=1}^{N} \frac{\partial E}{\partial a_n}\nabla a_n \\ &= -\sum_{n=1}^{N} \frac{\partial}{\partial a_n}\left[t_n \ln y_n + (1-t_n)\ln(1-y_n)\right]\nabla a_n \\ &= -\sum_{n=1}^{N} \{\frac{\partial(t_n \ln y_n)}{\partial y_n}\frac{\partial y_n}{\partial a_n} + \frac{\partial(1-t_n)\ln(1-y_n)}{\partial y_n}\frac{\partial y_n}{\partial a_n}\}\nabla a_n \\ &= -\sum_{n=1}^{N} [\frac{t_n}{y_n}\cdot y_n(1-y_n) + (1-t_n)\frac{-1}{1-y_n}\cdot y_n(1-y_n)]\nabla a_n \\ &= -\sum_{n=1}^{N} \left[t_n(1-y_n) - (1-t_n)y_n\right]\nabla a_n \\ &= \sum_{n=1}^{N} (y_n - t_n)\nabla a_n\end{aligned}$$

Where we have used the conclusion of problem 5.6. Now we calculate the second derivative.

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \{y_n(1-y_n)\nabla a_n \nabla a_n + (y_n - t_n)\nabla\nabla a_n\}$$

Similarly, we can drop the last term, which gives exactly what has been asked.

Problem 5.20 Solution(waiting for update)

We begin by writing down the error function.

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

Here we assume that the output of the network has K units in total and there are W weights parameters in the network. WE first calculate the first derivative:

$$\begin{aligned}
\nabla E &= \sum_{n=1}^{N} \frac{dE}{d\mathbf{a}_n} \cdot \nabla \mathbf{a}_n \\
&= -\sum_{n=1}^{N} \big[\frac{d}{d\mathbf{a}_n}(\sum_{k=1}^{K} t_{nk} \ln y_{nk})\big] \cdot \nabla \mathbf{a}_n \\
&= \sum_{n=1}^{N} \mathbf{c}_n \cdot \nabla \mathbf{a}_n
\end{aligned}$$

Note that here $\mathbf{c}_n = -dE/d\mathbf{a}_n$ is a vector with size $K \times 1$, $\nabla \mathbf{a}_n$ is a matrix with size $K \times W$. Moreover, the operator $\cdot$ means inner product, which gives ∇E as a vector with size $1 \times W$. According to (4.106), we can obtain the jth element of $\mathbf{c_n}$:

$$\begin{aligned}
c_{n,j} &= -\frac{\partial}{\partial a_j}(\sum_{k=1}^{K} t_{nk} \ln y_{nk}) \\
&= -\sum_{k=1}^{K} \frac{\partial}{\partial a_j}(t_{nk} \ln y_{nk}) \\
&= -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk}(I_{kj} - y_{nj}) \\
&= -\sum_{k=1}^{K} t_{nk} I_{kj} + \sum_{k=1}^{K} t_{nk} y_{nj} \\
&= -t_{nj} + y_{nj}(\sum_{k=1}^{K} t_{nk}) \\
&= y_{nj} - t_{nj}
\end{aligned}$$

Now we calculate the second derivative:

$$\nabla\nabla E \quad = \quad \sum_{n=1}^{N}(\frac{d\mathbf{c}_n}{d\mathbf{a}_n}\nabla\mathbf{a}_n)\cdot\nabla\mathbf{a_n} + \mathbf{c}_n\nabla\nabla\mathbf{a}_n$$

Here $d\mathbf{c}_n/d\mathbf{a}_n$ is a matrix with size $K\times K$. Therefore, the second term can be neglected as before, which gives:

$$\mathbf{H} = \sum_{n=1}^{N}(\frac{d\mathbf{c}_n}{d\mathbf{a}_n}\nabla\mathbf{a}_n)\cdot\nabla\mathbf{a_n}$$

Problem 5.21 Solution

We first write down the expression of Hessian Matrix in the case of K outputs.

$$\mathbf{H}_{N,K} = \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{b}_{n,k}\mathbf{b}_{n,k}^T$$

Where $\mathbf{b}_{n,k} = \nabla_{\mathbf{w}}\mathbf{a}_{n,k}$. Therefore, we have:

$$\mathbf{H}_{N+1,K} = \mathbf{H}_{N,K} + \sum_{k=1}^{K}\mathbf{b}_{N+1,k}\mathbf{b}_{N+1,k}^T = \mathbf{H}_{N,K} + \mathbf{B}_{N+1}\mathbf{B}_{N+1}^T$$

Where $\mathbf{B}_{N+1} = [\mathbf{b}_{N+1,1}, \mathbf{b}_{N+1,2}, ..., \mathbf{b}_{N+1,K}]$ is a matrix with size $W\times K$, and here W is the total number of the parameters in the network. By analogy with (5.88)-(5.89), we can obtain:

$$\mathbf{H}_{N+1,K}^{-1} = \mathbf{H}_{N,K}^{-1} - \frac{\mathbf{H}_{N,K}^{-1}\mathbf{B}_{N+1}\mathbf{B}_{N+1}^T\mathbf{H}_{N,K}^{-1}}{1+\mathbf{B}_{N+1}^T\mathbf{H}_{N,K}^{-1}\mathbf{B}_{N+1}} \qquad (*)$$

Furthermore, similarly, we have:

$$\mathbf{H}_{N+1,K+1} = \mathbf{H}_{N+1,K} + \sum_{n=1}^{N+1}\mathbf{b}_{n,K+1}\mathbf{b}_{n,K+1}^T = \mathbf{H}_{N+1,K} + \mathbf{B}_{K+1}\mathbf{B}_{K+1}^T$$

Where $\mathbf{B}_{K+1} = [\mathbf{b}_{1,K+1}, \mathbf{b}_{2,K+1}, ..., \mathbf{b}_{N+1,K+1}]$ is a matrix with size $W\times(N+1)$. Also, we can obtain:

$$\mathbf{H}_{N+1,K+1}^{-1} = \mathbf{H}_{N+1,K}^{-1} - \frac{\mathbf{H}_{N+1,K}^{-1}\mathbf{B}_{K+1}\mathbf{B}_{K+1}^T\mathbf{H}_{N+1,K}^{-1}}{1+\mathbf{B}_{K+1}^T\mathbf{H}_{N+1,K}^{-1}\mathbf{B}_{K+1}}$$

Where $\mathbf{H}_{N+1,K}^{-1}$ is defined by $(*)$. If we substitute $(*)$ into the expression above, we can obtain the relationship between $\mathbf{H}_{N+1,K+1}^{-1}$ and $\mathbf{H}_{N,K}^{-1}$.

Problem 5.22 Solution

We begin by handling the first case.

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{kj}^{(2)} \partial w_{k'j'}^{(2)}} &= \frac{\partial}{\partial w_{kj}^{(2)}}(\frac{\partial E_n}{\partial w_{k'j'}^{(2)}}) \\
&= \frac{\partial}{\partial w_{kj}^{(2)}}(\frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial w_{k'j'}^{(2)}}) \\
&= \frac{\partial}{\partial w_{kj}^{(2)}}(\frac{\partial E_n}{\partial a_{k'}} \frac{\partial \sum_{j'} w_{k'j'} z_{j'}}{\partial w_{k'j'}^{(2)}}) \\
&= \frac{\partial}{\partial w_{kj}^{(2)}}(\frac{\partial E_n}{\partial a_{k'}} z_{j'}) \\
&= \frac{\partial}{\partial w_{kj}^{(2)}}(\frac{\partial E_n}{\partial a_{k'}}) z_{j'} + \frac{\partial E_n}{\partial a_{k'}} \frac{\partial z_{j'}}{\partial w_{kj}^{(2)}} \\
&= \frac{\partial}{\partial a_k}(\frac{\partial E_n}{\partial a_{k'}}) \frac{\partial a_k}{\partial w_{kj}^{(2)}} z_{j'} + 0 \\
&= \frac{\partial}{\partial a_k}(\frac{\partial E_n}{\partial a_{k'}}) z_j z_{j'} \\
&= z_j z_{j'} M_{kk'}
\end{aligned}$$

Then we focus on the second case, and if here $j \neq j'$

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{j'i'}^{(1)}} &= \frac{\partial}{\partial w_{ji}^{(1)}}(\frac{\partial E_n}{\partial w_{j'i'}^{(1)}}) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}}(\sum_{k'} \frac{\partial E_n}{\partial a_{k'}} \frac{\partial a_{k'}}{\partial w_{j'i'}^{(1)}}) \\
&= \sum_{k'} \frac{\partial}{\partial w_{ji}^{(1)}}(\frac{\partial E_n}{\partial a_{k'}} w_{k'j'}^{(2)} h'(a_{j'}) x_{i'}) \\
&= \sum_{k'} h'(a_{j'}) x_{i'} \frac{\partial}{\partial w_{ji}^{(1)}}(\frac{\partial E_n}{\partial a_{k'}} w_{k'j'}^{(2)}) \\
&= \sum_{k'} h'(a_{j'}) x_{i'} \sum_k \frac{\partial}{\partial a_k}(\frac{\partial E_n}{\partial a_{k'}} w_{k'j'}^{(2)}) \frac{\partial a_k}{\partial w_{ji}^{(1)}} \\
&= \sum_{k'} h'(a_{j'}) x_{i'} \sum_k \frac{\partial}{\partial a_k}(\frac{\partial E_n}{\partial a_{k'}} w_{k'j'}^{(2)}) \cdot (w_{kj}^{(2)} h'(a_j) x_i) \\
&= \sum_{k'} h'(a_{j'}) x_{i'} \sum_k M_{kk'} w_{k'j'}^{(2)} \cdot w_{kj}^{(2)} h'(a_j) x_i \\
&= x_{i'} x_i h'(a_{j'}) h'(a_j) \sum_{k'} \sum_k w_{k'j'}^{(2)} \cdot w_{kj}^{(2)} M_{kk'}
\end{aligned}$$

When $j = j'$, similarly we have:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{ji'}^{(1)}} &= \sum_{k'} \frac{\partial}{\partial w_{ji}^{(1)}} (\frac{\partial E_n}{\partial a_{k'}} w_{k'j}^{(2)} h'(a_j) x_{i'}) \\
&= x_{i'} \sum_{k'} \frac{\partial}{\partial w_{ji}^{(1)}} (\frac{\partial E_n}{\partial a_{k'}} w_{k'j}^{(2)}) h'(a_j) + x_{i'} \sum_{k'} (\frac{\partial E_n}{\partial a_{k'}} w_{k'j}^{(2)}) \frac{\partial h'(a_j)}{\partial w_{ji}^{(1)}} \\
&= x_{i'} x_i h'(a_j) h'(a_j) \sum_{k'} \sum_{k} w_{k'j}^{(2)} \cdot w_{kj}^{(2)} M_{kk'} + x_{i'} \sum_{k'} (\frac{\partial E_n}{\partial a_{k'}} w_{k'j}^{(2)}) \frac{\partial h'(a_j)}{\partial w_{ji}^{(1)}} \\
&= x_{i'} x_i h'(a_j) h'(a_j) \sum_{k'} \sum_{k} w_{k'j}^{(2)} \cdot w_{kj}^{(2)} M_{kk'} + x_{i'} \sum_{k'} (\frac{\partial E_n}{\partial a_{k'}} w_{k'j}^{(2)}) h''(a_j) x_i \\
&= x_{i'} x_i h'(a_j) h'(a_j) \sum_{k'} \sum_{k} w_{k'j}^{(2)} \cdot w_{kj}^{(2)} M_{kk'} + h''(a_j) x_i x_{i'} \sum_{k'} \delta_{k'} w_{k'j}^{(2)}
\end{aligned}$$

It seems that what we have obtained is slightly different from (5.94) when $j = j'$. However this is not the case, since the summation over k' in the second term of our formulation and the summation over k in the first term of (5.94) is actually the same (i.e., they both represent the summation over all the output units). Combining the situation when $j = j'$ and $j \neq j'$, we can obtain (5.94) just as required. Finally, we deal with the third case. Similarly we first focus on $j \neq j'$:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{kj'}^{(2)}} &= \frac{\partial}{\partial w_{ji}^{(1)}} (\frac{\partial E_n}{\partial w_{kj'}^{(2)}}) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} (\frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial w_{kj'}^{(2)}}) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} (\frac{\partial E_n}{\partial a_k} \frac{\partial \sum_{j'} w_{kj'} z_{j'}}{\partial w_{kj'}^{(2)}}) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} (\frac{\partial E_n}{\partial a_k} z_{j'}) \\
&= z_{j'} \sum_{k'} \frac{\partial}{\partial a_{k'}} (\frac{\partial E_n}{\partial a_k}) \frac{\partial a_{k'}}{\partial w_{ji}^{(1)}} \\
&= z_{j'} \sum_{k'} M_{kk'} w_{k'j}^{(2)} h'(a_j) x_i \\
&= x_i h'(a_j) z_{j'} \sum_{k'} M_{kk'} w_{k'j}^{(2)}
\end{aligned}$$

Note that in (5.95), there are two typos: (i)$H_{kk'}$ should be $M_{kk'}$. (ii) j should exchange position with j' in the right side of (5.95). When $j = j'$, we have:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{kj}^{(2)}} &= \frac{\partial}{\partial w_{ji}^{(1)}} (\frac{\partial E_n}{\partial w_{kj}^{(2)}}) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} (\frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial w_{kj}^{(2)}}) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} (\frac{\partial E_n}{\partial a_k} \frac{\partial \sum_j w_{kj} z_j}{\partial w_{kj}^{(2)}}) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} (\frac{\partial E_n}{\partial a_k} z_j) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} (\frac{\partial E_n}{\partial a_k}) z_j + \frac{\partial E_n}{\partial a_k} \frac{\partial z_j}{w_{ji}^{(1)}} \\
&= x_i h'(a_j) z_j \sum_{k'} M_{kk'} w_{k'j}^{(2)} + \frac{\partial E_n}{\partial a_k} \frac{\partial z_j}{w_{ji}^{(1)}} \\
&= x_i h'(a_j) z_j \sum_{k'} M_{kk'} w_{k'j}^{(2)} + \delta_k h'(a_j) x_i
\end{aligned}$$

Combing these two situations, we obtain (5.95) just as required.

Problem 5.23 Solution

It is similar to the previous problem.

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{k'i'} \partial w_{kj}} &= \frac{\partial}{\partial w_{k'i'}} (\frac{\partial E_n}{\partial w_{kj}}) \\
&= \frac{\partial}{\partial w_{k'i'}} (\frac{\partial E_n}{\partial a_k} z_j) \\
&= z_j \frac{\partial w_{k'i'}}{\partial a_{k'}} \frac{\partial}{\partial a_{k'}} (\frac{\partial E_n}{\partial a_k}) \\
&= z_j x_{i'} M_{kk'}
\end{aligned}$$

And

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{k'i'}\partial w_{ji}} &= \frac{\partial}{\partial w_{k'i'}}(\sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial w_{ji}}) \\
&= \frac{\partial}{\partial w_{k'i'}}(\sum_k \frac{\partial E_n}{\partial a_k} w_{kj} h'(a_j) x_i) \\
&= \sum_k h'(a_j) x_i w_{kj} \frac{\partial}{\partial w_{k'i'}}(\frac{\partial E_n}{\partial a_k}) \\
&= \sum_k h'(a_j) x_i w_{kj} \frac{\partial}{\partial a_{k'}}(\frac{\partial E_n}{\partial a_k})\frac{a_{k'}}{w_{k'i'}} \\
&= \sum_k h'(a_j) x_i w_{kj} M_{kk'} x_{i'} \\
&= x_i x_{i'} h'(a_j) \sum_k w_{kj} M_{kk'}
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{k'i'} w_{ki}} &= \frac{\partial}{\partial w_{k'i'}}(\frac{\partial E_n}{\partial w_{ki}}) \\
&= \frac{\partial}{\partial w_{k'i'}}(\frac{\partial E_n}{\partial a_k} x_i) \\
&= x_i \frac{\partial}{\partial a_{k'}}(\frac{\partial E_n}{\partial a_k})\frac{\partial a_{k'}}{w_{k'i'}} \\
&= x_i x_{i'} M_{kk'}
\end{aligned}$$

Problem 5.24 Solution

It is obvious. According to (5.113), we have:

$$\begin{aligned}
\widetilde{a}_j &= \sum_i \widetilde{w}_{ji}\widetilde{x}_i + \widetilde{w}_{j0} \\
&= \sum_i \frac{1}{a} w_{ji} \cdot (a x_i + b) + w_{j0} - \frac{b}{a}\sum_i w_{ji} \\
&= \sum_i w_{ji} x_i + w_{j0} = a_j
\end{aligned}$$

Where we have used (5.115), (5.116) and (5.117). Currently, we have proved that under the transformation the hidden unit a_j is unchanged. If the activation function at the hidden unit is also unchanged, we have $\widetilde{z}_j = z_j$. Now we deal with the output unit $\widetilde{y}_k$:

$$\begin{aligned}
\widetilde{y}_k &= \sum_j \widetilde{w}_{kj}\widetilde{z}_j + \widetilde{w}_{k0} \\
&= \sum_j c w_{kj} \cdot z_j + c w_{k0} + d \\
&= c \sum_j \left[w_{kj} \cdot z_j + w_{k0} \right] + d \\
&= c y_k + d
\end{aligned}$$

Where we have used (5.114), (5.119) and (5.120). To be more specific, here we have proved that the linear transformation between $\widetilde{y}_k$ and y_k can be achieved by making transformation (5.119) and (5.120).

