机器学习导论
综合能力测试

151220023, 段建辉, djhbarca@163.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)
设多项分布的参数有n个$(\theta_1, \theta_2, \theta_3, \ldots, \theta_n)$，其中

$$\theta_i = p(x_i = i|\theta), p(x_i = k|\theta) = 1 - \sum_{i=1}^{k-1}\theta_i$$

那么我们有了多项式分布为:

$$f_X(x|\theta) = P_X(x|\theta) = \prod_{k=1}^{n}\theta_k^{x_k}$$

现在就是要给出每个x_i的概率，我们可以定义指示函数如下：$\mathbb{I}(\cdot)$:

$$\mathbb{I}(y = j) = \begin{cases} 1 & 若y等于j \\ 0 & 若y不等于j \end{cases}$$

设$T(x)$是一个$k-1$维的向量，那么我们有：

$$\begin{aligned} T(x)_i &= \mathbb{I}(x=i) \\ \text{期望为：}E[T(x)_i] &= p(x_i=i)=\theta_i \end{aligned}$$

那么我们有：

$$\begin{aligned} P_X(x|\theta) &= \theta_1^{\mathbb{I}(x_1=1)}\theta_2^{\mathbb{I}(x_2=2)}\theta_3^{\mathbb{I}(x_3=3)}\ldots\theta_k^{\mathbb{I}(x_k=k)} \\ &= \theta_1^{T(x)_1}\theta_2^{T(x)_2}\theta_3^{T(x)_3}\ldots\theta_k^{1-\sum_{i=1}^{k-1}T(x)_i} \\ &= \exp(T(x)_1\ln\theta_1+T(x)_2\ln\theta_2+T(x)_3\ln\theta_3+\cdots+(1-\sum_{i=1}^{k-1}T(x)_i)\ln\theta_k) \\ &= \exp(T(x)_1\ln\frac{\theta_1}{\theta_k}+T(x)_2\ln\frac{\theta_2}{\theta_k}+T(x)_3\ln\frac{\theta_3}{\theta_k}+\cdots+T(x)_{k-1}\ln\frac{\theta_{k-1}}{\theta_k}+\ln\theta_k) \\ &= \exp(\mathbf{T}(x)\ln\frac{\Theta}{\theta_k}-\ln\theta_k) \\ &= h(x)\exp\left(\eta(\theta)\cdot T(x)-A(\theta)\right) \end{aligned}$$

其中，Θ、 $\mathbf{T}(x)$为向量，θ_k为常数：

$$\begin{aligned} h(x) &= 1 \\ \eta(\theta) &= \ln\frac{\Theta}{\theta_k} \\ T(x) &= \mathbf{T}(x) \\ A(\theta) &= -\ln\theta_k \end{aligned}$$

那么多项式分布是属于指数分布族的。

(2)

对于多元高斯分布，我们有：

$$\begin{aligned} f_X(x|\theta) &= \frac{\exp\left(-\frac{1}{2}(x-\theta)^2\right)}{\sqrt{2\pi}} \\ &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)\exp(\theta x-\frac{1}{2}\theta^2) \end{aligned}$$

其中：

$$\begin{aligned} h(x) &= \left(\frac{1}{\sqrt{2\pi}}\right)\exp\left(-\frac{x^2}{2}\right) \\ \eta(\theta) &= \theta \\ T(x) &= x \\ A(\theta) &= \frac{1}{2}\theta^2 \end{aligned}$$

那么高斯分布同样属于指数分布族。

(3)

考虑到独立同分布采样，我们有$x=(x_1,x_2,\ldots,x_N)$，那么我们有似然为：

$$p(x|\boldsymbol{\theta})=\left(\prod_{n=1}^{N}h(x_n)\right)\exp\left(\eta^T\left(\sum_{n=1}^{N}T(x_n)\right)-NA(\eta)\right)$$

将先验带入似然之后可以得到：

$$p_\pi(\boldsymbol{\theta}|x,\boldsymbol{\chi},\nu)\propto\exp\left(\left(\boldsymbol{\chi}+\sum_{n=1}^{N}T(x_n)\right)^T\boldsymbol{\theta}-(\nu+N)A(\boldsymbol{\theta})\right)$$

先验和后验具有相同的形式，因为对于先验与后验的形式来说其中：

$$\begin{aligned} f(\boldsymbol{\chi},\nu) &\rightarrow 1 \\ \boldsymbol{\chi} &\rightarrow \boldsymbol{\chi}+\sum_{n=1}^{N}T(x_n) \\ \nu &\rightarrow \nu+N \end{aligned}$$

