# 机器学习导论
# 综合能力测试

学号：141130077

作者姓名：邱梓豪

邮箱：2957606241@qq.com

2017 年 6 月 18 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$\nu$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.** 此处用于写证明(中英文均可)

(1)令$X = (X_1, X_2, \cdots, X_K)$为试验中每个随机变量的取值集合，设共有M次试验，则其中$X_k$表示 M次试验中事件$X_k$发生的次数，设$p_k$表示第k个事件的发生概率，则多项分布的概率密度函数为：

$$P(\mathbf{x}|M,p_1,\cdots,p_K) = \frac{M!}{x_1!,x_2!,\cdots,x_m!}p_1^{x_1}p_2^{x_2}\cdots p_K^{x_K} \tag{1.3}$$

可将上式的连乘通过取对数改成连加：

$$P(\mathbf{x}|M,p_1,\cdots,p_K) = \frac{M!}{x_1!,x_2!,\cdots,x_m!}\exp[\sum_{k=1}^{K}x_k\log p_k] \tag{1.4}$$

因为有$\sum_{k=1}^{K} x_k = M$，$\sum_{k=1}^{K} p_k = 1$，所以有：

$$\begin{align}
P(\mathbf{x}|M, p_1, \cdots, p_K) &= \frac{M!}{x_1!, x_2!, \cdots, x_m!} \exp\{\sum_{k=1}^{K} x_k \log p_k\} \tag{1.5} \\
&= \frac{M!}{x_1!, x_2!, \cdots, x_m!} \exp\{\sum_{k=1}^{K-1} x_k \log p_k + (M - \sum_{k=1}^{K-1} x_k) \log(1 - \sum_{k=1}^{K-1} p_k)\} \tag{1.6} \\
&= \frac{M!}{x_1!, x_2!, \cdots, x_m!} \exp\{\sum_{k=1}^{K-1} x_k \log(\frac{p_k}{1 - \sum_{k=1}^{K-1} p_k}) + M \log(1 - \sum_{k=1}^{K-1} p_k)\} \tag{1.7}
\end{align}$$

所以有：

$$\eta_k = \log(\frac{p_k}{1 - \sum_{k=1}^{K-1} p_k}) = \log(\frac{p_k}{p_K}) \tag{1.8}$$

$$A(M, p_1, \cdots, p_k) = -M \log(1 - \sum_{k=1}^{K-1} p_k) \tag{1.9}$$

令$h(\mathbf{x}) = \frac{M!}{x_1!, x_2!, \cdots, x_m!}$，$\eta(M, p_1, \cdots, p_k) = [\eta_1, \cdots, \eta_K]$，$T(x) = [x_1, \cdots, x_K]^T$，则有：

$$P(\mathbf{x}|M, p_1, \cdots, p_K) = h(\mathbf{x}) \exp(\eta(M, p_1, \cdots, p_k) T(x) - A(M, p_1, \cdots, p_k)) \tag{1.10}$$

从上式可以看出，多项分布属于指数分布族。
(2)多元高斯分布如下：

$$p(x|D, \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp(-(x - \mu)^T \Sigma^{-1} (x - \mu)/2) \tag{1.11}$$

故：

$$\begin{align}
p(x|D, \mu, \Sigma) &= \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp(-(x - \mu)^T \Sigma^{-1} (x - \mu)/2) \tag{1.12} \\
&= (2\pi)^{-D/2} \exp\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) - \frac{1}{2} \log(\Sigma)\} \tag{1.13} \\
&= (2\pi)^{-D/2} \exp\{-\frac{1}{2}[(x - \mu)^T \Sigma^{-1} (x - \mu) + \log(\Sigma)]\} \tag{1.14} \\
&= (2\pi)^{-D/2} \exp\{-\frac{1}{2}[x^T \Sigma^{-1} x - 2\mu^T \Sigma^{-1} x + \mu^T \Sigma^{-1} \mu + \log(\Sigma)]\} \tag{1.15} \\
& \tag{1.16}
\end{align}$$

