机器学习导论
综合能力测试

151242041, 王昊庭, hatsuyukiw@gmail.com

2017 年 6 月 17 日

1 [40pts] Exponential Families

指数分布族 (Exponential Families) 是一类在机器学习和统计中非常常见的分布族, 具有良好的性质。在后文不引起歧义的情况下, 简称为指数族。

指数分布族是一组具有如下形式概率密度函数的分布族群:

$$f_X(x|\theta) = h(x)\exp\left(\eta(\theta)\cdot T(x) - A(\theta)\right) \tag{1.1}$$

其中, $\eta(\theta)$, $A(\theta)$ 以及函数 $T(\cdot)$, $h(\cdot)$ 都是已知的。

(1) **[10pts]** 试证明多项分布 (Multinomial distribution) 属于指数分布族。

(2) **[10pts]** 试证明多元高斯分布 (Multivariate Gaussian distribution) 属于指数分布族。

(3) **[20pts]** 考虑样本集 $\mathcal{D} = \{x_1, \cdots, x_n\}$ 是从某个已知的指数族分布中独立同分布地 (i.i.d.) 采样得到, 即对于 $\forall i \in [1, n]$, 我们有 $f(x_i|\boldsymbol{\theta}) = h(x_i)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}T(x_i) - A(\boldsymbol{\theta})\right)$.

对参数 $\boldsymbol{\theta}$, 假设其服从如下先验分布：

$$p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi} - \nu A(\boldsymbol{\theta})\right) \tag{1.2}$$

其中, $\boldsymbol{\chi}$ 和 ν 是 $\boldsymbol{\theta}$ 生成模型的参数。请计算其后验, 并证明后验与先验具有相同的形式。(**Hint**: 上述又称为"共轭"(Conjugacy), 在贝叶斯建模中经常用到)

Solution. 1. Suppose the parameters are $p_1, \cdots, p_k$, and $\sum_{i=1}^k p_i = 1$. Let

$$\begin{aligned}
h(X) &= \frac{n!}{\prod_{i=1}^k x_i!}\,, \\
\eta(\boldsymbol{\theta}) &= [\log p_1, \cdots, \log p_k]\,, \\
T(X) &= X\,, \\
A(\boldsymbol{\theta}) &= 0\,,
\end{aligned}$$

and the resulting distribution is multinomial distribution.

2. Suppose the parameters are $\theta = [\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and the dimension of the variable is k. Let

$$
\begin{aligned}
h(X) &= (2\pi)^{-\frac{k}{2}}\,, \\
\eta(\boldsymbol{\theta}) &= [\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}]\,, \\
T(X) &= [X, XX^{\mathrm{T}}]\,, \\
A(\boldsymbol{\theta}) &= \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\log|\boldsymbol{\Sigma}|\,,
\end{aligned}
$$

and the resulting distribution is multivariate Gaussian distribution.

3. Let $X = \{x_1, \cdots, x_n\}$, and $g(\boldsymbol{\theta}) = \exp(-A(\boldsymbol{\theta}))$,

$$
\begin{aligned}
p(\boldsymbol{\theta}|X, \boldsymbol{\chi}, \nu) &\propto p(X|\boldsymbol{\theta}, \boldsymbol{\chi}, \nu)p(\boldsymbol{\theta}|\boldsymbol{\chi}, \nu) \\
&= \left(\prod_{i=1}^{n} h(x_i)\right) \exp\left(\boldsymbol{\theta} \cdot \sum_{i=1}^{n} T(x_i) - nA(\boldsymbol{\theta})\right) f(\boldsymbol{\chi}, \nu) \exp\left(\boldsymbol{\theta}\boldsymbol{\chi} - \nu A(\boldsymbol{\theta})\right) \\
&= \left(\prod_{i=1}^{n} h(x_i)\right) f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\theta})^{(\nu+n)} \exp\left(\boldsymbol{\theta} \cdot (\sum_{i=1}^{n} T(x_i) + \boldsymbol{\chi})\right) \\
&\propto g(\boldsymbol{\theta})^{(\nu+n)} \exp\left(\boldsymbol{\theta} \cdot (\sum_{i=1}^{n} T(x_i) + \boldsymbol{\chi})\right) \\
&= \exp\left(\boldsymbol{\theta} \cdot (\sum_{i=1}^{n} T(x_i) + \boldsymbol{\chi}) - (\nu + n)A(\boldsymbol{\theta})\right)\,.
\end{aligned}
$$