所以先验和后验从公式上具有相同的形式。根据贝叶斯公式，我们有对于样本空间中的任意样本x_i有：

$$\begin{aligned} P(\theta|x_i) &= \frac{P(\theta x_i)}{P(x_i)} \\ &= \frac{P(x_i|\theta)p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi},\nu)}{P(x_i)} \\ &= \frac{h(x_i)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}T(x_i)-A(\boldsymbol{\theta})\right)f(\boldsymbol{\chi},\nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi}-\nu A(\boldsymbol{\theta})\right)}{P(x_i)} \\ &= \frac{h(x_i)f(\boldsymbol{\chi},\nu)\exp(\boldsymbol{\theta}^T(T(x_i)+\boldsymbol{\chi})-(1+\nu)A(\boldsymbol{\chi}))}{P(x_i)} \end{aligned}$$

这样$P(x_i)$可以根据θ的先验分布进行积分，是常数，且其中的$h(x_i)$、$T(x_i)$都为已知的，所以上式与先验分布的形式相同,证毕。

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)
由于标记服从Bernoulli分布，属于指数分布族，那么我们有后验概率分布为：

$$\begin{aligned}
P(Y = y|X = x) &\propto P(X = x|y = y)P(Y = y) \\
&= \prod_{i=1}^{d} P(X_i = x_i|Y = y)P(Y = y) \\
&= \prod_{i=1}^{d} h_i(x_i)\exp\left(\theta_{iy}\mathbf{T}_i(x_i) - A_i(\theta_{iy})\right)\pi^y(1-\pi)^{1-y} \\
&= \left(\prod_{i=1}^{d} h_i(x_i)\right)\exp\left(\sum_{i=1}^{d}(\theta_{iy}\mathbf{T}_i(x_i) - A_i(\theta_{iy})\right)\pi(1-\pi)^{1-y}
\end{aligned}$$

这样就可以计算后验概率，并区分出最后的Decision Boundary。首先设如下变量：

$$\phi_k = \left(\prod_{i=1}^{d} h_i(x_i)\right)\exp\left(\sum_{i=1}^{d}(\theta_{ik}\mathbf{T}_i(x_i) - A_i(\theta_{ik})\right)$$

其中，k为分类的Label，$k \in 0, 1$

那么计算分类边界：

$$\begin{aligned}
P(Y = 1|X = x) &= \frac{\pi\phi_1}{\pi\phi_1 + (1-\pi)\phi_0} \\
&= \frac{1}{1 + \frac{(1-\pi)\phi_0}{\pi\phi_1}} \\
&= \sigma\left(\sum_{i=1}^{d}(\theta_{i1} - \theta_{i0})\mathbf{T}_i(x_i) - \sum_{i=1}^{d}(A_i(\theta_{i1}) - A_i(\theta_{i0})) + \ln\frac{\pi}{1-\pi}\right)
\end{aligned}$$

根据题目中的条件，分类边界为$\{x\in\mathcal{X}:P(Y=1|X=x)=P(Y=0|X=x)\}$，那么可以得到$P(Y=1|X=x)=\frac{1}{2}$，将其带入得到：

$$\sum_{i=1}^{d}(\theta_{i1}-\theta_{i0})\mathbf{T}_i(x_i)=\sum_{i=1}^{d}(A_i(\theta_{i1})-A_i(\theta_{i0}))-\ln\frac{\pi}{1-\pi}$$

即为分类边界。

(2)

$P(X_i=x_i|Y=y)$服从高斯分布，那么有：

$$\begin{aligned}P(X_i=x_i|Y=y)&=\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\frac{-(x_i-\mu_{iy})^2}{2\sigma_i^2}\\&=\frac{1}{\sqrt{2\pi}}\exp\left(-\ln\sigma_i-\frac{x_i^2}{2\sigma_i^2}+\frac{\mu_{iy}x_i}{\sigma_i^2}-\frac{\mu_{iy}^2}{2\sigma_i^2}\right)\\&=h_i(x_i)\exp(\eta(\theta_{iy})T_i(x_i)-A_i(\theta_{iy}))\end{aligned}$$

其中的变量为：

$$\begin{aligned}\theta_{iy}&=[\mu_{iy},\theta_i^2]^T\\h_i(x_i)&=\frac{1}{\sqrt{2\pi}}\\\eta(\theta_{iy})&=\left[\frac{\mu_{iy}}{\sigma_i^2},-\frac{1}{2\sigma_i^2}\right]^T\\A_i(\theta_{iy})&=\frac{\mu_{iy}^2}{2\sigma_i^2}+\ln\sigma_i\\T_i(x_i)&=[x_i,x_i^2]\end{aligned}$$

