机器学习（西瓜书）

注　解

（第 10 章　降维与度量学习）

https://blog.csdn.net/jbb0523

前言

经常听人说南大周老师所著的《机器学习》（以下统称为西瓜书）是一本入门教材，是一本科普性质的教科书。在该书第十次印刷之际，周老师在“如何使用本书”中也提到“这是一本入门级教科书”。然而，本人读起来却感觉该书远不止“科普”“入门”那么简单，书中的很多公式需要思考良久方能推导，很多概念需要反复咀嚼才能消化。边读边想着是不是应该将自己学习时遇到的一些知识难点的解析分享出来，以帮助更多的人入门。自己的确也随手做过一些笔记，但由于怀疑这仅是自己的个别现象，毕竟读书期间，思考更多的是如何使用单片机、DSP、ARM、FGPA 等，而这些基本是不需要推导任何公式的，因此作罢。偶然间在周老师的新浪微博看到如下对话：

此时方知，可能“读不懂”并不是个别现象。因此决定写一本“西瓜书注解”或者称为“西瓜书读书笔记”，对自己研读西瓜书时遇到的“台阶”进行解释和推导，以帮助更多的人能够更快地进入到这个领域。另外，近期越来越强地意识到，扎扎实实地推导一些基础算法的公式，无论是对于理解算法本身机理还是进行学术研究，都是非常有必要的。

自己会根据个人学习进度和研究需要按章发布，不知道能不能坚持写完，加油！

毕竟自己也是一名初学者，所以可能一些概念解释并不完整、一些公式推导并不优美，甚至会存在错误，这是不可避免的，不接受漫骂，但欢迎将问题反馈给我，共同学习进步！

（网盘链接：https://pan.baidu.com/s/1QtEiNnk8jMzmbs0KPBN-_w）

第 10 章目录

第 10 章 降维与度量学习.....1
10.0 预备知识.....1
1、符号约定.....1
2、矩阵与单位阵、向量的乘法.....1
3、矩阵的 F 范数与迹.....1
10.1 k 近邻学习.....3
1、式(10.1)的解释.....3
2、式(10.2)的推导.....3
10.2 低维嵌入.....4
1、图 10.2 的解释.....4
2、式(10.3)的推导.....4
3、式(10.4)的推导.....5
4、式(10.5)的推导.....5
5、式(10.6)的推导.....6
6、式(10.10)的推导.....6
7、图 10.3 关于 MDS 算法的解释.....7
10.3 主成分分析.....9
1、式(10.14)的推导.....9
2、式(10.16)的解释.....12
3、式(10.17)的推导.....14
4、根据式(10.17)求解式(10.16).....14
10.4 核化线性降维.....15
1、式(10.19)的解释.....15
2、式(10.20)的解释.....15
3、式(10.21)的解释.....16
4、式(10.22)的解释.....16
5、式(10.24)的推导.....16
6、式(10.25)的解释.....16
10.5 流形学习.....17
1、等度量映射(Isomap)的解释.....17
2、式(10.28)的推导.....17
3、式(10.31)的推导.....19
10.6 度量学习.....20
1、式(10.34)的解释.....20
2、式(10.35)的解释.....21
3、式(10.36)的解释.....21
4、式(10.37)的解释.....22
5、式(10.38)的解释.....22
6、式(10.39)的解释.....22

第 10 章 降维与度量学习

10.0 预备知识

本章需要较多线性代数与矩阵分析基础。

1、符号约定

向量元素之间分号“;”表示列元素分隔符，如$\boldsymbol{\alpha}=(a_1;a_2;\ldots;a_i;\ldots;a_m)$表示$m\times 1$的列向量；而逗号“,”表示行元素分隔符，如$\boldsymbol{\alpha}=(a_1,a_2,\ldots,a_i,\ldots,a_m)$表示$1\times m$的行向量；这与 Matlab 软件中的语法习惯一致。

2、矩阵与单位阵、向量的乘法

(1)矩阵左乘对角阵相当于矩阵每行乘以对应对角阵的对角线元素，如：

$$\begin{bmatrix}\lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3\end{bmatrix}\begin{bmatrix}x_{11} & x_{12} & x_{13}\\ x_{21} & x_{22} & x_{23}\\ x_{31} & x_{32} & x_{33}\end{bmatrix}=\begin{bmatrix}\lambda_1x_{11} & \lambda_1x_{12} & \lambda_1x_{13}\\ \lambda_2x_{21} & \lambda_2x_{22} & \lambda_2x_{23}\\ \lambda_3x_{31} & \lambda_3x_{32} & \lambda_3x_{33}\end{bmatrix}$$