Thus,

$$
p(\boldsymbol{\theta}|X, \boldsymbol{\chi}, \nu) = p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi} + \sum_{i=1}^{n} T(x_i), \nu + n)\,.
$$

2 [40pts] Decision Boundary

考虑二分类问题, 特征空间 $X \in \mathcal{X} = \mathbb{R}^d$, 标记 $Y \in \mathcal{Y} = \{0, 1\}$. 我们对模型做如下生成式假设：

- Attribute conditional independence assumption: 对已知类别, 假设所有属性相互独立, 即每个属性特征独立地对分类结果发生影响；
- Bernoulli prior on label: 假设标记满足 Bernoulli 分布先验, 并记 $\Pr(Y = 1) = \pi$.

(1) **[20pts]** 假设 $P(X_i|Y)$ 服从指数族分布, 即

$$\Pr(X_i = x_i|Y = y) = h_i(x_i)\exp(\theta_{iy} \cdot T_i(x_i) - A_i(\theta_{iy}))$$

请计算后验概率分布 $\Pr(Y|X)$ 以及分类边界 $\{x \in \mathcal{X} : P(Y = 1|X = x) = P(Y = 0|X = x)\}$. (**Hint**: 你可以使用 sigmoid 函数 $\mathcal{S}(x) = 1/(1 + e^{-x})$ 进行化简最终的结果).

(2) **[20pts]** 假设 $P(X_i|Y = y)$ 服从高斯分布, 且记均值为 μ_{iy} 以及方差为 σ_i^2 (注意, 这里的方差与标记 Y 是独立的), 请证明分类边界与特征 X 是成线性的。

Solution. 1.

$$\begin{aligned}\frac{P(Y=1|X)}{P(Y=0|X)} &= \frac{P(X|Y=1)P(Y=1)}{P(X|Y=0)P(Y=0)} \\ &= \frac{\pi}{1-\pi}\exp\left(\sum_{i=1}^{d}\left((\theta_{i1}-\theta_{i0})\cdot T_i(x_i) - (A_i(\theta_{i1}) - A_i(\theta_{i0}))\right)\right).\end{aligned}$$

Thus,

$$P(Y=0|X) = \mathcal{S}\left(-\log\frac{\pi}{1-\pi} + \sum_{i=1}^{d}\left((\theta_{i0}-\theta_{i1})\cdot T_i(x_i) - (A_i(\theta_{i0}) - A_i(\theta_{i1}))\right)\right).$$

Let $P(Y=0|X) = \frac{1}{2}$, we have

$$\log\frac{\pi}{1-\pi} = \sum_{i=1}^{d}\left((\theta_{i0}-\theta_{i1})\cdot T_i(x_i) - A_i(\theta_{i0}) + A_i(\theta_{i1})\right),$$

which is the classification boundary.

2. Substitute the parameters in 1.(2) into the above classification boundary, and we have

$$\log\frac{\pi}{1-\pi} = \sum_{i=1}^{d}\left(\frac{x_i-\mu_{i1}}{2\sigma^2}\right)^2 - \sum_{i=1}^{d}\left(\frac{x_i-\mu_{i0}}{2\sigma^2}\right)^2.$$

Simplify the above equation, and we have

$$\log\frac{\pi}{1-\pi} = \sum_{i=1}^{d}\left(\frac{2(\mu_{i0}-\mu_{i1})x_i + \mu_{i1}^2 - \mu_{i0}^2}{2\sigma^2}\right),$$

which is linear in X.

3 [70pts] Theoretical Analysis of k-means Algorithm

给定样本集 $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, k-means 聚类算法希望获得簇划分 $\mathcal{C} = \{C_1, C_2, \cdots, C_k\}$, 使得最小化欧式距离

$$J(\gamma, \mu_1, \dots, \mu_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} ||\mathbf{x}_i - \mu_j||^2 \tag{3.1}$$

其中, $\mu_1, \dots, \mu_k$ 为 k 个簇的中心 (means), $\gamma \in \mathbb{R}^{n \times k}$ 为指示矩阵 (indicator matrix) 定义如下：若 $\mathbf{x}_i$ 属于第 j 个簇, 则 $\gamma_{ij} = 1$, 否则为 0.

