机器学习导论
综合能力测试

141220132, 银琦, 141220132@smail.nju.edu.cn

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)多项式分布为：$p(x) = \frac{\Gamma(\sum_{i=1}^n x_i+1)}{\prod_{i=1}^n \Gamma(x_i+1)}\prod_{i=1}^n \alpha_i^{x_i}$，该式可变形为：

$$\begin{aligned}
p(x) &= \frac{\Gamma(\sum_{i=1}^n x_i + 1)}{\prod_{i=1}^n \Gamma(x_i+1)} exp(log\prod_{i=1}^n \alpha_i^{x_i}) && (1.3)\\
&= \frac{\Gamma(\sum_{i=1}^n x_i + 1)}{\prod_{i=1}^n \Gamma(x_i+1)} exp(\sum_{i=1}^n x_i log\alpha_i) && (1.4)\\
&= \frac{\Gamma(\sum_{i=1}^n x_i + 1)}{\prod_{i=1}^n \Gamma(x_i+1)} exp((log\alpha_1 \quad log\alpha_2 \quad \cdots \quad log\alpha_n)\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} - 0) && (1.5)\\
&= h(x)exp\left(\boldsymbol{\theta}^{\mathbf{T}} T(x) - A(\boldsymbol{\theta})\right) && (1.6)
\end{aligned}$$

其中$h(x) = \frac{\Gamma(\sum_{i=1}^{n} x_i+1)}{\prod_{i=1}^{n}\Gamma(x_i+1)}$，$\boldsymbol{\theta} = \begin{pmatrix} log\alpha_1 \\ log\alpha_2 \\ \vdots \\ log\alpha_n \end{pmatrix}$，$T(x) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$，$A(\boldsymbol{\theta}) = 0$

(2)多元高斯分布为：$p(x) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} e^{\frac{-(\mathbf{x}-\mu)^{\mathbf{T}}\Sigma^{-1}(\mathbf{x}-\mu)}{2}}$，该式可变形为：

$$
\begin{aligned}
p(x) &= \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} exp(\frac{-(\mathbf{x}-\mu)^{\mathbf{T}}\Sigma^{-1}(\mathbf{x}-\mu)}{2}) && (1.7) \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}} exp(-\frac{1}{2}log|\Sigma| - \frac{(\mathbf{x}-\mu)^{\mathbf{T}}\Sigma^{-1}(\mathbf{x}-\mu)}{2}) && (1.8) \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}} exp(-\frac{1}{2}log|\Sigma| - \frac{1}{2}\Sigma^{-1}\mathbf{x}\mathbf{x}^{\mathbb{T}} + \mu^{\mathbb{T}}\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu^{\mathbb{T}}\Sigma^{-1}\mu) && (1.9) \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}} exp((\Sigma^{-1}\mu \quad -\tfrac{1}{2}\Sigma^{-1})(\begin{matrix}\mathbf{x} \\ \mathbf{x}\mathbf{x}^{\mathbb{T}}\end{matrix}) - \frac{1}{2}log|\Sigma| - \frac{1}{2}\mu^{\mathbb{T}}\Sigma^{-1}\mu) && (1.10) \\
&= h(x) exp\left(\boldsymbol{\theta}^{\mathbf{T}} T(x) - A(\boldsymbol{\theta})\right) && (1.11)
\end{aligned}
$$

其中$h(x) = (2\pi)^{-\frac{D}{2}}$，$\boldsymbol{\theta} = \begin{pmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1} \end{pmatrix}$，$T(x) = \begin{pmatrix} \mathbf{x} \\ \mathbf{x}\mathbf{x}^{\mathbb{T}} \end{pmatrix}$，$A(\boldsymbol{\theta}) = \frac{1}{2}log|\Sigma| + \frac{1}{2}\mu^{\mathbb{T}}\Sigma^{-1}\mu$。

(3)首先，假设单个观测的概率遵循指数族，使用其自然参数进行参数化：

$$
p_F(x|\boldsymbol{\theta}) = h(x)\exp\left(\boldsymbol{\theta}^{\mathrm{T}} T(x) - A(\boldsymbol{\theta})\right) \tag{1.12}
$$

对于数据$\mathcal{D} = (x_1, x_2, \cdots, x_n)$，似然计算如下：

$$
p(\mathcal{D}|\boldsymbol{\theta}) = \left(\prod_{i=1}^{n} h(x_i)\right) exp\left(\boldsymbol{\theta}^{\mathbb{T}} \sum_{i=1}^{n} T(x_i) - nA(\boldsymbol{\theta})\right) \tag{1.13}
$$

