机器学习导论
综合能力测试

141210016, 刘冰楠, bingnliu@outlook.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.
The most general form of exponential distribution family is

$$f_X(\mathbf{x} \mid \boldsymbol{\theta}) = h(\mathbf{x})\exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta})\right), \tag{1.3}$$

where $\mathbf{x}$ and $\boldsymbol{\theta}$ are vectors, $\boldsymbol{\eta}(\cdot)$ and $\mathbf{T}(\cdot)$ are vector valued functions.

1.1 Problem (1)

Parameters of *multinomial distribution* includes n and $p_1, p_2, \ldots, p_k$, where $n \geq 1$, $p_i \geq 0 \forall i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} = 1$. The probability mass function (PMF) of *multinomial*

distribution is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{\prod_{i=1}^{k}(x_i!)} \prod_{i=1}^{k} p_i^{x_i}. \tag{1.4}$$

Let $\mathbf{x} = (x_1, x_2, \ldots, x_k)^{\mathrm{T}}$, $\boldsymbol{\theta} = (\ln p_1, \ln p_2, \ldots, \ln p_k)^{\mathrm{T}}$. Now we rewrite Eq.(1.4):

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) =& \frac{n!}{\prod_{i=1}^{k}(x_i!)} \prod_{i=1}^{k} p_i^{x_i} \\ =& \frac{n!}{\prod_{i=1}^{k}(x_i!)} \exp \ln \left(\prod_{i=1}^{k} p_i^{x_i} \right) \\ =& \frac{n!}{\prod_{i=1}^{k}(x_i!)} \exp \left(\sum_{i=1}^{k} x_i \ln p_i \right) \\ =& \frac{n!}{\prod_{i=1}^{k}(x_i!)} \exp \left(\boldsymbol{\theta}^{\mathrm{T}} \mathbf{x} \right). \end{aligned} \tag{1.5}$$

Comparing to $h(\mathbf{x}) \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathrm{T}} \mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta})\right)$, we have:

表 1: Results for Problem (1)

$h(\mathbf{x})$	$\boldsymbol{\eta}(\boldsymbol{\theta})$	$\mathbf{T}(\mathbf{x})$	$A(\boldsymbol{\theta})$
$n!/\prod_{i=1}^{k}(x_i!)$	$(\ln p_1, \ln p_2, \ldots, \ln p_k)^{\mathrm{T}}$	$\mathbf{x}$	0

1.2 Problem (2)

Parameters of *multivariate Gaussian distribution* includes μ and $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a real valued symmetric matrix. The probability density function (PDF) of *multivariate distribution* is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{1}{(2\pi)^k |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right). \tag{1.6}$$

We first define *vectorization operation*. If $\mathbf{C}$ is a $m \times n$ matrix. Then

$$\mathrm{vec}(\mathbf{C}) = (c_{11}, c_{12}, \ldots, c_{1n}, \ldots, c_{m1}, c_{m2}, \ldots, c_{mn}) \tag{1.7}$$

Now we rewrite Eq.(1.6):

$$\begin{aligned} & P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) \\ =& \frac{1}{(2\pi)^k |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \\ =& \frac{1}{(2\pi)^k |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{x}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x} \right) \\ =& \frac{1}{(2\pi)^k} \exp\left(-\frac{1}{2} \mathbf{x}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x} + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \frac{1}{2} \ln |\boldsymbol{\Sigma}| \right) \\ =& \frac{1}{(2\pi)^k} \exp\left(-\frac{1}{2} \mathrm{vec}(\boldsymbol{\Sigma}^{-1})^{\mathrm{T}} \mathrm{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \frac{1}{2} \ln |\boldsymbol{\Sigma}| \right) \\ =& \frac{1}{(2\pi)^k} \exp\left((\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}; \mathrm{vec}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}))^{\mathrm{T}} (\mathbf{x}; \mathrm{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}})) - (\frac{1}{2} \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \frac{1}{2} \ln |\boldsymbol{\Sigma}|) \right). \end{aligned} \tag{1.8}$$

Comparing to $h(\mathbf{x})\exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{T}(\mathbf{x})-A(\boldsymbol{\theta})\right)$, we have:

表 2: Results for Problem (2)

$h(\mathbf{x})$	$\boldsymbol{\eta}(\boldsymbol{\theta})$	$\mathbf{T}(\mathbf{x})$	$A(\boldsymbol{\theta})$
$\frac{1}{(2\pi)^k}$	$(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu};\mathrm{vec}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}))$	$(\mathbf{x};\mathrm{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}}))$	$\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\frac{1}{2}\ln\lvert\boldsymbol{\Sigma}\rvert$

1.3 Problem (3)

According to *the bayes rule*, we have

$$h(\boldsymbol{\theta}\mid\mathbf{x})=\frac{p_{\pi}(\boldsymbol{\theta})f(D\mid\boldsymbol{\theta})}{p_z(\mathbf{x})},\tag{1.9}$$

where $p_z(\mathbf{x})$ is the normalization factor. We can omit the denominator since only the *kernel* of the probability density function is important.

$$\begin{aligned}
h(\boldsymbol{\theta}\mid\mathbf{x}) &= \frac{p_{\pi}(\boldsymbol{\theta})f(D\mid\boldsymbol{\theta})}{p_z(\mathbf{x})}\\
&\propto p_{\pi}(\boldsymbol{\theta})f(D\mid\boldsymbol{\theta})\\
&= p_{\pi}(\boldsymbol{\theta})\prod_{i=1}^{n}f(x_i\mid\boldsymbol{\theta})\\
&= p_{\pi}(\boldsymbol{\theta})\prod_{i=1}^{n}\left[h(x_i)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{T}(x_i)-A(\boldsymbol{\theta})\right)\right]\\
&= f(\boldsymbol{\chi},\nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi}-\nu A(\boldsymbol{\theta})\right)\prod_{i=1}^{n}\left[h(x_i)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{T}(x_i)-A(\boldsymbol{\theta})\right)\right]\\
&= f(\boldsymbol{\chi},\nu)\prod_{i=1}^{n}\ln(x_i)\exp\left[\boldsymbol{\theta}^{\mathrm{T}}\left(\boldsymbol{\chi}+\sum_{i=1}^{n}\mathbf{T}(x_i)\right)-(n+\nu)A(\boldsymbol{\theta})\right].
\end{aligned}\tag{1.10}$$