Problem 5.25 Solution

Since we know the gradient of the error function with respect to $\mathbf{w}$ is:

$$\nabla E = \mathbf{H}(\mathbf{w}-\mathbf{w}^*)$$

Together with (5.196), we can obtain:

$$\begin{aligned}\mathbf{w}^{(\tau)} &= \mathbf{w}^{(\tau-1)}-\rho\nabla E \\ &= \mathbf{w}^{(\tau-1)}-\rho\mathbf{H}(\mathbf{w}^{(\tau-1)}-\mathbf{w}^*)\end{aligned}$$

Multiplying both sides by $\mathbf{u}_j^T$, using $w_j = \mathbf{w}^T\mathbf{u}_j$, we can obtain:

$$\begin{aligned}w_j^{(\tau)} &= \mathbf{u}_j^T\big[\mathbf{w}^{(\tau-1)}-\rho\mathbf{H}(\mathbf{w}^{(\tau-1)}-\mathbf{w}^*)\big] \\ &= w_j^{(\tau-1)}-\rho\mathbf{u}_j^T\mathbf{H}(\mathbf{w}^{(\tau-1)}-\mathbf{w}^*) \\ &= w_j^{(\tau-1)}-\rho\eta_j\mathbf{u}_j^T(\mathbf{w}^{(\tau-1)}-\mathbf{w}^*) \\ &= w_j^{(\tau-1)}-\rho\eta_j(w_j^{(\tau-1)}-w_j^*) \\ &= (1-\rho\eta_j)w_j^{(\tau-1)}+\rho\eta_jw_j^*\end{aligned}$$

Where we have used (5.198). Then we use mathematical deduction to prove (5.197), beginning by calculating $w_j^{(1)}$:

$$\begin{aligned}w_j^{(1)} &= (1-\rho\eta_j)w_j^{(0)}+\rho\eta_jw_j^* \\ &= \rho\eta_jw_j^* \\ &= \big[1-(1-\rho\eta_j)\big]w_j^*\end{aligned}$$

Suppose (5.197) holds for τ, we now prove that it also holds for $\tau+1$.

$$\begin{aligned}w_j^{(\tau+1)} &= (1-\rho\eta_jw_j^{(\tau)}+\rho\eta_jw_j^* \\ &= (1-\rho\eta_j)\big[1-(1-\rho\eta_j)^\tau\big]w_j^*+\rho\eta_jw_j^* \\ &= \big\{(1-\rho\eta_j)\big[1-(1-\rho\eta_j)^\tau\big]+\rho\eta_j\big\}w_j^* \\ &= \big[1-(1-\rho\eta_j)^{\tau+1}\big]w_j^*\end{aligned}$$

Hence (5.197) holds for $\tau = 1,2,....$ Provided $|1-\rho\eta_j|<1$, we have $(1-\rho\eta_j)^\tau \to 0$ as $\tau\to\infty$ ans thus $\mathbf{w}^{(\tau)} = \mathbf{w}^*$. If τ is finite and $\eta_j >> (\rho\tau)^{-1}$, the above argument still holds since τ is still relatively large. Conversely, when $\eta_j << (\rho\tau)^{-1}$, we expand the expression above:

$$|w_j^{(\tau)}| = |\big[1-(1-\rho\eta_j)^\tau\big]w_j^*| \approx |\tau\rho\eta_jw_j^*| << |w_j^*|$$

We can see that $(\rho\tau)^{-1}$ works as the regularization parameter α in section 3.5.3.

Problem 5.26 Solution

Based on definition or by analogy with (5.128), we have:

$$\begin{aligned}
\Omega_n &= \frac{1}{2}\sum_k (\frac{\partial y_{nk}}{\partial \xi}|_{\xi=0})^2 \\
&= \frac{1}{2}\sum_k (\sum_i \frac{\partial y_{nk}}{\partial x_i}\frac{\partial x_i}{\partial \xi}|_{\xi=0})^2 \\
&= \frac{1}{2}\sum_k (\sum_i \tau_i \frac{\partial}{\partial x_i} y_{nk})^2
\end{aligned}$$

Where we have denoted

$$\tau_i = \frac{\partial x_i}{\partial \xi}|_{\xi=0}$$

And this is exactly the form given in (5.201) and (5.202) if the nth observation y_{nk} is denoted as y_k in short. Firstly, we define α_j and β_j as (5.205) shows, where z_j and a_j are given by (5.203). Then we will prove (5.204) holds:

$$\begin{aligned}
\alpha_j &= \sum_i \tau_i \frac{\partial z_j}{\partial x_i} = \sum_i \tau_i \frac{\partial h(a_j)}{\partial x_i} \\
&= \sum_i \tau_i \frac{\partial h(a_j)}{\partial a_j}\frac{\partial a_j}{\partial x_i} \\
&= h'(a_j)\sum_i \tau_i \frac{\partial}{\partial x_i} a_j = h'(a_j)\beta_j
\end{aligned}$$

Moreover,

$$\begin{aligned}
\beta_j &= \sum_i \tau_i \frac{\partial a_j}{\partial x_i} = \sum_i \tau_i \frac{\partial \sum_{i'} w_{ji'} z_{i'}}{\partial x_i} \\
&= \sum_i \tau_i \sum_{i'} \frac{\partial w_{ji'} z_{i'}}{\partial x_i} = \sum_i \tau_i \sum_{i'} w_{ji'} \frac{\partial z_{i'}}{\partial x_i} \\
&= \sum_{i'} w_{ji'} \sum_i \tau_i \frac{\partial z_{i'}}{\partial x_i} = \sum_{i'} w_{ji'} \alpha_{i'}
\end{aligned}$$

So far we have proved that (5.204) holds and now we aim to find a forward propagation formula to calculate Ω_n. We firstly begin by evaluating $\{\beta_j\}$ at the input units, and then use the first equation in (5.204) to obtain $\{\alpha_j\}$ at the input units, and then the second equation to evaluate $\{\beta_j\}$ at the first hidden layer, and again the first equation to evaluate $\{\alpha_j\}$ at the first hidden layer. We repeatedly evaluate $\{\beta_j\}$ and $\{\alpha_j\}$ in this way until reaching the output layer. Then we deal with (5.206):

$$\begin{aligned}
\frac{\partial \Omega_n}{\partial w_{rs}} &= \frac{\partial}{\partial w_{rs}}\{\frac{1}{2}\sum_k (\mathcal{G}y_k)^2\} = \frac{1}{2}\sum_k \frac{\partial(\mathcal{G}y_k)^2}{\partial w_{rs}} \\
&= \frac{1}{2}\sum_k \frac{\partial(\mathcal{G}y_k)^2}{\partial(\mathcal{G}y_k)}\frac{\partial(\mathcal{G}y_k)}{\partial w_{rs}} = \sum_k \mathcal{G}y_k \frac{\partial \mathcal{G}y_k}{\partial w_{rs}} \\
&= \sum_k \mathcal{G}y_k \mathcal{G}[\frac{\partial y_k}{\partial w_{rs}}] = \sum_k \alpha_k \mathcal{G}[\frac{\partial y_k}{\partial a_r}\frac{\partial a_r}{\partial w_{rs}}] \\
&= \sum_k \alpha_k \mathcal{G}[\delta_{kr} z_s] = \sum_k \alpha_k \{\mathcal{G}[\delta_{kr}]z_s + \mathcal{G}[z_s]\delta_{kr}\} \\
&= \sum_k \alpha_k \{\phi_{kr} z_s + \alpha_s \delta_{kr}\}
\end{aligned}$$

Provided with the idea in section 5.3, the backward propagation formula is easy to derive. We can simply replace E_n with y_k to obtain a backward equation, so we omit it here.

Problem 5.27 Solution

Following the procedure in section 5.5.5, we can obtain:

$$\Omega = \frac{1}{2}\int (\boldsymbol{\tau}^T \nabla y(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x}$$

Since we have $\boldsymbol{\tau} = \partial \mathbf{s}(\mathbf{x}, \boldsymbol{\xi}) / \partial \boldsymbol{\xi}$ and $\mathbf{s} = \mathbf{x} + \boldsymbol{\xi}$, so we have $\boldsymbol{\tau} = \mathbf{I}$. Therefore, substituting $\boldsymbol{\tau}$ into the equation above, we can obtain:

$$\Omega = \frac{1}{2}\int (\nabla y(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x}$$

Just as required.

Problem 5.28 Solution

The modifications only affect derivatives with respect to the weights in the convolutional layer. The units within a feature map (indexed m) have different inputs, but all share a common weight vector, $\mathbf{w}^{(m)}$. Therefore, we can write:

$$\frac{\partial E_n}{\partial w_i^{(m)}} = \sum_j \frac{\partial E_n}{\partial a_j^{(m)}} \frac{\partial a_j^{(m)}}{\partial w_i^{(m)}} = \sum_j \delta_j^{(m)} z_{ji}^{(m)}$$

Here $a_j^{(m)}$ denotes the activation of the jth unit in th mth feature map, whereas $w_i^{(m)}$ denotes the ith element of the corresponding feature vector and finally $z_{ij}^{(m)}$ denotes the ith input for the jth unit in the mth feature map. Note that $\delta_j^{(m)}$ can be computed recursively from the units in the following layer.

Problem 5.29 Solution

It is obvious. Firstly, we know that:

$$\frac{\partial}{\partial w_i}\{\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\} = -\pi_j \frac{w_i-\mu_j}{\sigma_j^2}\mathcal{N}(w_i|\mu_j,\sigma_j^2)$$

We now derive the error function with respect to w_i:

$$\begin{aligned}
\frac{\partial \widetilde{E}}{\partial w_i} &= \frac{\partial E}{\partial w_i} + \frac{\partial \lambda \Omega(\mathbf{w})}{\partial w_i} \\
&= \frac{\partial E}{\partial w_i} - \lambda \frac{\partial}{\partial w_i}\left\{\sum_i \ln\left(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\right)\right\} \\
&= \frac{\partial E}{\partial w_i} - \lambda \frac{\partial}{\partial w_i}\left\{\ln\left(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\right)\right\} \\
&= \frac{\partial E}{\partial w_i} - \lambda \frac{1}{\sum_{j=1}^{M}\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)} \frac{\partial}{\partial w_i}\left\{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\right\} \\
&= \frac{\partial E}{\partial w_i} + \lambda \frac{1}{\sum_{j=1}^{M}\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)} \left\{\sum_{j=1}^{M} \pi_j \frac{w_i-\mu_j}{\sigma_j^2}\mathcal{N}(w_i|\mu_j,\sigma_j^2)\right\} \\
&= \frac{\partial E}{\partial w_i} + \lambda \frac{\sum_{j=1}^{M}\pi_j \frac{w_i-\mu_j}{\sigma_j^2}\mathcal{N}(w_i|\mu_j,\sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k,\sigma_k^2)} \\
&= \frac{\partial E}{\partial w_i} + \lambda \sum_{j=1}^{M} \frac{\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k,\sigma_k^2)} \frac{w_i-\mu_j}{\sigma_j^2} \\
&= \frac{\partial E}{\partial w_i} + \lambda \sum_{j=1}^{M} \gamma_j(w_i) \frac{w_i-\mu_j}{\sigma_j^2}
\end{aligned}$$

Where we have used (5.138) and defined (5.140).

Problem 5.30 Solution

Is is similar to the previous problem. Since we know that:

$$\frac{\partial}{\partial \mu_j}\{\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\} = \pi_j \frac{w_i-\mu_j}{\sigma_j^2}\mathcal{N}(w_i|\mu_j,\sigma_j^2)$$

We can derive:

$$\begin{aligned}
\frac{\partial \widetilde{E}}{\partial \mu_j} &= \frac{\partial \lambda \Omega(\mathbf{w})}{\partial \mu_j} \\
&= -\lambda \frac{\partial}{\partial \mu_j} \left\{ \sum_i \ln \left(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2) \right) \right\} \\
&= -\lambda \sum_i \frac{\partial}{\partial \mu_j} \left\{ \ln \left(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2) \right) \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)} \frac{\partial}{\partial \mu_j} \left\{ \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2) \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)} \pi_j \frac{w_i - \mu_j}{\sigma_j^2} \mathcal{N}(w_i|\mu_j,\sigma_j^2) \\
&= \lambda \sum_i \frac{\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(w_i|\mu_k,\sigma_k^2)} \frac{\mu_j - w_i}{\sigma_j^2} = \lambda \sum_i \gamma_j(w_i) \frac{\mu_j - w_i}{\sigma_j^2}
\end{aligned}$$

Note that there is a typo in (5.142). The numerator should be $\mu_j - w_i$ instead of $\mu_i - w_j$. This can be easily seen through the fact that the mean and variance of the Gaussian Distribution should have the same subindex and since σ_j is in the denominator, μ_j should occur in the numerator instead of μ_i.

Problem 5.31 Solution

It is similar to the previous problem. Since we know that:

$$\frac{\partial}{\partial \sigma_j} \{\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\} = \left(-\frac{1}{\sigma_j} + \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right) \pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)$$

We can derive:

$$\begin{aligned}
\frac{\partial \widetilde{E}}{\partial \sigma_j} &= \frac{\partial \lambda \Omega(\mathbf{w})}{\partial \sigma_j} \\
&= -\lambda \frac{\partial}{\partial \sigma_j} \left\{ \sum_i \ln \left(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \right) \right\} \\
&= -\lambda \sum_i \frac{\partial}{\partial \sigma_j} \left\{ \ln \left(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \right) \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \frac{\partial}{\partial \sigma_j} \left\{ \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \frac{\partial}{\partial \sigma_j} \left\{ \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \right\} \\
&= \lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \left(\frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right) \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \\
&= \lambda \sum_i \frac{\pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)}{\sum_{k=1}^{M} \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2)} \left(\frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right) \\
&= \lambda \sum_i \gamma_j(w_i) \left(\frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right)
\end{aligned}$$

Just as required.

Problem 5.32 Solution

It is trivial. We begin by verifying (5.208) when $j \neq k$.

$$\begin{aligned}
\frac{\partial \pi_k}{\partial \eta_j} &= \frac{\partial}{\partial \eta_j} \left\{ \frac{exp(\eta_k)}{\sum_k exp(\eta_k)} \right\} \\
&= \frac{-exp(\eta_k) exp(\eta_j)}{\left[\sum_k exp(\eta_k)\right]^2} \\
&= -\pi_j \pi_k
\end{aligned}$$

And if now we have $j = k$:

$$\begin{aligned}
\frac{\partial \pi_k}{\partial \eta_k} &= \frac{\partial}{\partial \eta_k} \left\{ \frac{exp(\eta_k)}{\sum_k exp(\eta_k)} \right\} \\
&= \frac{exp(\eta_k) \left[\sum_k exp(\eta_k)\right] - exp(\eta_k) exp(\eta_k)}{\left[\sum_k exp(\eta_k)\right]^2} \\
&= \pi_k - \pi_k \pi_k
\end{aligned}$$

If we combine these two cases, we can easily see that (5.208) holds. Now we prove (5.147).

$$\begin{aligned}
\frac{\partial \widetilde{E}}{\partial \eta_j} &= \lambda \frac{\partial \Omega(\mathbf{w})}{\partial \eta_j} \\
&= -\lambda \frac{\partial}{\partial \eta_j} \left\{ \sum_i \ln \left\{ \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \right\} \right\} \\
&= -\lambda \sum_i \frac{\partial}{\partial \eta_j} \left\{ \ln \left\{ \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \right\} \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \frac{\partial}{\partial \eta_j} \left\{ \sum_{k=1}^{M} \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2) \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \sum_{k=1}^{M} \frac{\partial}{\partial \eta_j} \{\pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2)\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \sum_{k=1}^{M} \frac{\partial}{\partial \pi_k} \{\pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2)\} \frac{\partial \pi_k}{\partial \eta_j} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \sum_{k=1}^{M} \mathcal{N}(w_i|\mu_k, \sigma_k^2)(\delta_{jk}\pi_j - \pi_j \pi_k) \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \left\{ \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) - \pi_j \sum_{k=1}^{M} \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2)) \right\} \\
&= -\lambda \sum_i \left\{ \frac{\pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} - \frac{\pi_j \sum_{k=1}^{M} \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2))}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \right\} \\
&= -\lambda \sum_i \{\gamma_j(w_i) - \pi_j\} = \lambda \sum_i \{\pi_j - \gamma_j(w_i)\}
\end{aligned}$$

Just as required.

Problem 5.33 Solution

It is trivial. We set the attachment point of the lower arm with the ground as the origin of the coordinate. We first aim to find the vertical distance from the origin to the target point, and this is also the value of x_2.

$$\begin{aligned}
x_2 &= L_1 \sin(\pi - \theta_1) + L_2 \sin(\theta_2 - (\pi - \theta_1)) \\
&= L_1 \sin\theta_1 - L_2 \sin(\theta_1 + \theta_2)
\end{aligned}$$

Similarly, we calculate the horizontal distance from the origin to the target point.

$$\begin{aligned}
x_1 &= -L_1 \cos(\pi - \theta_1) + L_2 \cos(\theta_2 - (\pi - \theta_1)) \\
&= L_1 \cos\theta_1 - L_2 \cos(\theta_1 + \theta_2)
\end{aligned}$$

From these two equations, we can clearly see the 'forward kinematics' of the robot arm.

Problem 5.34 Solution

By analogy with (5.208), we can write:

$$\frac{\partial \pi_k(\mathbf{x})}{\partial a_j^{\pi}} = \delta_{jk}\pi_j(\mathbf{x}) - \pi_j(\mathbf{x})\pi_k(\mathbf{x})$$

Using (5.153), we can see that:

$$E_n = -\ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\}$$

Therefore, we can derive:

$$\begin{aligned}
\frac{\partial E_n}{\partial a_j^{\pi}} &= -\frac{\partial}{\partial a_j^{\pi}}\ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\frac{\partial}{\partial a_j^{\pi}}\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\sum_{k=1}^{K}\frac{\partial \pi_k}{\partial a_j^{\pi}}\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2) \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\sum_{k=1}^{K}\left[\delta_{jk}\pi_j(\mathbf{x}_n) - \pi_j(\mathbf{x}_n)\pi_k(\mathbf{x}_n)\right]\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2) \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\left\{\pi_j(\mathbf{x}_n)\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_j,\sigma_j^2) - \pi_j(\mathbf{x}_n)\sum_{k=1}^{K}\pi_k(\mathbf{x}_n)\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= \frac{1}{\sum_{k=1}^{K}\pi_k \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\left\{-\pi_j(\mathbf{x}_n)\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_j,\sigma_j^2) + \pi_j(\mathbf{x}_n)\sum_{k=1}^{K}\pi_k(\mathbf{x}_n)\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\}
\end{aligned}$$

And if we denoted (5.154), we will have:

$$\frac{\partial E_n}{\partial a_j^{\pi}} = -\gamma_j + \pi_j$$

Note that our result is slightly different from (5.155) by the subindex. But there are actually the same if we substitute index j by index k in the final expression.

