机器学习导论
综合能力测试

141220120, 徐世坚, xsj13260906215@gmail.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. 此处用于写证明(中英文均可)

(1)$P(x_1, x_2, ..., x_d|N, \boldsymbol{\mu}) = \frac{N!}{x_1!x_2!...x_d!}\prod_{i=1}^d \mu_i^{x_i}$

$= \frac{N!}{x_1!x_2!...x_d!}\exp(\sum_{i=1}^d x_i \ln\mu_i)$

$= \frac{1}{x_1!x_2!...x_d!}\exp(\sum_{i=1}^d x_i \ln\mu_i + \ln N!)$

令$\theta = (N, \boldsymbol{\mu})$，则

$h(\mathbf{x}) = \frac{1}{x_1!x_2!...x_d!}$, $\eta(\theta) = (\ln\boldsymbol{\mu}) = [\ln\mu_1, ..., \ln\mu_d]$，$T(\mathbf{x}) = \mathbf{x}$，$A(\theta) = -\ln N!$

$\therefore$多项分布属于指数分布族。

(2)$P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))$

$= (2\pi)^{-\frac{d}{2}}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}|)$

$= (2\pi)^{-\frac{d}{2}}\exp(-\frac{1}{2}(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}|)$

$= (2\pi)^{-\frac{d}{2}} \exp(-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - (\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\ln|\boldsymbol{\Sigma}|))$

$\therefore h(\mathbf{x}) = (2\pi)^{-\frac{d}{2}}$，$\eta(\theta) = [\Sigma^{-1}\boldsymbol{\mu}; -\frac{1}{2}\Sigma^{-1}]$，$T(\mathbf{x}) = [\mathbf{x}; \mathbf{x}\mathbf{x}^T]$，$A(\theta) = \frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} + \frac{1}{2}\ln|\Sigma|$

$\therefore$多元高斯分布属于指数分布。

$(3) P(\boldsymbol{\theta}|\mathbf{X}) \propto p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi}, \nu) f(\mathbf{X}|\boldsymbol{\theta})$

$= f(\boldsymbol{\chi}, \nu) \exp(\boldsymbol{\theta}^T\boldsymbol{\chi} - \nu A(\boldsymbol{\theta})) \prod_{i=1}^n f(x_i|\boldsymbol{\theta})$

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

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

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

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

$\therefore$后验与先验具有相同的形式。

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. 此处用于写解答(中英文均可)

(1)$Pr(Y=1|X=x) = \frac{Pr(X=x|Y=1)Pr(Y=1)}{Pr(X=x|Y=1)Pr(Y=1)+Pr(X=x|Y=0)Pr(Y=0)}$

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

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

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

同理可得：

$Pr(Y=0|X=x) = S(\sum_{i=1}^d(A_i(\theta_{i1})-A_i(\theta_{i0})) - \sum_{i=1}^d T_i(x_i)(\theta_{i1}-\theta_{i0}) - (\ln\pi-\ln(1-\pi)))$

为求分类边界，令$Pr(Y=1|X=x) = Pr(Y=0|X=x)$，得：

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

两边同取ln，得：

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

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

$\therefore$上式可以写成$\mathbf{w^T}T(\mathbf{x})+b=0$

(2)令$Pr(Y=1|X=x) = Pr(Y=0|X=x)$，得：

$\pi\prod_{i=1}^d(\frac{1}{\sqrt{2\pi}\sigma_i}\exp(-\frac{(x_i-\mu_{i1})^2}{2\sigma_i^2})) = (1-\pi)\prod_{i=1}^d(\frac{1}{\sqrt{2\pi}\sigma_i}\exp(-\frac{(x_i-\mu_{i0})^2}{2\sigma_i^2}))$

$\ln\frac{\pi}{1-\pi} + \sum_{i=1}^d[\ln\frac{1}{\sqrt{2\pi}\sigma_i} - \frac{(x_i-\mu_{i1})^2}{2\sigma_i^2}] = \sum_{i=1}^d[\ln\frac{1}{\sqrt{2\pi}\sigma_i} - \frac{(x_i-\mu_{i0})^2}{2\sigma_i^2}]$

$\ln\frac{\pi}{1-\pi} = \sum_{i=1}^d[\frac{(x_i-\mu_{i1})^2}{2\sigma_i^2} - \frac{(x_i-\mu_{i0})^2}{2\sigma_i^2}]$

$\sum_{i=1}^d\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}x_i - \sum_{i=1}^d\frac{\mu_{i0}^2-\mu_{i1}^2}{2\sigma_i^2} - \ln\frac{\pi}{1-\pi} = 0$

$\therefore$上式可以写成$\mathbf{w^T x}+b=0$

$\therefore$分类边界和X成线性关系。

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2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 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2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2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2(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)对任意的$\mathbf{x}_i$，设它原来属于第λ_i类，而在Step1中修改为第λ_i'类。则

$J^{'}(\gamma, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k) = \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}^{'}||\mathbf{x}_i - \boldsymbol{\mu}_j||^2$