Define

$$\begin{cases}
\boldsymbol{\chi}' & =\boldsymbol{\chi}+\sum_{i=1}^{n}\mathbf{T}(x_i)\\
\nu' & =\nu+n\\
g(\mathbf{x},y) & =f(\mathbf{x}-\sum_{i=1}^{n}\mathbf{T}(x_i),y-n)\prod_{i=1}^{n}h(x_i),
\end{cases}\tag{1.11}$$

then

$$h(\boldsymbol{\theta}\mid\mathbf{x})\propto g(\boldsymbol{\chi}',\nu')\exp(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi}'-\nu' A(\boldsymbol{\theta}))\tag{1.12}$$

Therefore, the posterior distribution has the same form as the prior distribution.

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.

Because $X \in \mathcal{X} = \mathbb{R}^d$, X is of dimension d.

From the **conditional independence** of attributes, we have:

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d \mid Y = y) \\ =&P(X_1 = x_1 \mid Y = y)P(X_2 = x_2 \mid Y = y) \ldots P(X_d = x_d \mid Y = y). \end{aligned} \tag{2.1}$$

From *Bernoulli distribution* on label, we know the probability mass function of label is:

$$P(Y = y) = \begin{cases} \pi, & y = 1 \\ 1 - \pi, & y = 0 \end{cases} \tag{2.2}$$

From the *Bayes' theorem*, we know

$$\begin{aligned} &P(Y = y \mid X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d) \\ =&\frac{P(Y = y)P(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d \mid Y = y)}{P(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d)}. \end{aligned} \tag{2.3}$$

When we compare the posterior probability that from the same X but different Y, we can **omit the denominator** in Eq.(2.3) since they are the same.

2.1 Problem (1)

$$\begin{cases} P(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d \mid Y = 1) & \propto \pi \prod_{i=1}^d h_i(x_i) \exp(\theta_{i1} \cdot T_i(x_i) - A_i(\theta_{i1})) \\ P(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d \mid Y = 0) & \propto (1 - \pi) \prod_{i=1}^d h_i(x_i) \exp(\theta_{i0} \cdot T_i(x_i) - A_i(\theta_{i0})) \end{cases} \tag{2.4}$$

Let $\mathbf{x} = (x_1, x_2, \ldots, x_d)^{\mathrm{T}}$, then

$$P(X = \mathbf{x} \mid Y = 1) \propto \quad \pi \prod_{i=1}^{d} h_i(x_i) \exp(\theta_{i1} \cdot T_i(x_i) - A_i(\theta_{i1})), \tag{2.5}$$

$$P(X = \mathbf{x} \mid Y = 0) \propto \quad (1-\pi) \prod_{i=1}^{d} h_i(x_i) \exp(\theta_{i0} \cdot T_i(x_i) - A_i(\theta_{i0})). \tag{2.6}$$

Let $P(X = \mathbf{x} \mid Y = 1) = P(X = \mathbf{x} \mid Y = 0)$, and we simplify the result:

$$\pi \prod_{i=1}^{d} h_i(x_i) \exp(\theta_{i1} \cdot T_i(x_i) - A_i(\theta_{i1})) = (1-\pi) \prod_{i=1}^{d} h_i(x_i) \exp(\theta_{i0} \cdot T_i(x_i) - A_i(\theta_{i0}))$$

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

$$\frac{1-\pi}{\pi} \exp\left[\sum_{i=1}^{d} (A_i(\theta_{i1}) - A_i(\theta_{i0})) \right] = \exp\left[\sum_{i=1}^{d} (\theta_{i1} - \theta_{i0}) T_i(x_i) \right]$$

And finally,

$$\ln \frac{1-\pi}{\pi} + \sum_{i=1}^{d} (A_i(\theta_{i1}) - A_i(\theta_{i0})) = \sum_{i=1}^{d} (\theta_{i1} - \theta_{i0}) T_i(x_i) \tag{2.7}$$

We see this is a hyperplane in space spanned by $(T_i(x_i))$. This explanation is intuitive enough that we do not want to bother to use the sigmoid function.

2.2 Problem (2)

First we use the result in Table.(2) to write the univariate Gaussian distribution in the form of exponential family. Then we use the result in Eq.(2.7) to get the classification boundary.

According to Table.(2),

$$P_{\mathrm{norm}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \tag{2.8}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left((-\frac{1}{2\sigma^2}; \frac{\mu}{\sigma^2})^{\mathrm{T}} (x^2; x) - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \ln \sigma^2 \right). \tag{2.9}$$

Then according to Eq.(2.7), the classification boundary is

$$\ln \frac{1-\pi}{\pi} + \sum_{i=1}^{d} \left(\frac{\mu_{i1}^2}{2\sigma_i^2} + \frac{1}{2} \ln \sigma_i^2 - \frac{\mu_{i0}^2}{2\sigma_i^2} - \frac{1}{2} \ln \sigma_i^2 \right) = \sum_{i=1}^{d} (-\frac{1}{2\sigma_i^2} + \frac{1}{2\sigma_i^2}; \frac{\mu_{i1}}{\sigma_i^2} - \frac{\mu_{i0}}{\sigma_i^2})^{\mathrm{T}} (x^2; x), \tag{2.10}$$

i.e.

$$\ln \frac{1-\pi}{\pi} + \sum_{i=1}^{d} \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2} = \sum_{i=1}^{d} \frac{\mu_{i1} - \mu_{i0}}{\sigma_i^2} x_i. \tag{2.11}$$

This is a linear equation of $(x_1, x_2, \ldots, x_d)$, representing a hyperplane in $\mathcal{X} = \mathbb{R}^d$.

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

给定样本集 $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, 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

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

其中, $\mu_1, \ldots, \mu_k$ 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

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.

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

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:

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

5 **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.