上式中$x^T \Sigma^{-1} x - 2\mu^T \Sigma^{-1} x$可以拆分为$\eta(\theta) T(x)$的形式，拆分如下：

$$x^T \Sigma^{-1} x - 2\mu^T \Sigma^{-1} x = \begin{bmatrix} \Sigma^{-1} \\ -2\Sigma^{-1}\mu \end{bmatrix}^T \begin{bmatrix} xx^T \\ x \end{bmatrix} \tag{1.17}$$

所以有：

$$\begin{align}
p(x|D, \mu, \Sigma) &= (2\pi)^{-D/2} e^{-1/2} \exp(x^T \Sigma^{-1} x - 2\mu^T \Sigma^{-1} x + \mu^T \Sigma^{-1} \mu + \log(\Sigma)) \tag{1.18} \\
&= h(x) exp(\eta(D, \mu, \Sigma) T(x) - A(D, \mu, \Sigma)) \tag{1.19}
\end{align}$$

其中，$h(x) = (2\pi)^{-D/2}e^{-1/2}, \eta(D,\mu,\Sigma) = \begin{bmatrix} \Sigma^{-1} \\ -2\Sigma^{-1}\mu \end{bmatrix}^T, T(x) = \begin{bmatrix} xx^T \\ x \end{bmatrix}, A(D,\mu,\Sigma) = -(\mu^T\Sigma^{-1}\mu + \log(\Sigma))$

从上式可以看出，多元高斯分布属于指数分布族。

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

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

样本集$\mathcal{D} = \{x_1, \cdots, x_n\}$产生的似然为：

$$\prod_{i=1}^n p(x_i|\boldsymbol{\theta}) = \prod_{i=1}^n \{f(x_i|\boldsymbol{\theta}) = h(x_i)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}T(x_i) - A(\boldsymbol{\theta})\right)\} \tag{1.20}$$

所以参数$\boldsymbol{\theta}$服从的后验分布：

$$p(\boldsymbol{\theta}|x_{1:n},\boldsymbol{\chi},\nu) \propto p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi},\nu)\prod_{i=1}^n p(x_i|\boldsymbol{\theta}) \tag{1.21}$$

$$= f(\boldsymbol{\chi},\nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi} - \nu A(\boldsymbol{\theta})\right)(\prod_{i=1}^n h(x_i))\exp(\boldsymbol{\theta}^{\mathrm{T}}\sum_{i=1}^n T(x_i) - nA(\boldsymbol{\theta})) \tag{1.22}$$

$$= f(\boldsymbol{\chi},\nu)\prod_{i=1}^n h(x_i)\exp(\boldsymbol{\theta}^{\mathrm{T}}(\sum_{i=1}^n T(x_i) + \boldsymbol{\chi}) - (\nu+n)A(\boldsymbol{\theta})) \tag{1.23}$$

$$\propto f(\boldsymbol{\chi},\nu)\exp(\boldsymbol{\theta}^{\mathrm{T}}(\sum_{i=1}^n T(x_i) + \boldsymbol{\chi}) - (\nu+n)A(\boldsymbol{\theta})) \tag{1.24}$$

如果做如下的参数替换，则该后验分布具有和先验分布同样的形式。

$$\hat{\boldsymbol{\chi}} = \sum_{i=1}^n T(x_i) + \boldsymbol{\chi} \tag{1.25}$$

$$\hat{\nu} = \nu + n \tag{1.26}$$

## 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$\mu_{iy}$以及方差为$\sigma_i^2$ (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.** 此处用于写解答(中英文均可)

(1)P(Y=1)=$\pi$，P(Y=0)=$1-\pi$，则Y=1时的后验概率分布：

$$P(Y=1|X=x) = \frac{P(X=x|Y=1)P(Y=1)}{P(X=x|Y=1)P(Y=1)+P(X=x|Y=0)P(Y=0)} \tag{2.1}$$

$$= \frac{\pi\prod_{i=1}^d h_i(x_i)\exp(\sum_{i=1}^d \theta_{i1}T_i(x_i) - \sum_{i=1}^d A_i(\theta_{i1}))}{\pi\prod_{i=1}^d h_i(x_i)\exp(\sum_{i=1}^d \theta_{i1}T_i(x_i) - \sum_{i=1}^d A_i(\theta_{i1})) + (1-\pi)\prod_{i=1}^d h_i(x_i)\exp(\sum_{i=1}^d \theta_{i0}T_i(x_i) - \sum_{i=1}^d A_i(\theta_{i0}))} \tag{2.2}$$