(2)矩阵右乘对角阵相当于矩阵每列乘以对应对角阵的对角线元素，如：

$$\begin{bmatrix}x_{11} & x_{12} & x_{13}\\ x_{21} & x_{22} & x_{23}\\ x_{31} & x_{32} & x_{33}\end{bmatrix}\begin{bmatrix}\lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3\end{bmatrix}=\begin{bmatrix}\lambda_1x_{11} & \lambda_2x_{12} & \lambda_3x_{13}\\ \lambda_1x_{21} & \lambda_2x_{22} & \lambda_3x_{23}\\ \lambda_1x_{31} & \lambda_2x_{32} & \lambda_3x_{33}\end{bmatrix}$$

(3)矩阵左乘行向量相当于矩阵每行乘以对应行向量的元素之和，如：

$$\begin{bmatrix}\lambda_1 & \lambda_2 & \lambda_3\end{bmatrix}\begin{bmatrix}x_{11} & x_{12} & x_{13}\\ x_{21} & x_{22} & x_{23}\\ x_{31} & x_{32} & x_{33}\end{bmatrix}$$
$$=\lambda_1\begin{bmatrix}x_{11} & x_{12} & x_{13}\end{bmatrix}+\lambda_2\begin{bmatrix}x_{21} & x_{22} & x_{23}\end{bmatrix}+\lambda_3\begin{bmatrix}x_{31} & x_{32} & x_{33}\end{bmatrix}$$
$$=(\lambda_1x_{11}+\lambda_2x_{21}+\lambda_3x_{31},\lambda_1x_{12}+\lambda_2x_{22}+\lambda_3x_{32},\lambda_1x_{13}+\lambda_2x_{23}+\lambda_3x_{33})$$

(4)矩阵右乘列向量相当于矩阵每列乘以对应列向量的元素之和，如：

$$\begin{bmatrix}x_{11} & x_{12} & x_{13}\\ x_{21} & x_{22} & x_{23}\\ x_{31} & x_{32} & x_{33}\end{bmatrix}\begin{bmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\end{bmatrix}$$
$$=\lambda_1\begin{bmatrix}x_{11}\\ x_{21}\\ x_{31}\end{bmatrix}+\lambda_2\begin{bmatrix}x_{12}\\ x_{22}\\ x_{32}\end{bmatrix}+\lambda_3\begin{bmatrix}x_{13}\\ x_{23}\\ x_{33}\end{bmatrix}=\sum_{i=1}^{3}\left(\lambda_i\begin{bmatrix}x_{1i}\\ x_{2i}\\ x_{3i}\end{bmatrix}\right)$$
$$=(\lambda_1x_{11}+\lambda_2x_{12}+\lambda_3x_{13};\lambda_1x_{21}+\lambda_2x_{22}+\lambda_3x_{23};\lambda_1x_{31}+\lambda_2x_{32}+\lambda_3x_{33})$$

综上，左乘是对矩阵的行操作，而右乘则是对矩阵的列操作，第(2)个和第(4)个结论后面推导过程中灵活应用较多。

3、矩阵的 F 范数与迹

(1)对于矩阵$\mathbf{A}\in\mathbb{R}^{m\times n}$，其 Frobenius 范数（简称 F 范数）$\|\mathbf{A}\|_F$定义为

$$\|\mathbf{A}\|_F=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right)^{\frac{1}{2}}$$

其中a_{ij}为矩阵$\mathbf{A}$第i行第j列的元素，即

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}$$

(2)若$\mathbf{A}=(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\ldots,\boldsymbol{\alpha}_j,\ldots,\boldsymbol{\alpha}_n)$，其中$\boldsymbol{\alpha}_j=(a_{1j};a_{2j};\ldots;a_{ij};\ldots;a_{mj})$为其列向量，$\mathbf{A}\in\mathbb{R}^{m\times n}$，$\boldsymbol{\alpha}_j\in\mathbb{R}^{m\times 1}$，则$\|\mathbf{A}\|_F^2=\sum_{j=1}^n\|\boldsymbol{\alpha}_j\|_2^2$；

