机器学习导论
作业四

151220097, 孙旭东, 248381185@qq.com

2018 年 5 月 29 日

1 [30pts] Kernel Methods

Mercer定理告诉我们对于一个二元函数$k(\cdot,\cdot)$，它是正定核函数当且仅当对任意N和$\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N$，它对应的核矩阵是半正定的. 假设$k_1(\cdot,\cdot)$和$k_2(\cdot,\cdot)$分别是关于核矩阵$K_1$和$K_2$的正定核函数. 另外，核矩阵$K$中的元素为$K_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)$. 请根据Mercer定理证明对应于以下核矩阵的核函数正定.

(1) **[10pts]** $K_3=a_1K_1+a_2K_2$, 其中$a_1,a_2\geq 0$.

(2) **[10pts]** $f(\cdot)$是任意实值函数，由$k_4(\mathbf{x},\mathbf{x}')=f(\mathbf{x})f(\mathbf{x}')$定义的$K_4$.

(3) **[10pts]** 由$k_5(\mathbf{x},\mathbf{x}')=k_1(\mathbf{x},\mathbf{x}')k_2(\mathbf{x},\mathbf{x}')$定义的$K_5$.

Solution.

(1) 要证明K_3对应的核函数正定，即证明K_3半正定。
由于K_1和K_2对应的核函数是正定核函数，所以K_1和K_2都是半正定矩阵，所以K_1和K_2是对称的，所以显然K_3也是对称的。
又因为K_1和K_2是半正定矩阵，所以还有：对任意非零向量$\mathbf{x}$，都有：

$$\mathbf{x}^{\mathrm{T}}K_1\mathbf{x}\geq 0 \tag{1.1}$$

$$\mathbf{x}^{\mathrm{T}}K_2\mathbf{x}\geq 0 \tag{1.2}$$

又因为$a_1,a_2\geq 0$，所以：

$$\mathbf{x}^{\mathrm{T}}a_1K_1\mathbf{x}\geq 0 \tag{1.3}$$

$$\mathbf{x}^{\mathrm{T}}a_2K_2\mathbf{x}\geq 0 \tag{1.4}$$

那么可以得到：

$$\mathbf{x}^{\mathrm{T}}K_3\mathbf{x}=\mathbf{x}^{\mathrm{T}}(a_1K_1+a_2K_2)\mathbf{x}=\mathbf{x}^{\mathrm{T}}a_1K_1\mathbf{x}+\mathbf{x}^{\mathrm{T}}a_2K_2\mathbf{x}\geq 0 \tag{1.5}$$

所以可得K_3是半正定矩阵，所以对应的核函数是正定的。

(2) 设K_{ij}表示K_4的i行，第j列的项。首先因为$K_{ij} = f(\mathbf{x}_i)f(\mathbf{x}_j) = f(\mathbf{x}_j)f(\mathbf{x}_i) = K_{ji}$，所以$K_4$是对称的。对于任意非零向量$\mathbf{x}$，用$x_i$表示$\mathbf{x}$的第$i$个项，可以有：

$$\mathbf{x}^{\mathrm{T}}K_4\mathbf{x} = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i K_{ij} x_j \tag{1.6}$$

又因为$k_4(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i)f(\mathbf{x}_j)$，所以为了方便证明令$t_i = f(\mathbf{x}_i)$，可得$K_{ij} = k_4(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i)f(\mathbf{x}_j) = t_i t_j$，结合上式可得：

$$\begin{aligned}\mathbf{x}^{\mathrm{T}}K_4\mathbf{x} &= \sum_{i=1}^{N}\sum_{j=1}^{N} x_i K_{ij} x_j = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i t_i t_j x_j \\ &= \sum_{i=1}^{N}(x_i t_i)^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} x_i t_i x_j t_j \\ &= (t_1 x_1 + t_2 x_2 + ... + t_n x_n)^2 \geq 0\end{aligned} \tag{1.7}$$