$$= \frac{\pi\exp(\sum_{i=1}^d \theta_{i1}T_i(x_i) - \sum_{i=1}^d A_i(\theta_{i1}))}{\pi\exp(\sum_{i=1}^d \theta_{i1}T_i(x_i) - \sum_{i=1}^d A_i(\theta_{i1})) + (1-\pi)\exp(\sum_{i=1}^d \theta_{i0}T_i(x_i) - \sum_{i=1}^d A_i(\theta_{i0}))} \tag{2.3}$$

令：

$$a_0 = \sum_{i=1}^d \theta_{i0}T_i(x_i) - \sum_{i=1}^d A_i(\theta_{i0}) \tag{2.4}$$

$$a_1 = \sum_{i=1}^d \theta_{i1}T_i(x_i) - \sum_{i=1}^d A_i(\theta_{i1}) \tag{2.5}$$

则有：

$$P(Y=1|X=x) = \frac{\pi e^{a_1}}{\pi e^{a_1} + (1-\pi)e^{a_0}} \tag{2.6}$$

$$= \frac{1}{1+\frac{1-\pi}{\pi}e^{a_0-a_1}} \tag{2.7}$$

$$= \frac{1}{1+e^{-\log\frac{\pi}{1-\pi}+(a_0-a_1)}} \tag{2.8}$$

又因为：

$$\begin{aligned}a_0-a_1&=(\sum_{i=1}^{d}\theta_{i0}T_i(x_i)-\sum_{i=1}^{d}A_i(\theta_{i0}))-(\sum_{i=1}^{d}\theta_{i1}T_i(x_i)-\sum_{i=1}^{d}A_i(\theta_{i1}))\\&=\sum_{i=1}^{d}(\theta_{i0}-\theta_{i1})T_i(x_i)-\sum_{i=1}^{d}[A_i(\theta_{i0})-A_i(\theta_{i1})]\end{aligned}$$

令$\beta_i=\theta_{i1}-\theta_{i0}$，$\gamma=\log(\frac{\pi}{1-\pi})+\sum_{i=1}^{d}[A_i(\theta_{i0})-A_i(\theta_{i1})]$，则有：

$$P(Y=1|X=x)=\frac{1}{1+e^{-\log\frac{\pi}{1-\pi}(a_0-a_1)}} \tag{2.9}$$

$$=\frac{1}{1+e^{-(\sum_{i=1}^{d}\beta_iT_i(x_i)+\gamma)}} \tag{2.10}$$

记sigmoid函数$\mathcal{S}(x)=1/(1+e^{-x})$，则上式可化简为：

$$P(Y=1|X=x)=S(\sum_{i=1}^{d}\beta_iT_i(x_i)+\gamma) \tag{2.11}$$

同理可得：

$$P(Y=0|X=x)=S(\sum_{i=1}^{d}\beta_i{}^{'}T_i(x_i)+\gamma^{'}) \tag{2.12}$$

其中：

$$\beta_i{}^{'}=\theta_{i0}-\theta_{i1} \tag{2.13}$$

$$\gamma^{'}=\log(\frac{1-\pi}{\pi})+\sum_{i=1}^{d}[A_i(\theta_{i1})-A_i(\theta_{i0})] \tag{2.14}$$

令$P(Y=1|X=x)=P(Y=0|X=x)$，可获得分类边界为：

$$P(Y=1|X=x)=P(Y=0|X=x) \tag{2.15}$$

$$\pi e^{a_1}=(1-\pi)e^{a_0} \tag{2.16}$$

$$\pi e^{\sum_{i=1}^{d}\theta_{i1}T_i(x_i)-\sum_{i=1}^{d}A_i(\theta_{i1})}=(1-\pi)e^{\sum_{i=1}^{d}\theta_{i0}T_i(x_i)-\sum_{i=1}^{d}A_i(\theta_{i0})} \tag{2.17}$$

$$\sum_{i=1}^{d}(\theta_{i1}-\theta_{i0})T_i(x_i)=\log(\frac{1-\pi}{\pi})+\sum_{i=1}^{d}[A_i(\theta_{i1})-A_i(\theta_{i0})] \tag{2.18}$$