3.1 Problem (1)

First we define J_i:

$$J_i(\gamma, \mu_1, \ldots, \mu_k) = \sum_{j=1}^{k} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 \tag{3.3}$$

3.1.1 Step 1

What Step 1 does is, for **fixed** $(\mu_1, \mu_2, \ldots, \mu_k)$, assign value to γ_{ij}, s.t.

$$\gamma_{ij} = \begin{cases} 1, & ||\mathbf{x}_i - \mu_j||^2 \leq ||\mathbf{x}_i - \mu_{j'}||^2, \forall j' \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

$\forall i = 1, 2, \ldots, n$, suppose only $\gamma_{iq} = 1$, then Eq.(3.3) becomes

$$J_i(\gamma, \mu_1, \ldots, \mu_k) = \|\mathbf{x}_i - \mu_q\|^2. \tag{3.5}$$

According to Eq.(3.4), γ_{iq} satisfies that q is the solution of

$$\min_{j=1,2,\ldots,k} \|\mathbf{x}_i - \mu_j\|^2. \tag{3.6}$$

Therefore after step 1, J_i reaches its minimum (w.r.t classification of $\mathbf{x}_i$). Thus the value of J_i becomes smaller or remains the same. Since $J = \sum_{i=1}^{n} J_i$, J also becomes smaller or remains the same after Step 1.

3.1.2 Step 2

What Step 2 does is, for **fixed** γ_{ij}, assign value to $\mu_1, \mu_2, \ldots, \mu_k$ s.t.

$$\mu_j = \frac{\sum_{i=1}^{n} \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^{n} \gamma_{ij}}. \tag{3.7}$$

Rewrite J as

$$\begin{aligned} J(\gamma, \mu_1, \ldots, \mu_k) &= \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 \\ &= \sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 \end{aligned} \tag{3.8}$$

and define J_j:

$$J_j(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2, \tag{3.9}$$

If we define D_j as the set composed of all $\mathbf{x}_i$ s.t. $\gamma_{ij} = 1$, then Eq.(3.9) can be written as

$$\begin{aligned} J_j(\gamma, \mu_1, \ldots, \mu_k) &= \sum_{\mathbf{x} \in D_j} \|\mathbf{x} - \mu_j\|^2 \\ &= \sum_{\mathbf{x} \in D_j} (\mathbf{x}^2 - 2\mu_j \mathbf{x} + \mu_j^2) \\ &= \mu_j^2 - 2(\sum_{\mathbf{x} \in D_j} \mathbf{x})\mu_j + \sum_{\mathbf{x} \in D_j} \mathbf{x}^2. \end{aligned} \tag{3.10}$$

We see J_j is a quadratic function of μ_j and J_j is minimized when

$$\mu_j = -\frac{-2 \sum_{\mathbf{x} \in D_j} \mathbf{x}}{2|D_j|} = \frac{\sum_{\mathbf{x} \in D_j} \mathbf{x}}{|D_j|}, \tag{3.11}$$

which is exactly the same with Eq.(3.7).

Therefore after step 2, J_j reaches its minimum (w.r.t μ_j). Thus the value of J_i becomes smaller or remains the same. Since $J = \sum_{j=1}^{k} J_j$, J also becomes smaller or remains the same after Step 2.

□

3.2 Problem (2)

I divide my proof into three parts.

3.2.1 A Way of Clustering is Determined By the Indicator Matirx

To define a way of clustering, we need to know

1. Location of centroids: $\mu_1, \mu_2, \ldots, \mu_k$;
2. classification of each sample: γ_{ij}.

Because μ_j satisfies $\mu_j = \frac{\sum_{i=1}^{n} \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^{n} \gamma_{ij}}$, if we only consider the clustering state **after** each iteration, the way of clustering is completely determined by classification of each sample. i.e. the way of clustering is completely determined by the indicator matrix. Thus the value of J is completely determined by the indicator matrix.

3.2.2 Sequences of Values of J Will Fall Into A Loop

During the iteration, what Algo.1 does is essentially changing the indicator matrix from one to another. And what the matrix will be after the iteration is **only** determined what it was before the iteration. Note, there are k^n different indicator matrices in total. So after at most k^n iterations, the value of J will fall into a loop. (or if the value of J does not change before k^n iterations, then Algo.1 stops)

3.2.3 Period of the Loop is 1

We have proved that after each iteration, the value of J **will not increase**. This is also true for the loop of values that J eventually falls into. Therefore period of the loop **must be** 1, i.e. there is only one element in the loop. In other words, the value of J will not change after it falls into the loop.

In conclusion, the value of J will stay the same after finite (at most n^k) iterations. Therefore Algo.1 will stop in finite iterations.

□

3.3 Problem (3)

Given the number of clusters k, we denote the minimal value of J as $J^{(k)}$ ($J^{(k)}$ must **exist** since $|\mathcal{S}^{(k)}| = k^n$ is finite), a specific way of clustering as s and the set of all ways of clustering as $\mathcal{S}^{(k)}$. The value of J corresponding to s is denoted as $J(s)$. Then the following relationship is true:

$$J(s) \leq J^{(k)}, \quad \forall s \in \mathcal{S}^{(k)}. \tag{3.12}$$

$\forall \ell \in \mathbb{N}^+$, suppose s_1 is the way of clustering corresponding to $J^{(\ell)}$, i.e. $J(s_1) = J^{(\ell)}$. If we keep all parameters unchanged, but add a **new** centroid $\mu_{\ell+1}$ s.t. it is **far enough from all samples** that no sample belongs to cluster $\ell+1$, or equivalently $\gamma_{i,\ell+1} = 0, \quad \forall i = 1, 2, \ldots, n)$. Now we have a new way of clustering, s_2. And $J(s_2) = J(s_1)$ because

$$\begin{aligned} J(s_2) &= \sum_{i=1}^{n} \sum_{j=1}^{\ell+1} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 \\ &= \sum_{i=1}^{n} \sum_{j=1}^{\ell} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 \\ &= J(s_1) \end{aligned} \tag{3.13}$$

Then according to Eq.(3.12), we have $J(s_2) \leq J^{(\ell+1)}$. Therefore

$$J^{(\ell)} = J(s_1) = J(s_2) \leq J^{(\ell+1)}, \qquad \forall \ell \in \mathbb{N}^+. \tag{3.14}$$

□

3.4 Problem (4)

3.4.1 Relationship Between Three Deviations

The answer is

$$T(X) = B(X) + \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \gamma_{ij}}{n} W_j(X). \tag{3.15}$$

Now we prove it.