所以可得K_4是半正定矩阵，所以对应的核函数正定。

(3) 为了方便证明，令$A = K_1$，$B = K_2$，$C = K_5$。然后再定义A_{ij}表示A的第i行第j列的元素，即$A_{ij} = k_1(\mathbf{x}_i, \mathbf{x}_j)$，同理有$B_{ij} = k_2(\mathbf{x}_i, \mathbf{x}_j)$，$C_{ij} = k_5(\mathbf{x}_i, \mathbf{x}_j)$。根据题目定义可得$C_{ij} = A_{ij}B_{ij}$。

因为k_1和k_2都是正定核函数，所以A和B都是半正定矩阵。所以A和B都是对称的，所以得到$k_5(\mathbf{x}_i, \mathbf{x}_j) = k_1(\mathbf{x}_i, \mathbf{x}_j)k_2(\mathbf{x}_i, \mathbf{x}_j) = k_1(\mathbf{x}_j, \mathbf{x}_i)k_2(\mathbf{x}_j, \mathbf{x}_i) = k_5(\mathbf{x}_j, \mathbf{x}_i)$，所以$C$是实对称矩阵。又因为$B$是半正定矩阵，所以根据半正定矩阵的性质可知，通过使用*Cholesky*分解，一定存在N阶实矩阵L，满足$B = LL^{\mathrm{T}}$，使用L_{ij}表示L的每个元素，根据定义可得$B_{ij} = \sum_{k=1}^{N} L_{ik}L_{jk}$。对于任意非零向量$\mathbf{x}$，用$x_i$表示$\mathbf{x}$的第$i$个项，可以有：

$$\mathbf{x}^{\mathrm{T}}C\mathbf{x} = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i C_{ij} x_j \tag{1.8}$$

再根据以上定义，有

$$\begin{aligned}\mathbf{x}^{\mathrm{T}}C\mathbf{x} &= \sum_{i=1}^{N}\sum_{j=1}^{N} x_i C_{ij} x_j \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} x_i A_{ij} B_{ij} x_j \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} x_i A_{ij} x_j \sum_{k=1}^{N} L_{ik} L_{jk} \\ &= \sum_{k=1}^{N}\sum_{i=1}^{N}\sum_{j=1}^{N} (x_i L_{ik}) A_{ij} (x_j L_{jk})\end{aligned} \tag{1.9}$$

然后定义向量$\mathbf{y}_k = (x_1 L_{1k}, x_2 L_{2k}, \cdots, x_n L_{nk})$，所以有

$$\mathbf{x}^{\mathrm{T}}C\mathbf{x} = \sum_{k=1}^{N} \mathbf{y}_k^{\mathrm{T}} A \mathbf{y}_k \tag{1.10}$$

又因为A是半正定矩阵，所以有$\mathbf{y}_k^{\mathrm{T}} A \mathbf{y}_k \geq 0$，所以可得

$$\mathbf{x}^{\mathrm{T}} C \mathbf{x} = \sum_{k=1}^{N} \mathbf{y}_k^{\mathrm{T}} A \mathbf{y}_k \geq 0 \tag{1.11}$$

所以可得C是半正定矩阵，所以K_5是半正定矩阵，所以对应的核函数正定。

2 [25pts] SVM with Weighted Penalty

考虑标准的SVM优化问题如下(即课本公式(6.35)),

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \cdots, m. \end{aligned} \tag{2.1}$$

注意到，在(2.1)中，对于正例和负例，其在目标函数中分类错误的“惩罚”是相同的. 在实际场景中，很多时候正例和负例错分的“惩罚”代价是不同的. 比如考虑癌症诊断问题，将一个确实患有癌症的人误分类为健康人，以及将健康人误分类为患有癌症，产生的错误影响以及代价不应该认为是等同的.

现在，我们希望对负例分类错误的样本(即false positive)施加$k > 0$倍于正例中被分错的样本的“惩罚”. 对于此类场景下，

(1) **[10pts]** 请给出相应的SVM优化问题.

(2) **[15pts]** 请给出相应的对偶问题，要求详细的推导步骤，尤其是如KKT条件等.

