Learning And Mining from DatA
http://lamda.nju.edu.cn

机器学习导论

习题课

詹德川
zhandc@lamda.nju.edu.cn
2017.06.21

Outline

- HW-final
  - PS1 - Exponential Families
  - PS2 - Decision Boundary
  - PS3 - Theoretical Analysis of *k*-means algorithm
  - PS4 - Kernel, Optimization and Learning

PS1 - 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。

PS1 - Exponential Families

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。

Solution. (1)

$$\begin{aligned} f_X(\mathbf{x}|\boldsymbol{\theta}) &= \frac{\Gamma(\sum_{i=1}^k x_i + 1)}{\prod_{i=1}^k \Gamma(x_i+1)} \prod_{i=1}^k \theta_i^{x_i} \\ &= \frac{\Gamma(\sum_{i=1}^k x_i + 1)}{\prod_{i=1}^k \Gamma(x_i+1)} \exp(\log \prod_{i=1}^k \theta_i^{x_i}) \\ &= \frac{\Gamma(\sum_{i=1}^k x_i + 1)}{\prod_{i=1}^k \Gamma(x_i+1)} \exp(\sum_{i=1}^k x_i \log\theta_i) \end{aligned} \tag{1.3}$$

显然，多项分布属于指数分布族，其中$h(x) = \frac{\Gamma(\sum_{i=1}^k x_i+1)}{\prod_{i=1}^k \Gamma(x_i+1)}$, $\eta(\theta) = [\log\theta_1, \cdots, \log\theta_k]$, $T(\mathrm{x}) = \mathrm{x}$, $A(\theta) = 0$.

PS1 - Exponential Families

Learning And Mining from DatA
http://lamda.nju.edu.cn

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。

(2) **[10pts]** 试证明多元高斯分布(Multivariate Gaussian distribution)属于指数分布族。

Solution:

$$\begin{aligned}
f_X(\mathbf{x}|\boldsymbol{\Sigma},\boldsymbol{\mu}) &= \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\\
&= \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right)\\
&= \frac{1}{(2\pi)^{k/2}}\exp\left(\mathrm{vec}(-\frac{1}{2}\boldsymbol{\Sigma}^{-1})^T\mathrm{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}}) + \mathrm{vec}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^T\mathrm{vec}(\mathbf{x}^{\mathrm{T}}) - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\ln|\boldsymbol{\Sigma}|\right)
\end{aligned} \tag{1.4}$$

则令 $h(x) = \frac{1}{(2\pi)^{k/2}}$, $\eta(\boldsymbol{\Sigma},\boldsymbol{\mu}) = \begin{bmatrix} -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\end{bmatrix}$, $T(\mathbf{x}) = \begin{bmatrix}\mathbf{x}\mathbf{x}^{\mathrm{T}} \\ \mathbf{x}^{\mathrm{T}}\end{bmatrix}$, $A(\boldsymbol{\Sigma},\boldsymbol{\mu}) = \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\ln|\boldsymbol{\Sigma}|$

即可。

PS1 - Exponential Families

(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: 后验概率可以写为

$$\begin{aligned}p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi},\nu,\mathbf{X}) &\propto p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi},\nu)p(\mathcal{D}|\boldsymbol{\theta})=p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi},\nu)\prod_{i=1}^{N}p(x_i|\boldsymbol{\theta})\\ &=f(\boldsymbol{\chi},\nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\chi}-\nu A(\boldsymbol{\theta})\right)\prod_{i=1}^{N}h(x_i)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}\sum_{i=1}^{n}T(x_i)-nA(\boldsymbol{\theta})\right)\end{aligned} \tag{1.6}$$

化简后，可得

$$p_\pi(\boldsymbol{\theta}|\boldsymbol{\chi},\nu,\mathbf{X})\propto(\prod_{i=1}^{n}h(x_i))f(\boldsymbol{\chi},\nu)\exp\left(\boldsymbol{\theta}^{\mathrm{T}}(\boldsymbol{\chi}+\sum_{i=1}^{n}T(x_i))-(\nu+n)A(\boldsymbol{\theta})\right) \tag{1.7}$$

PS2 - Decision Boundary

Learning And Mining from DatA
http://lamda.nju.edu.cn

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$\{x \in \mathcal{X} : P(Y=1|X=x) = P(Y=0|X=x)\}$

PS2 - Decision Boundary

(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$\{x \in \mathcal{X} : P(Y = 1|X = x) = P(Y = 0|X = x)\}$

Solution.