所以最后的分类边界就是：

$$\{x|\sum_{i=1}^{d}(\theta_{i1}-\theta_{i0})T_i(x_i)=\log(\frac{1-\pi}{\pi})+\sum_{i=1}^{d}[A_i(\theta_{i1})-A_i(\theta_{i0})]\} \tag{2.19}$$

(2)$P(X_i|Y=y)$服从高斯分布，则有：

$$P(X|Y,\mu,\Sigma)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)] \tag{2.20}$$

其中$\Sigma$是协方差矩阵，其定义如下：

$$\Sigma_{ij}=\begin{cases}\sigma_i{}^2 & i=j\\ \rho\sigma_i\sigma_j & i\neq j\end{cases} \tag{2.21}$$

其中$\rho$是两个不同分量的相关系数。设P(Y=1)=$\pi_1$，P(Y=0)=$\pi_0$，则有：

$$P(Y=1|X)=\frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)} \tag{2.22}$$

$$=\frac{\pi_1\exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]}{\pi_1\exp[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]+\pi_0\exp[-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)]} \tag{2.23}$$

记：

$$a_c=-\frac{1}{2}(x-\mu_c)^T\Sigma^{-1}(x-\mu_c) \quad c=1,2 \tag{2.24}$$

于是有：

$$P(Y=1|X)=\frac{\pi_1e^{a_1}}{\pi_1e^{a_1}+\pi_0e^{a_0}} \tag{2.25}$$

$$=\frac{1}{1+\frac{\pi_0}{\pi_1}e^{a_0-a_1}} \tag{2.26}$$

$$=\frac{1}{1+\exp[-\log(\frac{\pi_1}{\pi_0})+a_0-a_1]} \tag{2.27}$$

$$a_0-a_1=-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) \tag{2.28}$$

$$=-(\mu_1-\mu_0)^T\Sigma^{-1}x+\frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0) \tag{2.29}$$

所以：

$$P(Y=1|X=x)=\frac{1}{1+\exp[-\log(\frac{\pi_1}{\pi_0})+a_0-a_1]} \tag{2.30}$$

$$=S(\beta^Tx+\gamma) \tag{2.31}$$

其中：

$$\beta=\Sigma^{-1}(\mu_1-\mu_0) \tag{2.32}$$

$$\gamma=-\frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0)+\log(\frac{\pi_1}{\pi_0})=-\frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0)+\log(\frac{\pi}{1-\pi}) \tag{2.33}$$

同理可得：

$$P(Y=1|X=x)=S(\beta^{T'}x+\gamma') \tag{2.34}$$

其中：

$$\beta'=\Sigma^{-1}(\mu_0-\mu_1) \tag{2.35}$$

$$\gamma'=-\frac{1}{2}(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0+\mu_1)+\log(\frac{1-\pi}{\pi}) \tag{2.36}$$

故分类边界为：

$$P(Y=1|X=x)=P(Y=0|X=x) \tag{2.37}$$

$$(\beta^T-\beta^{T'})x=\gamma'-\gamma \tag{2.38}$$

$$\{x|(\beta^T-\beta^{T'})x=\gamma'-\gamma\} \tag{2.39}$$

# 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
Initialize $\mu_1, \ldots, \mu_k$.
repeat
    Step 1: Decide the class memberships of $\{\mathbf{x}_i\}_{i=1}^n$ by assigning each of them to
      its nearest cluster center.
        $$\gamma_{ij} = \begin{cases} 1, & ||\mathbf{x}_i - \mu_j||^2 \leq ||\mathbf{x}_i - \mu_{j'}||^2, \forall j' \\ 0, & \text{otherwise} \end{cases}$$
    Step 2: For each $j \in \{1, \cdots, k\}$, recompute $\mu_j$ using the updated $\gamma$ 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$\ell_2$-范数来度量距离(即欧式距离), 下面我们考虑使用$\ell_1$-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-$\ell_2$算法), 给出新的算法(命名为$k$-means-$\ell_1$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-$\ell_2$和$k$-means-$\ell_1$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.** 此处用于写解答(中英文均可)