则最经典的 k-means 聚类算法流程如算法1中所示 (与课本中描述稍有差别, 但实际上是等价的)。

Algorithm 1: k-means Algorithm

Initialize $\mu_1, \dots, \mu_k$.

repeat

Step 1: Decide the class memberships of $\{\mathbf{x}_i\}_{i=1}^n$ by assigning each of them to its nearest cluster center.

$$\gamma_{ij} = \begin{cases} 1, & ||\mathbf{x}_i - \mu_j||^2 \le ||\mathbf{x}_i - \mu_{j'}||^2, \forall j' \\ 0, & \text{otherwise} \end{cases}$$

Step 2: For each $j \in \{1, \cdots, k\}$, recompute μ_j using the updated γ to be the center of mass of all points in C_j:

$$\mu_j = \frac{\sum_{i=1}^n \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^n \gamma_{ij}}$$

until *the objective function J no longer changes*;

---

(1) **[10pts]** 试证明, 在算法1中, **Step 1** 和 **Step 2** 都会使目标函数 J 的值降低.

(2) **[10pts]** 试证明, 算法1会在有限步内停止。

(3) **[10pts]** 试证明, 目标函数 J 的最小值是关于 k 的非增函数, 其中 k 是聚类簇的数目。

(4) **[20pts]** 记 $\hat{\mathbf{x}}$ 为 n 个样本的中心点, 定义如下变量,

total deviation	$T(X) = \sum_{i=1}^n \|\mathbf{x}_i - \hat{\mathbf{x}}\|^2 / n$
intra-cluster deviation	$W_j(X) = \sum_{i=1}^n \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 / \sum_{i=1}^n \gamma_{ij}$
inter-cluster deviation	$B(X) = \sum_{j=1}^k \frac{\sum_{i=1}^n \gamma_{ij}}{n} \|\mu_j - \hat{\mathbf{x}}\|^2$

试探究以上三个变量之间有什么样的等式关系？基于此, 请证明, k-means 聚类算法可以认为是在最小化 intra-cluster deviation 的加权平均, 同时近似最大化 inter-cluster deviation.

(5) **[20pts]** 在公式(3.1)中, 我们使用 ℓ_2-范数来度量距离 (即欧式距离), 下面我们考虑使用 ℓ_1-范数来度量距离

$$J'(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} ||\mathbf{x}_i - \mu_j||_1 \tag{3.2}$$

- **[10pts]** 请仿效算法1(k-means-ℓ_2 算法), 给出新的算法 (命名为 k-means-ℓ_1 算法) 以优化公式3.2中的目标函数 J'.
- **[10pts]** 当样本集中存在少量异常点 (outliers) 时, 上述的 k-means-ℓ_2 和 k-means-ℓ_1 算法, 我们应该采用哪种算法？即, 哪个算法具有更好的鲁棒性？请说明理由。

Solution. (1) **Step 1**: Let $J_i = \sum_{j=1}^{k} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2$, then $J = \sum_{i=1}^{n} J_i$. If an instance $\mathbf{x}_i$ is not assigned to a new cluster, the contribution of $\mathbf{x}_i$, i.e., J_i is not changed. Suppose $\mathbf{x}$ was assigned to j' but reassigned to j during **Step 1**. The contribution J_i is decreased from $||\mathbf{x}_i - \mu_j||^2$ to $||\mathbf{x}_i - \mu_{j'}||^2$.

Step 2: From elementary geometry we know that the center of a point set minimizes the sum of the distances between it and the points in the set. **Step 2** minimizes the contribution of each cluster to the function J, thus it decreases J.

(2) There are only k^n possible assignments of points. For each assignment, **Step 2** guarantees that at the end of each loop, J is minimum with respect to the current assignment. Thus, at the end of each loop, J takes finitely many possible values. From (1), we know J is non-increasing, so after finitely many loops, J will no longer change, and the algorithm terminates.

(3) Suppose for $k = k_0 >= 1$, the minimum of J is achieved with γ and $\mu_j, j \in \{1, \cdots, k_0\}$. For $k = k_0 + 1$, Let $\gamma' = \gamma$, $\mu'_j = \mu_j, j \in \{1, \cdots, k_0\}$, and $\mu'_k = \mu_0$. It is obvious that with these assignments $J(k_0 + 1) = \min J(k_0)$. Thus we have $J(k_0 + 1) \leq J(k_0)$.