同理，若$\mathbf{A}=(\boldsymbol{\beta}_1;\boldsymbol{\beta}_2;\ldots;\boldsymbol{\beta}_i;\ldots;\boldsymbol{\beta}_m)$，其中$\boldsymbol{\beta}_i=(a_{i1},a_{i2},\ldots,a_{ij},\ldots,a_{in})$为其行向量，$\mathbf{A}\in\mathbb{R}^{m\times n}$，$\boldsymbol{\beta}_i\in\mathbb{R}^{1\times n}$，则$\|\mathbf{A}\|_F^2=\sum_{i=1}^m\|\boldsymbol{\beta}_i\|_2^2$。

证明：该结论是显而易见的，因为$\|\boldsymbol{\alpha}_j\|_2^2=\sum_{i=1}^m|a_{ij}|^2$，而$\|\mathbf{A}\|_F=\sum_{i=1}^m\sum_{j=1}^n|a_{ij}|^2$。

(3)若$\lambda_j\left(\mathbf{A}^\top\mathbf{A}\right)$表示$n$阶方阵$\mathbf{A}^\top\mathbf{A}$的第$j$个特征值，$\mathrm{tr}\left(\mathbf{A}^\top\mathbf{A}\right)$是$\mathbf{A}^\top\mathbf{A}$的迹（对角线元素之和）；$\lambda_i\left(\mathbf{A}\mathbf{A}^\top\right)$表示$m$阶方阵$\mathbf{A}\,\mathbf{A}^\top$的第$i$个特征值，$\mathrm{tr}\left(\mathbf{A}\mathbf{A}^\top\right)$是$\mathbf{A}\,\mathbf{A}^\top$的迹，则

$$\begin{aligned}\|\mathbf{A}\|_F^2&=\mathrm{tr}\left(\mathbf{A}^\top\mathbf{A}\right)=\sum_{j=1}^n\lambda_j\left(\mathbf{A}^\top\mathbf{A}\right)\\&=\mathrm{tr}\left(\mathbf{A}\mathbf{A}^\top\right)=\sum_{i=1}^m\lambda_i\left(\mathbf{A}\mathbf{A}^\top\right)\end{aligned}$$

证明：(a)先证$\|\mathbf{A}\|_F^2=\mathrm{tr}\left(\mathbf{A}^\top\mathbf{A}\right)$：

令$\mathbf{B}=\mathbf{A}^\top\mathbf{A}\in\mathbb{R}^{n\times n}$，$b_{ij}$表示$\mathbf{B}$第$i$行第$j$列元素，$\mathrm{tr}\left(\mathbf{B}\right)=\sum_{j=1}^n b_{jj}$，

$$\mathbf{B}=\mathbf{A}^\top\mathbf{A}=\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{i1} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{i2} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{1j} & a_{2j} & \cdots & a_{ij} & \cdots & a_{mj} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{in} & \cdots & a_{mn} \end{bmatrix}\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}$$

由矩阵运算规则，b_{jj}等于$\mathbf{A}^\top$的第j行与$\mathbf{A}$的第j列的内积（红色元素），因此

$$\mathrm{tr}\left(\mathbf{B}\right)=\sum_{j=1}^n b_{jj}=\sum_{j=1}^n\left(\sum_{i=1}^m|a_{ij}|^2\right)=\sum_{i=1}^m\sum_{j=1}^n|a_{ij}|^2=\|\mathbf{A}\|_F^2$$

以上第三个等号交换了求和号次序（类似于交换积分号次序），显然这不影响求和结果。

(b)同理，可证$\|\mathbf{A}\|_F^2=\mathrm{tr}\left(\mathbf{A}\mathbf{A}^\top\right)$：

令$\mathbf{C}=\mathbf{A}\mathbf{A}^{\top}\in\mathbb{R}^{m\times m}$，$c_{ij}$表示$\mathbf{C}$第$i$行第$j$列元素，$\mathrm{tr}\,(\mathbf{C})=\sum_{i=1}^{m}c_{ii}$，

$$\mathbf{C}=\mathbf{A}\mathbf{A}^{\top}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{i1} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{i2} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{1j} & a_{2j} & \cdots & a_{ij} & \cdots & a_{mj} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{in} & \cdots & a_{mn} \end{bmatrix}$$