Problem 5.35 Solution

We deal with the derivative of error function with respect to $\boldsymbol{\mu}_k$ instead, which will give a vector as result. Furthermore, the lth element of this vector will be what we have been required. Since we know that:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\{\pi_k \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\} = \frac{\mathbf{t}_n - \boldsymbol{\mu}_k}{\sigma_k^2}\pi_k \mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)$$

One thing worthy noticing is that here we focus on the isotropic case as stated in page 273 of the textbook. To be more precise, $\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)$ should be $\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2\mathbf{I})$. Provided with the equation above, we can further obtain:

$$\begin{aligned}
\frac{\partial E_n}{\partial \boldsymbol{\mu}_k} &= \frac{\partial}{\partial \boldsymbol{\mu}_k}\left\{-\ln\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\frac{\partial}{\partial \boldsymbol{\mu}_k}\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\cdot\frac{\mathbf{t}_n-\boldsymbol{\mu}_k}{\sigma_k^2}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2) \\
&= -\gamma_k\frac{\mathbf{t}_n-\boldsymbol{\mu}_k}{\sigma_k^2}
\end{aligned}$$

Hence noticing (5.152), the lth element of the result above is what we are required.

$$\frac{\partial E_n}{\partial a_{kl}^{\mu}} = \frac{\partial E_n}{\partial \mu_{kl}} = \gamma_k\frac{\mu_{kl}-\mathbf{t}_l}{\sigma_k^2}$$

Problem 5.36 Solution

Similarly, we know that:

$$\frac{\partial}{\partial \sigma_k}\{\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\} = \left\{-\frac{D}{\sigma_k}+\frac{||\mathbf{t}_n-\boldsymbol{\mu}_k||^2}{\sigma_k^3}\right\}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)$$

Therefore, we can obtain:

$$\begin{aligned}
\frac{\partial E_n}{\partial \sigma_k} &= \frac{\partial}{\partial \sigma_k}\left\{-\ln\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\frac{\partial}{\partial \sigma_k}\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\cdot\left\{-\frac{D}{\sigma_k}+\frac{||\mathbf{t}_n-\boldsymbol{\mu}_k||^2}{\sigma_k^3}\right\}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2) \\
&= -\gamma_k\left\{-\frac{D}{\sigma_k}+\frac{||\mathbf{t}_n-\boldsymbol{\mu}_k||^2}{\sigma_k^3}\right\}
\end{aligned}$$

Note that there is a typo in (5.157) and the underlying reason is that: $|\sigma_k^2\mathbf{I}_{D\times D}| = (\sigma_k^2)^D$

Problem 5.37 Solution

First we know two properties for the Gaussian distribution $\mathcal{N}(\mathbf{t}|\boldsymbol{\mu},\sigma^2\mathbf{I})$:

$$\mathbb{E}[\mathbf{t}] = \int \mathbf{t}\mathcal{N}(\mathbf{t}|\boldsymbol{\mu},\sigma^2\mathbf{I})d\mathbf{t} = \boldsymbol{\mu}$$

And

$$\mathbb{E}[||\mathbf{t}||^2] = \int ||\mathbf{t}||^2 \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}, \sigma^2\mathbf{I})d\mathbf{t} = L\sigma^2 + ||\boldsymbol{\mu}||^2$$

Where we have used $\mathbb{E}[\mathbf{t}^T\mathbf{A}\mathbf{t}] = \mathrm{Tr}[\mathbf{A}\sigma^2\mathbf{I}] + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$ by setting $\mathbf{A} = \mathbf{I}$. This property can be found in *Matrixcookbook* eq(378). Here L is the dimension of $\mathbf{t}$. Noticing (5.148), we can write:

$$\begin{aligned}
\mathbb{E}[\mathbf{t}|\mathbf{x}] &= \int \mathbf{t}p(\mathbf{t}|\mathbf{x})d\mathbf{t} \\
&= \int \mathbf{t}\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k, \sigma_k^2)d\mathbf{t} \\
&= \sum_{k=1}^{K} \pi_k \int \mathbf{t}\mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k, \sigma_k^2)d\mathbf{t} \\
&= \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k
\end{aligned}$$

Then we prove (5.160).

$$\begin{aligned}
s^2(\mathbf{x}) &= \mathbb{E}[||\mathbf{t} - \mathbb{E}[\mathbf{t}|\mathbf{x}]||^2|\mathbf{x}] = \mathbb{E}[\left(\mathbf{t}^2 - 2\mathbf{t}\mathbb{E}[\mathbf{t}|\mathbf{x}] + \mathbb{E}[\mathbf{t}|\mathbf{x}]^2\right)|\mathbf{x}] \\
&= \mathbb{E}[\mathbf{t}^2|\mathbf{x}] - \mathbb{E}[2\mathbf{t}\mathbb{E}[\mathbf{t}|\mathbf{x}]|\mathbf{x}] + \mathbb{E}[\mathbf{t}|\mathbf{x}]^2 = \mathbb{E}[\mathbf{t}^2|\mathbf{x}] - \mathbb{E}[\mathbf{t}|\mathbf{x}]^2 \\
&= \int ||\mathbf{t}||^2 \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2)d\mathbf{t} - ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= \sum_{k=1}^{K} \pi_k \int ||\mathbf{t}||^2 \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2)d\mathbf{t} - ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= \sum_{k=1}^{K} \pi_k (L\sigma_k^2 + ||\boldsymbol{\mu}_k||^2) - ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k ||\boldsymbol{\mu}_k||^2 - ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k ||\boldsymbol{\mu}_k||^2 - 2 \times ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 + 1 \times ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k ||\boldsymbol{\mu}_k||^2 - 2(\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l)(\sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k) + \left(\sum_{k=1}^{K} \pi_k\right) ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k ||\boldsymbol{\mu}_k||^2 - 2(\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l)(\sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k) + \sum_{k=1}^{K} \pi_k ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k ||\boldsymbol{\mu}_k - \sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= \sum_{k=1}^{K} \pi_k \left(L\sigma_k^2 + ||\boldsymbol{\mu}_k - \sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2\right)
\end{aligned}$$

Note that there is a typo in (5.160), i.e., the coefficient L in front of σ_k^2 is missing.

Problem 5.38 Solution

From (5.167) and (5.171), we can write down the expression for the predictive distribution:

$$\begin{aligned}p(t|\mathbf{x},D,\alpha,\beta) &= \int p(\mathbf{w}|D,\alpha,\beta)p(t|\mathbf{x},\mathbf{w},\beta)\,d\mathbf{w}\\ &\approx \int q(\mathbf{w}|D)p(t|\mathbf{x},\mathbf{w},\beta)\,d\mathbf{w}\\ &= \int \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}},\mathbf{A}^{-1})\mathcal{N}(t|\mathbf{g}^T\mathbf{w}-\mathbf{g}^T\mathbf{w}_{\mathrm{MAP}}+y(\mathbf{x},\mathbf{w}_{\mathrm{MAP}}),\beta^{-1})\,d\mathbf{w}\end{aligned}$$

Note here $p(t|\mathbf{x},\mathbf{w},\beta)$ is given by (5.171) and $q(\mathbf{w}|D)$ is the approximation to the posterior $p(\mathbf{w}|D,\alpha,\beta)$, which is given by (5.167). Then by analogy with (2.115), we first deal with the mean of the predictive distribution:

$$\begin{aligned}\text{mean} &= \mathbf{g}^T\mathbf{w}-\mathbf{g}^T\mathbf{w}_{\mathrm{MAP}}+y(\mathbf{x},\mathbf{w}_{\mathrm{MAP}})|_{\mathbf{w}=\mathbf{w}_{\mathrm{MAP}}}\\ &= y(\mathbf{x},\mathbf{w}_{\mathrm{MAP}})\end{aligned}$$

Then we deal with the covariance matrix:

$$\text{Covariance matrix} = \beta^{-1}+\mathbf{g}^T\mathbf{A}^{-1}\mathbf{g}$$

Just as required.

Problem 5.39 Solution

Using Laplace Approximation, we can obtain:

$$p(D|\mathbf{w},\beta)p(\mathbf{w}|\alpha) = p(D|\mathbf{w}_{\mathrm{MAP}},\beta)p(\mathbf{w}_{\mathrm{MAP}}|\alpha)\exp\left\{-(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})^T\mathbf{A}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})\right\}$$

Then using (5.174), (5.162) and (5.163), we can obtain:

$$\begin{aligned}p(D|\alpha,\beta) &= \int p(D|\mathbf{w},\beta)p(\mathbf{w},\alpha)\,d\mathbf{w}\\ &= \int p(D|\mathbf{w}_{\mathrm{MAP}},\beta)p(\mathbf{w}_{\mathrm{MAP}}|\alpha)\exp\left\{-(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})^T\mathbf{A}(\mathbf{w}-\mathbf{w}_{\mathrm{MAP}})\right\}d\mathbf{w}\\ &= p(D|\mathbf{w}_{\mathrm{MAP}},\beta)p(\mathbf{w}_{\mathrm{MAP}}|\alpha)\frac{(2\pi)^{W/2}}{|\mathbf{A}|^{1/2}}\\ &= \prod_{n=1}^{N}\mathcal{N}(t_n|y(\mathbf{x}_n,\mathbf{w}_{\mathrm{MAP}}),\beta^{-1})\mathcal{N}(\mathbf{w}_{\mathrm{MAP}}|\mathbf{0},\alpha^{-1}\mathbf{I})\frac{(2\pi)^{W/2}}{|\mathbf{A}|^{1/2}}\end{aligned}$$

If we take logarithm of both sides, we will obtain (5.175) just as required.

Problem 5.40 Solution

For a k-class classification problem, we need to use softmax activation function and also the error function is now given by (5.24). Therefore, the

Hessian matrix should be derived from (5.24) and the cross entropy in (5.184) will also be replaced by (5.24).

Problem 5.41 Solution

By analogy to Prob.5.39, we can write:

$$p(D|\alpha) = p(D|\mathbf{w}_{\text{MAP}})p(\mathbf{w}_{\text{MAP}}|\alpha)\frac{(2\pi)^{W/2}}{|\mathbf{A}|^{1/2}}$$

Since we know that the prior $p(\mathbf{w}|\alpha)$ follows a Gaussian distribution, i.e., (5.162), as stated in the text. Therefore we can obtain:

$$\begin{aligned}\ln p(D|\alpha) &= \ln p(D|\mathbf{w}_{\text{MAP}}) + \ln p(\mathbf{w}_{\text{MAP}}|\alpha) - \frac{1}{2}\ln|\mathbf{A}| + \text{const} \\ &= \ln p(D|\mathbf{w}_{\text{MAP}}) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + \frac{W}{2}\ln\alpha - \frac{1}{2}\ln|\mathbf{A}| + \text{const} \\ &= -E(\mathbf{w}_{\text{MAP}}) + \frac{W}{2}\ln\alpha - \frac{1}{2}\ln|\mathbf{A}| + \text{const}\end{aligned}$$

Just as required.

0.6 Kernel Methods

Problem 6.1 Solution

Recall that in section.6.1, a_n can be written as (6.4). We can derive:

$$\begin{aligned}a_n &= -\frac{1}{\lambda}\{\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n) - t_n\} \\ &= -\frac{1}{\lambda}\{w_1\phi_1(\mathbf{x}_n) + w_2\phi_2(\mathbf{x}_n) + ... + w_M\phi_M(\mathbf{x}_n) - t_n\} \\ &= -\frac{w_1}{\lambda}\phi_1(\mathbf{x}_n) - \frac{w_2}{\lambda}\phi_2(\mathbf{x}_n) - ... - \frac{w_M}{\lambda}\phi_M(\mathbf{x}_n) + \frac{t_n}{\lambda} \\ &= (c_n - \frac{w_1}{\lambda})\phi_1(\mathbf{x}_n) + (c_n - \frac{w_2}{\lambda})\phi_2(\mathbf{x}_n) + ... + (c_n - \frac{w_M}{\lambda})\phi_M(\mathbf{x}_n)\end{aligned}$$

Here we have defined:

$$c_n = \frac{t_n/\lambda}{\phi_1(\mathbf{x}_n) + \phi_2(\mathbf{x}_n) + ... + \phi_M(\mathbf{x}_n)}$$

From what we have derived above, we can see that a_n is a linear combination of $\boldsymbol{\phi}(\mathbf{x}_n)$. What's more, we first substitute $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$ into (6.7), and then we will obtain (6.5). Next we substitute (6.3) into (6.5) we will obtain (6.2) just as required.

Problem 6.2 Solution

By analogy to Eq (2.115), i.e.,

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$$

We can obtain:

$$p(a_{N+1}|\mathbf{t}_N) = N(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \qquad (*)$$

Where we have defined:

$$\mathbf{A} = \mathbf{k}^T\mathbf{C}_N^{-1}, \mathbf{b} = \mathbf{0}, \mathbf{L}^{-1} = c - \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{k}$$

And

$$\boldsymbol{\mu} = \mathbf{a}_N^{\star}, \boldsymbol{\Lambda} = \mathbf{H}$$

Therefore, the mean is given by:

$$\mathbf{A}\boldsymbol{\mu} + \mathbf{b} = \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{a}_N^{\star} = \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{C}_N(\mathbf{t}_N - \sigma_N) = \mathbf{k}^T(\mathbf{t}_N - \sigma_N)$$

Where we have used Eq (6.84). The covariance matrix is given by:

$$\begin{aligned}
\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T &= c - \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{k} + \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{H}^{-1}(\mathbf{k}^T\mathbf{C}_N^{-1})^T \\
&= c - \mathbf{k}^T(\mathbf{C}_N^{-1} - \mathbf{C}_N^{-1}\mathbf{H}^{-1}\mathbf{C}_N^{-1})\mathbf{k} \\
&= c - \mathbf{k}^T\Big(\mathbf{C}_N^{-1} - \mathbf{C}_N^{-1}(\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1}\mathbf{C}_N^{-1}\Big)\mathbf{k} \\
&= c - \mathbf{k}^T\Big(\mathbf{C}_N^{-1} - (\mathbf{C}_N\mathbf{W}_N\mathbf{C}_N + \mathbf{C}_N^{-1})^{-1}\Big)\mathbf{k}
\end{aligned}$$

Where we have used Eq (6.85) and the fact that $\mathbf{C}_N$ is symmetric. Then we use matrix identity (C.7) to further reduce the expression, which will finally give Eq (6.88).

Problem 6.27 Solution(Wait for update) This problem is really complicated.

What's more, I find that Eq (6.91) seems not right.

0.7 Sparse Kernel Machines

Problem 7.1 Solution

By analogy to Eq (2.249), we can obtain:

$$p(\mathbf{x}|t) = \begin{cases} \dfrac{1}{N_{+1}}\displaystyle\sum_{n=1}^{N_{+1}} \dfrac{1}{Z_k} \cdot k(\mathbf{x}, \mathbf{x}_n) & t = +1 \\ \dfrac{1}{N_{-1}}\displaystyle\sum_{n=1}^{N_{-1}} \dfrac{1}{Z_k} \cdot k(\mathbf{x}, \mathbf{x}_n) & t = -1 \end{cases}$$

where N_{+1} represents the number of samples with label $t = +1$ and it is the same for N_{-1}. Z_k is a normalization constant representing the volume of the hypercube. Since we have equal prior for the class, i.e.,

$$p(t) = \begin{cases} 0.5 & t = +1 \\ 0.5 & t = -1 \end{cases}$$

Based on Bayes' Theorem, we have $p(t|\mathbf{x}) \propto p(\mathbf{x}|t) \cdot p(t)$, yielding:

$$p(t|\mathbf{x}) = \begin{cases} \frac{1}{Z} \cdot \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \cdot k(\mathbf{x}, \mathbf{x}_n) & t = +1 \\ \frac{1}{Z} \cdot \frac{1}{N_{-1}} \sum_{n=1}^{N_{-1}} \cdot k(\mathbf{x}, \mathbf{x}_n) & t = -1 \end{cases}$$

Where $1/Z$ is a normalization constant to guarantee the integration of the posterior equal to 1. To classify a new sample $\mathbf{x}^\star$, we try to find the value $t^\star$ that can maximize $p(t|\mathbf{x})$. Therefore, we can obtain:

$$t^\star = \begin{cases} +1 & \text{if} \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \geq \frac{1}{N_{-1}} \sum_{n=1}^{N_{-1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \\ -1 & \text{if} \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \leq \frac{1}{N_{-1}} \sum_{n=1}^{N_{-1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \end{cases} \qquad (*)$$

If we now choose the kernel function as $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$,we have:

$$\frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} k(\mathbf{x}, \mathbf{x}_n) = \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \mathbf{x}^T \mathbf{x}_n = \mathbf{x}^T \tilde{\mathbf{x}}_{+1}$$

Where we have denoted:

$$\tilde{\mathbf{x}}_{+1} = \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \mathbf{x}_n$$

and similarly for $\tilde{\mathbf{x}}_{-1}$. Therefore, the classification criterion $(*)$ can be written as:

$$t^\star = \begin{cases} +1 & \text{if} \quad \tilde{\mathbf{x}}_{+1} \geq \tilde{\mathbf{x}}_{-1} \\ -1 & \text{if} \quad \tilde{\mathbf{x}}_{+1} \leq \tilde{\mathbf{x}}_{-1} \end{cases}$$

When we choose the kernel function as $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$, we can similarly obtain the classification criterion:

$$t^\star = \begin{cases} +1 & \text{if} \quad \tilde{\phi}(\mathbf{x}_{+1}) \geq \tilde{\phi}(\mathbf{x}_{-1}) \\ -1 & \text{if} \quad \tilde{\phi}(\mathbf{x}_{+1}) \leq \tilde{\phi}(\mathbf{x}_{-1}) \end{cases}$$

Where we have defined:

$$\tilde{\phi}(\mathbf{x}_{+1}) = \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \phi(\mathbf{x}_n)$$

Problem 7.2 Solution

Suppose we have find $\mathbf{w}_0$ and b_0, which can let all points satisfy Eq (7.5) and simultaneously minimize Eq (7.3). This hyperlane decided by $\mathbf{w}_0$ and b_0 is the optimal classification margin. Now if the constraint in Eq (7.5) becomes:

$$t_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b) \geq \gamma$$

We can conclude that if we perform change of variables: $\mathbf{w}_0->\gamma\mathbf{w}_0$ and $b->\gamma b$, the constraint will still satisfy and Eq (7.3) will be minimize. In other words, if the right side of the constraint changes from 1 to γ, The new hyperlane decided by $\gamma\mathbf{w}_0$ and γb_0 is the optimal classification margin. However, the minimum distance from the points to the classification margin is still the same.

Problem 7.3 Solution

Suppose we have $\mathbf{x}_1$ belongs to class one and we denote its target value $t_1 = 1$, and similarly $\mathbf{x}_2$ belongs to class two and we denote its target value $t_2 = -1$. Since we only have two points, they must have $t_i \cdot y(\mathbf{x}_i) = 1$ as shown in Fig. 7.1. Therefore, we have an equality constrained optimization problem:

$$\text{minimize} \frac{1}{2}||\mathbf{w}||^2 \quad \text{s.t.} \begin{cases} \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_1)+b = 1 \\ \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_2)+b = -1 \end{cases}$$

This is an convex optimization problem and it has been proved that global optimal exists.

Problem 7.4 Solution

Since we know that

$$\rho = \frac{1}{||\mathbf{w}||}$$

Therefore, we have:

$$\frac{1}{\rho^2} = ||\mathbf{w}||^2$$

In other words, we only need to prove that

$$||\mathbf{w}||^2 = \sum_{n=1}^{N} a_n$$

When we find th optimal solution, the second term on the right hand side of Eq (7.7) vanishes. Based on Eq (7.8) and Eq (7.10), we also observe that its dual is given by:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}||\mathbf{w}||^2$$

Therefore, we have:

$$\frac{1}{2}||\mathbf{w}||^2 = L(\mathbf{a}) = \tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}||\mathbf{w}||^2$$

Rearranging it, we will obtain what we are required.

Problem 7.5 Solution

We have already proved this problem in the previous one.

Problem 7.6 Solution

If the target variable can only choose from $\{-1, 1\}$, and we know that

$$p(t = 1|y) = \sigma(y)$$

We can obtain:

$$p(t = -1|y) = 1 - p(t = 1|y) = 1 - \sigma(y) = \sigma(-y)$$

Therefore, combining these two situations, we can derive:

$$p(t|y) = \sigma(yt)$$

Consequently, we can obtain the negative log likelihood:

$$-\ln p(\mathbf{D}) = -\ln \prod_{n=1}^{N} \sigma(y_n t_n) = -\sum_{n=1}^{N} \ln \sigma(y_n t_n) = \sum_{n=1}^{N} E_{LR}(y_n t_n)$$

Here $\mathbf{D}$ represents the dataset, i.e.,$\mathbf{D} = \{(\mathbf{x}_n, t_n); n = 1, 2, ..., N\}$, and $E_{LR}(yt)$ is given by Eq (7.48). With the addition of a quadratic regularization, we obtain exactly Eq (7.47).

