기계학습 Assignment2

2023 2학기

1 (20 points) 본 문항은 constrained optimization을 위한 Lagrangian multipler 에 관한 것이다. Bishop교재의 Appendix E를 중점적으로 참조하여 다음 물음에 답하시오.

(i) Equality constraint $g(\mathbf{x}) = 0$가 만족하는 다음 최적화 문제를 가정하자.

$$\begin{aligned} &\text{maximize } f(\mathbf{x}) \\ &\text{subject to } g(\mathbf{x}) = 0 \end{aligned} \tag{1}$$

표기를 단순화 하기 위해, 임의의 함수 $h(\mathbf{x})$의 gradient 벡터 ∇h를 다음과 같이 정의하자.

$$\nabla h = \frac{\partial h(\mathbf{x})}{\partial \mathbf{x}} \tag{2}$$

식 (1)의 최적화 문제의 solution을 $\mathbf{x}^*$라 할때, 다음과 같이 ∇f와 ∇g가 **서로 평행 (parallel)**이 되어야 함을 설명하시오[1].

$$\nabla f(\mathbf{x}^*) = \lambda \nabla g(\mathbf{x}^*) \tag{3}$$

여기서 $\lambda \in \mathbb{R}$로 임의의 실수이다.

(ii) Inequality constraint $g(\mathbf{x}) >= 0$를 갖는 다음 최적화 문제가 주어졌다고 하자.

$$\begin{aligned} &\text{maximize } f(\mathbf{x}) \\ &\text{subject to } g(\mathbf{x}) \geq 0 \end{aligned} \tag{4}$$

식 (4)의 최적화 문제의 solution이 $\mathbf{x}^*$이라고 하자. $\mathbf{x}^*$은 두 가지 case로 나뉘는데, 1) 제약 영역 경계에 위치하는 $g(\mathbf{x}^*) = 0$를 만족하는 *active* solution, 2) 제약 영역 안에 위치하는 $g(\mathbf{x}^*) > 0$를 만족하는 *inactive* soluion의 두 가지 경우로 나뉜다. 먼저, active 해인 $g(\mathbf{x}^*) = 0$인 경우에 다음과 같이 ∇f와 ∇g이 방향이 서로 **반대**이어야 함을 설명하시오.

$$\nabla f(\mathbf{x}) = -\lambda \nabla g(\mathbf{x}) \tag{5}$$

이 경우에는 방향이 반대임을 보장하기 위해 $\lambda > 0$이다.

다음으로, inactive 해인 $g(\mathbf{x}^*) > 0$인 경우에는 다음과 같이 unconstrained optimization문제와 equivalent함을 간단히 설명하시오.

$$\nabla f(\mathbf{x}) = 0 \tag{6}$$

(iii) 앞 문항의 식 (4)의 최적화 문제를 다음 식 (7)의 *Lagrangian function* 를 최적화하는 문제로 변환하고자 한다.

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}) \tag{7}$$

[1]두 벡터는 같은 방향이거나 반대 방향이 될 수 있다.

이때 다음 KKT (Karush-Kuhn-Tucker) 조건이 성립해야 함을 설명하시오.

$$\begin{aligned} g(\mathbf{x}) &\geq 0 \\ \lambda &\geq 0 \\ \lambda \cdot g(\mathbf{x}) &= 0 \end{aligned} \tag{8}$$

특히, 위의 KKT 조건이 앞에서 유도한 두 가지 경우에 대한 조건인 식 (5)과 식 (6)을 포함함을 보이시오.

2 (30 points) 본 문항에서는 **Lagrange dual problem**내용을 정리하고자 한다. m개의 inequality constraint가 있는 다음 최적화 문제를 가정하자.

$$\begin{aligned} &\text{maximize } f(\mathbf{x}) \\ &\text{subject to } g_i(\mathbf{x}) \geq 0 \text{ for } 1 \leq i \leq m \end{aligned} \tag{9}$$

위의 식을 **primal form**이라고 하고, 이를 **dual form**으로 변환하기 위해 다음 Lagrangian function을 사용한다.