First we state a **simple but important relationship**, which we will use several times later:

If a_i is an expression that only depends on i, b_j is an expression that only depends on j, and $f(\cdot), g(\cdot), h(\cdot)$ are arbitrary univariate functions, then from

$$\begin{cases} \gamma_{ij} & = 0, 1 \quad \forall j = 1, 2, \ldots, k \\ \sum_{j=1}^{k} \gamma_{ij} & = 1 \end{cases} \tag{3.16}$$

we have

$$f(g(a_i) + \sum_{j=1}^{k} \gamma_{ij} h(b_j)) = \sum_{j=1}^{k} \gamma_{ij} f(g(a_i) + h(b_j)) \tag{3.17}$$

Second we **split** $\|\mathbf{x}_i - \hat{\mathbf{x}}\|^2$ into three terms:

$$\begin{aligned} & \sum_{i=1}^{n} \|\mathbf{x}_i - \hat{\mathbf{x}}\|^2 \\ =& \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \gamma_{ij}\mu_j + \sum_{j=1}^{k} \gamma_{ij}\mu_j - \hat{\mathbf{x}}\|^2 \\ =& \sum_{i=1}^{n} \left[\|\mathbf{x}_i - \sum_{j=1}^{k} \gamma_{ij}\mu_j\|^2 + \|\sum_{j=1}^{k} \gamma_{ij}\mu_j - \hat{\mathbf{x}}\|^2 + 2(\sum_{j=1}^{k} \gamma_{ij}\mu_j - \hat{\mathbf{x}})^{\mathrm{T}} (\mathbf{x}_i - \sum_{j=1}^{k} \gamma_{ij}\mu_j) \right] \end{aligned} \tag{3.18}$$

We prove the third term in Eq.(3.18) is **zero**.

$$\begin{aligned} & \sum_{i=1}^{n} 2(\sum_{j=1}^{k} \gamma_{ij}\mu_j - \hat{\mathbf{x}})^{\mathrm{T}} (\mathbf{x}_i - \sum_{j=1}^{k} \gamma_{ij}\mu_j) \\ =& \sum_{\ell=1}^{k} \sum_{\mathbf{x} \in D_\ell} 2(\mu_\ell - \hat{\mathbf{x}})^{\mathrm{T}} (\mathbf{x} - \mu_\ell) \\ =& \sum_{\ell=1}^{k} 2(\mu_\ell - \hat{\mathbf{x}})^{\mathrm{T}} \sum_{\mathbf{x} \in D_\ell} (\mathbf{x} - \mu_\ell) \\ =& \sum_{\ell=1}^{k} 2(\mu_\ell - \hat{\mathbf{x}})^{\mathrm{T}} (\sum_{\mathbf{x} \in D_\ell} \mathbf{x} - \sum_{\mathbf{x} \in D_\ell} \mu_\ell) \\ =& \sum_{\ell=1}^{k} 2(\mu_\ell - \hat{\mathbf{x}})^{\mathrm{T}} (|D_\ell|\mu_\ell - |D_\ell|\mu_\ell) \\ =& 0 \end{aligned} \tag{3.19}$$

where we have used that when summing over all samples, we can go cluster by cluster.

Insert Eq.(3.19) into Eq.(3.18), we have:

$$\sum_{i=1}^{n}\|\mathbf{x}_i-\hat{\mathbf{x}}\|^2=\sum_{i=1}^{n}\|\mathbf{x}_i-\sum_{j=1}^{k}\gamma_{ij}\mu_j\|^2+\sum_{i=1}^{n}\|\sum_{j=1}^{k}\gamma_{ij}\mu_j-\hat{\mathbf{x}}\|^2 \tag{3.20}$$

The first term is

$$\begin{aligned}
&\sum_{i=1}^{n}\|\mathbf{x}_i-\sum_{j=1}^{k}\gamma_{ij}\mu_j\|^2\\
=&\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\|\mathbf{x}_i-\mu_j\|^2\\
=&\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}\|\mathbf{x}_i-\mu_j\|^2\\
=&\sum_{j=1}^{k}\frac{\sum_{i=1}^{n}\gamma_{ij}}{\sum_{i=1}^{n}\gamma_{ij}}\sum_{i=1}^{n}\gamma_{ij}\|\mathbf{x}_i-\mu_j\|^2\\
=&\sum_{j=1}^{k}\left(\sum_{i=1}^{n}\gamma_{ij}\right)\frac{\sum_{i=1}^{n}\gamma_{ij}\|\mathbf{x}_i-\mu_j\|^2}{\sum_{i=1}^{n}\gamma_{ij}}\\
=&\sum_{j=1}^{k}\left(\sum_{i=1}^{n}\gamma_{ij}\right)W_j(X)
\end{aligned} \tag{3.21}$$

The second term is

$$\begin{aligned}
&\sum_{i=1}^{n}\|\sum_{j=1}^{k}\gamma_{ij}\mu_j-\hat{\mathbf{x}}\|^2\\
=&\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\|\mu_j-\hat{\mathbf{x}}\|^2\\
=&\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}\|\mu_j-\hat{\mathbf{x}}\|^2\\
=&\sum_{j=1}^{k}\left(\sum_{i=1}^{n}\gamma_{ij}\right)\|\mu_j-\hat{\mathbf{x}}\|^2\\
=&n\sum_{j=1}^{k}\frac{\sum_{i=1}^{n}\gamma_{ij}}{n}\|\mu_j-\hat{\mathbf{x}}\|^2\\
=&nB(X)
\end{aligned} \tag{3.22}$$