Solution. 此处用于写解答*(*中英文均可*)*

(1) 定义$\mathbf{P}$表示正例的下标的集合，$\mathbf{N}$表示负例的下标的集合，相应的SVM优化问题为:

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i\in\mathbf{P}}^{m}\xi_i + Ck\sum_{i\in\mathbf{N}}^{m}\xi_i \\ s.t. \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \cdots, m. \end{aligned} \tag{2.2}$$

(2) 根据上式(2.2)，使用拉格朗日乘子法可得拉格朗日函数如下:

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) =& \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i\in\mathbf{P}}^{m}\xi_i + Ck\sum_{i\in\mathbf{N}}^{m}\xi_i \\ & + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b)) - \sum_{i=1}^{m}\mu_i\xi_i \end{aligned} \tag{2.3}$$

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

令$L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})$对$\mathbf{w}$，$b$，$\xi_i$的偏导为零可得

$$\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i \tag{2.4}$$

$$0 = \sum_{i=1}^{m}\alpha_i y_i \tag{2.5}$$

$$(\mathbb{I}(i \in \mathbf{P}) + k\mathbb{I}(i \in \mathbf{N}))C = \alpha_i + \mu_i \tag{2.6}$$

为了方便解答，这里用到了指示器函数，定义如下：

$$\mathbb{I}(\cdot)=\begin{cases}1 & \cdot\ is\ true\\ 0 & \cdot\ is\ false\end{cases} \tag{2.7}$$

现在将(2.4) − (2.6)代入式(2.3)即可得到式(2.2)的对偶问题，具体的：

$$\begin{aligned}L(\mathbf{w},b,\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\mu})=&\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i\in\mathbf{P}}^{m}\xi_i+Ck\sum_{i\in\mathbf{N}}^{m}\xi_i\\ &+\sum_{i=1}^{m}\alpha_i(1-\xi_i-y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b))-\sum_{i=1}^{m}\mu_i\xi_i\\ =&\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_jy_iy_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j+C\sum_{i\in\mathbf{P}}^{m}\xi_i+Ck\sum_{i\in\mathbf{N}}^{m}\xi_i\\ &+\sum_{i=1}^{m}\alpha_i-\sum_{i=1}^{m}\alpha_i\xi_i-\sum_{i=1}^{m}\alpha_i\alpha_jy_iy_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j-\sum_{i=1}^{m}\alpha_iy_ib-\sum_{i=1}^{m}\mu_i\xi_i\\ =&\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_jy_iy_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j+\sum_{i\in\mathbf{P}}^{m}(\alpha_i+\mu_i)\xi_i+\sum_{i\in\mathbf{N}}^{m}(\alpha_i+\mu_i)\xi_i\\ &+\sum_{i=1}^{m}\alpha_i-\sum_{i=1}^{m}\alpha_i\xi_i-\sum_{i=1}^{m}\alpha_i\alpha_jy_iy_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j-\sum_{i=1}^{m}\mu_i\xi_i\\ =&\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_jy_iy_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\end{aligned} \tag{2.8}$$

由KKT条件中的

$$\begin{aligned}&\alpha_i\geq 0\\ &\mu_i\geq 0\end{aligned} \tag{2.9}$$

可得约束条件

$$0\leq\alpha_i\leq C(\mathbb{I}(i\in\mathbf{P})+k\mathbb{I}(i\in\mathbf{N})) \tag{2.10}$$

综上可得，对偶问题为：

$$\begin{aligned}\max_{\alpha_i}\quad&\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_jy_iy_j\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\\ s.t.\quad&\sum_{i=1}^{m}\alpha_iy_i=0\\ &0\leq\alpha_i\leq C(\mathbb{I}(i\in\mathbf{P})+k\mathbb{I}(i\in\mathbf{N})),i=1,2,\cdots,m.\end{aligned} \tag{2.11}$$

KKT条件要求为

$$\begin{cases}\alpha_i\geq 0\\ \mu_i\geq 0\\ \xi_i\geq 0\\ y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b)-1+\xi_i\geq 0\\ \alpha_i(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b)-1+\xi_i)=0\\ \mu_i\xi_i=0\end{cases} \tag{2.12}$$