(1)
$$\Pr(\mathbf{X}|Y = y) = (\prod^{d} h_i(x_i))\exp(\sum^{d}(\theta_{iy} \cdot T_i(x_i) - A_i(\theta_{iy}))) \tag{2.1}$$

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

$$\begin{aligned}\Pr(Y = 1|\mathbf{X}) &= \frac{\pi}{\left(\pi + (1-\pi)\exp(\sum_{i=1}^{d}((\theta_{i0} - \theta_{i1}) \cdot T_i(x_i) + A_i(\theta_{i1}) - A_i(\theta_{i0})))\right)} \\ \Pr(Y = 0|\mathbf{X}) &= \frac{1-\pi}{\left(1 - \pi + \pi \cdot \exp(\sum_{i=1}^{d}((\theta_{i1} - \theta_{i0}) \cdot T_i(x_i) + A_i(\theta_{i0}) - A_i(\theta_{i1})))\right)}\end{aligned} \tag{2.3}$$

PS2 - Decision Boundary

(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$\{x \in \mathcal{X} : P(Y = 1|X = x) = P(Y = 0|X = x)\}$

Solution.

分类边界:

$$\begin{aligned}
&\pi^2 \cdot \exp(\sum_{i=1}^{d}((\theta_{i1} - \theta_{i0}) \cdot T_i(x_i) + A_i(\theta_{i0}) - A_i(\theta_{i1}))) \\
&= (1-\pi)^2 \exp(\sum_{i=1}^{d}((\theta_{i0} - \theta_{i1}) \cdot T_i(x_i) + A_i(\theta_{i1}) - A_i(\theta_{i0})))
\end{aligned} \tag{2.4}$$

化简可得:

$$\ln\frac{\pi}{1-\pi} = \sum_{i=1}^{d}((\theta_{i0} - \theta_{i1}) \cdot T_i(x_i) + A_i(\theta_{i1}) - A_i(\theta_{i0})) \tag{2.5}$$

PS2 - Decision Boundary

Learning And Mining from DatA
http://lamda.nju.edu.cn

(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.

(2) 因为 $P(X_i|Y=y)$ 服从高斯分布，由问题1中(2)结论可知:

$$h_i(x_i)=\frac{1}{\sqrt{2\pi}} \quad \theta_{iy}=\begin{bmatrix}\frac{\mu_{iy}}{\sigma_i^2}\\ \frac{1}{-2\sigma_i^2}\end{bmatrix} \quad T_i(x_i)=\begin{bmatrix}x_i\\ x_i^2\end{bmatrix} \quad A_i(\theta_{iy})=\frac{\mu_{iy}^2}{2\sigma_i^2}+\ln\sigma_i \tag{2.6}$$

从而分类边界为:

$$\ln\frac{\pi}{1-\pi}=\sum_{i=1}^{d}((\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2})x_i+A_i(\theta_{i1})-A_i(\theta_{i0})) \tag{2.7}$$

可见与 x_i 呈线性。

PS3 - Theoretical Analysis of *k*-means

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.

PS3 - Theoretical Analysis of k-means

Learning And Mining from DatA
http://lamda.nju.edu.cn

最经典的k-means聚类算法流程如算法1中所示

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.

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

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*;

PS3 - Theoretical Analysis of k-means

Learning And Mining from DatA
http://lamda.nju.edu.cn

(1) [**10pts**] 试证明，在算法$\boxed{1}$中，**Step 1**和**Step 2**都会使目标函数J的值降低.

Solution.

(1) 在 **Step 1** 中，$\forall i$，令 $$\hat{j} = \min_{1 \le j \le k} ||\mathbf{x}_i - \mu_j||^2$$

又因为γ_i是指示向量，

$$\sum_{j=1}^{k} \gamma_{ij} ||\mathbf{x}_i - \mu_j||^2 \ge ||\mathbf{x}_i - \mu_{\hat{j}}||^2$$

故 $J(\gamma, \mu_1, \ldots, \mu_k)$若在第一步发生改变，一定下降。

在 **Step 2** 中，令 $X_{\in j} = \{\mathbf{x}_i | \gamma_{ij} = 1\}$代表在簇 j 中点的集合，$n_j = |X_{\in j}|$ 为该集合大小，$\bar{\mathbf{x}}$ 为该集合的均值。则$\forall j, \forall \mathbf{a}$:

$$\begin{aligned}
\sum_{x \in X_{\in j}} ||\mathbf{x} - \mathbf{a}||^2 &= \sum_{x \in X_{\in j}} (\mathbf{x}^{\mathrm{T}}\mathbf{x} + \mathbf{a}^{\mathrm{T}}\mathbf{a} - 2\mathbf{x}^{\mathrm{T}}\mathbf{a}) \\
&= \sum_{x \in X_{\in j}} \mathbf{x}^{\mathrm{T}}\mathbf{x} + n_j \mathbf{a}^{\mathrm{T}}\mathbf{a} - 2n_j \bar{\mathbf{x}}^T \mathbf{a}
\end{aligned} \tag{3.3}$$