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) \tag{10}$$

여기서 $\lambda_i \geq 0$, $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_m]^T$이다. 주어진 제약을 모두 만족하는 $\mathbf{x}$를 *feasible point* 라고 하고, 다음과 같이 $\mathcal{A}$를 모든 feasible point들이 집합으로 정의하자.

$$\mathcal{A} = \{\mathbf{x} | g_i(\mathbf{x}) \geq 0 \ \forall i \in \{1, \cdots, m\}\} \tag{11}$$

Bishop 교재외에 다른 자료도 참조하여 다음 물음에 답하시오.

(i) 정의에 따라, 다음과 같이 모든 feasible $\mathbf{x} \in \mathcal{A}$와 $\boldsymbol{\lambda}$에 대해서, $L(\mathbf{x}, \boldsymbol{\lambda})$는 $f(\mathbf{x})$의 upper bound가 된다.

$$L(\mathbf{x}, \boldsymbol{\lambda}) \geq f(\mathbf{x}) \text{ for } \mathbf{x} \in \mathcal{A} \tag{12}$$

식 (9)의 Primal form의 해(solution)을 $\mathbf{x}^*$라고 할때, 다음이 성립함을 보이시오.

$$f(\mathbf{x}^*) = \max_{\mathbf{x}} \min_{\boldsymbol{\lambda} \geq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\lambda}) \tag{13}$$

*Hint: $g_i(\mathbf{x}) \geq 0$를 만족하는 경우와 그렇지 않는 경우, 전자의 경우에서도 $g_i(\mathbf{x}) > 0$와 $g_i(\mathbf{x}) = 0$의 경우로 나누어 각 case별로 $\max_{\lambda_i} \lambda_i g_i(\mathbf{x})$의 값을 정리해보자.

(ii) Dual form에서는 다음 *Lagrangian dual function*을 사용한다.

$$\tilde{L}(\boldsymbol{\lambda}) = \max_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) \tag{14}$$

Primal form의 해(solution)을 $\mathbf{x}^* \in \mathcal{A}$라고 할때, 임의의 $\boldsymbol{\lambda} \geq \mathbf{0}$에 대해서, 다음식과 같이 dual function이 $f(\mathbf{x}^*)$의 upper bound가 됨을 보이시오.

$$\tilde{L}(\boldsymbol{\lambda}) \geq f(\mathbf{x}^*) \tag{15}$$

*Hint: $\mathbf{x} \in \mathcal{A}$에 대해서, 식 (12)과 식 (14)을 결합해보자.

(iii) Lagrangian dual problem (또는 dual form)은 $\tilde{L}(\boldsymbol{\lambda})$을 최적화 (최소화)하여 $f(\mathbf{x}^*)$의 best upper bound를 구하는 것으로 다음 최적화하는 문제를 가리킨다.

$$\begin{aligned} &\text{minimize } \tilde{L}(\boldsymbol{\lambda}) \\ &\text{subject to } \boldsymbol{\lambda} \geq \mathbf{0} \end{aligned} \tag{16}$$

식 (12)에 따라 마찬가지로 다음이 성립한다.

$$\min_{\boldsymbol{\lambda}} \tilde{L}(\boldsymbol{\lambda}) \geq f(\mathbf{x}^*) \tag{17}$$

Best upper bound와 최적해간의 차이를 **Duality gap**라고 하며, 다음과 같이 정의된다.

$$f(\mathbf{x}^*) - \min_{\boldsymbol{\lambda}} \tilde{L}(\boldsymbol{\lambda}) \tag{18}$$

일반적으로, 식 (18)의 duality gap은 nonzero이지만, **어떠한 최적화 문제의 경우에는 zero**가 되며, 이때 dual form과 primal form의 최적값은 서로 동일하다.
이러한 **zero duality gap**의 특징을 가지는 최적화 문제 유형에 대한 조건 중, 주어진 제약식에 대한 feasible point의 존재 유무를 가지고 판단하는 **Slater's condition**이 간단하면서도 잘 알려져있다. Slater's condition이란 무엇인가? Slater's condition을 만족하는 경우와 그렇지 않는 경우의 예시를 각각 제시하시오.