Insert Eq.(3.21) and Eq.(3.22) into Eq.(3.20), we have:

$$\sum_{i=1}^{n}\|\mathbf{x}_i-\hat{\mathbf{x}}\|^2=\sum_{j=1}^{k}\left(\sum_{i=1}^{n}\gamma_{ij}\right)W_j(X)+nB(X) \tag{3.23}$$

Therefore,

$$T(X)=\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i-\hat{\mathbf{x}}\|^2=\sum_{j=1}^{k}\frac{\sum_{i=1}^{n}\gamma_{ij}}{n}W_j(X)+B(X) \tag{3.24}$$

□

3.4.2 Proof of the Problem (4).2

According to Eq.(3.24), we have

$$\begin{aligned} T(X) &= \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \gamma_{ij}}{n} W_j(X) + B(X) \\ &= \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \gamma_{ij}}{n} \frac{\sum_{i=1}^{n} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2}{\sum_{i=1}^{n} \gamma_{ij}} + B(X) \\ &= \sum_{j=1}^{k} \frac{1}{n} \sum_{i=1}^{n} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 + B(X) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 + B(X) \\ &= \frac{J(\gamma, \mu_1, \ldots, \mu_k)}{n} + B(X) \end{aligned} \tag{3.25}$$

It is shown in the derivation of Eq.(3.25) that

$$\sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \gamma_{ij}}{n} W_j(X) = \frac{J(\gamma, \mu_1, \ldots, \mu_k)}{n}, \tag{3.26}$$

which is to say, the weighted average of intra-cluster deviation equals J/n.

Conclusion: Algo.1 seeks to minimize J, thus minimize the weighted average of intra-cluster deviation (n is constant). And Since $T(X)$ is constant for given $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, inter-cluster deviation $B(X)$ is maximized when J is minimized.

Remark: Algo.1 tries to minimize J, but it may fall into local minimum. For a general distance metric, it might fall into a point that is not even a local minimum. See (SHOKRI Z. SELIM and M. A. ISMAIL, 1984)[1] for details.

3.5 Problem (5)

3.5.1 Problem (5).1

Step 1 is **does not depend on specific distance metric**, so we only need to replace ℓ_2 norm with ℓ_1 norm in Step 1. However, the assignment of μ_j in Step 2 **depends** on the form of distance metric.

Here we derive the value of μ_j that can help J reach its minimum, **given** the indicator matrix γ. i.e.

$$\min_{\mu_1, \mu_2, \ldots, \mu_k} \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|_1, \tag{3.27}$$

which is equivalent to

$$\min_{\mu_j} \sum_{i=1}^{n} \gamma_{ij} \|\mathbf{x}_i - \mu_j\|_1. \tag{3.28}$$

Algorithm 2: k-means-ℓ_2 Algorithm

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:

$$\mu_j = \text{median}(D_j),$$

where D_j is the set composed of all x_i with $\gamma_{ij} = 1$.

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

Suppose $\mathbf{x} \in \mathbb{R}^d$, and we use $a^{(i)}$ to denote the i^{th} component of vector $\mathbf{a}$. Then

$$\begin{aligned} &\sum_{i=1}^n \gamma_{ij} \|\mathbf{x}_i - \mu_j\|_1 \\ =&\sum_{i=1}^n \gamma_{ij} \sum_{\nu=1}^d |x_i^{(\nu)} - \mu_j^{(\nu)}| \\ =&\sum_{\nu=1}^d \sum_{i=1}^n \gamma_{ij} |x_i^{(\nu)} - \mu_j^{(\nu)}|. \end{aligned} \tag{3.29}$$

Therefore optimization problem (3.28) can be written as

$$\min_{\mu_j} \sum_{\nu=1}^d \sum_{i=1}^n \gamma_{ij} |x_i^{(\nu)} - \mu_j^{(\nu)}|. \tag{3.30}$$

We can **strengthen** our objective here — we demand $\forall \nu = 1, 2, \ldots, d$,

$$\min_{\mu_j} \sum_{i=1}^n \gamma_{ij} |x_i^{(\nu)} - \mu_j^{(\nu)}|. \tag{3.31}$$

Solution to optimization problem (3.31) is easy to find:

$$\mu_j = \text{median}(D_j), \tag{3.32}$$

where D_j is the set composed of all x_i with $\gamma_{ij} = 1$.

3.5.2 Problem (5).2

When outliers exist, **k-means-ℓ_1 algorithm is more robust** because its step 2 is more robust.

In step 2, we determine the location of the centroid of each cluster. Step 2 in k-means-ℓ_2 algorithm will **take average of all points** in that cluster, which is severely affected

by outliers. While step 2 in k-means-ℓ_1 algorithm only find the median of points in that cluster. And taking **median itself is naturally insensitive to outliers**.

4 [50pts] Kernel, Optimization and Learning

给定样本集 $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, $\mathcal{F} = \{\mathbf{\Phi}_1 \cdots, \mathbf{\Phi}_d\}$ 为非线性映射族。考虑如下的优化问题

$$\min_{\mathbf{w}, \boldsymbol{\mu} \in \Delta_q} \quad \frac{1}{2} \sum_{k=1}^{d} \frac{1}{\mu_k} \|\mathbf{w}_k\|_2^2 + C \sum_{i=1}^{m} \max \left\{ 0, 1 - y_i \left(\sum_{k=1}^{d} \mathbf{w}_k \cdot \mathbf{\Phi}_k(\mathbf{x}_i) \right) \right\} \tag{4.1}$$

其中, $\Delta_q = \{\boldsymbol{\mu} | \mu_k \geq 0, k = 1, \cdots, d; \|\boldsymbol{\mu}\|_q = 1\}$.

(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, $\mathbf{Y} = \mathrm{diag}([y_1, \cdots, y_m])$, $\mathbf{K}_k$ 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.