而当 $\mathbf{a} = \bar{\mathbf{x}}$时上式取得最小值。故 $J(\gamma, \mu_1, \ldots, \mu_k)$ 若在第二步发生改变，一定下降。

PS3 - Theoretical Analysis of *k*-means

(2) **[10pts]** 试证明, 算法1会在有限步内停止。

Solution.

(2) 由于不同的 $J(\gamma, \mu_1, \ldots, \mu_k)$ 对应不同的γ ，且每次更新时 J均下降(不下降时终止), 故γ不会与先前重复。又由于 $\gamma \in \{0,1\}^{n\times k}$ 最多有 2^{kn} 种可能取值，所以算法会在有限步终止。

(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。

Solution.

(3) 令 $k = k_0$ 时取得 J 的最小值的指示矩阵 γ不变, $k = k_1$时将 μ_{k+1} 设为 $\mathcal{D}$ 中任意一个点, 则 J 的值必不上升。从而J 的最小值不上升。

PS3 - Theoretical Analysis of k-means

(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，同时最大化inter-cluster deviation.

PS3 - Theoretical Analysis of *k*-means

Learning And Mining from DatA
http://lamda.nju.edu.cn

Solution.

$$
\begin{aligned}
n_j W_j(X) + n B_j(X) &= \sum_{x \in X_{\in j}} \|\mathbf{x} - \mu_j\|^2 + n_j \|\mu_j - \hat{\mathbf{x}}\|^2 \\
&= \sum_{x \in X_{\in j}} \mathbf{x}^{\mathrm{T}}\mathbf{x} - n_j ||\boldsymbol{\mu}_j||^2 + n_j(||\boldsymbol{\mu}_j||^2 + ||\hat{\mathbf{x}}||^2 - 2\boldsymbol{\mu}_j^T \hat{\mathbf{x}}) \\
&= \sum_{x \in X_{\in j}} \mathbf{x}^{\mathrm{T}}\mathbf{x} + n_j ||\hat{\mathbf{x}}||^2 - 2n_i \boldsymbol{\mu}_j^T \hat{\mathbf{x}} \\
&= \sum_{x \in X_{\in j}} \|\mathbf{x} - \hat{\mathbf{x}}\|^2
\end{aligned}
\tag{3.4}
$$

从而:

$$
\sum_{j=1}^{k} \frac{n_j}{n} W_j(X) + B(X) = T(X) \tag{3.5}
$$

由于算法迭代过程中$T(X)$ 不变，而目标函数 J 的值下降, 即 $\sum_{j=1}^{k} \frac{n_j}{n} W_j(X)$ 的值下降, 所以 $B(X)$ 上升。所以可以认为 “是在最小化intra-cluster deviation的加权平均, 同时近似最大化inter-cluster deviation”。

PS3 - Theoretical Analysis of k-means

(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。

PS3 - Theoretical Analysis of *k*-means

Algorithm 2: k-means-ℓ_1 Algorithm

1 Initialize $\mu_1, \ldots, \mu_k$.

2 **repeat**

3 **Step 1:** Decide the class memberships:

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

4 **Step 2:** For each $j \in \{1, \cdots, k\}$, recompute μ_j using the updated γ :

$$\forall d, \mu_j[d] = \text{median of } \{\mathbf{x_i} | \gamma_i j = 1\}$$

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

当样本集中存在少数异常点时, k-median算法具有更好的鲁棒性。
因为采用mean时，与这些异常点最近的簇的中心点很可能受到较大影响，从而偏离本应在的中心, 而采用median时，该中心点受到异常点影响之后，不会发生太大变化。

PS4 - Kernel, Optimization and Learning

- 给定样本集$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, $\mathcal{F} = \{\Phi_1 \cdots, \Phi_d\}$为非线性映射族。

考虑如下的优化问题

$$\min_{\mathbf{w}_1,\cdots,\mathbf{w}_d;\boldsymbol{\mu}\in\Delta_q} \quad \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2 + C\sum_{i=1}^{m}\max\left\{0, 1 - y_i\left(\sum_{k=1}^{d}\mathbf{w}_k \cdot \Phi_k(\mathbf{x}_i)\right)\right\} \tag{4.1}$$

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$\boldsymbol{\Phi}_k$定义的核函数(kernel).