3 (20 points) 본 문항에서는 2번의 Lagrangian dual problem 변환 방법을 간단한 예시를 통해 적용해보고자 한다. 다음 각 constrained optimization 문제에 대해서 Lagrangian dual problem로 변환하고 해당 dual form상에서 최적화를 수행하시오.

(i) 아래 최적화 문제의 primal form 기술하고 dual form을 유도한 후에 해당 Lagrangian dual problem의 최적화를 수행하시오.

$$\begin{aligned} &\text{maximize } f(\mathbf{x}) = 1 - \|\mathbf{x}\|^2 \\ &\text{subject to } g(\mathbf{x}) = \mathbf{1}^T\mathbf{x} - 1 \geq 0 \end{aligned} \tag{19}$$

여기서 $\mathbf{x} \in \mathbb{R}^D$이다.

(ii) 앞문항과 동일하게 최적화를 수행 하시오.

$$\begin{aligned} &\text{maximize } f(\mathbf{x}) = 1 - \|\mathbf{x} - \mathbf{1}\|^2 \\ &\text{subject to } g(\mathbf{x}) = \mathbf{1}^T\mathbf{x} - 1 \geq 0 \end{aligned} \tag{20}$$

*Hint: KKT 조건도 활용하도록 하자. $D = 2$인 경우에 대해서 2D 좌표상에서 목적함수와 제약함수의의 위치를 확인해봐도 좋다.

(iii) 위의 최적화 문제가 Slater's condition을 만족하는 지 간단히 설명하시오.

4 (20 points) 본 문항은 Support vector machine (SVM)의 margin maximum solution에 관한 것이다.

Binary classification을 위한 학습데이터셋이 $\mathcal{D} = \{(\mathbf{x}_1, t_1), \cdots, (\mathbf{x}_N, t_N)\}$로 주어지고, 이때 $t_i \in \{-1, 1\}$로, $t_i = 1$인 경우는 positive샘플 $t_i = -1$인 경우는 negative샘플이라고 하자. Binary classification 위해 linear model을 사용한다고 가정한다.

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

학습 단계에서는 $\mathcal{D}$상의 모든 데이터들을 정확히 분류하도록 다음 조건을 만족하면서, parameters $\mathbf{w}$와 b를 결정해야 한다.

$$t_n y(\mathbf{x}_n) > 0 \ \forall n \in \{1, \cdots, N\} \tag{21}$$

위의 조건을 만족하는 $\mathbf{w}$와 b의 집합을 *feasible parameters* $\mathcal{A}_\theta$라고 하고, 구체적으로 $\mathcal{A}_\theta$는 다음과 같이 정의된다.

$$\mathcal{A}_\theta = \{(\mathbf{w}, b) | t_n y(\mathbf{x}_n) > 0 \ \forall n \in \{1, \cdots, N\}\} \tag{22}$$

다음 물음에 답하시오.

(i) $\mathbf{x}_n \in \mathbb{R}^D$ 에서 decision surface $y(\mathbf{x})$에 이르는 거리 (distance)가 $|y(\mathbf{x}_n)|/\|\mathbf{w}\|$임을 증명하시오. 또한, $(\mathbf{w}, b) \in \mathcal{A}_\theta$에 대해서, distance가 $t_n y(\mathbf{x}_n)/\|\mathbf{w}\|$이 됨을 간단히 보이시오.

(ii) SVM에서의 margin개념에 대해서 설명하고 generalization error를 감소시키는 측면에서 maximum margin 방법의 장점을 기술하시오.