由矩阵运算规则，c_{ii}等于$\mathbf{A}$的第i行与$\mathbf{A}^{\top}$的第i列的内积（红色元素），因此

$$\mathrm{tr}\,(\mathbf{C})=\sum_{i=1}^{m}c_{ii}=\sum_{i=1}^{m}\left(\sum_{j=1}^{n}|a_{ij}|^{2}\right)=\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^{2}=\|\mathbf{A}\|_{F}^{2}$$

有关方阵的特征值之和等于对角线元素之和，可以参见线性代数教材，如同济大学主编的《线性代数(第五版)》第五章第 2 节“方阵的特征值与特征向量”（第 117 页）：

设 n 阶矩阵 $\boldsymbol{A}=(a_{ij})$ 的特征值为 $\lambda_1,\lambda_2,\cdots,\lambda_n$，不难证明

(i) $\lambda_1+\lambda_2+\cdots+\lambda_n=a_{11}+a_{22}+\cdots+a_{nn}$；

(ii) $\lambda_1\lambda_2\cdots\lambda_n=|\boldsymbol{A}|$.

10.1 *k* 近邻学习

1、式(10.1)的解释

式(10.1)为：$P(err)=1-\sum_{c\in\mathcal{Y}}P(c|\boldsymbol{x})P(c|\boldsymbol{z})$

首先，$P(c|\boldsymbol{x})$表示样本$\boldsymbol{x}$为类别c的后验概率，$P(c|\boldsymbol{z})$表示样本$\boldsymbol{z}$为类别c的后验概率；
其次，$P(c|\boldsymbol{x})P(c|\boldsymbol{z})$表示样本$\boldsymbol{x}$和样本$\boldsymbol{z}$同时为类别$c$的概率；

再次，$\sum_{c\in\mathcal{Y}}P(c|\boldsymbol{x})P(c|\boldsymbol{z})$表示样本$\boldsymbol{x}$和样本$\boldsymbol{z}$类别相同的概率；这一点可以进一步解释，设$\mathcal{Y}=\{c_1,c_2,\cdots,c_N\}$，则该求和式子变为：

$$P(c_1|\boldsymbol{x})P(c_1|\boldsymbol{z})+P(c_2|\boldsymbol{x})P(c_2|\boldsymbol{z})+\cdots+P(c_N|\boldsymbol{x})P(c_N|\boldsymbol{z})$$

即样本$\boldsymbol{x}$和样本$\boldsymbol{z}$同时为c_1的概率，加上同时为c_2的概率，……，加上同时为c_N的概率，即样本$\boldsymbol{x}$和样本$\boldsymbol{z}$类别相同的概率；

最后，$P(err)$表示样本$\boldsymbol{x}$和样本$\boldsymbol{z}$类别不相同的概率，即 1 减去二者类别相同的概率。

2、式(10.2)的推导

式(10.2)推导关键在于理解第二行的“约等$(\simeq)$”关系和第三行的“小于等于$(\leqslant)$”关系。

第二行的“约等$(\simeq)$”关系的依据在于该式前面一段话：“假设样本独立同分布，且对任意$\boldsymbol{x}$和任意小正数δ，在$\boldsymbol{x}$附近δ距离范围内总能找到一个训练样本”，这意味着对于任意测试样本在训练集中都可以找出一个与其非常像（任意小正数δ）的近邻，这里还有一个假设书中未提及：$P(c|\boldsymbol{x})$必须是连续函数（对于连续函数$f(x)$和任意小正数δ，$f(x)\simeq f(x+\delta)$），即对于两个非常像的样本$\boldsymbol{z}$与$\boldsymbol{x}$有$P(c|\boldsymbol{x})\simeq P(c|\boldsymbol{z})$，即

$$\sum_{c\in\mathcal{Y}}P(c|\boldsymbol{x})P(c|\boldsymbol{z})\simeq\sum_{c\in\mathcal{Y}}P^{2}(c|\boldsymbol{x})$$

第三行的“小于等于($\leqslant$)”关系更简单：由于$c^* \in \mathcal{Y}$，所以$P^2(c^*|\boldsymbol{x}) \leqslant \sum_{c\in\mathcal{Y}} P^2(c|\boldsymbol{x})$，也就是“小于等于($\leqslant$)”左边只是右边的一部分，所以肯定是小于等于的关系；

第四行就是数学公式$a^2 - b^2 = (a+b)(a-b)$

第五行是由于$1 + P(c^*|\boldsymbol{x}) \leqslant 2$，这是由于概率值$P(c^*|\boldsymbol{x}) \leqslant 1$