3 [30pts+10*pts] Nearest Neighbor

假设数据集$\{\mathbf{x}_1, ..., \mathbf{x}_n\}$是从一个以$\mathbf{0}$为中心的$p$维单位球中独立均匀采样而得到的$n$个样本点. 这个球可以表示为:

$$B = \{\mathbf{x} : \|\mathbf{x}\|^2 \le 1\} \subset \mathbb{R}^p. \tag{3.1}$$

其中，$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$，$\langle \mathbf{x}, \mathbf{x} \rangle$是$\mathbb{R}^p$空间中向量的内积. 在本题中，我们将探究原点$O$与其最近邻(1-NN)的距离$d^*$，以及这个距离$d^*$与$p$之间的关系. 在这里，我们将原点$O$以及其1-NN之间的距离定义为:

$$d^* \coloneqq \min_{1 \le i \le n} \|\mathbf{x}_i\|, \tag{3.2}$$

不难发现d^*是一个随机变量，因为$\mathbf{x}_i$是随机产生的.

(1) **[5pts]** 当$p = 1$且$t \in [0, 1]$时，请计算$\Pr(d^* \le t)$，即随机变量d^*的累积分布函数(Cumulative Distribution Function, **CDF**).

(2) **[10pts]** 请写出d^*的**CDF**的一般公式，即当$p \in \{1, 2, 3, ...\}$时d^*对应的取值. 提示：半径为r的p维球体积是：

$$V_p(r) = \frac{(r\sqrt{\pi})^p}{\Gamma(p/2 + 1)}, \tag{3.3}$$

其中，$\Gamma(1/2) = \sqrt{\pi}$，$\Gamma(1) = 1$，且有$\Gamma(x + 1) = x\Gamma(x)$对所有的$x > 0$成立；并且对于$n \in \mathbb{N}^*$，有$\Gamma(n + 1) = n!$.

(3) **[10pts]** 请求解随机变量d^*的中位数，即使得$\Pr(d^* \le t) = 1/2$成立时的t值. 答案是与n和p相关的函数.

(4) **[附加题10pts]** 请通过**CDF**计算使得原点O距其最近邻的距离d^*小于1/2的概率至少0.9的样本数n的大小. 提示：答案仅与p相关. 你可能会用到$\ln(1 - x)$的泰勒展开式：

$$\ln(1 - x) = -\sum_{i=1}^{\infty} \frac{x^i}{i}, \quad \text{for } -1 \le x < 1. \tag{3.4}$$

(5) **[5pts]** 在解决了以上问题后，你关于n和p以及它们对1-NN的性能影响有什么理解.

Solution. *此处用于写解答(中英文均可)*

(1) *因为$p = 1$，所以可得$d^* \coloneqq \min_{1 \le i \le n} x_i$。要计算$\Pr(d^* \le t)$，可以先计算$\Pr(d^* > t)$。定义事件$E_i$表示$x_i > t$，考虑到$d^* \coloneqq \min_{1 \le i \le n} x_i$，所以$d^* > t$等价于对任意$i = 1, 2, \cdots, n$，都有$x_i > t$。所以$Pr(d^* > t) = Pr(E_1 \wedge E_2 \wedge \cdots \wedge E_n)$，再考虑到$n$个样本独立均匀的采样得到的，所以：*

$$\begin{aligned} Pr(d^* > t) =& Pr(E_1 \wedge E_2 \wedge \cdots \wedge E_n) \\ =& Pr(E_1)Pr(E_2) \cdots Pr(E_n) \\ =& Pr(x_1 > t)Pr(x_2 > t) \cdots Pr(x_n > t) \\ =& (1 - t)^n \end{aligned} \tag{3.5}$$

所以可得

$$Pr(d^* \le t) = 1 - Pr(d^* > t) = 1 - (1 - t)^n \tag{3.6}$$

(2) 先考虑$Pr(\|\mathbf{x}_i\| \le t)$。根据独立均匀采样，球体的定义以及$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$，可得满足$\|\mathbf{x}_i\| \le t$的$\mathbf{x}_i$构成了一个半径为$t$的$p$维球体，所以可得：

$$Pr(\|\mathbf{x}_i\| > t) = 1 - Pr(\|\mathbf{x}_i\| \le t) = 1 - \frac{V_p(t)}{V_p(1)} \tag{3.7}$$