(iii) Maximum margin solution이 다음식으로 정리됨을 설명하시오 (Bishop교재 Eq. (7.3)).

$$\arg \max_{(\mathbf{w}, b) \in \mathcal{A}_\theta} \left\{ \frac{1}{\|\mathbf{w}\|} \min_n \left[t_n y(\mathbf{x}_n) \right] \right\} \tag{23}$$

(iv) Bishop교재 Eq. (7.4)에서처럼, $\mathbf{x}_n$를 decision surface상에서 가장 가까운 points인 support vectors 라고 할때, 다음과 같이 $t_n y(\mathbf{x}_n)$를 constant value로 고정하는 이유를 기술하시오.

$$t_n y(\mathbf{x}_n) = 1 \tag{24}$$

5 (40 points) Linear SVM의 primal form은 다음과 같다.

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{subject to } t_n \left(\mathbf{w}^T \mathbf{x}_n + b \right) - 1 \geq 0 \ \forall n \in \{1, \cdots, N\} \end{aligned} \tag{25}$$

Design matrix $\mathbf{X}$와 design target vector $\mathbf{t}$는 아래와 같이 정의된다.

$$\begin{aligned} \mathbf{X} &= \begin{bmatrix} \mathbf{x}_1{}^T \\ \vdots \\ \mathbf{x}_N{}^T \end{bmatrix} \\ \mathbf{t} &= [t_1, \cdots, t_N]^T \end{aligned} \tag{26}$$

다음 물음에 답하시오.

(i) 위의 최적화 문제의 식 (25)에 대한 primal form이 아래와 같이 정리됨을 보이시오.

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - (\mathbf{a} \odot \mathbf{t})^T (\mathbf{X}\mathbf{w} + b\mathbf{1}) + \mathbf{a}^T \mathbf{1} \tag{27}$$

여기서 $\mathbf{a} = [a_1, \cdots, a_N]^T$ 로 Lagrangian multipliers이며 $\odot$는 element-wise multiplication로 $\mathbf{a} \odot \mathbf{t}$는 다음과 같다.

$$\mathbf{a} \odot \mathbf{t} = [a_1 \cdot t_1, \cdots, a_N \cdot t_N]^T$$

*Hint: minimize문제라는 것과, Lagrangian multiplier의 조건에 대해 유의할 것.

(ii) 식 (27)에 대한 $\mathbf{w}$와 b의 gradient를 유도하시오.

$$\frac{\partial L(\mathbf{w}, b, \mathbf{a})}{\partial \mathbf{w}} = \tag{28}$$

$$\frac{\partial L(\mathbf{w}, b, \mathbf{a})}{\partial b} = \tag{29}$$

$$\tag{30}$$

(iii) 앞문항의 gradients를 zero로 취하여 Bishop교재 Eq. (7.8)과 Eq. (7.9)의 다음을 유도하시오.

$$\begin{aligned} \mathbf{w} &= \mathbf{X}^T(\mathbf{a} \odot \mathbf{t}) \\ 0 &= \mathbf{a}^T\mathbf{t} \end{aligned} \tag{31}$$

(iv) $\mathbf{a}^T\mathbf{t} = 0$ 조건은 어떤 의미를 갖는지 간단히 기술하시오. *Hint: $t_i = 1$인 경우와 $t_i = -1$인 경우로 나누어서 해당 식을 재구성해보자.

(v) 식 (31)을 이용하여, Linear SVM의 dual form $\tilde{L}(\mathbf{a}) = \min_{(\mathbf{w},b)\in\mathcal{A}_\theta} L(\mathbf{w}, b, \mathbf{a})$이 다음과 같이 정리됨을 유도하시오.

$$\tilde{L}(\mathbf{a}) = -\frac{1}{2}(\mathbf{a} \odot \mathbf{t})^T \mathbf{X}\mathbf{X}^T (\mathbf{a} \odot \mathbf{t}) + \mathbf{a}^T\mathbf{1} \tag{32}$$

(vi) 위의 dual form은 *Feasible* parameters $(\mathbf{w}, b) \in \mathcal{A}_\theta$에 대한 KKT 조건에 따라 다음 식과 같이 Bishop교재 Eq. (7.14-16)의 추가로 만족되어야 한다.