经过以上推导，本节最后给出一个惊人的结论：最近邻分类器虽简单，但它的泛化错误率不超过贝叶斯最优分类器的错误率的两倍！

然而这是一个没啥实际用途的结论，因为这个结论必须满足两个假设条件，且不说$P(c|\boldsymbol{x})$是连续函数（第一个假设）是否满足，单就“对任意$\boldsymbol{x}$和任意小正数δ，在$\boldsymbol{x}$附近δ距离范围内总能找到一个训练样本”（第二个假设）是不可能满足的，这也就有了 10.2 节开头一段的讨论，抛开“任意小正数δ”不谈，具体到$\delta = 0.001$都是不现实的。

10.2 低维嵌入

1、图 10.2 的解释

只要注意一点就行：在图(a)三维空间中，红色线是弯曲的，但去掉高度这一维（竖着的坐标轴）后，红色线变成直线，而直线更容易学习。

2、式(10.3)的推导

已知$\mathbf{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_i, \ldots, \boldsymbol{z}_m\} \in \mathbb{R}^{d'\times m}$，其中$\boldsymbol{z}_i = (z_{i1}; z_{i2}; \ldots; z_{id'}) \in \mathbb{R}^{d'\times 1}$；降维后的内积矩阵$\mathbf{B} = \mathbf{Z}^\top\mathbf{Z} \in \mathbb{R}^{m\times m}$，其中第$i$行第$j$列元素$b_{ij}$，特别的

$$b_{ii} = \boldsymbol{z}_i^\top\boldsymbol{z}_i = \|\boldsymbol{z}_i\|^2,\ b_{jj} = \boldsymbol{z}_j^\top\boldsymbol{z}_j = \|\boldsymbol{z}_j\|^2,\ b_{ij} = \boldsymbol{z}_i^\top\boldsymbol{z}_j$$

MDS 算法的目标是$\|\boldsymbol{z}_i - \boldsymbol{z}_j\| = dist_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$，即保持样本的欧氏距离在$d'$维空间和原始$d$维空间相同($d' \leqslant d$)。

$$\begin{aligned}
dist_{ij}^2 &= \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2 = (z_{i1} - z_{j1})^2 + (z_{i2} - z_{j2})^2 + \ldots + (z_{id'} - z_{jd'})^2 \\
&= (z_{i1}^2 - 2z_{i1}z_{j1} + z_{j1}^2) + (z_{i2}^2 - 2z_{i2}z_{j2} + z_{j2}^2) + \ldots + (z_{id'}^2 - 2z_{id'}z_{jd'} + z_{jd'}^2) \\
&= (z_{i1}^2 + z_{i2}^2 + \ldots + z_{id'}^2) + (z_{j1}^2 + z_{j2}^2 + \ldots + z_{jd'}^2) \\
&- 2(z_{i1}z_{j1} + z_{i2}z_{j2} + \ldots + z_{id'}z_{jd'}) \\
&= \|\boldsymbol{z}_i\|^2 + \|\boldsymbol{z}_j\|^2 - 2\boldsymbol{z}_i^\top\boldsymbol{z}_j \\
&= b_{ii} + b_{jj} - 2b_{ij}
\end{aligned}$$

本章矩阵运算非常多，刚刚是从矩阵元素层面的推导；实际可发现上式运算结果基本与标量运算规则相同，因此后面会尽可能不再从元素层面推导。具体来说：

$$\begin{aligned}
dist_{ij}^2 &= \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2 = (\boldsymbol{z}_i - \boldsymbol{z}_j)^\top(\boldsymbol{z}_i - \boldsymbol{z}_j) \\
&= \boldsymbol{z}_i^\top\boldsymbol{z}_i - \boldsymbol{z}_i^\top\boldsymbol{z}_j - \boldsymbol{z}_j^\top\boldsymbol{z}_i + \boldsymbol{z}_j^\top\boldsymbol{z}_j \\
&= \boldsymbol{z}_i^\top\boldsymbol{z}_i + \boldsymbol{z}_j^\top\boldsymbol{z}_j - 2\boldsymbol{z}_i^\top\boldsymbol{z}_j \\
&= \|\boldsymbol{z}_i\|^2 + \|\boldsymbol{z}_j\|^2 - 2\boldsymbol{z}_i^\top\boldsymbol{z}_j \\
&= b_{ii} + b_{jj} - 2b_{ij}
\end{aligned}$$