Solution.

(1)首先，引入松弛变量 $\{\epsilon_i\}_{i=1}^{m}$, 对(4.1)化简，可得：

$$\begin{aligned}
&\min_{\mathbf{w}_1,\cdots,\mathbf{w}_d;\boldsymbol{\mu}\in\Delta_q} \quad \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2 + C\sum_{i=1}^{m}\max\left\{0, 1-y_i\left(\sum_{k=1}^{d}\mathbf{w}_k\cdot\boldsymbol{\Phi}_k(\mathbf{x}_i)\right)\right\} \\
&= \min_{\boldsymbol{\mu}\in\Delta_q}\min_{\mathbf{w}_1,\cdots,\mathbf{w}_d} \quad \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2 + C\sum_{i=1}^{m}\max\left\{0, 1-y_i\left(\sum_{k=1}^{d}\mathbf{w}_k\cdot\boldsymbol{\Phi}_k(\mathbf{x}_i)\right)\right\} \qquad (4.3) \\
&= \min_{\boldsymbol{\mu}\in\Delta_q}\min_{\mathbf{w}_1,\cdots,\mathbf{w}_d;\epsilon_i\geq 1-y_i\left(\sum_{k=1}^{d}\mathbf{w}_k\cdot\Phi_k(\mathbf{x}_i)\right),\epsilon_i\geq 0} \quad \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|_2^2 + C\sum_{i=1}^{m}\epsilon_i
\end{aligned}$$

对4.3使用拉格朗日乘子法（先将$\boldsymbol{\mu}$看作常量），引入拉格朗日乘子$\boldsymbol{\alpha},\boldsymbol{\beta}$且对$\forall\alpha_i,\beta_i$ 有$\alpha_i\geq 0,\beta_i\geq 0$，于是可得拉格朗日函数为：

$$L(\mathbf{w},\epsilon,\alpha,\beta) = \frac{1}{2}\sum_{k=1}^{d}\frac{1}{\mu_k}\|\mathbf{w}_k\|^2 + C\sum_{i=1}^{m}\epsilon_i + \sum_{i=1}^{m}\alpha_i\left(1-y_i\left(\sum_{k=1}^{d}\mathbf{w}_k\cdot\Phi_k(\mathbf{x}_i)\right)-\epsilon_i\right) - \sum_{i=1}^{m}\beta_i\epsilon_i \qquad (4.4)$$

PS4 - Kernel, Optimization and Learning

对4.4式对$\mathbf{w}, \epsilon$ 求导，可得：

$$
\begin{aligned}
\nabla_{\mathbf{w}_k} L(\mathbf{w}, \epsilon, \alpha, \beta) &= \frac{\mathbf{w}_k}{\mu_k} - \sum_{i=1}^{m} \alpha_i y_i \Phi_k(\mathbf{x}_i) = 0 \\
\mathbf{w}_k &= \mu_k \sum_{i=1}^{m} \alpha_i y_i \Phi_k(\mathbf{x}_i) \qquad (4.5) \\
\nabla_{\epsilon_i} L(\mathbf{w}, \epsilon, \alpha, \beta) &= C - \alpha_i - \beta_i = 0 \\
C &= \alpha_i + \beta_i \qquad (4.6)
\end{aligned}
$$

将4.5,4.6 带入4.4中，有：

$$
\begin{aligned}
L(\mathbf{w}, \epsilon, \alpha, \beta) &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{k=1}^{d} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mu_k \Phi_k(\mathbf{x}_i) \Phi_k(\mathbf{x}_j) \\
&= \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{1} - \frac{1}{2} \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{y}^{\mathrm{T}} \left(\sum_{k=1}^{d} \mu_k \mathbf{K}_k \right) \mathbf{y} \boldsymbol{\alpha} \\
&= 2\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{1} - \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{y}^{\mathrm{T}} \left(\sum_{k=1}^{d} \mu_k \mathbf{K}_k \right) \mathbf{y} \boldsymbol{\alpha} \qquad (4.7)
\end{aligned}
$$

PS4 - Kernel, Optimization and Learning

因此4.3中问题可化为：

$$\min_{\boldsymbol{\mu}\in\Delta_q}\max_{0\leq\boldsymbol{\alpha}\leq\mathbf{C}} \quad 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}-\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{y}^{\mathrm{T}}\left(\sum_{k=1}^{d}\mu_k\mathbf{K}_k\right)\mathbf{y}\boldsymbol{\alpha} \tag{4.8}$$