根据p维球体的定义，可得：

$$\frac{V_p(t)}{V_p(1)} = \frac{t^p}{1} \tag{3.8}$$

结合上一问可得

$$\begin{aligned} Pr(d^* > t) =& Pr(E_1 \wedge E_2 \wedge \cdots \wedge E_n) \\ =& Pr(E_1)Pr(E_2)\cdots Pr(E_n) \\ =& Pr(\|\mathbf{x}_1\| > t)Pr(\|\mathbf{x}_2\| > t)\cdots Pr(\|\mathbf{x}_n\| > t) \\ =& (1-t^p)^n \end{aligned} \tag{3.9}$$

所以可得

$$Pr(d^* \le t) = 1 - Pr(d^* > t) = 1 - (1-t^p)^n \tag{3.10}$$

(3) 因为$Pr(d^* \le t) = \frac{1}{2}$，所以有$1-(1-t^p)^n = \frac{1}{2}$，可得

$$\begin{aligned} & 1-(1-t^p)^n = 1/2 \\ \Leftrightarrow\ & 1/2 = (1-t^p)^n \\ \Leftrightarrow\ & \sqrt[n]{1/2} = 1 - t^p \\ \Leftrightarrow\ & t^p = 1 - \sqrt[n]{1/2} \\ \Leftrightarrow\ & t = \sqrt[p]{1 - \sqrt[n]{1/2}} \end{aligned} \tag{3.11}$$

所以可得$t = \sqrt[p]{1 - \sqrt[n]{1/2}}$。

(4) 根据第(2)题答案可得$Pr(d^* \le \frac{1}{2}) = 1 - (1-(1/2)^p)^n$，令$1-(1-(1/2)^p)^n \ge 0.9$，可得

$$\begin{aligned} & 1-(1-(1/2)^p)^n \ge 0.9 \\ \Leftrightarrow\ & 0.1 \ge (1-(1/2)^p)^n \\ \Leftrightarrow\ & n \ge \log_{1-(1/2)^p} 0.1 \\ \Leftrightarrow\ & n \ge \frac{\ln 0.1}{\ln(1-(1/2)^p)} \\ \Leftrightarrow\ & n \ge \frac{\ln 10}{\sum_{i=1}^{\infty} \frac{(1/2)^{pi}}{i}} \end{aligned} \tag{3.12}$$

所以n应该大于等于$\frac{\ln 10}{\sum_{i=1}^{\infty} \frac{(1/2)^{pi}}{i}}$。

(5) 用$Pr(d^* \le t)$表示原点O以及其1-*NN*之间距离d^*小于等于某个固定值t且$t \in [0,1]$的概率。
在t的值确定的情况下，如果考虑$Pr(d^* \le t)$的大小，那么：
在p固定的情况下，n越大，$Pr(d^* \le t)$就越大。
在n固定的情况下，p越大，$Pr(d^* \le t)$就越小。

在$Pr(d^* \leq t)$值确定的情况下，如果考虑t的大小，那么：
在p固定的情况下，n越大，t就越小。
在n固定的情况下，p越大，t就越大。
在$Pr(d^* \leq t)$和t的值都确定的情况下，可以得到：
p越大，n就越大。
总体而言，n越大，p越小，性能越好。

4 [15pts] Principal Component Analysis

一些经典的降维方法，例如PCA，可以将均值为**0**的高维数据通过对其协方差矩阵的特征值计算，取较高特征值对应的特征向量的操作而后转化为维数较低的数据. 在这里，我们记U_k为$d\times k$的矩阵，这个矩阵是由原数据协方差矩阵最高的k个特征值对应的特征向量组成的.

在这里我们有两种方法来求出低维的对应于$\mathbf{x}\in\mathbb{R}^d$的重构向量$\mathbf{w}\in\mathbb{R}^k$:

A. 利用$U_k\mathbf{w}$重构出对应的$\mathbf{x}$时, 最小化重构平方误差;

B. 将$\mathbf{x}$投影在由U_k的列向量张成的空间中.

在这里，我们将探究这两种方法的关系.

(1) **[5pts]** 写出方法A中最小化重构平方误差的目标函数的表示形式.