那么将上述的分类边界公式中的θ替换为这里的η，并带入上式可以得到：

$$\begin{aligned}\sum_{i=1}^{d}\left([\frac{\mu_{i1}-\mu_{i0}}{\sigma_i^2},0]^T\right)[x_i,x_i^2]&=\sum_{i=1}^{d}\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}-\ln\frac{\pi}{1-\pi}\\\sum_{i=1}^{d}\frac{\mu_{i1}-\mu_{i0}}{\sigma_i^2}x_i&=\sum_{i=1}^{d}\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}-\ln\frac{\pi}{1-\pi}\end{aligned}$$

显然此时为线性关系，综上得证。

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.

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

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)

针对**Step 1**：

因为第一步是确定集合中每一个向量最近的cluster，通过使得$||\mathbf{x}_i - \mu_j||^2$最小来更新$\gamma$。因此针对$\gamma$一定是可以使得x分类至一个目前欧氏距离最小的cluster。

目标函数J中的$||\mathbf{x}_i - \mu_j||^2$，这一步找到的对应$\gamma_{ij}$向量为1的值一定是最小的欧氏距离，因为其满足：

$$\gamma_{ij} = ||\mathbf{x}_i - \mu_j||^2 \leq ||\mathbf{x}_i - \mu_{j'}||^2, \forall j'$$

目标函数求和也只是针对γ向量的$\gamma_{ij} = 1$，此时更新的γ定为目前迭代轮次最优分类向量，因此求和时可以使得目标函数J的值降低。

针对**Step 2**：

第二步是重新计算这一轮通过**Step 1**所产生新的cluster的质心。由公式可以知道：

$<1>$ $\sum_{i=1}^{n} \gamma_{ij}\mathbf{x}_i$相当于对所有目前分类距离的cluster的向量距离求和。

$<2>$ $\sum_{i=1}^{n} \gamma_{ij}$分母的式子说明最后该计算表达式的值是取所有cluster的均值并更新μ_j

这样最后的μ_j是现在新cluster的质心，根据三角不等式我们可以知道，cluster中其余向量到质心之和为所有距离之和最短的。即我们最后可以最小化$||\mathbf{x}_i - \mu_j||^2$，因此同(1)我们最后可以使得目标函数$J$当前迭代轮次最小化。

(2)

从Algorithm 1可以发现，每次迭代都是朝向缩小聚类的欧氏距离方向逼近，是一个严格坐标下降的算法。并且γ只有n^2个有限状态，这样就无法在更新γ的时候出现相同的γ值。接下来对μ求偏导并令结果为0可以得到：

$$\mu_i = \frac{1}{N_i} \sum_{i=1}^{n} \mathbf{x}_i$$

其中,N_i为μ_i所在cluster元素个数

保持μ不变最小化J关于γ的函数，然后再保持γ不变最小化J关于μ的函数。因此，J是单调递减的，它的函数值一定收敛, 也就是当前cluster的均值就是当前方向的最优解，所以每

一次迭代目标测度函数都会减小，又因为其有下界则一定最后可以到达收敛的局部范围之内。
因为目标函数不是凸函数，那么只能确保找到局部最优解，但是最后也一定会收敛到这个最优解。即为算法1会在有限步内停止。

(3)
要想使得目标函数取得最小，那么就是选择最好的聚类中心，于是将目标函数对μ求偏导：

$$J_{\mu}^{'} = \frac{\partial J}{\partial \mu_j} = 2\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}||\mathbf{x}_i - \mu_j||$$

令其为零得到：

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

其中,N_j为μ_i所在cluster元素个数

那么最优解的μ即为对单个cluster进行均值求解，样本距离cluster中心服从高斯分布。将上述结果代入目标函数J之后可以得到最小值表达式：

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

因为$\sum_{j=1}^{k}\gamma_i = 1$，也即只有一个cluster被$\mathbf{x}_i$分类。
那么随着k值的增加，cluster数目增加，若果函数目前没有收敛，那么k值增加之后，目标函数最小值公式中N_j会整体变小，$\frac{1}{N_j}$会随之整体变大，那么 $\mathbf{x}_i - \frac{1}{N_j}\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}\mathbf{x}_i$就会减小。
如果此时已经收敛，k增加之后，目标函数依旧会减小。
如果想要目标函数最小，也就是k一直增加直至$k = n$的时候，此时$J = 0$，此时很大可能会过拟合。
因此整体上目标函数最小值不是随着k增加的一个递增函数，而是一个对于k而言的非增函数。

(4)