由极大极小定理(Minimax theorem)，公式4.8 可被写作：

$$\begin{aligned}
&\min_{\boldsymbol{\mu}\in\Delta_q}\max_{0\leq\boldsymbol{\alpha}\leq\mathbf{C}} \quad 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}-\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{y}^{\mathrm{T}}\left(\sum_{k=1}^{d}\mu_k\mathbf{K}_k\right)\mathbf{y}\boldsymbol{\alpha}\\
=&\max_{0\leq\boldsymbol{\alpha}\leq\mathbf{C}}\min_{\boldsymbol{\mu}\in\Delta_q} \quad 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}-\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{y}^{\mathrm{T}}\left(\sum_{k=1}^{d}\mu_k\mathbf{K}_k\right)\mathbf{y}\boldsymbol{\alpha}\\
=&\max_{0\leq\boldsymbol{\alpha}\leq C} 2\boldsymbol{\alpha}^{\mathrm{T}}1-\max_{\mu\in\delta_q}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{y}^{\mathrm{T}}\left(\sum_{k=1}^{p}\mu_k\mathbf{K}_k\right)\mathbf{y}\boldsymbol{\alpha}
\end{aligned} \tag{4.9}$$

由赫尔德不等式，以及题中条件($\Delta_q=\{\boldsymbol{\mu}|\mu_k\geq 0,k=1,\cdots,d;\|\boldsymbol{\mu}\|_q=1\}$),可得：

$$\sum_{k=1}^{d}\mu_k\mathbf{K}_k\leq\|\mu\|_q\cdot\|[\mathbf{K}_1,\cdots,\mathbf{K}_d]\|_p=\|[\mathbf{K}_1,\cdots,\mathbf{K}_d]\|_p \tag{4.10}$$

因此该对偶问题最终形式为：

$$\max_{\boldsymbol{\alpha}} \quad 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \left\| \begin{matrix} \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha} \end{matrix} \right\|_p \tag{4.11}$$

$$\text{s.t.} \quad \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C}$$

证毕。

极大极小定理(Minimax theorem)

- Formally, **von Neumann's minimax theorem** states:

Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be compact convex sets. If $f : X \times Y \to \mathbb{R}$ is a continuous function that is convex-concave, i.e.

$f(\cdot, y) : X \to \mathbb{R}$ is convex for fixed y, and
$f(x, \cdot) : Y \to \mathbb{R}$ is concave for fixed x.

Then we have that

$$\min_{x \in X} \max_{y \in Y} f(x, y) = \max_{y \in Y} \min_{x \in X} f(x, y).$$

PS4 - Kernel, Optimization and Learning

Learning And Mining from DatA
http://lamda.nju.edu.cn

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

$$\min_{\mathbf{w}_1, \cdots, \mathbf{w}_d; \boldsymbol{\mu} \in \Delta_q} \quad \frac{1}{2} \sum_{k=1}^{d} \frac{1}{\mu_k} \|\mathbf{w}_k\|_2^2 + C \sum_{i=1}^{m} \max\left\{0, 1 - y_i \left(\sum_{k=1}^{d} \mathbf{w}_k \cdot \Phi_k(\mathbf{x}_i) \right) \right\} \tag{4.1}$$

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。

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \left\| \begin{matrix} \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_1\mathbf{Y}\boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{K}_d\mathbf{Y}\boldsymbol{\alpha} \end{matrix} \right\|_p \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \end{aligned} \tag{4.2}$$

(2) **[10pts]** 考虑在优化问题4.2中, 当$p = 1$时, 试化简该问题。

PS4 - Kernel, Optimization and Learning

(2)当$p=1$时，由共轭关系$\frac{1}{p}+\frac{1}{q}=1$, 可得$q=\infty$,于是 $\boxed{4.10}$,可写为：

$$\sum_{k=1}^{d} \mu_k \mathbf{K}_k \leq \|\boldsymbol{\mu}\|_\infty \cdot [\mathbf{K}_1, \cdots, \mathbf{K}_d]\|_1 \tag{4.12}$$

对题中条件$\Delta_q = \{\boldsymbol{\mu} | \mu_k \geq 0, k=1,\cdots,d; \|\boldsymbol{\mu}\|_q = 1\}$)，对于$q=\infty$的情况，当$\mu_k=1$时使公式$\boxed{4.12}$最大。因此，对偶问题可化为：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & 2\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{y}^{\mathrm{T}}(\sum_{k=1}^{d} \mathbf{K}_k)\mathbf{y}\boldsymbol{\alpha} \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \end{aligned} \tag{4.13}$$

Q & A

Thanks !