对于式1.2先验分布：

$$
p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi} - \nu A(\boldsymbol{\theta})\right) \propto \exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi} - \nu A(\boldsymbol{\theta})\right) \tag{1.14}
$$

所以计算后验如下：

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\chi}, \nu) &\propto p(\mathcal{D}|\boldsymbol{\theta}) p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi}, \nu) && (1.15) \\
&= \left(\prod_{i=1}^{n} h(x_i)\right) exp\left(\boldsymbol{\theta}^{\mathbb{T}} \sum_{i=1}^{n} T(x_i) - nA(\boldsymbol{\theta})\right) f(\boldsymbol{\chi}, \nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi} - \nu A(\boldsymbol{\theta})\right) && (1.16) \\
&\propto exp\left(\boldsymbol{\theta}^{\mathbb{T}} \sum_{i=1}^{n} T(x_i) - nA(\boldsymbol{\theta})\right) \exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi} - \nu A(\boldsymbol{\theta})\right) && (1.17) \\
&\propto exp\left(\boldsymbol{\theta}^{\mathbb{T}} \left(\sum_{i=1}^{n} T(x_i) + \boldsymbol{\chi}\right) - (n+\nu)A(\boldsymbol{\theta})\right) && (1.18)
\end{aligned}
$$

即：

$$
p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\chi}, \nu) = p_\pi\left(\boldsymbol{\theta}| \sum_{i=1}^{n} T(x_i) + \boldsymbol{\chi}, n+\nu\right) \tag{1.19}
$$

所以后验与先验具有相同的形式。

注：参考资料：`https://en.wikipedia.org/wiki/Exponential_family`

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)由题可知：

$$\Pr(X_i = x_i|Y = 0) = h_i(x_i)\exp(\theta_{i0} \cdot T_i(x_i) - A_i(\theta_{i0})) \tag{2.1}$$
$$\Pr(X_i = x_i|Y = 1) = h_i(x_i)\exp(\theta_{i1} \cdot T_i(x_i) - A_i(\theta_{i1})) \tag{2.2}$$

因为$p(Y = 0|x) = 1 - p(Y = 1|x)$所以要计算$\Pr(Y|X)$只需计算$p(Y = 1|x)$即可，由贝叶斯定理：

$$\begin{aligned}
\Pr(Y = 1|X = x) &= \frac{\Pr(X = x|Y = 1)\Pr(Y = 1)}{\Pr(X = x)} &(2.3)\\
&= \frac{\sum_{i=1}^d \Pr(X_i = x_i|Y = 1)\Pr(Y = 1)}{\sum_{i=1}^d \Pr(X_i = x_i|Y = 1)\Pr(Y = 1) + \sum_{i=1}^d \Pr(X_i = x_i|Y = 0)\Pr(Y = 0)} &(2.4)\\
&= \frac{\sum_{i=1}^d \Pr(X_i = x_i|Y = 1)\pi}{\sum_{i=1}^d \Pr(X_i = x_i|Y = 1)\pi + \sum_{i=1}^d \Pr(X_i = x_i|Y = 0)(1-\pi)} &(2.5)\\
&= \frac{1}{1 + \frac{\sum_{i=1}^d \Pr(X_i=x_i|Y=0)}{\sum_{i=1}^d \Pr(X_i=x_i|Y=1)}\frac{1-\pi}{\pi}} &(2.6)\\
&= \frac{1}{1 + \exp\left(log\left(\frac{\sum_{i=1}^d \Pr(X_i=x_i|Y=0)}{\sum_{i=1}^d \Pr(X_i=x_i|Y=1)}\frac{1-\pi}{\pi}\right) + log\left(\frac{1-\pi}{\pi}\right)\right)} &(2.7)
\end{aligned}$$

将式2.1、2.2代入式2.7可得：

$$\begin{aligned}
\Pr(Y = 1|X = x) &= \frac{1}{1 + e^{\sum_{i=1}^d(\theta_{i0}\cdot T_i(x_i) - A_i(\theta_{i0})) - \sum_{i=1}^d(\theta_{i1}\cdot T_i(x_i) - A_i(\theta_{i1})) + log\left(\frac{1-\pi}{\pi}\right)}} &(2.8)\\
&= \frac{1}{1 + e^{\sum_{i=1}^d(\theta_{i0} - \theta_{i1})T_i(x_i) - \sum_{i=1}^d(A_i(\theta_{i0}) - A_i(\theta_{i1})) + log\left(\frac{1-\pi}{\pi}\right)}} &(2.9)\\
&= \mathcal{S}\left(\sum_{i=1}^d(\theta_{i1} - \theta_{i0})T_i(x_i) - \sum_{i=1}^d(A_i(\theta_{i1}) - A_i(\theta_{i0})) + log\left(\frac{\pi}{1-\pi}\right)\right) &(2.10)
\end{aligned}$$