Problem 7.7 Solution

The derivatives are easy to obtain. Our main task is to derive Eq (7.61) using Eq (7.57)-(7.60).

$$
\begin{aligned}
L &= C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(\mu_n\xi_n+\widehat{\mu}_n\widehat{\xi}_n)\\
&\quad -\sum_{n=1}^{N}a_n(\epsilon+\xi_n+y_n-t_n)-\sum_{n=1}^{N}\widehat{a}_n(\epsilon+\widehat{\xi}_n+y_n-t_n)\\
&= C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n+\mu_n)\xi_n-\sum_{n=1}^{N}(\widehat{a}_n+\widehat{\mu}_n)\widehat{\xi}_n\\
&\quad -\sum_{n=1}^{N}a_n(\epsilon+y_n-t_n)-\sum_{n=1}^{N}\widehat{a}_n(\epsilon+y_n-t_n)\\
&= C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}C\xi_n-\sum_{n=1}^{N}C\widehat{\xi}_n\\
&\quad -\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon-\sum_{n=1}^{N}(a_n-\widehat{a}_n)(y_n-t_n)\\
&= \frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon-\sum_{n=1}^{N}(a_n-\widehat{a}_n)(y_n-t_n)\\
&= \frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n-\widehat{a}_n)(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b-t_n)-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon+\sum_{n=1}^{N}\\
&= \frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n-\widehat{a}_n)(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b)-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n\\
&= \frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n-\widehat{a}_n)\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n\\
&= \frac{1}{2}||\mathbf{w}||^2-||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n\\
&= -\frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n
\end{aligned}
$$

Just as required.

Problem 7.8 Solution

This obviously follows from the KKT condition, described in Eq (7.67) and (7.68).

Problem 7.9 Solution

The prior is given by Eq (7.80).

$$p(\mathbf{w}|\boldsymbol{\alpha})=\prod_{i=1}^{M}\mathcal{N}(0,\alpha_i^{-1})=\mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1})$$

Where we have defined:

$$\mathbf{A}=diag(\alpha_i)$$

The likelihood is given by Eq (7.79).

$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) &= \prod_{n=1}^{N} p(t_n|\mathbf{x}_n,\mathbf{w},\beta^{-1}) \\ &= \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w},\beta^{-1}\mathbf{I}) \end{aligned}$$

Where we have defined:

$$\mathbf{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1),\boldsymbol{\phi}(\mathbf{x}_2),...,\boldsymbol{\phi}(\mathbf{x}_n)]^T$$

Our definitions of $\mathbf{\Phi}$ and $\mathbf{A}$ as consistent with the main text. Therefore, according to Eq (2.113)-Eq (2.117), we have:

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\boldsymbol{\alpha},\beta) = \mathcal{N}(\mathbf{m},\mathbf{\Sigma})$$

Where we have defined:

$$\mathbf{\Sigma} = (\mathbf{A}+\beta\mathbf{\Phi}^T\mathbf{\Phi})^{-1}$$

And

$$\mathbf{m} = \beta\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}$$

Just as required.

Problem 7.10&7.11 Solution

It is quite similar to the previous problem. We begin by writting down the prior:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M}\mathcal{N}(0,\alpha_i^{-1}) = \mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1})$$

Then we write down the likelihood:

$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) &= \prod_{n=1}^{N} p(t_n|\mathbf{x}_n,\mathbf{w},\beta^{-1}) \\ &= \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w},\beta^{-1}\mathbf{I}) \end{aligned}$$

Since we know that:

$$p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta) = \int p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\alpha})\,d\mathbf{w}$$

First as required by Prob.7.10, we will solve it by completing the square. We begin by write down the expression for $p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)$:

$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta) &= \int \mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1})\mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w},\beta^{-1}\mathbf{I})d\mathbf{w} \\ &= (\frac{\beta}{2\pi})^{N/2}\cdot\frac{1}{(2\pi)^{M/2}}\cdot\prod_{m=1}^{M}\alpha_i^{1/2}\cdot\int exp\{-E(\mathbf{w})\}d\mathbf{w} \end{aligned}$$

Where we have defined:

$$E(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} + \frac{\beta}{2}||\mathbf{t}-\mathbf{\Phi}\mathbf{w}||^2$$

We expand $E(\mathbf{w})$ with respect to $\mathbf{w}$:

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}\Big\{\mathbf{w}^T(\mathbf{A}+\beta\mathbf{\Phi}^T\mathbf{\Phi})\mathbf{w} - 2\beta\mathbf{t}^T(\mathbf{\Phi}\mathbf{w}) + \beta\mathbf{t}^T\mathbf{t}\Big\} \\ &= \frac{1}{2}\Big\{\mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w} - 2\mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{w} + \beta\mathbf{t}^T\mathbf{t}\Big\} \\ &= \frac{1}{2}\Big\{(\mathbf{w}-\mathbf{m})^T\mathbf{\Sigma}^{-1}(\mathbf{w}-\mathbf{m}) + \beta\mathbf{t}^T\mathbf{t} - \mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{m}\Big\} \end{aligned}$$

Where we have used Eq (7.82) and Eq (7.83). Substituting $E(\mathbf{w})$ into the integral, we will obtain:

$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta) &= (\frac{\beta}{2\pi})^{N/2}\cdot\frac{1}{(2\pi)^{M/2}}\cdot\prod_{m=1}^{M}\alpha_i^{1/2}\cdot\int exp\{-E(\mathbf{w})\}d\mathbf{w} \\ &= (\frac{\beta}{2\pi})^{N/2}\cdot\frac{1}{(2\pi)^{M/2}}\cdot\prod_{m=1}^{M}\alpha_i^{1/2}\cdot(2\pi)^{M/2}\cdot|\mathbf{\Sigma}|^{1/2}exp\Big\{-\frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-\mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{m})\Big\} \\ &= (\frac{\beta}{2\pi})^{N/2}\cdot|\mathbf{\Sigma}|^{1/2}\cdot\prod_{m=1}^{M}\alpha_i^{1/2}\cdot exp\Big\{-\frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-\mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{m})\Big\} \\ &= (\frac{\beta}{2\pi})^{N/2}\cdot|\mathbf{\Sigma}|^{1/2}\cdot\prod_{m=1}^{M}\alpha_i^{1/2}\cdot exp\Big\{-E(\mathbf{t})\Big\} \end{aligned}$$

We further expand $E(\mathbf{t})$:

$$\begin{aligned} E(\mathbf{t}) &= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-\mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{m}) \\ &= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-(\beta\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t})^T\mathbf{\Sigma}^{-1}(\beta\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t})) \\ &= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-\beta^2\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}) \\ &= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-\beta^2\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}) \\ &= \frac{1}{2}\mathbf{t}^T(\beta\mathbf{I}-\beta^2\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T)\mathbf{t} \\ &= \frac{1}{2}\mathbf{t}^T\Big[\beta\mathbf{I}-\beta\mathbf{\Phi}(\mathbf{A}+\beta\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\beta\Big]\mathbf{t} \\ &= \frac{1}{2}\mathbf{t}^T(\beta^{-1}\mathbf{I}+\mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T)^{-1}\mathbf{t} = \frac{1}{2}\mathbf{t}^T\mathbf{C}^{-1}\mathbf{t} \end{aligned}$$

Note that in the last step we have used matrix identity Eq (C.7). Therefore, as we know that the pdf is Gaussian and the exponential term has been given by $E(\mathbf{t})$, we can easily write down Eq (7.85) considering those normalization constant.

What's more, as required by Prob.7.11, the evaluation of the integral can be easily performed using Eq(2.113)- Eq(2.117).

Problem 7.12 Solution

According to the previous problem, we can explicitly write down the log marginal likelihood in an alternative form:

$$\ln p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta) \quad = \quad \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{i=1}^{M}\ln\alpha_i - E(\mathbf{t})$$

We first derive:

$$\begin{aligned}\frac{dE(\mathbf{t})}{d\alpha_i} &= -\frac{1}{2}\frac{d}{d\alpha_i}(\mathbf{m}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}) \\ &= -\frac{1}{2}\frac{d}{d\alpha_i}(\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}) \\ &= -\frac{1}{2}\frac{d}{d\alpha_i}(\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}) \\ &= -\frac{1}{2}Tr\Big[\frac{d}{d\boldsymbol{\Sigma}^{-1}}(\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t})\cdot\frac{d\boldsymbol{\Sigma}^{-1}}{d\alpha_i}\Big] \\ &= \frac{1}{2}\beta^2 Tr[\boldsymbol{\Sigma}(\boldsymbol{\Phi}^T\mathbf{t})(\boldsymbol{\Phi}^T\mathbf{t})^T\boldsymbol{\Sigma}\cdot\mathbf{I}_i] = \frac{1}{2}m_{ii}^2\end{aligned}$$

In the last step, we have utilized the following equation:

$$\frac{d}{d\mathbf{X}}Tr(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}) = -\mathbf{X}^{-T}\mathbf{A}^T\mathbf{B}^T\mathbf{X}^{-T}$$

Moreover, here $\mathbf{I}_i$ is a matrix with all elements equal to zero, expect the i-th diagonal element, and the i-th diagonal element equals to 1. Then we utilize matrix identity Eq (C.22) to derive:

$$\begin{aligned}\frac{d\ln|\boldsymbol{\Sigma}|}{d\alpha_i} &= -\frac{d\ln|\boldsymbol{\Sigma}^{-1}|}{d\alpha_i} \\ &= -Tr\Big[\boldsymbol{\Sigma}\frac{d}{d\alpha_i}(\mathbf{A}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})\Big] \\ &= -\Sigma_{ii}\end{aligned}$$

Therefore, we can obtain:

$$\frac{d\ln p}{d\alpha_i} = \frac{1}{2\alpha_i} - \frac{1}{2}m_i^2 - \frac{1}{2}\Sigma_{ii}$$

Set it to zero and obtain:

$$\alpha_i = \frac{1-\alpha_i\Sigma_{ii}}{m_i} = \frac{\gamma_i}{m_i^2}$$

Then we calculate the derivatives of $\ln p$ with respect to β beginning by:

$$\begin{aligned}
\frac{d\ln|\mathbf{\Sigma}|}{d\beta} &= -\frac{d\ln|\mathbf{\Sigma}^{-1}|}{d\beta} \\
&= -Tr\Big[\mathbf{\Sigma}\frac{d}{d\beta}(\mathbf{A}+\beta\mathbf{\Phi}^T\mathbf{\Phi})\Big] \\
&= -Tr\Big[\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{\Phi}\Big]
\end{aligned}$$

Then we continue:

$$\begin{aligned}
\frac{dE(\mathbf{t})}{d\beta} &= \frac{1}{2}\mathbf{t}^T\mathbf{t}-\frac{1}{2}\frac{d}{d\beta}(\mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{m}) \\
&= \frac{1}{2}\mathbf{t}^T\mathbf{t}-\frac{1}{2}\frac{d}{d\beta}(\beta^2\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}) \\
&= \frac{1}{2}\mathbf{t}^T\mathbf{t}-\frac{1}{2}\frac{d}{d\beta}(\beta^2\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}) \\
&= \frac{1}{2}\mathbf{t}^T\mathbf{t}-\beta\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}-\frac{1}{2}\beta^2\frac{d}{d\beta}(\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}) \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t}-2\beta\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}-\beta^2\frac{d}{d\beta}(\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t})\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t}-2\mathbf{t}^T(\mathbf{\Phi}\mathbf{m})-\beta^2\frac{d}{d\beta}(\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t})\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t}-2\mathbf{t}^T(\mathbf{\Phi}\mathbf{m})-\beta^2 Tr\big[\frac{d}{d\mathbf{\Sigma}^{-1}}(\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t})\cdot\frac{d\mathbf{\Sigma}^{-1}}{d\beta}\big]\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t}-2\mathbf{t}^T(\mathbf{\Phi}\mathbf{m})+\beta^2 Tr\big[\mathbf{\Sigma}(\mathbf{\Phi}^T\mathbf{t})(\mathbf{\Phi}^T\mathbf{t})^T\mathbf{\Sigma}\cdot\mathbf{\Phi}^T\mathbf{\Phi}\big]\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t}-2\mathbf{t}^T(\mathbf{\Phi}\mathbf{m})+Tr\big[\mathbf{m}\mathbf{m}^T\cdot\mathbf{\Phi}^T\mathbf{\Phi}\big]\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t}-2\mathbf{t}^T(\mathbf{\Phi}\mathbf{m})+Tr\big[\mathbf{\Phi}\mathbf{m}\mathbf{m}^T\cdot\mathbf{\Phi}^T\big]\Big\} \\
&= \frac{1}{2}||\mathbf{t}-\mathbf{\Phi}\mathbf{m}||^2
\end{aligned}$$

Therefore, we have obtained:

$$\frac{d\ln p}{d\beta} = \frac{1}{2}\Big(\frac{N}{\beta}-||\mathbf{t}-\mathbf{\Phi}\mathbf{m}||^2-Tr[\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{\Phi}]\Big)$$

Using Eq (7.83), we can obtain:

$$\begin{aligned}\mathbf{\Sigma\Phi}^T\mathbf{\Phi} &= \mathbf{\Sigma\Phi}^T\mathbf{\Phi} + \beta^{-1}\mathbf{\Sigma A} - \beta^{-1}\mathbf{\Sigma A} \\ &= \mathbf{\Sigma}(\beta\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{A})\beta^{-1} - \beta^{-1}\mathbf{\Sigma A} \\ &= \mathbf{I}\beta^{-1} - \beta^{-1}\mathbf{\Sigma A} \\ &= (\mathbf{I} - \mathbf{\Sigma A})\beta^{-1}\end{aligned}$$

Setting the derivative equal to zero, we can obtain:

$$\beta^{-1} = \frac{||\mathbf{t} - \mathbf{\Phi m}||^2}{N - Tr(\mathbf{I} - \mathbf{\Sigma A})} = \frac{||\mathbf{t} - \mathbf{\Phi m}||^2}{N - \sum_i \gamma_i}$$

Just as required.

Problem 7.13 Solution

This problem is quite confusing. In my point of view, the posterior should be denoted as $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \{a_i, b_i\}, a_\beta, b_\beta)$, where a_β, b_β controls the Gamma distribution of β, and a_i, b_i controls the Gamma distribution of α_i. What we should do is to maximize the marginal likelihood $p(\mathbf{t}|\mathbf{X}, \{a_i, b_i\}, a_\beta, b_\beta)$ with respect to $\{a_i, b_i\}, a_\beta, b_\beta$. Now we do not have a point estimation for the hyperparameters β and α_i. We have a distribution (controled by the hyper priors, i.e., $\{a_i, b_i\}, a_\beta, b_\beta$) instead.

Problem 7.14 Solution

We begin by writing down $p(t|\mathbf{x}, \mathbf{w}, \beta^*)$. Using Eq (7.76) and Eq (7.77), we can obtain:

$$p(t|\mathbf{x}, \mathbf{w}, \beta^*) = \mathcal{N}(t|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}), (\beta^*)^{-1})$$

Then we write down $p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha^*, \beta^*)$. Using Eq (7.81), (7.82) and (7.83), we can obtain:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha^*, \beta^*) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{\Sigma})$$

Where $\mathbf{m}$ and $\mathbf{\Sigma}$ are evaluated using Eq (7.82) and (7.83) given $\alpha = \alpha^*$ and $\beta = \beta^*$. Then we utilize Eq (7.90) and obtain:

$$\begin{aligned}p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \alpha^*, \beta^*) &= \int \mathcal{N}(t|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}), (\beta^*)^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{\Sigma})\,d\,\mathbf{w} \\ &= \int \mathcal{N}(t|\boldsymbol{\phi}(\mathbf{x})^T\mathbf{w}, (\beta^*)^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{\Sigma})\,d\,\mathbf{w}\end{aligned}$$

Using Eq (2.113)-(2.117), we can obtain:

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \alpha^*, \beta^*) = \mathcal{N}(\mu, \sigma^2)$$

Where we have defined:

$$\mu = \mathbf{m}^T\boldsymbol{\phi}(\mathbf{x})$$

And

$$\sigma^2 = (\beta^*)^{-1} + \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\Sigma} \boldsymbol{\phi}(\mathbf{x})$$

Just as required.

Problem 7.15 Solution

We just follow the hint.

$$\begin{aligned}
L(\boldsymbol{\alpha}) &= -\frac{1}{2}\{N\ln 2\pi + \ln|\mathbf{C}| + \mathbf{t}^T\mathbf{C}^{-1}\mathbf{t}\} \\
&= -\frac{1}{2}\Big\{N\ln 2\pi + \ln|\mathbf{C}_{-i}| + \ln|1+\alpha_i^{-1}\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| \\
&\quad +\mathbf{t}^T(\mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i})\mathbf{t}\Big\} \\
&= L(\boldsymbol{\alpha}_{-i}) - \frac{1}{2}\ln|1+\alpha_i^{-1}\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| + \frac{1}{2}\mathbf{t}^T\frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i}\mathbf{t} \\
&= L(\boldsymbol{\alpha}_{-i}) - \frac{1}{2}\ln|1+\alpha_i^{-1}s_i| + \frac{1}{2}\frac{q_i^2}{\alpha_i + s_i} \\
&= L(\boldsymbol{\alpha}_{-i}) - \frac{1}{2}\ln\frac{\alpha_i + s_i}{\alpha_i} + \frac{1}{2}\frac{q_i^2}{\alpha_i + s_i} \\
&= L(\boldsymbol{\alpha}_{-i}) + \frac{1}{2}\Big[\ln\alpha_i - \ln(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\Big] = L(\boldsymbol{\alpha}_{-i}) + \lambda(\alpha_i)
\end{aligned}$$

Where we have defined $\lambda(\alpha_i)$, s_i and q_i as shown in Eq (7.97)-(7.99).