(1)这里可以将指示矩阵$\gamma$的表示形式进行转化，以方便说明。将$\gamma$表示成当前集合的划分$X^{(t)}$，包括k个子集${X_1}^{(t)},\cdots,{X_k}^{(t)}$，每个簇的中心为：$\mu_1,\cdots,\mu_k$，则欧式距离可表示成：

$$D(X^{(t)})=\sum_{j=1}^{k}\sum_{i=1}^{n}||\mathbf{x}_i-\mu_j||^2 \tag{3.3}$$

设第t轮的指示函数为$A^{(t)}$，则此时$x_i$所属的簇为${A_{x_i}}^{(t)}$。

Step1的作用是根据远近得出每个样本点所属的簇，所以Step1之后，每个样本点离其所属簇的中心点的距离是其到每个中心点距离中最近的，所以有：

$$D(X^{(t)})=\sum_{j=1}^{k}\sum_{i=1}^{n}||\mathbf{x}_i-{\mu_j}^{(t)}||^2\geq\sum_{j=1}^{k}\sum_{i=1}^{n}||\mathbf{x}_i-\mu^{(t)}_{A^{(t+1)}_{(X_i)}}||^2 \tag{3.4}$$

Step2是重新计算每个簇的中心，将簇中所有样本点的平均值作为新的中心，这显然也会降低目标函数J的值，注意到如下等式：

$$(x_1-y)^2+(x_2-y)^2\geq 2(x_1-y)(x_2-y) \tag{3.5}$$

上面等号成立当且仅当$y=\frac{x_1+x_2}{2}$，所以有：

$$D(X^{(t)})=\sum_{j=1}^{k}\sum_{i=1}^{n}||\mathbf{x}_i-\mu^{(t)}_{A^{(t+1)}_{(X_i)}}||^2\geq\sum_{j=1}^{k}\sum_{i=1}^{n}||\mathbf{x}_i-{\mu_j}^{(t+1)}||^2=D(X^{(t+1)}) \tag{3.6}$$

所以Step1和Step2都能使J的值降低，故每轮迭代，J的值都会降低。

(2)每一轮结束后会开始下一轮当且仅当这轮的J值比上一轮小。所以要证明算法会在有限步停止，只要证不存在对原集合无限的划分即可。这里的聚类问题可以看成将n个不同的球放入k个不同的盒子，且盒子不能为空，总的放法为$k!S(n,k)$，其中S(n,k)为第二类斯特林数。由此可以看出划分的方法是有限的，因此算法不可能一直进行，一定会在有限步内停止。

(3)利用归纳法证明：假设当k≤t时，J的最小值关于k非增，现在任取一点作为一个新的cluster的中心，算法此时未停止，所以只要证下一轮的J值会更小，就证明了J的最小值是关于k的非增函数。注意到被选作中心的那个点，这个点在J中的项就变成了0，因此J的最小值在加入

新的中心后会下降。这就证明了J的最小值是关于k的非增函数。

(4)由题中所给的变量可得：

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}W_j(X)+nB(X)=\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}||x_i-\mu_j||^2+\gamma_{ij}||\mu_j-\hat{x}||^2 \tag{3.7}$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}(||x_i-\mu_j||^2+||\mu_j-\hat{x}||^2) \tag{3.8}$$

$$=\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}({x_i}^2+\hat{x}^2-2x_i\hat{x}+2x_i\hat{x}+2{\mu_j}^2-2x_i\mu_j-2\hat{x}\mu_j) \tag{3.9}$$

$$=\sum_{j=1}^{k}(\sum_{i=1}^{n}\gamma_{ij}(||x_i-\hat{x}||^2))+R \tag{3.10}$$

$$=n\sum_{i=1}^{n}(||x_i-\hat{x}||^2)+R \tag{3.11}$$

$$=n^2T(X)+R \tag{3.12}$$

上式中R表示(3.9)中其余各项之和。

上式中$n^2T(X)$是一个常数，表示数据集总的偏差，$\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}W_j(X)$即为题中的目标函数J，是算法运行中不断最小化的量，因此可以看出，nB(X)在算法中被近似最大化，说它近似是因为这个等式中的R这里看成是一个不变量，而实际上R是会变化的。