4.1 Problem (1)

Introduce *slack variable* $\boldsymbol{\xi}$, optimization problem (4.1) is equivalent to

$$\begin{aligned} \min_{\mathbf{w}, \boldsymbol{\mu} \in \Delta_q, \boldsymbol{\xi}} \quad & \frac{1}{2} \sum_{k=1}^{d} \frac{1}{\mu_k} \|\mathbf{w}_k\|_2^2 + C \sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & y_i \left(\sum_{k=1}^{d} \mathbf{w}_k \cdot \mathbf{\Phi}_k(\mathbf{x}_i) \right) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{4.3}$$

where $\Delta_q = \{\boldsymbol{\mu} | \mu_k \geq 0, k = 1, \cdots, d; \|\boldsymbol{\mu}\|_q = 1\}$.
Rewrite (4.3) as a **standard** constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\mu},\boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & 1 - \xi_i - y_i\left(\sum_{k=1}^{d}\mathbf{w}_k \cdot \boldsymbol{\Phi}_k(\mathbf{x}_i)\right) \leq 0 \\ & -\xi_i \leq 0 \\ & -\mu_k \leq 0 \\ & \left(\sum_{k=1}^{d}|\mu_k|^q\right)^{1/q} = 1 \end{aligned} \tag{4.4}$$

Using *KKT Multipliers*, Lagrangian of (4.4) is

$$\begin{aligned} & \mathcal{L}(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda) \\ = \; & \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i\left[1 - \xi_i - y_i\left(\sum_{k=1}^{d}\mathbf{w}_k \cdot \boldsymbol{\Phi}_k(\mathbf{x}_i)\right)\right] \\ & -\sum_{i=1}^{m}\beta_i\xi_i - \sum_{k=1}^{d}\gamma_k\mu_k + \lambda\left[\left(\sum_{k=1}^{d}|\mu_k|^q\right)^{1/q} - 1\right] \end{aligned} \tag{4.5}$$

With $\mathcal{L}(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda)$, we can write down an equivalent form of optimization problem (4.4) (this is also the **primal** problem):

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\mu},\boldsymbol{\xi}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\lambda} \quad & \mathcal{L}(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda) \\ \text{s.t.} \quad & \alpha_i \geq 0 \\ & \beta_i \geq 0 \\ & \gamma_k \geq 0 \end{aligned} \tag{4.6}$$

where $i = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, d$.

Exchange order of min max, we get the **dual problem**:

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\lambda} \min_{\mathbf{w},\boldsymbol{\mu},\boldsymbol{\xi}} \quad & \mathcal{L}(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda) \\ \text{s.t.} \quad & \alpha_i \geq 0 \\ & \beta_i \geq 0 \\ & \gamma_k \geq 0 \end{aligned} \tag{4.7}$$

where $i = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, d$.

Then our task is to explicitly write down the *dual function*

$$\Gamma(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda) = \min_{\mathbf{w},\boldsymbol{\mu},\boldsymbol{\xi}} \quad \mathcal{L}(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{4.8}$$

We take derivatives of $\mathcal{L}$ w.r.t. $\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\xi}$ and let them to be zero:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mathbf{w}_k} & = \frac{1}{\mu_k}\mathbf{w}_k - \sum_{i=1}^m \alpha_i y_i \mathbf{\Phi}_k(\mathbf{x}_i) = 0 \\ \frac{\partial \mathcal{L}}{\partial \mu_k} & = -\frac{1}{2\mu_k^2}\left\|\mathbf{w}_k\right\|_2^2 + \lambda\left(\frac{|\mu_k|}{\|\boldsymbol{\mu}\|_q}\right)^{q-1}\operatorname{sgn}(\mu_k) - \gamma_k = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_i} & = C - \alpha_i - \beta_i = 0 \end{cases} \tag{4.9}$$

i.e.

$$\mathbf{w}_k = \mu_k \sum_{i=1}^m \alpha_i y_i \mathbf{\Phi}_k(\mathbf{x}_i) \tag{4.10}$$

$$\left(\frac{|\mu_k|}{\|\boldsymbol{\mu}\|_q}\right)^{q-1} = \frac{1}{\lambda}\left[\frac{1}{2\mu_k^2}\left\|\mathbf{w}_k\right\|_2^2 + \gamma_k\right] \tag{4.11}$$

$$\alpha_i + \beta_i = C_i \tag{4.12}$$

Note that when calculating $\partial\mathcal{L}/\partial\mu_k$ we have implicitly used the fact that $\mu_k \geq 0$.

Above results need some further investigation.

First from Eq.(4.7) and Eq.(4.12), we can combine $\alpha_i \geq 0$ and $\beta_i \geq 0$:

$$0 \leq \alpha_i \leq C, \quad \forall i = 1, 2, \ldots, m. \tag{4.13}$$

Second from Eq.(4.10) we can derive

$$\begin{aligned} \|\mathbf{w}_k\|^2 &= \mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k \\ &= \left(\mu_k \sum_{i=1}^m \alpha_i y_i \mathbf{\Phi}_k(\mathbf{x}_i)\right)^{\mathrm{T}} \left(\mu_k \sum_{i=1}^m \alpha_i y_i \mathbf{\Phi}_k(\mathbf{x}_i)\right) \\ &= \mu_k^2 \sum_{i=1}^m \sum_{j=1}^m \alpha_i y_i \mathbf{\Phi}_k^{\mathrm{T}}(\mathbf{x}_i)\mathbf{\Phi}_k(x_j) y_j \alpha_j \\ &= \mu_k^2 \sum_{i=1}^m \sum_{j=1}^m \alpha_i y_i K_{ij}^{(k)} y_j \alpha_j \\ &= \mu_k^2 \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_k\mathbf{Y}\boldsymbol{\alpha}. \end{aligned} \tag{4.14}$$

Third, insert Eq.(4.14) into Eq.(4.11) and simplify the result, we have

$$\begin{cases} \mu_1^q = & A_1(\mu_1^q + \mu_2^q + \cdots + \mu_d^q) \\ \mu_2^q = & A_2(\mu_1^q + \mu_2^q + \cdots + \mu_d^q) \\ & \vdots \\ \mu_d^q = & A_d(\mu_1^q + \mu_2^q + \cdots + \mu_d^q) \end{cases} \tag{4.15}$$

where $A_k = \left[\frac{1}{\lambda}\left(\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_k\mathbf{Y}\boldsymbol{\alpha} + \gamma_k\right)\right]^{\frac{q}{q-1}} = \left[\frac{1}{\lambda}\left(\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_k\mathbf{Y}\boldsymbol{\alpha} + \gamma_k\right)\right]^p$.