$= \sum_{i=1}^{n}||\mathbf{x}_i - \boldsymbol{\mu}_{\lambda_i'}||^2$

$\leq \sum_{i=1}^{n}||\mathbf{x}_i - \boldsymbol{\mu}_{\lambda_i}||^2$

$= \sum_{i=1}^{n}\sum_{j=1}^{k}\gamma_{ij}||\mathbf{x}_i - \boldsymbol{\mu}_j||^2$

$= J(\gamma, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k)$

$\therefore$Step1使得目标函数J的值降低(非增)。

对于Step2，从J的表达式可知，它计算的是所有类的类内平方距离的和。所以考虑任意一个类C_j，它的类内平方距离为：

$\sum_{\mathbf{x}_i \in C_j}||\mathbf{x}_i - \boldsymbol{\mu}_j||^2$

为了使平方距离最小，对$\boldsymbol{\mu}_j$进行求导:

$\frac{\partial \sum_{\mathbf{x}_i \in C_j}||\mathbf{x}_i - \boldsymbol{\mu}_j||^2}{\partial \boldsymbol{\mu}_j}$

令偏导为0，可得：

$\sum_{\mathbf{x}_i \in C_j}\mathbf{x_i} = \sum_{\mathbf{x}_i \in C_j}\boldsymbol{\mu}_j$

得：$\boldsymbol{\mu}_j = \frac{1}{|C_j|}\sum_{\mathbf{x}_i \in C_j}\mathbf{x}_i = \frac{\sum_{i=1}^{n}\gamma_{ij}\mathbf{x}_i}{\sum_{i=1}^{n}\gamma_{ij}}$

即Step2中的调整就是最优的解

$\therefore$Step2使得目标函数J的值降低（非增）。

(2)因为有n个样本，k个类别，所以所有的可能的划分个数为k^n个。

算法每一轮迭代，如果做了调整，那么一定产生的是一个新的划分，该划分对应的目标函数比之前的都要小。而如果该轮迭代没有做调整，则目标函数的值不变，算法终止。

当算法终止时，所遍历的划分个数一定是一个有限值，且小于k^n。所以，算法会在有限步内停止。

(3)假设当前有k类，且算法已经停止，即当前的J的值为最小值。则当增加一个新的类时(增加一个新的$\boldsymbol{\mu}_{k+1}$)，算法会继续进行。

如果在Step1和Step2中没有发生变动，则目标函数J的值将不变。而如果Step1和Step2有进行调整，则由前面的结论可知，目标函数J的值将会降低。这样继续迭代得到新的最小值。

所以目标函数J的最小值是关于k的非增函数。同时可发现，当$k = n$时，J的值最小，为0，即每个样例自成一类，但此时的分类无意义。

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

$= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}(||\mathbf{x}_i - \boldsymbol{\mu}_j||^2 + ||\boldsymbol{\mu}_j - \hat{\mathbf{x}}||^2)$
$= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}(\mathbf{x}_i^T\mathbf{x}_i - 2\mathbf{x}_i^T\boldsymbol{\mu}_j + 2\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j - 2\boldsymbol{\mu}_j^T\hat{\mathbf{x}} + \mathbf{x}^T\mathbf{x})$
$= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}(\mathbf{x}_i^T\mathbf{x}_i - 2\mathbf{x}_i\hat{\mathbf{x}} + 2\mathbf{x}_i\hat{\mathbf{x}} - 2\mathbf{x}_i^T\boldsymbol{\mu}_j + 2\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j - 2\boldsymbol{\mu}_j^T\hat{\mathbf{x}} + \mathbf{x}^T\mathbf{x})$
$= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}||\mathbf{x}_i - \hat{\mathbf{x}}||^2 + \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}2(\mathbf{x}_i\hat{\mathbf{x}} - \mathbf{x}_i^T\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^T\boldsymbol{\mu}_j - \boldsymbol{\mu}_j^T\hat{\mathbf{x}})$
$= nT(X) + \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}2(\mathbf{x}_i\hat{\mathbf{x}} - \mathbf{x}_i^T\boldsymbol{\mu}_j + \boldsymbol{\mu}_j^T\boldsymbol{\mu}_j - \boldsymbol{\mu}_j^T\hat{\mathbf{x}})$
因为$\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}W_j(\mathbf{X}) = J$，而如果右边近似看成常数的话($n$和$T(X)$均为常数，变化的是最右边的求和部分)，则在最小化intra-cluster deviation的加权平均(即目标函数J)时，相应的$nB(\mathbf{X})$会增大，即近似最大化inter-cluster deviation.
(5)$k - means - \ell_1$算法如下：

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:

$$\mu_j = the \quad median \quad of \quad all \quad x_i \in Cluster_j$$

until *the objective function J no longer changes*;

应该采用k-means-ℓ_1算法。考虑一个类不幸分到了一个或多个异常点，则在计算簇的中心$\boldsymbol{\mu}$时，如果采用ℓ_2范数来度量距离，则计算得到的$\boldsymbol{\mu}$离最优的中心会产生较大的偏差，这些偏差在后面的迭代中会累积，最终导致分类结果很差；而如果采用ℓ_1范数来度量距离，则对中心的计算影响很小甚至没有影响，因为这时的中心是类中样本的中位数，距离对中心的计算影响不大，这样的话最终的效果会好很多。