$$\begin{aligned} a_n &\geq 0 \\ t_n y(\mathbf{x}_n) - 1 &\geq 0 \\ a_n\left(t_n y(\mathbf{x}_n) - 1\right) &= 0 \end{aligned} \tag{33}$$

support vectors들은 $a_n > 0$인 데이터 샘플들의 집합으로 다음과 같이 정의된다.

$$\mathcal{S} = \{\mathbf{x}_n | a_n > 0\} \tag{34}$$

Bishop교재 Eq. (7.13)를 참고하여, $y(\mathbf{x})$를 $\mathcal{S}$를 이용하여 다시 기술하시오. 이때 새로운 new example벡터를 분류하는 test단계에서 $y(\mathbf{x})$의 효율성에 대해서 간단히 서술하시오.

6 (30 points) Overlapped class분포상에서는 Linear SVM의 margin 제약에서 벗어나는 샘플들이 존재하며, 이를 위해 데이터별로 slack variable ξ_n을 도입하여, 다음과 같이 확장된 primal form을 고려하자 (Bishop교재 Eq. (7.20-21)).

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n \\ &\text{subject to } t_n\left(\mathbf{w}^T\mathbf{x}_n + b\right) - 1 + \xi_n \geq 0 \ \forall n \in \{1, \cdots, N\} \\ &\xi_n \geq 0 \ \forall n \in \{1, \cdots, N\} \end{aligned} \tag{35}$$

(i) 식 (35)에 대한 Lagrangian function이 다음과 같이 정리됨을 보이시오.

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \boldsymbol{\xi}^T\mathbf{1} - (\mathbf{a} \odot \mathbf{t})^T(\mathbf{X}\mathbf{w} + b\mathbf{1}) + \mathbf{a}^T\mathbf{1} - (\mathbf{a} + \boldsymbol{\mu})^T\boldsymbol{\xi} \tag{36}$$

여기서 $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_N]^T$, $\mathbf{a} = [a_1, \cdots, a_N]^T$와 $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_N]^T$로 Lagrangian multiplers이다.

(ii) 식 (36)에 대한 $\mathbf{w}$와 b, $\boldsymbol{\xi}$의 gradient를 유도하시오.

$$\frac{\partial L(\mathbf{w}, b, \mathbf{a})}{\partial \mathbf{w}} = \tag{37}$$

$$\frac{\partial L(\mathbf{w}, b, \mathbf{a})}{\partial b} = \tag{38}$$

$$\frac{\partial L(\mathbf{w}, b, \mathbf{a})}{\partial \boldsymbol{\xi}} = \tag{39}$$

$$\tag{40}$$

원래의 *기본* Linear SVM과 마찬가지로, 위의 gradients를 zero로 하여, 다음과 같이 Bishop교재 Eq. (7.29-31)가 됨을 유도하시오.

$$\begin{aligned} \mathbf{w} &= \mathbf{X}^T(\mathbf{a} \odot \mathbf{t}) \\ 0 &= \mathbf{a}^T\mathbf{t} \\ \mathbf{a} + \boldsymbol{\mu} &= C \cdot \mathbf{1} \end{aligned} \tag{41}$$

(iii) 식 (41)을 이용하여, *Class-overlapped 케이스*에 대한 Linear SVM의 dual form $\tilde{L}(\mathbf{a}) = \min_{(\mathbf{w},b)\in\mathcal{A}_\theta} L(\mathbf{w}, b, \mathbf{a})$이 *다음과 같이 기본 Linear SVM의 dual form과 동일함*을 증명하시오.

$$\tilde{L}(\mathbf{a}) = -\frac{1}{2}(\mathbf{a} \odot \mathbf{t})^T \mathbf{X}\mathbf{X}^T (\mathbf{a} \odot \mathbf{t}) + \mathbf{a}^T\mathbf{1} \tag{42}$$