(4) Let $n_j = \sum_{i=1}^{n} \gamma_{ij}$. We have

$$\begin{aligned}
n_j W_j(X) + n_j \|\mu_j - \hat{\mathbf{x}}\|^2 &= \sum_{i=1}^{n} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 + n_j \|\mu_j - \hat{\mathbf{x}}\|^2 \\
&= \sum_{i=1}^{n} \gamma_{ij} \left(\|\mathbf{x}_i - \mu_j\|^2 + \|\mu_j - \hat{\mathbf{x}}\|^2 \right) \\
&= \sum_{i=1}^{n} \gamma_{ij} \|\mathbf{x}_i - \hat{\mathbf{x}}\|^2
\end{aligned}$$

Sum the above equation for all j, and we have

$$\frac{\sum_{j=1}^{k} n_j W_j(X)}{n} + B(X) = T(X) \,.$$

As $T(X)$ is a constant, and $\frac{\sum_{j=1}^{k} n_j W_j(X)}{n}$ is $\frac{J}{n}$, we are minimizing the weighted average of intra-cluster deviation and maximizing inter-cluster deviation.

(5) We need to revise both **Step 1** and **Step 2**. **Step 1** should consider ℓ_1 distance. Similarly, **Step 2** should update each cluster center to the point that minimizes the sum of ℓ_1 distances.

Algorithm 2: k-means Algorithm with ℓ_1 Distance

1 Initialize $\mu_1, \ldots, \mu_k$.

2 **repeat**

3 **Step 1**: Decide the class memberships of $\{\mathbf{x}_i\}_{i=1}^n$ by assigning each of them to its nearest cluster center.

$$\gamma_{ij} = \begin{cases} 1, & ||\mathbf{x}_i - \mu_j||_1 \leq ||\mathbf{x}_i - \mu_{j'}||_1, \forall j' \\ 0, & \text{otherwise} \end{cases}$$

4 **Step 2**: For each $j \in \{1, \cdots, k\}$, recompute μ_j using the updated γ to be the center of mass of all points in C_j:

$$\mathbf{c}_p = \text{sort}([\mathbf{x}_{ip} \text{ if } \gamma_{ij}])$$
$$\mu_{jp} = c_p[\frac{\text{len}(c_p)}{2} + 1]$$

5 **until** *the objective function J no longer changes*;

We should use the k-means-ℓ_1 algorithm if outliers are present. The influence of outliers is more significant under ℓ_2 distance than under ℓ_1 distance. Thus, k-means-ℓ_1 is more robust.

4 [50pts] Kernel, Optimization and Learning

给定样本集 $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$, $\mathcal{F} = \{\Phi_1 \cdots, \mathbf{\Phi}_d\}$ 为非线性映射族。考虑如下的优化问题

$$\min_{\mathbf{w}, \mu \in \Delta_q} \quad \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|^2 + C\sum_{i=1}^{m}\max\left\{0, 1 - y_i\left(\sum_{k=1}^{d}\mathbf{w}_k \cdot \mathbf{\Phi}_k(x_i)\right)\right\} \tag{4.1}$$

其中, $\Delta_q = \{\mu : \mu \geq 0, \|\mu\|_q = 1\}$.

(1) **[30pts]** 请证明, 下面的问题4.2是优化问题4.1的对偶问题。

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \left\|\begin{matrix}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha}\end{matrix}\right\|_p \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \end{aligned} \tag{4.2}$$

其中, p 和 q 满足共轭关系, 即 $\frac{1}{p} + \frac{1}{q} = 1$. 同时, $\mathbf{K}_k$ 是由 $\mathbf{\Phi}_k$ 定义的核函数 (kernel).

(2) **[20pts]** 考虑在优化问题4.2中, 当 $p = 1$ 时, 试化简该问题。

Solution. (1) Rewrite the optimization problem:

$$\begin{aligned} \min_{\mathbf{w}, \mu \in \Delta_q} \quad & \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i\left(\sum_{k=1}^{d}\mathbf{w}_k \cdot \mathbf{\Phi}_k(x_i)\right) \geq 1 - \xi_i \\ & \xi_i \geq 0 \\ & \mu_i \geq 0 \\ & \|\mu\|_q = 1 \end{aligned} \tag{4.3}$$

The Lagrangian form of 4.3 is

$$\begin{aligned} L(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \delta) = & \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|^2 + C\sum_{i=1}^{m}\xi_i \\ & + \sum_{i=1}^{m}\alpha_i\left(1 - \xi_i - y_i\left(\sum_{k=1}^{d}\mathbf{w}_k \cdot \mathbf{\Phi}_k(x_i)\right)\right) \\ & - \sum_{i=1}^{m}\beta_i\xi_i - \sum_{i=1}^{d}\gamma_k\mu_k + \delta(\|\boldsymbol{\mu}\|_q - 1)\,. \end{aligned} \tag{4.4}$$