(2) **[10pts]** 证明通过方法A得到的重构向量就是$U_k^{\mathrm{T}}\mathbf{x}$，也就是$\mathbf{x}$在$U_k$列向量空间中的投影 (通过方法B得到的重构向量). 这里, 有$U_k^{\mathrm{T}}U_k = I_k$成立, 其中的$I_k$是$k\times k$的单位矩阵.

Solution. *此处用于写解答(中英文均可)*

(1) 假设共计有m个样例高维向量$\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_m$，他们对应的低维向量为$\mathbf{w}_1,\mathbf{w}_2,\cdots,\mathbf{w}_m$。所有的$\mathbf{x}_i$的均值为$\mathbf{0}$。用$U_k$和$\mathbf{w}_i$重构出$\hat{\mathbf{x}}_i$，即

$$\hat{\mathbf{x}}_i = U_k\mathbf{w}_i \tag{4.1}$$

为了最小化重构平方误差，可得目标函数

$$\min_{\hat{\mathbf{x}}_i} \quad \sum_{i=1}^{m}\|\hat{\mathbf{x}}_i-\mathbf{x}_i\|_2^2 \tag{4.2}$$

即

$$\min_{\mathbf{w}_i} \quad \sum_{i=1}^{m}\|U_k\mathbf{w}_i-\mathbf{x}_i\|_2^2 \tag{4.3}$$

或者写为

$$\min_{\mathbf{w}_i} \quad \sum_{i=1}^{m}(U_k\mathbf{w}_i-\mathbf{x}_i)^{\mathrm{T}}(U_k\mathbf{w}_i-\mathbf{x}_i) \tag{4.4}$$

(2) 令$f(\mathbf{w}_1,\mathbf{w}_2,\cdots,\mathbf{w}_m)=\sum_{i=1}^{m}\|U_k\mathbf{w}_i-\mathbf{x}_i\|_2^2$。为了得到(4.3)的解，即让$f$对$\mathbf{w}_1,\cdots,\mathbf{w}_m$求偏导，然后令导数为*0*，即

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{w}_1} &= 2U_k^{\mathrm{T}}(U_k\mathbf{w}_1-\mathbf{x}_1)=0\\ \frac{\partial f}{\partial \mathbf{w}_2} &= 2U_k^{\mathrm{T}}(U_k\mathbf{w}_2-\mathbf{x}_2)=0\\ &\cdots\cdots\\ \frac{\partial f}{\partial \mathbf{w}_m} &= 2U_k^{\mathrm{T}}(U_k\mathbf{w}_m-\mathbf{x}_m)=0\end{aligned} \tag{4.5}$$

可得，对任意的$i=1,2,\cdots,m$，都有

$$\begin{aligned}&2U_k^{\mathrm{T}}(U_k\mathbf{w}_i-\mathbf{x}_i)=0\\ \Leftrightarrow\ &U_k^{\mathrm{T}}U_k\mathbf{w}_i=U_k^{\mathrm{T}}\mathbf{x}_i\\ \Leftrightarrow\ &I_k\mathbf{w}_i=U_k^{\mathrm{T}}\mathbf{x}_i\\ \Leftrightarrow\ &\mathbf{w}_i=U_k^{\mathrm{T}}\mathbf{x}_i\end{aligned} \tag{4.6}$$

根据以上可得对于$\mathbf{x}_i$，通过方法A得到的重构向量$\mathbf{w}_i$就是$U_k^{\mathrm{T}}\mathbf{x}_i$。
假设U_k写成列向量组合的形式为

$$U_k = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_k \end{bmatrix} \tag{4.7}$$

其中$\mathbf{v}_i$为d维向量。那么对于$\mathbf{x}_i$投射在U_k列向量空间中的投影$\mathbf{w}_i$，有

$$\mathbf{w}_i = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}}\mathbf{x}_i \\ \mathbf{v}_2^{\mathrm{T}}\mathbf{x}_i \\ \cdots \\ \mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i \end{bmatrix} \tag{4.8}$$

所以投影向量为$\mathbf{w}_i = U_k^{\mathrm{T}}\mathbf{x}$。
综上可得，通过方法A得到的重构向量就是$U_k^{\mathrm{T}}\mathbf{x}$，也就是$\mathbf{x}$在$U_k$列向量空间中的投影 *(*通过方法$B$得到的重构向量*)*。