Eq.(4.15) is actually **a homogeneous linear system** of $\mu_1^q, \mu_2^q, \ldots, \mu_d^q$. If we would like a nonzero $\boldsymbol{\mu}$ to exist (which is required in constraints), determinant of the coefficient matrix should be zero:

$$\left|\begin{bmatrix} A_1 - 1 & A_1 & \ldots & A_1 \\ A_2 & A_2 - 1 & \ldots & A_2 \\ \vdots & & & \\ A_d & A_d & \ldots & A_d - 1 \end{bmatrix}\right| = 0 \tag{4.16}$$

after some algebra we have (see Appendix 4.1.3 for details):

$$\sum_{k=1}^{d} A_k = 1 \tag{4.17}$$

i.e.

$$\left\| \begin{bmatrix} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_1 \mathbf{Y} \boldsymbol{\alpha} + 2\gamma_1 \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_d \mathbf{Y} \boldsymbol{\alpha} + 2\gamma_d \end{bmatrix} \right\|_p = 2\lambda \tag{4.18}$$

Insert Eq.(4.10)—Eq.(4.18) into $\mathcal{L}(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda)$, we get the dual function $\Gamma(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda)$:

$$\begin{aligned}
&\Gamma(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda) \\
=&\sum_{i=1}^{m} \alpha_i + \left(C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i \xi_i - \sum_{i=1}^{m} \beta_i \xi_i \right) \\
&+ \frac{1}{2} \sum_{k=1}^{d} \mu_k \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha} - \sum_{i=1}^{m} \alpha_i y_i \left(\sum_{k=1}^{d} \mathbf{w}_k^{\mathrm{T}} \boldsymbol{\Phi}_k(\mathbf{x}_i) \right) - \sum_{k=1}^{d} \gamma_k \mu_k + \lambda(\|\boldsymbol{\mu}\|_q - 1) \\
=&\sum_{i=1}^{m} \alpha_i \\
&+ \frac{1}{2} \sum_{k=1}^{d} \mu_k \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha} - \sum_{i=1}^{m} \alpha_i y_i \left(\sum_{k=1}^{d} (\mu_k \sum_{j=1}^{m} \alpha_j y_j \boldsymbol{\Phi}_k(x_j)) \boldsymbol{\Phi}_k(\mathbf{x}_i) \right) - \sum_{k=1}^{d} \gamma_k \mu_k + \lambda(\|\boldsymbol{\mu}\|_q - 1) \\
=&\sum_{i=1}^{m} \alpha_i + \frac{1}{2} \sum_{k=1}^{d} \mu_k \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha} - \sum_{k=1}^{d} \mu_k \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha} - \sum_{k=1}^{d} \gamma_k \mu_k + \lambda(\|\boldsymbol{\mu}\|_q - 1) \\
=&\sum_{i=1}^{m} \alpha_i - \sum_{k=1}^{d} \mu_k \left(\frac{1}{2} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha} + \gamma_k \right) + \lambda(\|\boldsymbol{\mu}\|_q - 1) \\
=&\sum_{i=1}^{m} \alpha_i - \sum_{k=1}^{d} \lambda^{-\frac{1}{q-1}} \|\boldsymbol{\mu}\|_q \left(\frac{1}{2} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha} + \gamma_k \right)^{\frac{1}{q-1}} \left(\frac{1}{2} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha} + \gamma_k \right) + \lambda(\|\boldsymbol{\mu}\|_q - 1) \\
=&\sum_{i=1}^{m} \alpha_i - \lambda^{-\frac{1}{q-1}} \|\boldsymbol{\mu}\|_q \sum_{k=1}^{d} \left(\frac{1}{2} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha} + \gamma_k \right)^{p} + \lambda(\|\boldsymbol{\mu}\|_q - 1) \\
=&\sum_{i=1}^{m} \alpha_i - \lambda^{-\frac{1}{q-1}} \|\boldsymbol{\mu}\|_q \lambda^p + \lambda(\|\boldsymbol{\mu}\|_q - 1) \\
=&\sum_{i=1}^{m} \alpha_i - \lambda \\
=&\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{1} - \left\| \begin{bmatrix} \frac{1}{2} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_1 \mathbf{Y} \boldsymbol{\alpha} + \gamma_1 \\ \vdots \\ \frac{1}{2} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_d \mathbf{Y} \boldsymbol{\alpha} + \gamma_d \end{bmatrix} \right\|_p
\end{aligned} \tag{4.19}$$

Insert Eq.(4.13) and Eq.(4.19) into Eq.(4.7), and note maximize $\mathcal{L}$ is equivalent to

maximize $2\mathcal{L}$, we get

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \quad & 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \left\| \begin{bmatrix} \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha} + 2\gamma_1 \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha} + 2\gamma_d \end{bmatrix} \right\|_p \\ \text{s.t.} \quad & 0 \le \alpha_i \le C \\ & \gamma_k \ge 0 \end{aligned} \tag{4.20}$$

where $i = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, d$.

This can be further simplified. According to **positive semi-definiteness** of $\mathbf{K}_k$, we have

$$\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_k\mathbf{Y}\boldsymbol{\alpha} \ge 0, \quad \forall k = 1, 2, \ldots, d. \tag{4.21}$$

And γ_k appears "**independently**" in the objective function and $\gamma_k \ge 0$. So at the optimal point of the dual problem, we must have

$$\gamma_k = 0, \quad \forall k = 1, 2, \ldots, d. \tag{4.22}$$

So finally, the dual problem is

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \left\| \begin{bmatrix} \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha} \end{bmatrix} \right\|_p \\ \text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{C} \end{aligned} \tag{4.23}$$

4.1.1 Make Predictions After Training

For a new sample $\mathbf{x}$, we make prediction y using the decision function:

$$y = \sum_{k=1}^{d} \mathbf{w}_k^{\mathrm{T}} \boldsymbol{\Phi}_k(\mathbf{x}). \tag{4.24}$$

Insert Eq.(4.10) into Eq.(4.24), we have

$$y = \sum_{k=1}^{d} \mu_k \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{\Phi}_k^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\Phi}_k(\mathbf{x}). \tag{4.25}$$

Since $\mathbf{x}, \mathbf{x}_i, y_i$ are already known and α_i has been solved in optimization. Now we **only need** to calculate μ_k from linear equations system (4.15), for $k = 1, 2, \ldots, d$.