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}W_j(X) + nB(X) &= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}||x_i - \mu_j||^2 + \gamma_{ij}||\mu_j - \hat{x}||^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}(||x_i - \mu_j||^2 + ||\mu_j - \hat{x}||^2) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}(x_i^2 + \hat{x}^2 - 2x_i\hat{x} + 2x_i\hat{x} + 2\mu_j^2 - 2x_i\mu_j - 2\hat{x}\mu_j) \\
&= \sum_{j=1}^{k}\left(\sum_{i=1}^{n}\gamma_{ij}(||x_i - \hat{x}||^2)\right) + REM
\end{aligned}$$

其中，REM是同一cluster中剩余的总和

那么就有：

$$\begin{aligned}\sum_{j=1}^{k}\left(\sum_{i=1}^{n}\gamma_{ij}(||x_i-\hat{x}||^2)\right)+REM &= n\sum_{i=1}^{n}(||x_i-\hat{x}||^2)+REM \\ &= n^2T(X)+REM\end{aligned}$$

那么就有三个式子的关系为：

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

因为

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

那么根据前一问的证明我们有，随着迭代次数的增加，k-means算法就是在不断最小化目标函数J的过程，所以也是不断最小化$\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}W_j(X)$,所以也是在不断最小化$W_j(x)$，即最小化intra-cluster deviation。
又因为$n^2T(X)+REM$是个常数，那么因为不断最小化$\sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}W_j(X)$，因此对于$nB(X)$而言就是不断在被最大化，即最大化inter-cluster deviation。

(5)

Algorithm 2: k-means Algorithm 2

1 Initialize $\mu_1,\ldots,\mu_k$.
2 **repeat**
3 **Step 1**: Determine γ breaking ties arbitrarily.

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

4 **Step 2**: For each $j\in\{1,\cdots,k\}$, recompute μ_j using the updated γ Remove j from C if $\sum_{i=1}^{n}\gamma_{ij}=0$.
5 Otherwise:

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

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

应该采用k-means-ℓ_2算法。
因为如果采用L1范数，每个样本的分类最后的值只有-1或者1,两种可能，当1与-1数目相同的时候导数为0。L1范数相当于是一维的分类，将样本分在中值的两边，然后根据当前样本数目在更新μ_j的时候寻找整个cluster的median，因此这个时候如果出现了异常点，对L1范数的一维median有着相对比较大的影响。

而对于L2范数而言，二维的欧氏距离是一个几何距离，寻找的是几何质心，因此这个时候如果出现少量的异常点，对于整个cluster的质心影响偏移不是很大。
因此，k-means-ℓ_2算法在有异常点的时候鲁棒性更强。

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

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

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.
(1)
我们将公式重写之后为:

$$\begin{aligned} \min_{\mathbf{w}, \mu \in \Delta_q} \quad & \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2 + C\sum_{i=1}^{m}\xi_i \\ s.t. \quad & y_i\left(\sum_{k=1}^{d}\mathbf{w}_k^T \cdot \mathbf{\Phi}_k(\mathbf{x}_i)\right) \geqslant 1 - \xi_i \\ & \xi_i \geqslant 0, i = 1, 2, 3, \ldots, m \\ & ||\mu||_q = 1 \end{aligned}$$

我们由拉格朗日乘子法可以得到:

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

其中，$\alpha \geqslant 0, \mu \geqslant 0$是朗格朗日乘子。
那么令$L(\mathbf{w}, \mu, \alpha, \xi, \beta)$对$\mathbf{w}, \xi, \beta$求偏导并令偏导为0之后可以得到:

$$\begin{aligned} \mathbf{w}_k &= \frac{\sum_{i=1}^{m}\alpha_i\left(y_i(\sum_{k=1}^{d}\Phi_k(\mathbf{x}_i))\right)}{\sum_{k=1}^{d}\frac{1}{\mu_k}} \\ C &= \frac{\sum_{i=1}^{m}(\beta_i + \alpha_i)}{m} \\ 0 &= \sum_{i=1}^{m}\xi_i \end{aligned}$$

将上述结果代回原式可以得到对偶问题：

$$\begin{aligned}\max_{\alpha,\mu} \quad & \sum_{i=1}^{m} \alpha_i \left(1 - \xi_i - y_i(\sum_{k=1}^{d}(\mathbf{w}_k^T - \frac{1}{2}) \cdot \mathbf{\Phi}_k(\mathbf{x}_i))\right) - \sum_{i=1}^{m} \beta_i \xi_i \\ s.t. \quad & \sum_{i=1}^{m} \xi_i = 0\end{aligned}$$