(5)k-means-l1算法（更准确地说叫k-median）如下：

```
Algorithm 2: k-meidan Algorithm
Initialize $\mu_1, \ldots, \mu_k$.
repeat
    Step 1: Decide the class memberships of $\{\mathbf{x}_i\}_{i=1}^{n}$ by assigning each of them to its nearest cluster center.
        $\gamma_{ij}=\begin{cases}1, & ||\mathbf{x}_i-\mu_j||_1 \le ||\mathbf{x}_i-\mu_{j'}||_1, \forall j' \\ 0, & \text{otherwise}\end{cases}$
    Step 2: For each $j \in \{1, \cdots, k\}$, recompute $\mu_j$ using the updated $\gamma$:
        $\mu_j = median\{x_j|\gamma_{ij}=1\}$
until the objective function $J^{'}$ no longer changes;
```

若样本集中存在少量异常点，那么使用k-means-l1的效果会更好，因为这种算法的簇的中心是该簇所有数据点的“中位数”，中位数受极端情况的影响比平均数小。

# 4 [50pts] Kernel, Optimization and Learning

给定样本集$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, $\mathcal{F} = \{\Phi_1 \cdots, \mathbf{\Phi}_d\}$为非线性映射族。考虑如下的优化问题

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

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

(1) **[40pts]** 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) **[10pts]** 考虑在优化问题4.2中, 当$p = 1$时, 试化简该问题。

**Solution.** 此处用于写解答(中英文均可)

(1)原式中的损失函数为hinge损失，为了简化表达，可以引入松弛变量$\xi_i \geq 0$，这样原问题就可以化归到如下优化问题：

$$\min_{\mathbf{w},\mu\in\Delta_q} \quad \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2 + C\sum_{i=1}^{m}\xi_i \tag{4.3}$$

$$s.t. \quad y_i\left(\sum_{k=1}^{d}\mathbf{w}_k\cdot\mathbf{\Phi}_k(\mathbf{x}_i)\right) \geq 1-\xi_i \tag{4.4}$$

$$\xi_i \geq 0 \quad i = 1, 2, \cdots, m \tag{4.5}$$

$$\|\boldsymbol{\mu}\|_q = 1 \tag{4.6}$$

对上式使用拉格朗日乘子法可得到其对偶问题。首先该问题的拉格朗日函数可写为：

$$L(\mathbf{w}, \mu, \xi, \alpha, \beta, \gamma) = \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i(1-\xi_i-y_i\left(\sum_{k=1}^{d}\mathbf{w}_k\cdot\mathbf{\Phi}_k(\mathbf{x}_i)\right)) - \sum_{i=1}^{m}\beta_i\xi_i + \gamma(\|\boldsymbol{\mu}\|_q - 1) \tag{4.7}$$

其中$\alpha_i \geq 0$，$\beta_i \geq 0$，$\gamma \geq 0$是拉格朗日乘子。

令$L(\mathbf{w},\mu,\xi,\alpha,\beta,\gamma)$对$\mathbf{w}$，$\mu$，$\xi$偏导为0可得：

$$\mathbf{w} = \sum_{i=1}^{m} \mu_k \alpha_i y_i \mathbf{\Phi}_k(\mathbf{x}_i) \tag{4.8}$$

$$\frac{\partial L}{\partial \mu_k} = 0 = \frac{\partial}{\partial \mu_k}(\frac{\|\mathbf{w}_k\|_2^2}{2\mu_k}) + \frac{\partial}{\partial \mu_k}[\gamma(\|\boldsymbol{\mu}\|_q - 1)] = -\frac{\|\mathbf{w}_k\|_2^2}{2{\mu_k}^2} + \gamma\frac{\partial}{\partial \mu_k}\|\boldsymbol{\mu}\|_q \tag{4.9}$$

$$= -\frac{\|\mathbf{w}_k\|_2^2}{2{\mu_k}^2} + \gamma\frac{1}{q}(\sum_i |\mu_i|^q)^{\frac{1}{q}-1} \cdot q|\mu_k|^{q-1} \cdot sgn(\mu_k) \quad (chain \quad rule) \tag{4.10}$$