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) **[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. 此处用于写解答(中英文均可)

(1)优化问题4.1的表达式中采用hinge损失。引入"松弛变量"$\epsilon_i\geq 0$，则优化问题4.1重写为：

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

引入拉格朗日乘子$\alpha_i\geq 0,\beta_i\geq 0,\gamma_i\geq 0,h\geq 0$得：

$L=\frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2+C\sum_{i=1}^{m}\epsilon_i+\sum_{i=1}^{m}\alpha_i(1-\epsilon_i-y_i(\sum_{k=1}^{d}\mathbf{w}_k\cdot\mathbf{\Phi}_k(\mathbf{x}_i)))-\sum_{i=1}^{m}\beta_i\epsilon_i-\sum_{k=1}^{d}\gamma_k\mu_k+h(\|\boldsymbol{\mu}\|_q-1)$

分别对$\mathbf{w}_k$, ϵ_i 和μ_k求导，得：

$\frac{\partial L}{\partial\mathbf{w}_k}=0\Rightarrow\frac{\mathbf{w}_k}{\mu_k}=\sum_{i=1}^{m}\alpha_i y_i\mathbf{\Phi}_k(\mathbf{x}_i)$

$\frac{\partial L}{\partial\epsilon_i}=0\Rightarrow C=\alpha_i+\beta_i$

$\frac{\partial L}{\partial\mu_k}=0\Rightarrow\gamma_k=h\cdot\mu_k^{q-1}-\frac{\|\boldsymbol{w}_k\|_2^2}{2\mu_k^2}$

将上面的三个式子带入拉格朗日函数，得：

$$L=\sum_{i=1}^{m}\alpha_i-\sum_{k=1}^{d}h\mu_k^q+h((\sum_{k=1}^{d}\mu_k^q)^{1/q}-1) \tag{4.4}$$

$\because \|\boldsymbol{\mu}\|_q = 1$
$\therefore (\sum_{i=1}^{d} \mu_k^q)^{1/q} = 1$
$\therefore \sum_{i=1}^{d} \mu_k^q = 1$且$(\sum_{i=1}^{d} \mu_k^q)^{1/p} = 1$
代入上式，得：
$L = \sum_{i=1}^{m} \alpha_i - h$
$= \sum_{i=1}^{m} \alpha_i - h \cdot (\sum_{i=1}^{d} \mu_k^q)^{1/p}$
$= \sum_{i=1}^{m} \alpha_i - (\sum_{i=1}^{d} h^p \mu_k^q)^{1/p}$
$\because h \cdot \mu_k^{q-1} - \frac{\|\boldsymbol{w}_k\|_2^2}{2\mu_k^2} = \gamma_k$
$\therefore h \cdot \mu_k^q - \frac{\|\boldsymbol{w}_k\|_2^2}{2\mu_k} = \gamma_k \mu_k = 0$ ($\gamma_k \mu_k = 0$由KKT条件得)
$\therefore h \cdot \mu_k^{q-1} = \frac{\|\boldsymbol{w}_k\|_2^2}{2\mu_k^2}$
$\therefore h^p \cdot (\mu_k^{q-1})^p = (\frac{\|\boldsymbol{w}_k\|_2^2}{2\mu_k^2})^p$
$\because \frac{1}{p} + \frac{1}{q} = 1$
$\therefore q = p(q-1)$
$\therefore h^p \cdot \mu_k^q = (\frac{\|\boldsymbol{w}_k\|_2^2}{2\mu_k^2})^p$
代入L中即得：
$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \left(\sum_{k=1}^{d} (\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{\Phi}_k(\mathbf{x}_i) \boldsymbol{\Phi}_k(\mathbf{x}_j))^p \right)^{1/p}$
上式整理可得原问题的对偶问题：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & 2\sum_{i=1}^{m} \alpha_i - 1 \cdot \left(\sum_{k=1}^{d} \left(\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{\Phi}_k(\mathbf{x}_i) \boldsymbol{\Phi}_k(\mathbf{x}_j) \right)^p \right)^{1/p} \\ \text{s.t.} \quad & 0 \leq \boldsymbol{\alpha} \leq \mathbf{C} \end{aligned} \tag{4.5}$$

将上式做一些形式上的化简，即得优化问题4.2：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & 2\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{1} - \left\| \begin{matrix} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_1 \mathbf{Y} \boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Y}^{\mathrm{T}} \mathbf{K}_d \mathbf{Y} \boldsymbol{\alpha} \end{matrix} \right\|_p \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \end{aligned} \tag{4.6}$$

(2)当$p = 1$，化简可得：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & 2\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{1} - \sum_{i=1}^{d} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{\Phi}_k(\mathbf{x}_i) \boldsymbol{\Phi}_k(\mathbf{x}_i) \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \end{aligned} \tag{4.7}$$