Take the partial derivatives of $\mathbf{w}$, $\boldsymbol{\mu}$, and $\boldsymbol{\xi}$, and let them equal to 0:

$$\begin{aligned} \frac{\mathbf{w}_k}{\mu_k} &= \sum_{i=1}^{m}\alpha_i y_i \mathbf{\Phi}_k(x_i)\,, \\ \frac{1}{2}\frac{\|\mathbf{w}_k\|^2}{\mu_k^2} + \gamma_k &= \delta\left(\frac{\mu_k}{\|\boldsymbol{\mu}\|_q}\right)^{q-1}\,, \\ \alpha_i + \beta_i &= C\,. \end{aligned} \tag{4.5}$$

We first eliminate $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ with 4.5,

$$\begin{aligned} L(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \delta) = -\frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|^2 + \sum_{i=1}^{m}\alpha_i \\ -\sum_{i=1}^{d}\gamma_k\mu_k + \delta(\|\boldsymbol{\mu}\|_q - 1)\,. \end{aligned} \tag{4.6}$$

Then eliminate $\boldsymbol{\gamma}$ with 4.5.

$$\begin{aligned} L(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \delta) &= -\frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|^2 + \sum_{i=1}^{m}\alpha_i \\ &\quad + -\frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|^2 - \delta\sum_{k=1}^{d}\frac{\mu_k^q}{\|\mu\|_q^{q-1}} + \delta\|\mu\|_q - \delta \\ &= \sum_{i=1}^{m}\alpha_i + \delta\frac{(\sum_{k=1}^{d}\mu_k^q) - \|\mu\|_q^q}{\|\mu\|_q^{q-1}} - \delta \\ &= \sum_{i=1}^{m}\alpha_i - \delta\,. \end{aligned} \tag{4.7}$$

We manipulate the second equation of 4.5 to evaluate δ,

$$\left(\frac{\|\mathbf{w}_k\|^2}{\mu_k^2}\right)^{\frac{q}{q-1}} = \left(2\delta\left(\frac{\mu_k}{\|\boldsymbol{\mu}\|_q}\right)^{q-1} - \gamma_k\right)^{\frac{q}{q-1}}\,.$$

We are maximizing the Lagrangian, and the Lagrangian is monotonically decreasing in δ. Therefore, we need to minimize δ. $\boldsymbol{\gamma}$ is dependent only on δ, and δ is monotonically increasing in γ_k. Thus we can safely set γ_k to 0. Then, the above equation simplifies into

$$\left(\frac{\|\mathbf{w}_k\|^2}{\mu_k^2}\right)^{\frac{q}{q-1}} = (2\delta)^{\frac{q}{q-1}}\left(\frac{\mu_k}{\|\boldsymbol{\mu}\|_q}\right)^{q}$$

Sum it over k, and take the $\frac{q-1}{q}$-th root,

$$\left(\sum_{k=1}^{d}\left(\frac{\|\mathbf{w}_k\|^2}{\mu_k^2}\right)^{\frac{q}{q-1}}\right)^{\frac{q-1}{q}} = 2\delta\|\boldsymbol{\mu}\|_q^{1-q}\left(\sum_{k=1}^{d}\mu_k^q\right)^{\frac{q-1}{q}} = 2\delta\,,$$

which is equivalent to

$$\frac{1}{2}\left\|\begin{matrix}\frac{\|\mathbf{w}_1\|^2}{\mu_1^2}\\ \vdots \\ \frac{\|\mathbf{w}_d\|^2}{\mu_d^2}\end{matrix}\right\|_p = \delta\,. \tag{4.8}$$

Substitute the first equation of 4.5 into 4.8,

$$\frac{1}{2}\left\|\begin{matrix}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha}\\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha}\end{matrix}\right\|_p = \delta\,. \tag{4.9}$$

With 4.5 and 4.9, the Lagrangian 4.4 is simplified into

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \left\| \begin{matrix} \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha} \end{matrix} \right\|_p \\ \text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{C} \end{aligned} \tag{4.10}$$

which is exactly 4.2.

(2) When $p = 1$, the optimization problem degenerates to

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}} \left(\sum_{k=1}^{d} \mathbf{K}_k \right) \mathbf{Y}\boldsymbol{\alpha} \\ \text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{C} \end{aligned} \tag{4.11}$$

(3) 第一小题中, 将 γ_k 置为 0 所得到的优化问题与原问题之对偶问题等价的想法来自计科 (19') 的卢以宁小姐.