$$= -\frac{\|\mathbf{w}_k\|_2^2}{2{\mu_k}^2} + \gamma\frac{1}{q}(\sum_i |\mu_i|^q)^{\frac{1}{q}-1} \cdot q|\mu_k|^{q-1} \quad (\mu_k > 0) \tag{4.11}$$

$$= -\frac{\|\mathbf{w}_k\|_2^2}{2{\mu_k}^2} + \gamma((\sum_i |\mu_i|^q)^{-\frac{1}{q}})^{q-1} \cdot |\mu_k|^{q-1} \tag{4.12}$$

$$= -\frac{\|\mathbf{w}_k\|_2^2}{2{\mu_k}^2} + \gamma(\frac{|\mu_k|}{\|\boldsymbol{\mu}\|_q})^{q-1} = -\frac{\|\mathbf{w}_k\|_2^2}{2{\mu_k}^2} + \gamma(\mu_k)^{q-1} \tag{4.13}$$

$$C = \alpha_i + \beta_i \tag{4.14}$$

将上面各式带入式(4.7)即可得原问题的对偶问题：

$$\frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i\left(\sum_{k=1}^{d}\mathbf{w}_k \cdot \mathbf{\Phi}_k(\mathbf{x}_i)\right)) - \sum_{i=1}^{m}\beta_i\xi_i \tag{4.15}$$

$$=\frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}{w_k}^T w_k + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m}\alpha_i y_i(\sum_{k=1}^{d}\mathbf{w}_k \cdot \mathbf{\Phi}_k(\mathbf{x}_i)) - \sum_{i=1}^{m}\beta_i\xi_i \tag{4.16}$$

$$=\frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}{w_k}^T w_k + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i y_i(\sum_{k=1}^{d}\mathbf{w}_k \cdot \mathbf{\Phi}_k(\mathbf{x}_i)) \tag{4.17}$$

$$=\alpha^T I + \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}{w_k}^T w_k - \sum_{i=1}^{m}\alpha_i y_i(\sum_{k=1}^{d}\mathbf{w}_k \cdot \mathbf{\Phi}_k(\mathbf{x}_i)) \tag{4.18}$$

$$=\alpha^T I + \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\sum_i \mu_k\alpha_i y_i \mathbf{\Phi}_k(\mathbf{x}_i)^T \sum_j \mu_k \alpha_j y_j \mathbf{\Phi}_k(\mathbf{x}_j) - \sum_{i=1}^{m}\alpha_i y_i(\sum_{k=1}^{d}\mu_k\sum_{j=1}^{m}\alpha_j y_j \mathbf{\Phi}_k(\mathbf{x}_j)^T \mathbf{\Phi}_k(\mathbf{x}_i)) \tag{4.19}$$

$$=\alpha^T I + \frac{1}{2}\sum_{k=1}^{d}\mu_k\alpha^T Y^T \mathbf{\Phi}_k(\mathbf{x}_i)^T \mathbf{\Phi}_k(\mathbf{x}_j) Y\alpha - \sum_{k=1}^{d}\mu_k\alpha^T Y^T \mathbf{\Phi}_k(\mathbf{x}_i)^T \mathbf{\Phi}_k(\mathbf{x}_j) Y\alpha \tag{4.20}$$

$$=\alpha^T I - \frac{1}{2}\sum_{k=1}^{d}\mu_k\alpha^T Y^T \mathbf{\Phi}_k(\mathbf{x}_i)^T \mathbf{\Phi}_k(\mathbf{x}_j) Y\alpha \tag{4.21}$$

$$=\alpha^T I - \frac{1}{2}\sum_{k=1}^{d}\mu_k\alpha^T Y^T K_k Y\alpha \tag{4.22}$$

为了继续推导，这里使用Lp空间中一条基本Hölder不等式：
设$\frac{1}{p}+\frac{1}{q}=1$，令$a_1,\cdots,a_n$，$b_1,\cdots,b_n$是非负实数，那么：

$$\sum_{i=1}^{n}a_i b_i \leq (\sum_{i=1}^{n}{a_i}^p)^{\frac{1}{p}}(\sum_{i=1}^{n}{b_i}^q)^{\frac{1}{q}} \tag{4.23}$$