Problem 7.16 Solution

We first calculate the first derivative of Eq(7.97) with respect to α_i:

$$\frac{\partial\lambda}{\partial\alpha_i} = \frac{1}{2}[\frac{1}{\alpha_i} - \frac{1}{\alpha_i + s_i} - \frac{q_i^2}{(\alpha_i + s_i)^2}]$$

Then we calculate the second derivative:

$$\frac{\partial^2\lambda}{\partial\alpha_i^2} = \frac{1}{2}[-\frac{1}{\alpha_i^2} + \frac{1}{(\alpha_i + s_i)^2} + \frac{2q_i^2}{(\alpha_i + s_i)^3}]$$

Next we aim to prove that when α_i is given by Eq (7.101), i.e., setting the first derivative equal to 0, the second derivative (i.e., the expression above) is negative. First we can obtain:

$$\alpha_i + s_i = \frac{s_i^2}{q_i^2 - s_i} + s_i = \frac{s_i q_i^2}{q_i^2 - s_i}$$

Therefore, substituting $\alpha_i + s_i$ and α_i into the second derivative, we can obtain:

$$\begin{aligned}
\frac{\partial^2 \lambda}{\partial \alpha_i^2} &= \frac{1}{2}[-\frac{(q_i^2 - s_i)^2}{s_i^4} + \frac{(q_i^2 - s_i)^2}{s_i^2 q_i^4} + \frac{2q_i^2(q_i^2 - s_i)^3}{s_i^3 q_i^6}] \\
&= \frac{1}{2}[-\frac{q_i^4(q_i^2 - s_i)^2}{q_i^4 s_i^4} + \frac{s_i^2(q_i^2 - s_i)^2}{s_i^4 q_i^4} + \frac{2s_i(q_i^2 - s_i)^3}{s_i^4 q_i^4}] \\
&= \frac{1}{2}\frac{(q_i^2 - s_i)^2}{q_i^4 s_i^4}[-q_i^4 + s_i^2 + 2s_i(q_i^2 - s_i)] \\
&= \frac{1}{2}\frac{(q_i^2 - s_i)^2}{q_i^4 s_i^4}[-(q_i^2 - s_i)^2] \\
&= -\frac{1}{2}\frac{(q_i^2 - s_i)^4}{q_i^4 s_i^4} < 0
\end{aligned}$$

Just as required.

Problem 7.17 Solution

We just follow the hint. According to Eq (7.102), Eq (7.86) and matrix identity (C.7), we have:

$$\begin{aligned}
Q_i &= \boldsymbol{\varphi}_i^T \mathbf{C}^{-1} \mathbf{t} \\
&= \boldsymbol{\varphi}_i^T (\beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T)^{-1}\mathbf{t} \\
&= \boldsymbol{\varphi}_i^T (\beta\mathbf{I} - \beta\mathbf{I}\boldsymbol{\Phi}(\mathbf{A} + \boldsymbol{\Phi}^T\beta\mathbf{I}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\beta\mathbf{I})\mathbf{t} \\
&= \boldsymbol{\varphi}_i^T (\beta - \beta^2\boldsymbol{\Phi}(\mathbf{A} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T)\mathbf{t} \\
&= \boldsymbol{\varphi}_i^T (\beta - \beta^2\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T)\mathbf{t} \\
&= \beta\boldsymbol{\varphi}_i^T\mathbf{t} - \beta^2\boldsymbol{\varphi}_i^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}
\end{aligned}$$

Similarly, we can obtain:

$$\begin{aligned}
S_i &= \boldsymbol{\varphi}_i^T \mathbf{C}^{-1} \boldsymbol{\varphi}_i \\
&= \boldsymbol{\varphi}_i^T (\beta - \beta^2\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T)\boldsymbol{\varphi}_i \\
&= \beta\boldsymbol{\varphi}_i^T\boldsymbol{\varphi}_i - \beta^2\boldsymbol{\varphi}_i^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\varphi}_i
\end{aligned}$$

Just as required.

Problem 7.18 Solution

We begin by deriving the first term in Eq (7.109) with respect to $\mathbf{w}$. This can be easily evaluate based on Eq (4.90)-(4.91).

$$\frac{\partial}{\partial \mathbf{w}}\Big\{\sum_{n=1}^{N} t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\Big\} = \sum_{n=1}^{N}(t_n - y_n)\boldsymbol{\phi}_n = \boldsymbol{\Phi}^T(\mathbf{t} - \mathbf{y})$$

Since the derivative of the second term in Eq (7.109) with respect to $\mathbf{w}$ is rather simple to obtain. Therefore, The first derivative of Eq (7.109) with respect to $\mathbf{w}$ is:

$$\frac{\partial \ln p}{\partial \mathbf{w}} = \mathbf{\Phi}^T(\mathbf{t}-\mathbf{y}) - \mathbf{A}\mathbf{w}$$

For the Hessian matrix, we can first obtain:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\Big\{\mathbf{\Phi}^T(\mathbf{t}-\mathbf{y})\Big\} &= \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}}\Big\{(t_n - y_n)\boldsymbol{\phi}_n\Big\} \\
&= -\sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}}\Big\{y_n \cdot \boldsymbol{\phi}_n\Big\} \\
&= -\sum_{n=1}^{N} \frac{\partial \sigma(\mathbf{w}^T\boldsymbol{\phi}_n)}{\partial \mathbf{w}} \cdot \boldsymbol{\phi}_n^T \\
&= -\sum_{n=1}^{N} \frac{\partial \sigma(a)}{\partial a} \cdot \frac{\partial a}{\partial \mathbf{w}} \cdot \boldsymbol{\phi}_n^T
\end{aligned}$$

Where we have defined $a = \mathbf{w}^T\boldsymbol{\phi}_n$. Then we can utilize Eq (4.88) to derive:

$$\frac{\partial}{\partial \mathbf{w}}\Big\{\mathbf{\Phi}^T(\mathbf{t}-\mathbf{y})\Big\} = -\sum_{n=1}^{N} \sigma(1-\sigma)\cdot\boldsymbol{\phi}_n \cdot \boldsymbol{\phi}_n^T = -\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi}$$

Where $\mathbf{B}$ is a diagonal $N \times N$ matrix with elements $b_n = y_n(1-y_n)$. Therefore, we can obtain the Hessian matrix:

$$\mathbf{H} = \frac{\partial}{\partial \mathbf{w}}\Big\{\frac{\partial \ln p}{\partial \mathbf{w}}\Big\} = -(\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi} + \mathbf{A})$$

Just as required.

Problem 7.19 Solution

We begin from Eq (7.114).

$$\begin{aligned}
p(\mathbf{t}|\alpha) &= p(\mathbf{t}|\mathbf{w}^*)p(\mathbf{w}^*|\alpha)(2\pi)^{M/2}|\mathbf{\Sigma}|^{1/2} \\
&= \Big[\prod_{n=1}^{N} p(t_n|x_n,\mathbf{w})\Big]\Big[\prod_{i=1}^{M}\mathcal{N}(w_i|0,\alpha_i^{-1})\Big](2\pi)^{M/2}|\mathbf{\Sigma}|^{1/2}\Big|_{\mathbf{w}=\mathbf{w}^*} \\
&= \Big[\prod_{n=1}^{N} p(t_n|x_n,\mathbf{w})\Big]\cdot\mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A})\cdot(2\pi)^{M/2}|\mathbf{\Sigma}|^{1/2}\Big|_{\mathbf{w}=\mathbf{w}^*}
\end{aligned}$$

We further take logarithm for both sides.

$$\begin{aligned}
\ln p(\mathbf{t}|\alpha) &= \Big[\sum_{n=1}^{N} \ln p(t_n|x_n,\mathbf{w}) + \ln\mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A}) + \frac{M}{2}\ln 2\pi + \frac{1}{2}\ln|\mathbf{\Sigma}|\Big]\Big|_{\mathbf{w}=\mathbf{w}^*} \\
&= \Big[\sum_{n=1}^{N}\big[t_n \ln y_n + (1-t_n)\ln(1-y_n)\big] - \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} - \frac{1}{2}\ln|\mathbf{A}| + \frac{1}{2}\ln|\mathbf{\Sigma}| + const\Big]\Big|_{\mathbf{w}=\mathbf{w}^*} \\
&= \Big[\sum_{n=1}^{N}\big[t_n \ln y_n + (1-t_n)\ln(1-y_n)\big] - \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}\Big] + \Big[\frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\ln|\mathbf{A}| + const\Big]\Big|_{\mathbf{w}=\mathbf{w}^*}
\end{aligned}$$

Using the Chain rule, we can obtain:

$$\frac{\partial \ln p(\mathbf{t}|\alpha)}{\partial \alpha_i}\Big|_{\mathbf{w}=\mathbf{w}^*} = \frac{\partial \ln p(\mathbf{t}|\alpha)}{\partial \mathbf{w}} \frac{\partial \mathbf{w}}{\partial \alpha_i}\Big|_{\mathbf{w}=\mathbf{w}^*}$$

Observing Eq (7.109), (7.110) and that (7.110) will equal 0 at $\mathbf{w}^*$, we can conclude that the first term on the right hand side of $\ln p(\mathbf{t}|\alpha)$ will have zero derivative with respect to $\mathbf{w}$ at $\mathbf{w}^*$. Therefore, we only need to focus on the second term:

$$\frac{\partial \ln p(\mathbf{t}|\alpha)}{\partial \alpha_i}\Big|_{\mathbf{w}=\mathbf{w}^*} = \frac{\partial}{\partial \alpha_i}\left[\frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\ln|\mathbf{A}|\right]\Big|_{\mathbf{w}=\mathbf{w}^*}$$

It is rather easy to obtain:

$$\frac{\partial}{\partial \alpha_i}[-\frac{1}{2}\ln|\mathbf{A}|] = -\frac{1}{2}\frac{\partial}{\partial \alpha_i}[\sum_i \ln \alpha_i^{-1}] = \frac{1}{2\alpha_i}$$

Then we follow the same procedure as in Prob.7.12, we can obtain:

$$\frac{\partial}{\partial \alpha_i}[\frac{1}{2}\ln|\mathbf{\Sigma}|] = -\frac{1}{2}\Sigma_{ii}$$

Therefore, we obtain:

$$\frac{\partial \ln p(\mathbf{t}|\alpha)}{\partial \alpha_i} = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii}$$

Note: here I draw a different conclusion as the main text. I have also verified my result in another way. You can write the prior as the product of $\mathcal{N}(w_i|0,\alpha_i^{-1})$ instead of $\mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A})$. In this form, since we know that:

$$\frac{\partial}{\partial \alpha_i}\sum_{i=1}^{M} \ln \mathcal{N}(w_i|0,\alpha_i^{-1}) = \frac{\partial}{\partial \alpha_i}(\frac{1}{2}\ln \alpha_i - \frac{\alpha_i}{2}w_i^2) = \frac{1}{2\alpha_i} - \frac{1}{2}(w_i^*)^2$$

The above expression can be used to replace the derivative of $-1/2\mathbf{w}^T\mathbf{A}\mathbf{w} - 1/2\ln|\mathbf{A}|$. Since the derivative of the likelihood with respect to α_i is not zero at $\mathbf{w}^*$, (7.115) seems not right anyway.

0.8 Graphical Models

Problem 8.1 Solution

We are required to prove:

$$\int_{\mathbf{x}} p(\mathbf{x})d\mathbf{x} = \int_{\mathbf{x}} \prod_{k=1}^{K} p(x_k|pa_k)d\mathbf{x} = 1$$

only depends on only one node (except the root), i.e., its parent. Thus we can easily change a undirected tree to a directed one by matching the potential function with the corresponding conditional PDF, as shown in the example.

Moreover, we can choose any node in the undirected tree to be the root and then work outwards to obtain a directed tree. Therefore, in an undirected tree with n nodes, there is n corresponding directed trees in total.

Problem 8.19-8.29 Solution (Waiting for update)

I am quite confused by the deduction in Eq(8.66). I do not understand the sum-prodcut algorithm and the max-sum algorithm very well.

0.9 Mixture Models and EM

Problem 9.1 Solution

For each r_{nk} when n is fixed and $k = 1,2,...,K$, only one of them equals 1 and others are all 0. Therefore, there are K possible choices. When N data are given, there are K^N possible assignments for $\{r_{nk}; n = 1,2,...,N; k = 1,2,...,K\}$. For each assignments, the optimal $\{\boldsymbol{\mu}_k; k = 1,2,...,K\}$ are well determined by Eq (9.4).

As discussed in the main text, by iteratively performing E-step and M-step, the distortion measure in Eq (9.1) is gradually minimized. The worst case is that we find the optimal assignment and $\{\boldsymbol{\mu}_k\}$ in the last iteration. In other words, K^N iterations are required. However, it is guaranteed to converge because the assignments are finite and the optimal $\{\boldsymbol{\mu}_k\}$ is determined once the assignment is given.

Problem 9.2 Solution

By analogy to Eq (9.1), we can write down:

$$J_N = J_{N-1} + \sum_{k=1}^{K} r_{Nk} ||\mathbf{x}_N - \boldsymbol{\mu}_k||^2$$

In the E-step, we still assign the N-th data $\mathbf{x}_N$ to the closet center and suppose that this cloest center is $\boldsymbol{\mu}_m$. Therefore, the expression above will reduce to:

$$J_N = J_{N-1} + ||\mathbf{x}_n - \boldsymbol{\mu}_m||^2$$

In the M-step, we set the derivative of J_N with respect to $\boldsymbol{\mu}_k$ to 0, where $k = 1,2,...,K$. We can observe that for those $\boldsymbol{\mu}_k$, $k \neq m$, we have:

$$\frac{\partial J_N}{\partial \boldsymbol{\mu}_k} = \frac{\partial J_{N-1}}{\partial \boldsymbol{\mu}_k}$$

In other words, we will only update $\boldsymbol{\mu}_m$ in the M-step by setting the derivative of J_N equal to 0. Utilizing Eq (9.4), we can obtain:

$$\begin{aligned}
\boldsymbol{\mu}_m^{(N)} &= \frac{\sum_{n=1}^{N-1} r_{nk}\mathbf{x}_n + \mathbf{x}_N}{\sum_{n=1}^{N-1} r_{nk} + 1} \\
&= \frac{\frac{\sum_{n=1}^{N-1} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N-1} r_{nk}} + \frac{\mathbf{x}_N}{\sum_{n=1}^{N-1} r_{nk}}}{1 + \frac{1}{\sum_{n=1}^{N-1} r_{nk}}} \\
&= \frac{\boldsymbol{\mu}_m^{(N-1)} + \frac{\mathbf{x}_N}{\sum_{n=1}^{N-1} r_{nk}}}{1 + \frac{1}{\sum_{n=1}^{N-1} r_{nk}}} \\
&= \boldsymbol{\mu}_m^{(N-1)} + \frac{\frac{\mathbf{x}_N}{\sum_{n=1}^{N-1} r_{nk}} - \frac{\boldsymbol{\mu}_m^{(N-1)}}{\sum_{n=1}^{N-1} r_{nk}}}{1 + \frac{1}{\sum_{n=1}^{N-1} r_{nk}}} \\
&= \boldsymbol{\mu}_m^{(N-1)} + \frac{\mathbf{x}_N - \boldsymbol{\mu}_m^{(N-1)}}{1 + \sum_{n=1}^{N-1} r_{nk}}
\end{aligned}$$

So far we have obtained a sequential on-line update formula just as required.

Problem 9.3 Solution

We simply follow the hint.

$$\begin{aligned}
p(\mathbf{x}) &= \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) \\
&= \sum_{\mathbf{z}} \prod_{k=1}^{K} \Big[(\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\Big]^{z_k}
\end{aligned}$$

Note that we have used 1-of-K coding scheme for $\mathbf{z} = [z_1, z_2, ..., z_K]^T$. To be more specific, only one of $z_1, z_2, ..., z_K$ will be 1 and all others will equal 0. Therefore, the summation over $\mathbf{z}$ actually consists of K terms and the k-th term corresponds to z_k equal to 1 and others 0. Moreover, for the k-th term, the product will reduce to $\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Therefore, we can obtain:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} \prod_{k=1}^{K} \Big[(\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\Big]^{z_k} = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Just as required.

Problem 9.4 Solution

According to Bayes' Theorem, we can write:

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Taking logarithm on both sides, we can write:

$$\ln p(\boldsymbol{\theta}|\mathbf{X}) \propto \ln p(\mathbf{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$$

Further utilizing Eq (9.29), we can obtain:

$$\begin{aligned}\ln p(\boldsymbol{\theta}|\mathbf{X}) &\propto \ln\Big\{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\Big\} + \ln p(\boldsymbol{\theta}) \\ &= \ln\Big\{\big[\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\big]\cdot p(\boldsymbol{\theta})\Big\} \\ &= \ln\Big\{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})\Big\}\end{aligned}$$

In other words, in thise case, the only modification is that the term $p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$ in Eq (9.29) will be replaced by $p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Therefore, in the E-step, we still need to calculate the posterior $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})$ and then in the M-step, we are required to maximize $Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$. In this case, by analogy to Eq (9.30), we can write down $Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$:

$$\begin{aligned}Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) &= \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})\ln\Big[p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})\Big] \\ &= \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})\Big[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})\Big] \\ &= \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})\ln p(\boldsymbol{\theta}) \\ &= \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})\cdot\sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old}) \\ &= \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \\ &= Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) + \ln p(\boldsymbol{\theta})\end{aligned}$$

Just as required.

Problem 9.5 Solution

Notice that the condition on $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$ can be omitted here, and we only need to prove $p(\mathbf{Z}|\mathbf{X})$ can be written as the product of $p(\mathbf{z}_n|\mathbf{x}_n)$. Correspondingly, the small dots representing $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$ can also be omitted in Fig 9.6. Observing Fig 9.6 and based on definition, we can write :

$$p(\mathbf{X},\mathbf{Z}) = p(\mathbf{x}_1,\mathbf{z}_1)p(\mathbf{z}_1)...p(\mathbf{x}_N,\mathbf{z}_N)p(\mathbf{z}_N) = p(\mathbf{x}_1,\mathbf{z}_1)...p(\mathbf{x}_N,\mathbf{z}_N)$$

Moreover, since there is no link from $\mathbf{z}_m$ to $\mathbf{z}_n$, from $\mathbf{x}_m$ to $\mathbf{x}_n$, and from $\mathbf{z}_m$ to $\mathbf{x}_n$ $(m \neq n)$, we can obtain:

$$p(\mathbf{Z}) = p(\mathbf{z}_1)...p(\mathbf{z}_N), \quad p(\mathbf{X}) = p(\mathbf{x}_1)...p(\mathbf{x}_N)$$

These can also be verified by calculating the marginal distribution from $p(\mathbf{X},\mathbf{Z})$, for example:

$$p(\mathbf{Z}) = \sum_{\mathbf{X}} p(\mathbf{X},\mathbf{Z}) = \sum_{\mathbf{x}_1,...,\mathbf{x}_N} p(\mathbf{x}_1,\mathbf{z}_1)...p(\mathbf{x}_N,\mathbf{z}_N) = p(\mathbf{z}_1)...p(\mathbf{z}_N)$$

According to Bayes' Theorem, we have

$$\begin{aligned} p(\mathbf{Z}|\mathbf{X}) &= \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{p(\mathbf{X})} \\ &= \frac{\left[\prod_{n=1}^N p(\mathbf{x}_n|\mathbf{z}_n)\right]\left[\prod_{n=1}^N p(\mathbf{z}_n)\right]}{\prod_{n=1}^N p(\mathbf{x}_n)} \\ &= \prod_{n=1}^N \frac{p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n)}{p(\mathbf{x}_n)} \\ &= \prod_{n=1}^N p(\mathbf{z}_n|\mathbf{x}_n) \end{aligned}$$

Just as required. The essence behind the problem is that in the directed graph, there are only links from $\mathbf{z}_n$ to $\mathbf{x}_n$. The deeper reason is that (i) the mixture model is given by Fig 9.4, and (ii) we assume the data $\{\mathbf{x}_n\}$ is i.i.d, and thus there is no link from $\mathbf{x}_m$ to $\mathbf{x}_n$.