于是，

$$\Pr(Y=0|X=x) = 1-\Pr(Y=1|X=x) \tag{2.11}$$

$$= \mathcal{S}\left(\sum_{i=1}^{d}(\theta_{i0}-\theta_{i1})T_i(x_i) - \sum_{i=1}^{d}(A_i(\theta_{i0})-A_i(\theta_{i1})) + log\left(\frac{1-\pi}{\pi}\right)\right) \tag{2.12}$$

对于分类边界：$P(Y=1|X=x) = P(Y=0|X=x)$

所以令式2.10与式2.12相等，有：

$$\sum_{i=1}^{d}(\theta_{i0}-\theta_{i1})T_i(x_i) - \sum_{i=1}^{d}(A_i(\theta_{i0})-A_i(\theta_{i1})) + log\left(\frac{1-\pi}{\pi}\right) = 0 \tag{2.13}$$

综上，后验概率分布：

$$\Pr(Y|X) = \begin{cases} \mathcal{S}\left(\sum_{i=1}^{d}(\theta_{i0}-\theta_{i1})T_i(x_i) - \sum_{i=1}^{d}(A_i(\theta_{i0})-A_i(\theta_{i1})) + log\left(\frac{1-\pi}{\pi}\right)\right) & Y=0 \\ \mathcal{S}\left(\sum_{i=1}^{d}(\theta_{i1}-\theta_{i0})T_i(x_i) - \sum_{i=1}^{d}(A_i(\theta_{i1})-A_i(\theta_{i0})) + log\left(\frac{\pi}{1-\pi}\right)\right) & Y=1 \end{cases}$$

分类边界：

$$\sum_{i=1}^{d}(\theta_{i0}-\theta_{i1})T_i(x_i) - \sum_{i=1}^{d}(A_i(\theta_{i0})-A_i(\theta_{i1})) + log\left(\frac{1-\pi}{\pi}\right) = 0 \tag{2.14}$$

(2)高斯分布：$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$，由于$P(X_i|Y=y)$服从高斯分布，所以：

$$P(X_i=x_i|Y=y) = \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(\frac{-(x_i-\mu_{iy})^2}{2\sigma_i^2}\right) \tag{2.15}$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-log\sigma_i - \frac{x_i^2}{2\sigma_i^2} + \frac{\mu_{iy}x_i}{\sigma_i^2} - \frac{\mu_{iy}^2}{2\sigma_i^2}\right) \tag{2.16}$$

$$= h_i(x_i)\exp\left(\theta_{iy}^{\mathbb{T}}\cdot T_i(x_i) - A(\theta_{iy})\right) \tag{2.17}$$

其中$h_i(x_i) = \frac{1}{\sqrt{2\pi}}$，$\theta_{iy} = \begin{pmatrix} \frac{\mu_{iy}}{\sigma_i^2} \\ -\frac{1}{2\sigma_i^2} \end{pmatrix}$，$T_i(x_i) = \begin{pmatrix} x_i \\ x_i^2 \end{pmatrix}$，$A(\theta_{iy}) = \frac{\mu_{iy}^2}{2\sigma_i^2} + log\sigma_i$

将式2.17得到的$\theta_{iy}, T_i(x_i), A(\theta_{iy})$代入2.14中，可得到分类边界：

$$\sum_{i=1}^{d}\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}x_i - \sum_{i=1}^{d}\frac{\mu_{i0}^2-\mu_{i1}^2}{2\sigma_i^2} + log\left(\frac{1-\pi}{\pi}\right) = 0 \tag{2.18}$$

从式2.18中可看出，分类边界与特征X显然是成线性的。

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.