Denote the coefficient matrix of (4.15) by M:

$$M = \begin{bmatrix} A_1 - 1 & A_1 & \ldots & A_1 \\ A_2 & A_2 - 1 & \ldots & A_2 \\ \vdots & & & \\ A_d & A_d & \ldots & A_d - 1 \end{bmatrix} \tag{4.26}$$

And denote **principle minors** of order k of M by M_k respectively. From Appendix (4.1.3) we know $\forall k = 1, 2, \ldots, d$,

$$\det(M_k) = -1 + \sum_{\ell=1}^{k} A_\ell = -\sum_{\ell=k+1}^{d} A_\ell. \tag{4.27}$$

We already know that $A_k \geq 0, \quad \forall k = 1, 2, \ldots, d$.

- If each $A_k > 0$, then $\text{rank}(M) = d - 1$, so $\text{nullity}(M) = 1$, i.e. solution space of μ_k^q has a dimension of 1. Then we only need to pick the $\boldsymbol{\mu}$ that satisfies $\|\boldsymbol{\mu}\|_q = 1$.
- If some A_k is zero, then dimension of solution space of mu_k^q will be larger than 1. We will have **multiple set of $\boldsymbol{\mu}$ that are feasible** to be used to make predictions.

4.1.2 Remarks

About Constraint $\|\boldsymbol{\mu}\|_q = 1$

When dealing with constraint

$$\|\boldsymbol{\mu}\|_q = C, \tag{4.28}$$

where C is a constant and $C = 1$ for this specific problem. We can first rewrite it as

$$\sum_{k=1}^{d} \mu_k^q = C^q, \tag{4.29}$$

In some problems, which will make Lagrangian and its derivatives **much simpler**. For example, Eq.(4.11) becomes

$$\frac{1}{2\mu_k^2} \|\mathbf{w}_k\|_2^2 + \gamma_k = \lambda q |\mu_k|^{q-1}. \tag{4.30}$$

If γ_k does not appear in Eq.(4.30), this will simplify later derivations.

About Constraint $\mu_k \geq 0$

Constraint $\|\boldsymbol{\mu}\|_q = 1$ tells us that feasible $\boldsymbol{\mu}$ is a hyper-sphere in $\mathbb{R}^d$, which is symmetric about the origin. **During derivation of the dual problem**, We can avoid introducing $\mu_k \geq 0$ into Lagrangian but implicitly use this property. Because for a negative μ_k we can just flip its sign.

4.1.3 Appendix

Consider the following matrix $M_{n\times n}$:

$$M = \begin{bmatrix} a_{11} - \lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} - \lambda & \ldots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \ldots & a_{nn} - \lambda \end{bmatrix} = \begin{bmatrix} a_{11} - \lambda & a_{12} - 0 & \ldots & a_{1n} - 0 \\ a_{21} - 0 & a_{22} - \lambda & \ldots & a_{2n} - 0 \\ \vdots & & & \\ a_{n1} - 0 & a_{n2} - 0 & \ldots & a_{nn} - \lambda \end{bmatrix} \tag{4.31}$$

Each row vector of the matrix can be written as sum of two row vectors (we use the first row as an example): $(a11 - 1, a_{12}, \ldots, a1n) = (a11, a_{12}, \ldots, a1n) + (-1, 0, 0, \ldots, 0)$.

Using linearity of determinant, $\det(M)$ equals sum of 2^d determinants, which is a polynomial of order n of variable λ. Among those monomials, only coefficients of λ^n and λ^{n-1} are nonzero:

$$\begin{bmatrix} -\lambda & 0 & \dots & 0 \\ 0 & -\lambda & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & -\lambda \end{bmatrix}, \begin{bmatrix} a_{11}-\lambda & 0 & \dots & 0 \\ 0 & -\lambda & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & -\lambda \end{bmatrix}, \dots, \begin{bmatrix} -\lambda & 0 & \dots & 0 \\ 0 & -\lambda & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & a_{nn}-\lambda \end{bmatrix} \tag{4.32}$$

Therefore

$$\det(M) = (-\lambda)^n + (-\lambda)^{n-1} \sum_{i=1}^{n} a_{ii} = -\lambda + \sum_{i=1}^{n} a_{ii} \tag{4.33}$$

If we apply this result to Eq.(4.16), we will get Eq.(4.18)

4.2 Problem (2)

When $p = 1$, a vector $\mathbf{a}$'s l_p norm is

$$\|a\|_1 = \left(\sum_{i=1}^{n} |a_i|^1 \right)^{\frac{1}{1}} = \sum_{i=1}^{n} |a_i|. \tag{4.34}$$

Therefore

$$\left\| \begin{bmatrix} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_1 \mathbf{Y} \boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_d \mathbf{Y} \boldsymbol{\alpha} \end{bmatrix} \right\|_p = \sum_{k=1}^{d} |\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha}| = \sum_{k=1}^{d} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \left(\sum_{k=1}^{d} \mathbf{K}_k \right) \mathbf{Y} \boldsymbol{\alpha} \tag{4.35}$$

where the non-negativeness of $\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_k \mathbf{Y} \boldsymbol{\alpha}$ has been said in Eq.(4.21).

Therefore, the dual problem when $p = 1$ is

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & 2\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{1} - \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \left(\sum_{k=1}^{d} \mathbf{K}_k \right) \mathbf{Y} \boldsymbol{\alpha} \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \end{aligned} \tag{4.36}$$

参考文献

[1] Shokri Z Selim and Mohamed A Ismail. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*, (1):81–87, 1984.