Problem 9.6 Solution

By analogy to Eq (9.19), we calculate the derivative of Eq (9.14) with respect to $\boldsymbol{\Sigma}$:

$$\frac{\partial \ln p}{\partial \boldsymbol{\Sigma}} = \frac{\partial}{\partial \boldsymbol{\Sigma}}\{\sum_{n=1}^N \ln a_n\} = \sum_{n=1}^N \frac{1}{a_n}\frac{\partial a_n}{\partial \boldsymbol{\Sigma}} \tag{$*$}$$

Where we have defined:

$$a_n = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})$$

Recall that in Prob.2.34, we have proved:

$$\frac{\partial \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}_{nk}\boldsymbol{\Sigma}^{-1}$$

Where we have defined:

$$\mathbf{S}_{nk} = (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

Therefore, we can obtain:

$$
\begin{aligned}
\frac{\partial a_n}{\partial \mathbf{\Sigma}} &= \frac{\partial}{\partial \mathbf{\Sigma}}\Big\{\sum_{k=1}^{K}\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma})\Big\} \\
&= \sum_{k=1}^{K}\frac{\partial}{\partial \mathbf{\Sigma}}\Big\{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma})\Big\} \\
&= \sum_{k=1}^{K}\pi_k\frac{\partial}{\partial \mathbf{\Sigma}}\Big\{\exp\big[\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma})\big]\Big\} \\
&= \sum_{k=1}^{K}\pi_k\cdot\exp\big[\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma})\big]\cdot\frac{\partial}{\partial \mathbf{\Sigma}}\Big[\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma})\Big] \\
&= \sum_{k=1}^{K}\pi_k\cdot\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma})\cdot(-\frac{1}{2}\mathbf{\Sigma}^{-1}+\frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{S}_{nk}\mathbf{\Sigma}^{-1})
\end{aligned}
$$

Substitute the equation above into $(*)$, we can obtain:

$$
\begin{aligned}
\frac{\partial \ln p}{\partial \mathbf{\Sigma}} &= \sum_{n=1}^{N}\frac{1}{a_n}\frac{\partial a_n}{\partial \mathbf{\Sigma}} \\
&= \sum_{n=1}^{N}\frac{\sum_{k=1}^{K}\pi_k\cdot\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma})\cdot(-\frac{1}{2}\mathbf{\Sigma}^{-1}+\mathbf{\Sigma}^{-1}\mathbf{S}_{nk}\mathbf{\Sigma}^{-1})}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\mathbf{\Sigma})} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\cdot(-\frac{1}{2}\mathbf{\Sigma}^{-1}+\frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{S}_{nk}\mathbf{\Sigma}^{-1}) \\
&= -\frac{1}{2}\Big\{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\Big\}\mathbf{\Sigma}^{-1}+\frac{1}{2}\mathbf{\Sigma}^{-1}\Big\{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\mathbf{S}_{nk}\Big\}\mathbf{\Sigma}^{-1}
\end{aligned}
$$

If we set the derivative equal to 0, we can obtain:

$$
\mathbf{\Sigma} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\mathbf{S}_{nk}}{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})}
$$

Problem 9.7 Solution

We begin by calculating the derivative of Eq (9.36) with respect to $\boldsymbol{\mu}_k$:

$$
\begin{aligned}
\frac{\partial \ln p}{\partial \boldsymbol{\mu}_k} &= \frac{\partial}{\partial \boldsymbol{\mu}_k}\Big\{\sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\left[\ln\pi_k+\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma}_k)\right]\Big\} \\
&= \frac{\partial}{\partial \boldsymbol{\mu}_k}\Big\{\sum_{n=1}^{N}z_{nk}\left[\ln\pi_k+\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma}_k)\right]\Big\} \\
&= \sum_{n=1}^{N}\frac{\partial}{\partial \boldsymbol{\mu}_k}\Big\{z_{nk}\ \ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma}_k)\Big\} \\
&= \sum_{\mathbf{x}_n\in C_k}\frac{\partial}{\partial \boldsymbol{\mu}_k}\Big\{\ \ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma}_k)\Big\}
\end{aligned}
$$

Where we have used $\mathbf{x}_n \in C_k$ to represent the data point $\mathbf{x}_n$ which are assigned to the k-th cluster. Therefore, $\boldsymbol{\mu}_k$ is given by the mean of those $x_n \in C_k$ just as the case of a single Gaussian. It is exactly the same for the covariance. Next, we maximize Eq (9.36) with respect to π_k by enforcing a Lagrange multiplier:

$$L = \ln p + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

We calculate the derivative of L with respect to π_k and set it to 0:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{z_{nk}}{\pi_k} + \lambda = 0$$

We multiply both sides by π_k and sum over k making use of the constraint Eq (9.9), yielding $\lambda = -N$. Substituting it back into the expression, we can obtain:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk}$$

Just as required.

Problem 9.8 Solution

Since $\gamma(z_{nk})$ is fixed, the only dependency of Eq (9.40) on $\boldsymbol{\mu}_k$ occurs in the Gaussian, yielding:

$$\begin{aligned}
\frac{\partial \mathbb{E}_z[\ln p]}{\partial \boldsymbol{\mu}_k} &= \frac{\partial}{\partial \boldsymbol{\mu}_k}\Big\{\sum_{n=1}^{N} \gamma(z_{nk}) \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\Big\} \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \frac{\partial \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k} \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \Big[-\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big]
\end{aligned}$$

Setting the derivative equal to 0, we obtain exactly Eq (9.16), and consequently Eq (9.17) just as required. Note that there is a typo in Eq (9.16), $\boldsymbol{\Sigma}_k$ shoule be $\boldsymbol{\Sigma}_k^{-1}$.

Problem 9.9 Solution

We first calculate the derivative of Eq (9.40) with respect to $\boldsymbol{\Sigma}_k$:

$$\begin{aligned}
\frac{\partial \mathbb{E}_z}{\partial \boldsymbol{\Sigma}_k} &= \frac{\partial}{\partial \boldsymbol{\Sigma}_k}\Big\{\sum_{n=1}^{N} \gamma(z_{nk}) \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\Big\} \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \frac{\partial \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\Sigma}_k} \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \Big[-\frac{1}{2}\boldsymbol{\Sigma}_k^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_k^{-1}\mathbf{S}_{nk}\boldsymbol{\Sigma}_k^{-1}\Big]
\end{aligned}$$

As in Prob 9.6, we have defined:

$$\mathbf{S}_{nk} = (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

Setting the derivative equal to 0 and rearranging it, we obtain:

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\,\mathbf{S}_{nk}}{\sum_{n=1}^{N} \gamma(z_{nk})} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\,\mathbf{S}_{nk}}{N_k}$$

Where N_k is given by Eq (9.18). So now we have obtained Eq (9.19) just as required. Next to maximize Eq (9.40) with respect to π_k, we still need to introduce Lagrange multiplier to enforce the summation of pi_k over k equal to 1, as in Prob 9.7:

$$L = \mathbb{E}_z + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

We calculate the derivative of L with respect to π_k and set it to 0:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\gamma(z_{nk})}{\pi_k} + \lambda = 0$$

We multiply both sides by π_k and sum over k making use of the constraint Eq (9.9), yielding $\lambda = -N$ (you can see Eq (9.20)- Eq (9.22) for more details). Substituting it back into the expression, we can obtain:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk}) = \frac{N_k}{N}$$

Just as Eq (9.22).

Problem 9.10 Solution

According to the property of PDF, we know that:

$$p(\mathbf{x}_b|\mathbf{x}_a) = \frac{p(\mathbf{x}_a, \mathbf{x}_b)}{p(\mathbf{x}_a)} = \frac{p(\mathbf{x})}{p(\mathbf{x}_a)} = \sum_{k=1}^{K} \frac{\pi_k}{p(\mathbf{x}_a)} \cdot p(\mathbf{x}|k)$$

Note that here $p(\mathbf{x}_a)$ can be viewed as a normalization constant used to guarantee that the integration of $p(\mathbf{x}_b|\mathbf{x}_a)$ equal to 1. Moreover, similarly, we can also obtain:

$$p(\mathbf{x}_a|\mathbf{x}_b) = \sum_{k=1}^{K} \frac{\pi_k}{p(\mathbf{x}_b)} \cdot p(\mathbf{x}|k)$$

Problem 9.11 Solution

According to the problem description, the expectation, i.e., Eq(9.40), can now be written as:

$$\mathbb{E}_z[\ln p] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \Big\{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \epsilon \mathbf{I}) \Big\}$$

In the M-step, we are required to maximize the expression above with respect to $\boldsymbol{\mu}_k$ and π_k. In Prob.9.8, we have already proved that $\boldsymbol{\mu}_k$ should be given by Eq (9.17):

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \qquad (*)$$

Where N_k is given by Eq (9.18). Moreover, in this case, by analogy to Eq (9.16), $\gamma(z_{nk})$ is slightly different:

$$\gamma(z_{nk}) = \frac{\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\epsilon\mathbf{I})}{\sum_j \pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\epsilon\mathbf{I})}$$

When $\epsilon \to 0$, we can obtain:

$$\sum_j \pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\epsilon\mathbf{I}) \approx \pi_m\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m,\epsilon\mathbf{I}), \quad \text{where} \quad m = \text{argmin}_j||\mathbf{x}_n - \boldsymbol{\mu}_j||^2$$

To be more clear, the summation is dominated by the max of $\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\epsilon\mathbf{I})$, and this term is further determined by the exponent, i.e., $-||\mathbf{x}_n - \boldsymbol{\mu}_j||^2$. Therefore, $\gamma(z_{nk})$ is given by exactly Eq (9.2), i.e., we have $\gamma(z_{nk}) = r_{nk}$. Combining with $(*)$, we can obtain exactly Eq (9.4). Next, according to Prob.9.9, π_k is given by Eq(9.22):

$$\pi_k = \frac{N_k}{N} = \frac{\sum_{n=}^{N}\gamma(z_{nk})}{N} = \frac{r_{nk}}{N}$$

In other words, π_k equals the fraction of the data points assigned to the k-th cluster.

Problem 9.12 Solution

First we calculate the mean $\boldsymbol{\mu}_k$:

$$\begin{aligned}\boldsymbol{\mu}_k &= \int \mathbf{x}p(\mathbf{x})d\mathbf{x}\\ &= \int \mathbf{x}\sum_{k=1}^{K}\pi_k\, p(\mathbf{x}|k)d\mathbf{x}\\ &= \sum_{k=1}^{K}\pi_k \int \mathbf{x}\, p(\mathbf{x}|k)d\mathbf{x}\\ &= \sum_{k=1}^{K}\pi_k\, \boldsymbol{\mu}_k\end{aligned}$$

Then we deal with the covariance matrix. For an arbitrary random variable $\mathbf{x}$, according to Eq (2.63) we have:

$$\begin{aligned}\text{cov}[\mathbf{x}] &= \mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T]\\ &= \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T\end{aligned}$$

Since $\mathbb{E}[\mathbf{x}]$ is already obtained, we only need to solve $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$. First we only focus on the k-th component and rearrange the expression above, yielding:

$$\mathbb{E}_k[\mathbf{x}\mathbf{x}^T] = \mathrm{cov}_k[\mathbf{x}] + \mathbb{E}_k[\mathbf{x}]\mathbb{E}_k[\mathbf{x}]^T = \boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T$$

We further use Eq (2.62), yielding:

$$\begin{aligned}
\mathbb{E}[\mathbf{x}\mathbf{x}^T] &= \int \mathbf{x}\mathbf{x}^T \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}|k) \, d\mathbf{x} \\
&= \sum_{k=1}^{K} \pi_k \int \mathbf{x}\mathbf{x}^T \, p(\mathbf{x}|k) \, d\mathbf{x} \\
&= \sum_{k=1}^{K} \pi_k \, \mathbb{E}_k[\mathbf{x}\mathbf{x}^T] \\
&= \sum_{k=1}^{K} \pi_k \, (\boldsymbol{\mu}_k \boldsymbol{\mu}_k^T + \boldsymbol{\Sigma}_k)
\end{aligned}$$

Therefore, we obtain Eq (9.50) just as required.

Problem 9.13 Solution

First, let's make this problem more clear. In a mixture of Bernoulli distribution, whose complete-data log likelihood is given by Eq (9.54) and whose model parameters are π_k and $\boldsymbol{\mu}_k$. If we want to obtain those parameters, we can adopt EM algorithm. In the E-step, we calculate $\gamma(z_{nk})$ as shown in Eq (9.56). In the M-step, we update π_k and $\boldsymbol{\mu}_k$ according to Eq (9.59) and Eq (9.60), where N_k and $\bar{\mathbf{x}_k}$ are defined in Eq (9.57) and Eq (9.58). Now let's back to this problem. The expectation of $\mathbf{x}$ is given by Eq (9.49):

$$\mathbb{E}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k^{(opt)} \boldsymbol{\mu}_k^{(opt)}$$

Here $\pi_k^{(opt)}$ and $\boldsymbol{\mu}_k^{(opt)}$ are the parameters obtained when EM is converged.

Using Eq (9.58) and Eq(9.59), we can obtain:

$$
\begin{aligned}
\mathbb{E}[\mathbf{x}] &= \sum_{k=1}^{K} \pi_k^{(opt)} \boldsymbol{\mu}_k^{(opt)} \\
&= \sum_{k=1}^{K} \pi_k^{(opt)} \frac{1}{N_K^{(opt)}} \sum_{n=1}^{N} \gamma(z_{nk})^{(opt)} \mathbf{x}_n \\
&= \sum_{k=1}^{K} \frac{N_k^{(opt)}}{N} \frac{1}{N_K^{(opt)}} \sum_{n=1}^{N} \gamma(z_{nk})^{(opt)} \mathbf{x}_n \\
&= \sum_{k=1}^{K} \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk})^{(opt)} \mathbf{x}_n \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\gamma(z_{nk})^{(opt)} \mathbf{x}_n}{N} \\
&= \sum_{n=1}^{N} \frac{\mathbf{x}_n}{N} \sum_{k=1}^{K} \gamma(z_{nk})^{(opt)} \\
&= \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n = \bar{\mathbf{x}}
\end{aligned}
$$

If we set all $\boldsymbol{\mu}_k$ equal to $\widehat{\boldsymbol{\mu}}$ in initialization, in the first E-step, we can obtain:

$$
\gamma(z_{nk})^{(1)} = \frac{\pi_k^{(0)} p(\mathbf{x}_n | \boldsymbol{\mu}_k = \widehat{\boldsymbol{\mu}})}{\sum_{j=1}^{K} \pi_j^{(0)} p(\mathbf{x}_n | \boldsymbol{\mu}_j = \widehat{\boldsymbol{\mu}})} = \frac{\pi_k^{(0)}}{\sum_{j=1}^{K} \pi_j^{(0)}} = \pi_k^{(0)}
$$

Note that here $\widehat{\boldsymbol{\mu}}$ and $\pi_k^{(0)}$ are the initial values. In the subsequent M-step, according to Eq (9.57)-(9.60), we can obtain:

$$
\boldsymbol{\mu}_k^{(1)} = \frac{1}{N_k^{(1)}} \sum_{n=1}^{N} \gamma(z_{nk})^{(1)} \mathbf{x}_n = \frac{\sum_{n=1}^{N} \gamma(z_{nk})^{(1)} \mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})^{(1)}} = \frac{\sum_{n=1}^{N} \pi_k^{(0)} \mathbf{x}_n}{\sum_{n=1}^{N} \pi_k^{(0)}} = \frac{\sum_{n=1}^{N} \mathbf{x}_n}{N}
$$

And

$$
\pi_k^{(1)} = \frac{N_k^{(1)}}{N} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})^{(1)}}{N} = \frac{\sum_{n=1}^{N} \pi_k^{(0)}}{N} = \pi_k^{(0)}
$$

In other words, in this case, after the first EM iteration, we find that the new $\boldsymbol{\mu}_k^{(1)}$ are all identical, which are all given by $\bar{\mathbf{x}}$. Moreover, the new $\pi_k^{(1)}$ are identical to their corresponding initial value $\pi_k^{(0)}$. Therefore, in the second EM iteration, we can similarly conclude that:

$$
\boldsymbol{\mu}_k^{(2)} = \boldsymbol{\mu}_k^{(1)} = \bar{\mathbf{x}}\,, \quad \pi_k^{(2)} = \pi_k^{(1)} = \pi_k^{(0)}
$$

In other words, the EM algorithm actually stops after the first iteration.

Problem 9.14 Solution

Let's follow the hint.

$$\begin{aligned}
p(\mathbf{x},\mathbf{z}|\boldsymbol{\mu},\pi) &= p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu})\cdot p(\mathbf{z}|\pi) \\
&= \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k)^{z_k}\cdot\prod_{k=1}^{K}\pi_k^{z_k} \\
&= \prod_{k=1}^{K}\Big[\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)\Big]^{z_k}
\end{aligned}$$

Then we marginalize over $\mathbf{z}$, yielding:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z}|\boldsymbol{\mu},\pi) = \sum_{\mathbf{z}}\prod_{k=1}^{K}\Big[\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)\Big]^{z_k}$$

The summation over $\mathbf{z}$ is made up of K terms and the k-th term corresponds to $z_k = 1$ and other z_j, where $j \neq k$, equals 0. Therefore, the k-th term will simply reduce to $\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)$. Hence, performing the summation over $\mathbf{z}$ will finally give Eq (9.47) just as required. To be more clear, we summarize the aforementioned statement:

$$\begin{aligned}
p(\mathbf{x}|\boldsymbol{\mu}) &= \sum_{\mathbf{z}}\prod_{k=1}^{K}\Big[\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)\Big]^{z_k} \\
&= \prod_{k=1}^{K}\Big[\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)\Big]^{z_k}\Big|_{z_1=1} + \ldots + \prod_{k=1}^{K}\Big[\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)\Big]^{z_k}\Big|_{z_K=1} \\
&= \pi_1 p(\mathbf{x}|\boldsymbol{\mu}_1) + \ldots + \pi_K p(\mathbf{x}|\boldsymbol{\mu}_K) \\
&= \sum_{k=1}^{K}\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)
\end{aligned}$$

Problem 9.15 Solution

Noticing that π_k doesn't depend on any μ_{ki}, we can omit the first term in the open brace when calculating the derivative of Eq (9.55) with respect to μ_{ki}:

$$\begin{aligned}
\frac{\partial \mathbb{E}_z[\ln p]}{\partial \mu_{ki}} &= \frac{\partial}{\partial \mu_{ki}}\sum_{n=1}^{N}\sum_{k=1}^{K}\Big\{\gamma(z_{nk})\sum_{i=1}^{D}\big[x_{ni}\ln\mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\big]\Big\} \\
&= \frac{\partial}{\partial \mu_{ki}}\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{i=1}^{D}\Big\{\gamma(z_{nk})\big[x_{ni}\ln\mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\big]\Big\} \\
&= \sum_{n=1}^{N}\frac{\partial}{\partial \mu_{ki}}\Big\{\gamma(z_{nk})\big[x_{ni}\ln\mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\big]\Big\} \\
&= \sum_{n=1}^{N}\gamma(z_{nk})\left(\frac{x_{ni}}{\mu_{ki}} - \frac{1-x_{ni}}{1-\mu_{ki}}\right) \\
&= \sum_{n=1}^{N}\gamma(z_{nk})\frac{x_{ni}-\mu_{ki}}{\mu_{ki}(1-\mu_{ki})}
\end{aligned}$$

Setting the derivative equal to 0, we can obtain:

$$\mu_{ki} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{N} \gamma(z_{nk})} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}$$

Where N_k is defined as Eq (9.57). If we group all the μ_{ki} as a column vector, i.e., $\boldsymbol{\mu}_k = [\mu_{k1}, \mu_{k2}, ..., \mu_{kD}]^T$, we will obtain Eq (9.59) just as required.