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

  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. 此处用于写解答(中英文均可)

(1)由于在Step1之前我们重新指定了中心点，所以在Step1中，每个点被x_i重新划分了类别，且为离它最近的中心的类别，所以目标函数J的值降低了；由于重新修改了每个点所属的类别，所以在Step2中重新估计每个类的中心点的位置，假设重新计算之前的中心点为$\mu^{'}$，重新计算之后的中心点为μ，则：

$$\begin{aligned} \sum_{x\in\mathcal{D}} \|x-\mu^{'}\|^2 - \sum_{x\in\mathcal{D}} \|x-\mu\|^2 &= \sum_{x\in\mathcal{D}} (2x-\mu^{'}-\mu)(\mu-\mu^{'}) && (3.3) \\ &= |\mathcal{D}|(2\mu-\mu^{'}-\mu)(\mu-\mu^{'}) && (3.4) \\ &= |\mathcal{D}|\|\mu-\mu^{'}\|^2 && (3.5) \end{aligned}$$

显然式3.5大于等于0，即：$\sum_{x\in\mathcal{D}} \|x-\mu^{'}\|^2 > \sum_{x\in\mathcal{D}} \|x-\mu\|^2$，所以重新计算中心点后目标函数$J$的值降低了。

(2)由(1)可知，目标函数J的值在算法的每轮迭代中是单调递减的，所以对于指示矩阵γ，不可能出现两种相同的状态，否则J就与之前某次的值一样，与单调递减矛盾，而γ的大小是$n\times k$，且γ_{ij}的值只可能为0或1，所以指示矩阵的状态共有$2^{n\times k}$个，即算法最多进行$2^{n\times k}$步，所以算法会在有限步内停止。

换个角度来看，共有k个簇，n个点，所以点的分类情况有k^n，同样的，每种情况最多只出现一次，所以算法最多进行k^n步，即会在有限步停止。

(3)若$k>n$，那么每个点各自作为自己的中心点，此时目标函数J的值始终为0；若$k<n$，当有k个簇时，记目标函数J的最小值为$Min(k)$，显然$J(\gamma,\mu_1,\ldots,\mu_k)\geq Min(k)$，此时增加一个簇，从样本集$\mathcal{D}$中选取一个原来不是中心点的点$s$作为新簇的中心点，即$s=\mu_{k+1}$，此时$Min(k)\geq J(\gamma,\mu_1,\ldots,\mu_{k+1})$，而$J(\gamma,\mu_1,\ldots,\mu_{k+1})\geq Min(k+1)$，即$Min(k)\geq Min(k+1)$，所以目标函数$J$的最小值是关于$k$的非增函数。

(4)观察三个式子的形式，计算如下：

$$
\begin{aligned}
\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}W_j(X)+nB(X) &= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}\|\mathbf{x_i}-\mu_j\|^2+\sum_{j=1}^{k}\sum_{i=1}^{n}\|\mu_j-\hat{\mathbf{x}}\|^2 && (3.6)\\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}(\|\mathbf{x_i}-\mu_j\|^2+\|\mu_j-\hat{\mathbf{x}}\|^2) && (3.7)\\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}({\mathbf{x_i}}^2+\mu_j^2-2\mathbf{x_i}\mu_j+\mu_j^2+\hat{\mathbf{x}}^2-2\mu_j\hat{\mathbf{x}}) && (3.8)\\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}({\mathbf{x_i}}^2+\hat{\mathbf{x}}^2-2\mathbf{x_i}\hat{\mathbf{x}}+2\mathbf{x_i}\hat{\mathbf{x}}+2\mu_j^2-2\mathbf{x_i}\mu_j-2\mu_j\hat{\mathbf{x}}) && (3.9)\\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}(\|\mathbf{x_i}-\hat{\mathbf{x}}\|^2+2\mathbf{x_i}\hat{\mathbf{x}}+2\mu_j^2-2\mathbf{x_i}\mu_j-2\mu_j\hat{\mathbf{x}}) && (3.10)\\
&= n\sum_{i=1}^{n}(\|\mathbf{x_i}-\hat{\mathbf{x}}\|^2)+\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}(2\mathbf{x_i}\hat{\mathbf{x}}+2\mu_j^2-2\mathbf{x_i}\mu_j-2\mu_j\hat{\mathbf{x}}) && (3.11)\\
&= n^2T(X)+\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}(2\mathbf{x_i}\hat{\mathbf{x}}+2\mu_j^2-2\mathbf{x_i}\mu_j-2\mu_j\hat{\mathbf{x}}) && (3.12)
\end{aligned}
$$

令$A=\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}(2\mathbf{x_i}\hat{\mathbf{x}}+2\mu_j^2-2\mathbf{x_i}\mu_j-2\mu_j\hat{\mathbf{x}})$，则可得到等式：

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}W_j(X)+nB(X)=n^2T(X)+A \tag{3.13}$$

可以看出，$\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}W_j(X)=J(\gamma,\mu_1,\dots,\mu_k)$，所以$k-means$聚类算法可以认为是在最小化intra-cluster deviation的加权平均，由于$n^2T(X)$为定值，但是存在其它求和项A，所以当最小化intra-cluster deviation的加权平均时，近似最大化inter-cluster deviation。

(5)

i.