等号成立条件为：

$$\exists c_1, c_2 \in R, c_1{>}0, c_2{>}0, \quad s.t. \quad c_1 {a_k}^p = c_2 {b_k}^q \quad k \in N$$

对式(4.22)中的$\sum_{k=1}^{d} \mu_k \alpha^T Y^T K_k Y \alpha$使用Hölder不等式，令$\alpha^T Y^T K_k Y \alpha = Z_k$，则有：

$$\begin{aligned}
\sum_{k=1}^{d} \mu_k \alpha^T Y^T K_k Y \alpha &= \sum_{k=1}^{d} \mu_k Z_k && (4.24)\\
&\leq (\sum_{k=1}^{d} {\mu_k}^q)^{\frac{1}{q}} \cdot (\sum_{k=1}^{d} {Z_k}^p)^{\frac{1}{p}} && (4.25)\\
&= \|\boldsymbol{\mu}\|_q \cdot \|Z\|_p && (4.26)\\
&= \|Z\|_p && (4.27)
\end{aligned}$$

注意到$Z_k = \alpha^T Y^T K_k Y \alpha = \frac{1}{{\mu_k}^2} \|\mathbf{w}_k\|_2^2$，所以，利用式(4.13)可以得到如下结果：

$$\begin{aligned}
\gamma(\frac{|\mu_k|}{\|\boldsymbol{\mu}\|_q})^{q-1} &= \frac{\|\mathbf{w}_k\|_2^2}{2{\mu_k}^2} && (4.28)\\
(\gamma(\frac{|\mu_k|}{\|\boldsymbol{\mu}\|_q})^{q-1})^p &= (\frac{\|\mathbf{w}_k\|_2^2}{2{\mu_k}^2})^p && (4.29)\\
(\gamma(\frac{1}{\|\boldsymbol{\mu}\|_q})^{q-1})^p \cdot ((\mu_k)^{q-1})^p &= (\frac{\|\mathbf{w}_k\|_2^2}{2{\mu_k}^2})^p && (4.30)\\
(\gamma(\frac{1}{\|\boldsymbol{\mu}\|_q})^{q-1})^p \cdot ((\mu_k)^{q-1})^p &= (\frac{1}{2})^p \cdot (\frac{\|\mathbf{w}_k\|_2^2}{{\mu_k}^2})^p && (4.31)\\
(\gamma(\frac{1}{\|\boldsymbol{\mu}\|_q})^{q-1})^p \cdot ((\mu_k)^{q-1})^p &= (\frac{1}{2})^p \cdot (Z_k)^p && (4.32)
\end{aligned}$$

从上式可以看出由于$(\gamma(\frac{1}{\|\boldsymbol{\mu}\|_q})^{q-1})^p{>}0$，$(\frac{1}{2})^p{>}0$，所以Hölder不等式取等的条件成立，于是有：

$$\sum_{k=1}^{d} \mu_k \alpha^T Y^T K_k Y \alpha = \sum_{k=1}^{d} \mu_k Z_k = \|Z\|_p \qquad (4.33)$$

其中，$Z_k = \alpha^T Y^T K_k Y \alpha$

所以原问题的对偶问题即为：

$$\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad 2\alpha^T \mathbf{I} - \sum_{k=1}^{d} \|Z\|_p \quad &= \quad 2\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{I} - \left\| \begin{matrix} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_1 \mathbf{Y} \boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_d \mathbf{Y} \boldsymbol{\alpha} \end{matrix} \right\|_p && (4.34)\\
\text{s.t.} \quad \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C}
\end{aligned}$$

其中，$\mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C}$是因为要满足$C = \alpha_i + \beta_i$，$\beta_i \geq 0$和$\alpha_i \geq 0$。

(2)当p=1时，按照p范数的定义，我觉得原式可以化为：

$$\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & 2\alpha^T \mathbf{I} - \sum_{k=1}^{d} \alpha^T Y^T K_k Y \alpha && (4.35)\\
\text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} && (4.36)
\end{aligned}$$