Problem 9.16 Solution

We follow the hint beginning by introducing a Lagrange multiplier:

$$L = \mathbb{E}_z[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})] + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

We calculate the derivative of L with respect to π_k and then set it equal to 0:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\gamma(z_{nk})}{\pi_k} + \lambda = 0 \qquad (*)$$

Here $\mathbb{E}_z[\ln p]$ is given by Eq (9.55). We first multiply both sides of the expression by π_k and then adopt summation with respect to k, which gives:

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) + \sum_{k=1}^{K} \lambda \pi_k = 0$$

Noticing that $\sum_{k=1}^{K} \pi_k$ equals 1, we can obtain:

$$\lambda = -\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})$$

Finally, substituting it back into $(*)$ and rearranging it, we can obtain:

$$\pi_k = -\frac{\sum_{k=1}^{K} \gamma(z_{nk})}{\lambda} = \frac{\sum_{k=1}^{K} \gamma(z_{nk})}{\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})} = \frac{N_k}{N}$$

Where N_k is defined by Eq (9.57) and N is the summation of N_k over k, and also equal to the number of data points.

Problem 9.17 Solution

The incomplete-data log likelihood is given by Eq (9.51), and $p(\mathbf{x}_n|\boldsymbol{\mu}_k)$ lies in the interval [0, 1], which can be easily verified by its definition, i.e., Eq (9.44). Therefore, we can obtain:

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \{ \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k) \} \leq \sum_{n=1}^{N} \ln \{ \sum_{k=1}^{K} \pi_k \times 1 \} \leq \sum_{n=1}^{N} \ln 1 = 0$$

Where we have used the fact that the logarithm is monotonic increasing, and that the summation of π_k over k equals 1. Moreover, if we want to achieve the equality, we need $p(\mathbf{x}_n|\boldsymbol{\mu}_k)$ equal to 1 for all $n = 1,2,...,N$. However, this is hardly possible.

To illustrate this, suppose that $p(\mathbf{x}_n|\boldsymbol{\mu}_k)$ equals 1 for all data points. Without loss of generality, consider two data points $\mathbf{x}_1 = [x_{11}, x_{12}, ..., x_{1D}]^T$ and $\mathbf{x}_2 = [x_{21}, x_{22}, ..., x_{2D}]^T$, whose i-th entries are different. We further assume $x_{1i} = 1$ and $x_{2i} = 0$ since x_i is a binary variable. According to Eq (9.44), if we want $p(\mathbf{x}_1|\boldsymbol{\mu}_k) = 1$, we must have $\mu_i = 1$ (otherwise it muse be less than 1). However, this will lead $p(\mathbf{x}_2|\boldsymbol{\mu}_k)$ equal to 0 since there is a term $1-\mu_i = 0$ in the product shown in Eq (9.44).

Therefore, when the data set is pathological, we will achieve this singularity point by adopting EM. Note that in the main text, the author states that the condition should be pathological initialization. This is also true. For instance, in the extreme case, when the data set is not pathological, if we initialize one π_k equal to 1 and others all 0, and some of μ_i to 1 and others 0, we may also achieve the singularity.

Problem 9.18 Solution

In Prob.9.4, we have proved that if we want to maximize the posterior by EM, the only modification is that in the M-step, we need to maximize $Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}) = Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta})$. Here $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$ has already been given by $\mathbb{E}_z[\ln p]$, i.e., Eq (9.55). Therefore, we derive for $\ln p(\boldsymbol{\theta})$. Note that $\ln p(\boldsymbol{\theta})$ is made up of two parts:(i) the prior for $\boldsymbol{\mu}_k$ and (ii) the prior for $\boldsymbol{\pi}$, we begin by dealing with the first part. Here we assume the Beta prior for μ_{ki}, where k is fixed, is the same, i.e.,:

$$p(\mu_{ki}|a_k,b_k) = \frac{\Gamma(a_k+b_k)}{\Gamma(a_k)\Gamma(b_k)}\mu_{ki}^{a_k-1}(1-\mu_{ki})^{b_k-1}, \quad i = 1,2,...,D$$

Therefore, the contribution of this Beta prior to $\ln p(\boldsymbol{\theta})$ should be given by:

$$\sum_{k=1}^{K}\sum_{i=1}^{D}(a_i-1)\ln\mu_{ki} + (b_i-1)\ln(1-\mu_{ki})$$

One thing worthy mentioned is that since we will maximize $Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$ with respect to $\boldsymbol{\pi},\boldsymbol{\mu}_k$, we can omit the terms which do not depend on $\boldsymbol{\pi},\boldsymbol{\mu}_k$, such as $\Gamma(a_k+b_k)/\Gamma(a_k)\Gamma(b_k)$. Then we deal with the second part. According to Eq (2.38), we can obtain:

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)...\Gamma(\alpha_K)}\prod_{k=1}^{K}\pi_k^{\alpha_k-1}$$

Therefore, the contribution of the Dirichlet prior to $\ln p(\boldsymbol{\theta})$ should be given by:

$$\sum_{k=1}^{K}(\alpha_k-1)\ln\pi_k$$

Therefore, now $Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$ can be written as:

$$Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})=\mathbb{E}_z[\ln p]+\sum_{k=1}^{K}\sum_{i=1}^{D}\Big[(a_i-1)\ln\mu_{ki}+(b_i-1)\ln(1-\mu_{ki})\Big]+\sum_{k=1}^{K}(\alpha_k-1)\ln\pi_k$$

Similarly, we calculate the derivative of $Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$ with respect to μ_{ki}. This can be simplified by reusing the deduction in Prob.9.15:

$$\begin{aligned}\frac{\partial Q^{'}}{\partial\mu_{ki}} &= \frac{\partial\mathbb{E}_z[\ln p]}{\partial\mu_{ki}}+\frac{a_i-1}{\mu_{ki}}-\frac{b_i-1}{1-\mu_{ki}}\\ &= \sum_{n=1}^{N}\gamma(z_{nk})(\frac{x_{ni}}{\mu_{ki}}-\frac{1-x_{ni}}{1-\mu_{ki}})+\frac{a_i-1}{\mu_{ki}}-\frac{b_i-1}{1-\mu_{ki}}\\ &= \frac{\sum_{n=1}^{N}x_{ni}\cdot\gamma(z_{nk})+a_i-1}{\mu_{ki}}-\frac{\sum_{n=1}^{N}(1-x_{ni})\gamma(z_{nk})+b_i-1}{1-\mu_{ki}}\\ &= \frac{N_k\bar{x}_{ki}+a_i-1}{\mu_{ki}}-\frac{N_k-N_k\bar{x}_{ki}+b_i-1}{1-\mu_{ki}}\end{aligned}$$

Note that here $\bar{x}_{ki}$ is defined as the i-th entry of $\bar{x}_k$ defined in Eq (9.58). To be more clear, we have used Eq (9.57) and Eq (9.58) in the last step:

$$\sum_{n=1}^{N}x_{ni}\cdot\gamma(z_{nk})=N_k\cdot\Big[\frac{1}{N_k}\sum_{n=1}^{N}x_{ni}\cdot\gamma(z_{nk})\Big]=N_k\cdot\bar{x}_{ki}$$

Setting the derivative equal to 0 and rearranging it, we can obtain:

$$\mu_{ki}=\frac{N_k\bar{x}_{ki}+a_i-1}{N_k+a_i-1+b_i-1}$$

Next we maximize $Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$ with respect to $\boldsymbol{\pi}$. By analogy to Prob.9.16, we introduce Lagrange multiplier:

$$L\propto\mathbb{E}_z+\sum_{k=1}^{K}(\alpha_k-1)\ln\pi_k+\lambda(\sum_{k=1}^{K}\pi_k-1)$$

Note that the second term on the right hand side of $Q^{'}$ in its definition has been omitted, since that term can be viewed as a constant with regard to $\boldsymbol{\pi}$. We then calculate the derivative of L with respect to π_k by taking advantage of Prob.9.16:

$$\frac{\partial L}{\partial\pi_k}=\sum_{n=1}^{N}\frac{\gamma(z_{nk})}{\pi_k}+\frac{\alpha_k-1}{\pi_k}+\lambda=0$$

Similarly, We first multiply both sides of the expression by π_k and then adopt summation with respect to k, which gives:

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma(z_{nk})+\sum_{k=1}^{K}(\alpha_k-1)+\sum_{k=1}^{K}\lambda\pi_k=0$$

Noticing that $\sum_{k=1}^{K} \pi_k$ equals 1, we can obtain:

$$\lambda = -\sum_{k=1}^{K} N_k - \sum_{k=1}^{K} (\alpha_k - 1) = -N - \alpha_0 + K$$

Here we have used Eq (2.39). Substituting it back into the derivative, we can obtain:

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) + \alpha_k - 1}{-\lambda} = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}$$

It is not difficult to show that if N is large, the update formula for $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$ in this case (MAP), will reduce to the results given in the main text (MLE).

Problem 9.19 Solution

We first introduce a latent variable $\mathbf{z} = [z_1, z_2, ..., z_K]^T$, only one of which equals 1 and others all 0. The conditional distribution of $\mathbf{x}$ is given by:

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k)^{z_k}$$

The distribution of the latent variable is given by:

$$p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

If we follow the same procedure as in Prob.9.14, we can show that Eq (9.84) holds. In other words, the introduction of the latent variable is valid. Therefore, according to Bayes' Theorem, we can obtain:

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \prod_{n=1}^{N} p(\mathbf{z}_n|\boldsymbol{\pi}) p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k p(\mathbf{x}|\boldsymbol{\mu})\right]^{z_{nk}}$$

We further use Eq (9.85), which gives:

$$\begin{aligned} \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi}) &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \left[\pi_k \prod_{d=1}^{D} \prod_{j=1}^{M} \mu_{kij}^{x_{nij}}\right] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[\ln \pi_k + \sum_{d=1}^{D} \sum_{j=1}^{M} x_{nij} \ln \mu_{kij}\right] \end{aligned}$$

Similarly, in the E-step, the responsibilities are evaluated using Bayes' theorem, which gives:

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)}$$

Next, in the M-step, we are required to maximize $\mathbb{E}_z[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\pi})]$ with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\mu}_k$, where $\mathbb{E}_z[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\pi})]$ is given by:

$$\mathbb{E}_z[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\Big[\ln\pi_k + \sum_{i=1}^{D}\sum_{j=1}^{M}x_{nij}\ln\mu_{kij}\Big]$$

Notice that there exists two constraints: (i) the summation of π_k over k equals 1, and (ii) the summation of μ_{kij} over j equals 1 for any k and i, we need to introduce Lagrange multiplier:

$$L = \mathbb{E}_z[\ln p] + \lambda(\sum_{k=1}^{K}\pi_k - 1) + \sum_{k=1}^{K}\sum_{i=1}^{D}\eta_{ki}(\sum_{j=1}^{M}\mu_{kij} - 1)$$

First we maximize L with respect to π_k. This is actually identical to the case in the main text. To be more clear, we calculate the derivative of L with respect to π_k:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N}\frac{\gamma(z_{nk})}{\pi_k} + \lambda$$

As in Prob.9.16, we can obtain:

$$\pi_k = \frac{N_k}{N}$$

Where N_k is defined as:

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

N is the summation of N_k over k, and also equals the number of data points. Then we calculate the derivative of L with respect to μ_{kij}:

$$\frac{\partial L}{\partial \mu_{kij}} = \sum_{n=1}^{N}\frac{\gamma(z_{nk})x_{nij}}{\mu_{kij}} + \eta_{ki}$$

We set it to 0 and multiply both sides by μ_{kij}, which gives:

$$\sum_{n=1}^{N}\gamma(z_{nk})x_{nij} + \eta_{ki}\mu_{kij} = 0$$

By analogy to deriving π_k, an intuitive idea is to perform summation for the above expression over j and hence we can use the constraint $\sum_j \mu_{kij} = 1$.

$$\eta_{ki} = -\sum_{j=1}^{M}\sum_{n=1}^{N}\gamma(z_{nk})x_{nij} = -\sum_{n=1}^{N}\gamma(z_{nk})\Big[\sum_{j=1}^{M}x_{nij}\Big] = -\sum_{n=1}^{N}\gamma(z_{nk}) = -N_k$$

Where we have used the fact that $\sum_j x_{nij} = 1$. Substituting back into the derivative, we can obtain:

$$\mu_{kij} = -\frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{nij}}{\eta_{ki}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_{nij}$$

Problem 9.20 Solution

We first calculate the derivative of Eq (9.62) with respect to α and set it to 0:

$$\frac{\partial E[\ln p]}{\partial \alpha} = \frac{M}{2} \frac{1}{2\pi} \frac{2\pi}{\alpha} - \frac{\mathbb{E}[\mathbf{w}^T \mathbf{w}]}{2} = 0$$

We rearrange the equation above, which gives:

$$\alpha = \frac{M}{\mathbb{E}[\mathbf{w}^T \mathbf{w}]} \qquad (*)$$

Therefore, we now need to calculate the expectation $\mathbb{E}[\mathbf{w}^T \mathbf{w}]$. Notice that the posterior has already been given by Eq (3.49):

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N)$$

To calculate $\mathbb{E}[\mathbf{w}^T \mathbf{w}]$, here we write down an property for a Gaussian random variable: if $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$, we have:

$$\mathbb{E}[\mathbf{x}^T \mathbf{A} \mathbf{x}] = \mathrm{Tr}[\mathbf{A}\mathbf{\Sigma}] + \mathbf{m}^T \mathbf{A} \mathbf{m}$$

This property has been shown in Eq(378) in 'the Matrix Cookbook'. Utilizing this property, we can obtain:

$$\mathbb{E}[\mathbf{w}^T \mathbf{w}] = \mathrm{Tr}[\mathbf{S}_N] + \mathbf{m}_N^T \mathbf{m}_N$$

Substituting it back into $(*)$, we obtain what is required.

Problem 9.21 Solution

We calculate the derivative of Eq (9.62) with respect to β and set it equal to 0:

$$\frac{\partial \ln p}{\partial \beta} = \frac{N}{2} \frac{1}{2\pi} \frac{2\pi}{\beta} - \frac{1}{2} \sum_{n=1}^{N} \mathbb{E}[(t_n - \mathbf{w}^T \boldsymbol{\phi}_n)^2] = 0$$

Rearranging it, we obtain:

$$\beta = \frac{N}{\sum_{n=1}^{N} \mathbb{E}[(t_n - \mathbf{w}^T \boldsymbol{\phi}_n)^2]}$$

Therefore, we are required to calculate the expectation. To be more clear, this expectation is with respect to the posterior defined by Eq (3.49):

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N)$$

We expand the expectation:

$$\begin{aligned}
\mathbb{E}[(t_n - \mathbf{w}^T \boldsymbol{\phi}_n)^2] &= \mathbb{E}[t_n^2 - 2t_n \cdot \mathbf{w}^T \boldsymbol{\phi}_n + \mathbf{w}^T \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \mathbf{w}] \\
&= \mathbb{E}[t_n^2] - \mathbb{E}[2t_n \cdot \mathbf{w}^T \boldsymbol{\phi}_n] + \mathbb{E}[\mathbf{w}^T (\boldsymbol{\phi}_n \boldsymbol{\phi}_n^T) \mathbf{w}] \\
&= t_n^2 - 2t_n \cdot \mathbb{E}[\boldsymbol{\phi}_n^T \mathbf{w}] + \mathrm{Tr}[\boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \mathbf{S}_N] + \mathbf{m}_N^T \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \mathbf{m}_N \\
&= t_n^2 - 2t_n \boldsymbol{\phi}_n^T \cdot \mathbb{E}[\mathbf{w}] + \mathrm{Tr}[\boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \mathbf{S}_N] + \mathbf{m}_N^T \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \mathbf{m}_N \\
&= t_n^2 - 2t_n \boldsymbol{\phi}_n^T \mathbf{m}_N + \mathrm{Tr}[\boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \mathbf{S}_N] + \mathbf{m}_N^T \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \mathbf{m}_N \\
&= (t_n - \mathbf{m}_N^T \boldsymbol{\phi}_N)^2 + \mathrm{Tr}[\boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \mathbf{S}_N]
\end{aligned}$$

Substituting it back into the derivative, we can obtain:

$$\begin{aligned}
\frac{1}{\beta} &= \frac{1}{N} \sum_{n=1}^{N} \left\{ (t_n - \mathbf{m}_N^T \boldsymbol{\phi}_N)^2 + \mathrm{Tr}[\boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \mathbf{S}_N] \right\} \\
&= \frac{1}{N} \left\{ ||\mathbf{t} - \boldsymbol{\Phi} \mathbf{m}_N||^2 + \mathrm{Tr}[\boldsymbol{\Phi}^T \boldsymbol{\Phi} \mathbf{S}_N] \right\}
\end{aligned}$$

Note that in the last step, we have performed vectorization. Here the j-th row of $\boldsymbol{\Phi}$ is given by $\boldsymbol{\phi}_j$, identical to the definition given in Chapter 3.