다만 Bishop교재의 KKT 조건인 Eq. (7.23-28)로 부터, 기본 Linear SVM와의 유일한 차이는 다음과 같이 a_n에 대한 upper bound 조건만이 추가된다는 점을 증명하시오.

$$\mathbf{0} \le \mathbf{a} \le C \cdot \mathbf{1} \tag{43}$$

7 (30 points) K component로 구성된 Mixture of Gaussians에서 data sample은 *one-hot* latent variable $\mathbf{z}$를 이용하여 두 단계에서 생성 (generation) 과정을 따른다고 볼 수 있다.

$$\begin{aligned} \text{Sample } z_k &\sim P(z_k = 1|\boldsymbol{\pi}) = \pi_k \\ \text{Sample } \mathbf{x} &\sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned}$$

Latent variable $\mathbf{z}$을 통해, $p(\mathbf{x})$는 다음과 같이 $p(\mathbf{x}, \mathbf{z})$의 margizalization로 기술될 수 있다 (Bishop Eq. (9.10-12)).

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

학습데이터셋은 label이 없이 $\mathcal{D} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$로 주어지고, Design matrix $\mathbf{X}$와 마찬가지로, latent design matrix $\mathbf{Z}$는 아래로 정의된다고 하자.

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1{}^T \\ \vdots \\ \mathbf{z}_N{}^T \end{bmatrix}$$

교재의 Eq. (9.14)와 같이 데이터셋에 대한 log-likelihood $\ln p(\mathcal{D})$는 다음과 같다.

$$\ln p(\mathcal{D}) = \ln p(\mathbf{X}|\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \tag{44}$$

log-likelihood를 maximize하기 위한 EM algorithm 유도를 위해 다음 물음에 답하시오.

(i) $\mathbf{x}_n$에 대한 responsibilities $\gamma(z_{nk})$를 다음과 같이 정의하자.

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{45}$$

Log-likelihood 식 (44)의 $\boldsymbol{\mu}$에 대한 gradient인 $\partial \ln p(\mathcal{D})/\partial \boldsymbol{\mu}_k$ 를 zero로 둠으로써, 다음식을 유도하시오 (Bishop교재 Eq. (9.17)).

$$\begin{aligned} \boldsymbol{\mu}_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\ N_k &= \sum_{n=1}^{N} \gamma(z_{nk}) \end{aligned} \tag{46}$$

이때, 우변의 $\gamma(z_{nk}) = h(\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma})$는 $\boldsymbol{\mu}_k$가 포함되어 있으므로, optimal $\boldsymbol{\mu}_k$는 closed-form으로 구해지지 않고, *fixed-point algorithm*처럼 iterative하게 식 (46)을 반복적으로 수행하여 얻어져야 함을 설명하시오.

(ii) $\partial \ln p(\mathcal{D})/\partial \boldsymbol{\Sigma}_k^{-1} = 0$로 둠으로써, 다음식을 유도하시오.

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \tag{47}$$

*Hint: Assignment 1의 Problem 1.(v) 및 Problem 14의 전개에서 사용한 유도과정을 참조

(iii) Bishop교재 Eq. (9.20-21)의 과정을 통해, 다음식을 유도하시오.

$$\pi_k = \frac{N_k}{N} \tag{48}$$

(iv) 위의 유도내용을 바탕으로 Bishop교재의 Eq. (9.23-28) EM 알고리즘 과정을 서술하시오.

8 (15 points) General EM알고리즘과 Variational inference에서 기반이 되는 다음 식을 유도하시오 (Bishop교재의 Eq. (9.70)).

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

여기서 $q(\mathbf{Z})$는 $\mathbf{Z}$에 대한 임의의 variational 분포이다.

제출 방법

각 문항별 수식 유도 및 풀이 과정을 정리한 **문제 풀이 답안지** 파일을 email로 제출해야 한다. (수기 작성시에는 별도 용지의 스캔본 제출 보다는 전자펜으로 답안을 작성하여 생성된 파일로 제출하는 것을 권장함).