Algorithm 2: k-means-ℓ_1 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 μ_j using the updated γ to be the center of mass of all points in C_j:

$$\mu_j = median(x_j | \gamma_{ij} = 1)$$

until *the objective function J no longer changes*;

该算法中，Step1显然是对公式3.2的优化；

Step2中，对于每一个簇j，都要求$\sum_{i=1}^m \|\mathbf{x}_i - \mu_j\|_1$最小，该式对$\mu$求导，得到的值为1或者-1，上式最小在导数为0处取得，此时1和-1的数量一样多，所以μ取中位数可以保证这一点，即μ 取当前簇中的中位数时，使得目标函数J'最小，所以该算法是对公式3.2的优化。

ii.当样本集中存在少量异常点时，应该采用k-means-ℓ_1算法，即k-means-ℓ_1算法具有更好的鲁棒性。k-means-ℓ_1更稳定是因为它的新中心点是当前簇下所有点中最中间位置的一个，不考虑他们之间的距离的远近，即使异常点离簇中心很远，也不会影响簇中心的选择，减小了异常点的影响。

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4.15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4.15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)对式4.1，引入松弛变量$\xi \geq 0$，可将式4.1重写为：

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

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

$$\xi_i \geq 0, i = 1, 2, 3, \cdots, m \tag{4.5}$$

$$\|\mu\|_q - 1 = 0 \tag{4.6}$$

通过拉格朗日乘子法可得到拉格朗日函数为：

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

其中$\alpha_i \geq 0, \lambda_i \geq 0, t \geq 0$是拉格朗日乘子。

令$L(\mathbf{w}, \alpha, \xi, \lambda)$对$\mathbf{w}, \xi$的偏导为零可得：

$$\|\mathbf{w}_k\|_2 = \sum_{i=1}^{m} \alpha_i y_i \mu_k \mathbf{\Phi}_k(\mathbf{x}_i) \tag{4.8}$$

$$C = \alpha_i + \lambda_i \tag{4.9}$$

$$\frac{1}{2} \sum_{k=1}^{d} \|w_k\|_2^2 = t \sum_{k=1}^{d} \frac{\mu_k |\mu_k|^q}{\|\boldsymbol{\mu}\|_q^{q-1}} \tag{4.10}$$

将4.8,4.9,4.10代入4.7可得到：

$$
\begin{aligned}
&2\sum_{i=1}^{m}\alpha_i-\sum_{k=1}^{d}\sum_{i=1}^{m}\sum_{j=1}^{m}\mu_k\alpha_i\alpha_j y_i y_j\mathbf{\Phi}_k(\mathbf{x}_i)\mathbf{\Phi}_k(\mathbf{x}_j) &(4.11)\\
\leq&2\sum_{i=1}^{m}\alpha_i-(\sum_{k=1}^{d}\sum_{i=1}^{m}\sum_{j=1}^{m}\mu_k^q)^{\frac{1}{q}}(\sum_{k=1}^{d}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i\alpha_j y_i y_j\mathbf{K}_k)^p)^{\frac{1}{p}} &(4.12)\\
=&2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}-\|\boldsymbol{\mu}\|_q\left\|\begin{matrix}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha}\\ \vdots\\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha}\end{matrix}\right\|_p &(4.13)\\
=&2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}-\left\|\begin{matrix}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha}\\ \vdots\\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha}\end{matrix}\right\|_p &(4.14)
\end{aligned}
$$

即对偶问题为：

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}}\quad & 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}-\left\|\begin{matrix}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha}\\ \vdots\\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha}\end{matrix}\right\|_p\\
\text{s.t.}\quad & \mathbf{0}\leq\boldsymbol{\alpha}\leq\mathbf{C}
\end{aligned}
\tag{4.15}
$$

(2)当$p=1$时，$q=\infty$，此时对偶问题为：

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}}\quad & 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}-\sum_{k=1}^{d}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_k\mathbf{Y}\boldsymbol{\alpha}\\
\text{s.t.}\quad & \mathbf{0}\leq\boldsymbol{\alpha}\leq\mathbf{C}
\end{aligned}
\tag{4.16}
$$