Problem 9.22 Solution

First let's expand the complete-data log likelihood using Eq (7.79), Eq (7.80) and Eq (7.76).

$$\begin{aligned}
\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) &= \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) + \ln p(\mathbf{w}|\boldsymbol{\alpha}) \\
&= \sum_{n=1}^{N} \ln p(t_n | x_n, \mathbf{w}, \beta^{-1}) + \sum_{i=1}^{M} \ln \mathcal{N}(w_i | 0, \alpha_i^{-1}) \\
&= \sum_{n=1}^{N} \ln \mathcal{N}(t_n | \mathbf{w}^T \boldsymbol{\phi}_n, \beta^{-1}) + \sum_{i=1}^{M} \ln \mathcal{N}(w_i | 0, \alpha_i^{-1}) \\
&= \frac{N}{2} \ln \frac{\beta}{2\pi} - \frac{\beta}{2} \sum_{n=1}^{N} (t_n - \mathbf{w}^T \boldsymbol{\phi}_n)^2 + \frac{1}{2} \sum_{i=1}^{M} \ln \frac{\alpha_i}{2\pi} - \sum_{i=1}^{M} \frac{\alpha_i}{2} w_i^2
\end{aligned}$$

Therefore, the expectation of the complete-data log likelihood with respect to the posterior of $\mathbf{w}$ equals:

$$\mathbb{E}_{\mathbf{w}}[\ln p] = \frac{N}{2} \ln \frac{\beta}{2\pi} - \frac{\beta}{2} \sum_{n=1}^{N} \mathbb{E}_{\mathbf{w}}[(t_n - \mathbf{w}^T \boldsymbol{\phi}_n)^2] + \frac{1}{2} \sum_{i=1}^{M} \ln \frac{\alpha_i}{2\pi} - \sum_{i=1}^{M} \frac{\alpha_i}{2} \mathbb{E}_{\mathbf{w}}[w_i^2]$$

We calculate the derivative of $\mathbb{E}_{\mathbf{w}}[\ln p]$ with respect to α_i and set it to 0:

$$\frac{\partial \mathbb{E}_{\mathbf{w}}[\ln p]}{\partial \alpha_i} = \frac{1}{2} \frac{1}{2\pi} \frac{2\pi}{\alpha_i} - \frac{1}{2} \mathbb{E}_{\mathbf{w}}[w_i^2] = 0$$

Rearranging it, we can obtain:

$$\alpha_i = \frac{1}{\mathbb{E}_{\mathbf{w}}[w_i^2]} = \frac{1}{\mathbb{E}_{\mathbf{w}}[\mathbf{w}\mathbf{w}^T]_{(i,i)}}$$

Here the subscript (i,i) represents the entry on the i-th row and i-th column of the matrix $\mathbb{E}_{\mathbf{w}}[\mathbf{w}\mathbf{w}^T]$. So now, we are required to calculate the expectation. To be more clear, this expectation is with respect to the posterior defined by Eq (7.81):

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\boldsymbol{\alpha},\beta) = \mathcal{N}(\mathbf{m},\boldsymbol{\Sigma})$$

Here we use Eq (377) described in 'the Matrix Cookbook'. We restate it here: if $\mathbf{w} \sim \mathcal{N}(\mathbf{m},\boldsymbol{\Sigma})$, we have:

$$\mathbb{E}[\mathbf{w}\mathbf{w}^T] = \boldsymbol{\Sigma} + \mathbf{m}\mathbf{m}^T$$

According to this equation, we can obtain:

$$\alpha_i = \frac{1}{\mathbb{E}_{\mathbf{w}}[\mathbf{w}\mathbf{w}^T]_{(i,i)}} = \frac{1}{(\boldsymbol{\Sigma} + \mathbf{m}\mathbf{m}^T)_{(i,i)}} = \frac{1}{\Sigma_{ii} + m_i^2}$$

Now We calculate the derivative of $\mathbb{E}_{\mathbf{w}}[\ln p]$ with respect to β and set it to 0:

$$\frac{\partial \mathbb{E}_{\mathbf{w}}[\ln p]}{\partial \beta} = \frac{N}{2}\frac{1}{2\pi}\frac{2\pi}{\beta} - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{w}}[(t_n - \mathbf{w}^T\boldsymbol{\phi}_n)^2] = 0$$

Rearranging it, we obtain:

$$\beta^{(new)} = \frac{N}{\sum_{n=1}^{N}\mathbb{E}_{\mathbf{w}}[(t_n - \mathbf{w}^T\boldsymbol{\phi}_n)^2]}$$

Therefore, we are required to calculate the expectation. By analogy to the deduction in Prob.9.21, we can obtain:

$$\begin{aligned}\frac{1}{\beta^{(new)}} &= \frac{1}{N}\sum_{n=1}^{N}\Big\{(t_n - \mathbf{m}^T\boldsymbol{\phi}_N)^2 + \mathrm{Tr}[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\boldsymbol{\Sigma}]\Big\} \\ &= \frac{1}{N}\Big\{||\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}||^2 + \mathrm{Tr}[\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}]\Big\}\end{aligned}$$

To make it consistent with Eq (9.68), let's first prove a statement:

$$(\beta^{-1}\mathbf{A} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})\boldsymbol{\Sigma} = \beta^{-1}\mathbf{I}$$

This can be easily shown by substituting $\boldsymbol{\Sigma}$, i.e., Eq(7.83), back into the expression:

$$(\beta^{-1}\mathbf{A} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})\,\boldsymbol{\Sigma} = (\beta^{-1}\mathbf{A} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})\,(\mathbf{A} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1} = \beta^{-1}\mathbf{I}$$

Now we start from this statement and rearrange it, which gives:

$$\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\Sigma} = \beta^{-1}\mathbf{I} - \beta^{-1}\mathbf{A}\boldsymbol{\Sigma} = \beta^{-1}(\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma})$$

Substituting back into the expression for $\beta^{(new)}$:

$$\begin{aligned}\frac{1}{\beta^{(new)}} &= \frac{1}{N}\Big\{||\mathbf{t}-\mathbf{\Phi m}||^2+\mathrm{Tr}[\mathbf{\Phi}^T\mathbf{\Phi\Sigma}]\Big\} \\ &= \frac{1}{N}\Big\{||\mathbf{t}-\mathbf{\Phi m}||^2+\mathrm{Tr}[\beta^{-1}(\mathbf{I}-\mathbf{A\Sigma})]\Big\} \\ &= \frac{1}{N}\Big\{||\mathbf{t}-\mathbf{\Phi m}||^2+\beta^{-1}\mathrm{Tr}[\mathbf{I}-\mathbf{A\Sigma}]\Big\} \\ &= \frac{1}{N}\Big\{||\mathbf{t}-\mathbf{\Phi m}||^2+\beta^{-1}\sum_i(1-\alpha_i\Sigma_{ii})\Big\} \\ &= \frac{||\mathbf{t}-\mathbf{\Phi m}||^2+\beta^{-1}\sum_i\gamma_i}{N}\end{aligned}$$

Here we have defined $\gamma_i = 1-\alpha_i\Sigma_{ii}$ as in Eq (7.89). Note that there is a typo in Eq (9.68), $\mathbf{m}_N$ should be $\mathbf{m}$.

Problem 9.23 Solution

Some clarifications must be made here, Eq (7.87)-(7.88) only gives the same stationary points, i.e., the same $\alpha^\star$ and $\beta^\star$, as those given by Eq (9.67)-(9.68). However, the hyper-parameters estimated at some specific iteration may not be the same by those two different methods.

When convergence is reached, Eq (7.87) can be written as:

$$\alpha^\star = \frac{1-\alpha^\star\Sigma_{ii}}{m_i^2}$$

Rearranging it, we can obtain:

$$\alpha^\star = \frac{1}{m_i^2+\Sigma_{ii}}$$

This is identical to Eq (9.67). When convergence is reached, Eq (9.68) can be written as:

$$(\beta^\star)^{-1} = \frac{||\mathbf{t}-\mathbf{\Phi m}||^2+(\beta^\star)^{-1}\sum_i\gamma_i}{N}$$

Rearranging it, we can obtain:

$$(\beta^\star)^{-1} = \frac{||\mathbf{t}-\mathbf{\Phi m}||^2}{N-\sum_i\gamma_i}$$

This is identical to Eq (7.88).

Problem 9.24 Solution

We substitute Eq (9.71) and Eq (9.72) into Eq (9.70):

$$\begin{aligned}
L(q,\boldsymbol{\theta}) + \mathrm{KL}(q||p) &= \sum_{\mathbf{Z}} q(\mathbf{Z})\Big\{\ln\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} - \ln\frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})}\Big\} \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z})\Big\{\ln\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}\Big\} \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z})\ \ln p(\mathbf{X}|\boldsymbol{\theta}) \\
&= \ln p(\mathbf{X}|\boldsymbol{\theta})
\end{aligned}$$

Note that in the last step, we have used the fact that $\ln p(\mathbf{X}|\boldsymbol{\theta})$ doesn't depend on $\mathbf{Z}$, and that the summation of $q(\mathbf{Z})$ over $\mathbf{Z}$ equal to 1 because $q(\mathbf{Z})$ is a PDF.

Problem 9.25 Solution

We calculate the derivative of Eq (9.71) with respect to $\boldsymbol{\theta}$, given $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})$:

$$\begin{aligned}
\frac{\partial L(q,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} &= \frac{\partial}{\partial\boldsymbol{\theta}}\Big\{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})\ln\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})}\Big\} \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\Big\{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})\ln p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})\Big\} \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\Big\{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\Big\} \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})\frac{\partial \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})\frac{1}{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}\frac{\partial p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})\frac{1}{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}\frac{\partial p(\mathbf{X}|\boldsymbol{\theta})\cdot p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \\
&= \sum_{\mathbf{Z}} \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})}{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}\Big[p(\mathbf{X}|\boldsymbol{\theta})\frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} + p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})\frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big]
\end{aligned}$$

We evaluate this derivative at $\boldsymbol{\theta} = \boldsymbol{\theta}^{\text{old}}$:

$$\begin{aligned}
\frac{\partial L(q,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{\text{old}}} &= \Big\{ \sum_{\mathbf{Z}} \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})})}{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})} \Big[p(\mathbf{X}|\boldsymbol{\theta}) \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) \frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big] \Big\}\Big|_{\boldsymbol{\theta}^{\text{old}}} \\
&= \sum_{\mathbf{Z}} \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})})}{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}^{(\text{old})})} \Big[p(\mathbf{X}|\boldsymbol{\theta}^{(\text{old})}) \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})}) \frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \Big] \\
&= \sum_{\mathbf{Z}} \frac{1}{p(\mathbf{X}|\boldsymbol{\theta}^{(\text{old})})} \Big[p(\mathbf{X}|\boldsymbol{\theta}^{(\text{old})}) \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})}) \frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \Big] \\
&= \sum_{\mathbf{Z}} \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + \sum_{\mathbf{Z}} \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})})}{p(\mathbf{X}|\boldsymbol{\theta}^{(\text{old})})} \cdot \frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \\
&= \sum_{\mathbf{Z}} \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + \frac{1}{p(\mathbf{X}|\boldsymbol{\theta}^{(\text{old})})} \cdot \frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \\
&= \sum_{\mathbf{Z}} \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \\
&= \Big\{ \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) \Big\}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \\
&= \frac{\partial 1}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \\
&= \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}}
\end{aligned}$$

This problem can be much easier to prove if we view it from the perspective of KL divergence. Note that when $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})})$, the KL divergence vanishes, and that in general KL divergence is less or equal to zero. Therefore, we must have:

$$\frac{\partial KL(q||p)}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} = 0$$

Otherwise, there exists a point $\boldsymbol{\theta}$ in the neighborhood near $\boldsymbol{\theta}^{(\text{old})}$ which leads the KL divergence less than 0. Then using Eq (9.70), it is trivial to prove.

Problem 9.26 Solution

From Eq (9.18), we have:

$$N_k^{\text{old}} = \sum_n \gamma^{\text{old}}(z_{nk})$$

If now we just re-evaluate the responsibilities for one data point $\mathbf{x}_m$, we can obtain:

$$\begin{aligned}
N_k^{\text{new}} &= \sum_{n \neq m} \gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk}) \\
&= \sum_n \gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \\
&= N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})
\end{aligned}$$

Similarly, according to Eq (9.17), we can obtain:

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k^{\text{new}}}\sum_{n\neq m}\gamma^{\text{old}}(z_{nk})\mathbf{x}_n + \frac{\gamma^{\text{new}}(z_{mk})\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \frac{1}{N_k^{\text{new}}}\sum_{n}\gamma^{\text{old}}(z_{nk})\mathbf{x}_n + \frac{\gamma^{\text{new}}(z_{mk})\mathbf{x}_m}{N_k^{\text{new}}} - \frac{\gamma^{\text{old}}(z_{mk})\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \frac{N_k^{\text{old}}}{N_k^{\text{new}}}\frac{1}{N_k^{\text{old}}}\sum_{n}\gamma^{\text{old}}(z_{nk})\mathbf{x}_n + \frac{\gamma^{\text{new}}(z_{mk})\mathbf{x}_m}{N_k^{\text{new}}} - \frac{\gamma^{\text{old}}(z_{mk})\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \frac{N_k^{\text{old}}}{N_k^{\text{new}}}\boldsymbol{\mu}_k^{\text{old}} + \Big[\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})\Big]\frac{\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \boldsymbol{\mu}_k^{\text{old}} - \frac{N_k^{\text{new}} - N_k^{\text{old}}}{N_k^{\text{new}}}\boldsymbol{\mu}_k^{\text{old}} + \Big[\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})\Big]\frac{\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \boldsymbol{\mu}_k^{\text{old}} - \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}}\boldsymbol{\mu}_k^{\text{old}} + \Big[\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})\Big]\frac{\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \boldsymbol{\mu}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}}\cdot\Big(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}\Big)
\end{aligned}$$

Just as required.

Problem 9.27 Solution

By analogy to the previous problem, we use Eq (9.24)-Eq(9.27), beginning by first deriving an update formula for mixing coefficients π_k:

$$\begin{aligned}
\pi_k^{\text{new}} &= \frac{N_k^{\text{new}}}{N} = \frac{1}{N}\Big\{N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})\Big\} \\
&= \pi_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N}
\end{aligned}$$

Here we have used the conclusion (the update formula for N_k^{new}) in the previous problem. Next we deal with the covariance matrix $\boldsymbol{\Sigma}$. By analogy to the previous problem, we can obtain:

$$\begin{aligned}
\boldsymbol{\Sigma}_k^{new} &= \frac{1}{N_k^{\text{new}}}\sum_{n\neq m}\gamma^{\text{old}}(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}})^T \\
&\quad +\frac{1}{N_k^{\text{new}}}\gamma^{\text{new}}(z_{mk})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})^T \\
&\approx \frac{1}{N_k^{\text{new}}}\sum_{n\neq m}\gamma^{\text{old}}(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{old}})^T \\
&\quad +\frac{1}{N_k^{\text{new}}}\gamma^{\text{new}}(z_{mk})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})^T \\
&= \frac{1}{N_k^{\text{new}}}\sum_{n}\gamma^{\text{old}}(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{old}})^T \\
&\quad +\frac{1}{N_k^{\text{new}}}\gamma^{\text{new}}(z_{mk})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})^T \\
&\quad -\frac{1}{N_k^{\text{new}}}\gamma^{\text{old}}(z_{mk})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{old}})^T \\
&= \frac{1}{N_k^{\text{new}}}N_k^{\text{old}}\boldsymbol{\Sigma}_k^{\text{old}}+\frac{1}{N_k^{\text{new}}}\gamma^{\text{new}}(z_{mk})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})^T \\
&\quad -\frac{1}{N_k^{\text{new}}}\gamma^{\text{old}}(z_{mk})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{old}})^T \\
&= \left\{1+\frac{N_k^{\text{old}}-N_k^{\text{new}}}{N_k^{\text{new}}}\right\}\boldsymbol{\Sigma}_k^{\text{old}} \\
&\quad +\frac{1}{N_k^{\text{new}}}\gamma^{\text{new}}(z_{mk})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})^T \\
&\quad -\frac{1}{N_k^{\text{new}}}\gamma^{\text{old}}(z_{mk})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{old}})^T \\
&= \left\{1+\frac{\gamma^{\text{old}}(z_{mk})-\gamma^{\text{new}}(z_{mk})}{N_k^{\text{new}}}\right\}\boldsymbol{\Sigma}_k^{\text{old}} \\
&\quad +\frac{\gamma^{\text{new}}(z_{mk})}{N_k^{\text{new}}}(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})^T \\
&\quad -\frac{\gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}}(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{old}})^T \\
&= \boldsymbol{\Sigma}_k^{\text{old}} \\
&\quad +\frac{\gamma^{\text{new}}(z_{mk})}{N_k^{\text{new}}}\left\{(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{new}})^T-\boldsymbol{\Sigma}^{\text{old}}\right\} \\
&\quad -\frac{\gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}}\left\{(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_m-\boldsymbol{\mu}_k^{\text{old}})^T-\boldsymbol{\Sigma}_k^{\text{old}}\right\}
\end{aligned}$$

One important thing worthy mentioned is that in the second step, there is an approximate equal sign. Note that in the previous problem, we have shown that if we only recompute the data point $\mathbf{x}_m$, all the center $\boldsymbol{\mu}_k$ will also change from $\boldsymbol{\mu}_k^{\text{old}}$ to $\boldsymbol{\mu}_k^{\text{new}}$, and the update formula is given by Eq (9.78). However, for the convenience of computing, we have made an approximation here. Other approximation methods can also be applied here. For instance, you can replace $\boldsymbol{\mu}_k^{\text{new}}$ with $\boldsymbol{\mu}_k^{\text{old}}$ whenever it occurs.

The complete solution should be given by substituting Eq (9.78) into the right side of the first equal sign and then rearranging it, in order to construct a relation between $\boldsymbol{\Sigma}_k^{\text{new}}$ and $\boldsymbol{\Sigma}_k^{\text{old}}$. However, this is too complicated.

0.10 Variational Inference

Problem 10.1 Solution

This problem is very similar to Prob.9.24. We substitute Eq (10.3) and Eq (10.4) into Eq (10.2):

$$\begin{aligned} L(q) + \text{KL}(q||p) &= \int_{\mathbf{Z}} q(\mathbf{Z}) \Big\{ \ln \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} - \ln \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \Big\} d\mathbf{Z} \\ &= \int_{\mathbf{Z}} q(\mathbf{Z}) \Big\{ \ln \frac{p(\mathbf{X},\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} \Big\} d\mathbf{Z} \\ &= \int_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}) d\mathbf{Z} \\ &= \ln p(\mathbf{X}) \end{aligned}$$

Note that in the last step, we have used the fact that $\ln p(\mathbf{X})$ doesn't depend on $\mathbf{Z}$, and that the integration of $q(\mathbf{Z})$ over $\mathbf{Z}$ equal to 1 because $q(\mathbf{Z})$ is a PDF.

Problem 10.2 Solution

To be more clear, we are required to solve:

$$\begin{cases} m_1 = \mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (m_2 - \mu_2) \\ m_2 = \mu_2 - \Lambda_{22}^{-1} \Lambda_{21} (m_1 - \mu_1) \end{cases}$$

To obtain the equation above, we need to substitute $\mathbb{E}[z_i] = m_i$, where $i = 1, 2$, into Eq (10.13) and Eq (10.14). Here the unknown parameters are m_1 and m_2. It is trivial to notice that $m_i = \mu_i$ is a solution for the equation above.

Let's solve this equation from another perspective. Firstly, if any (or both) of Λ_{11}^{-1} and Λ_{22}^{-1} equals 0, we can obtain $m_i = \mu_i$ directly from Eq (10.13)-(10.14). When none of Λ_{11}^{-1} and Λ_{22}^{-1} equals 0, we substitute m_1, i.e., the first