上式第三个等号化简是由于内积$\boldsymbol{z}_i^\top\boldsymbol{z}_j$和$\boldsymbol{z}_j^\top\boldsymbol{z}_i$均为标量，因此转置等于本身。

3、式(10.4)的推导

首先解释两个条件：

(1)令降维后的样本$\mathbf{Z}$被中心化，即$\sum_{i=1}^{m}\boldsymbol{z}_i=\mathbf{0}$

注意$\mathbf{Z}\in\mathbb{R}^{d'\times m}$，$d'$是样本维度（属性个数），$m$是样本个数，易知$\mathbf{Z}$的每一行有$m$个元素（每行表示样本集的一维属性），$\mathbf{Z}$的每一列有$d'$个元素（每列表示一个样本）。

式$\sum_{i=1}^{m}\boldsymbol{z}_i=\mathbf{0}$中的$\boldsymbol{z}_i$明显表示的是第$i$列，$m$列相加得到一个零向量$\mathbf{0}_{d'\times 1}$，意思是样本集合中所有样本的每一维属性之和均等于 0，因此**被中心化**的意思是将样本集合$\mathbf{Z}$的每一行（属性）减去该行的均值。

(2)显然，矩阵$\mathbf{B}$的行与列之各均为零，即$\sum_{i=1}^{m}b_{ij}=\sum_{j=1}^{m}b_{ij}=0$。

注意$b_{ij}=\boldsymbol{z}_i^{\top}\boldsymbol{z}_j$（也可以写为$b_{ij}=\boldsymbol{z}_j^{\top}\boldsymbol{z}_i$，其实就是对应元素相乘，再求和）

$$\sum_{i=1}^{m}b_{ij}=\sum_{i=1}^{m}\boldsymbol{z}_j^{\top}\boldsymbol{z}_i=\boldsymbol{z}_j^{\top}\sum_{i=1}^{m}\boldsymbol{z}_i=\boldsymbol{z}_j^{\top}\cdot\mathbf{0}_{d'\times 1}=0$$

$$\sum_{j=1}^{m}b_{ij}=\sum_{j=1}^{m}\boldsymbol{z}_i^{\top}\boldsymbol{z}_j=\boldsymbol{z}_i^{\top}\sum_{j=1}^{m}\boldsymbol{z}_j=\boldsymbol{z}_i^{\top}\cdot\mathbf{0}_{d'\times 1}=0$$

接下来我们推导式(10.4)，将式(10.3)的$dist_{ij}^2$表达式代入：

$$\begin{aligned}\sum_{i=1}^{m}dist_{ij}^2&=\sum_{i=1}^{m}\left(\|\boldsymbol{z}_i\|^2+\|\boldsymbol{z}_j\|^2-2\boldsymbol{z}_i^{\top}\boldsymbol{z}_j\right)\\&=\sum_{i=1}^{m}\|\boldsymbol{z}_i\|^2+\sum_{i=1}^{m}\|\boldsymbol{z}_j\|^2-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\top}\boldsymbol{z}_j\end{aligned}$$

根据定义：

$$\sum_{i=1}^{m}\|\boldsymbol{z}_i\|^2=\sum_{i=1}^{m}\boldsymbol{z}_i^{\top}\boldsymbol{z}_i=\sum_{i=1}^{m}b_{ii}=\operatorname{tr}(\mathbf{B})$$

$$\sum_{i=1}^{m}\|\boldsymbol{z}_j\|^2=\|\boldsymbol{z}_j\|^2\sum_{i=1}^{m}1=m\|\boldsymbol{z}_j\|^2=m\boldsymbol{z}_j^{\top}\boldsymbol{z}_j=mb_{jj}$$

根据前面结果：

$$\sum_{i=1}^{m}\boldsymbol{z}_i^{\top}\boldsymbol{z}_j=\left(\sum_{i=1}^{m}\boldsymbol{z}_i^{\top}\right)\boldsymbol{z}_j=\mathbf{0}_{1\times d'}\cdot\boldsymbol{z}_j=0$$

代入上式即得：

$$\begin{aligned}\sum_{i=1}^{m}dist_{ij}^2&=\sum_{i=1}^{m}\|\boldsymbol{z}_i\|^2+\sum_{i=1}^{m}\|\boldsymbol{z}_j\|^2-2\sum_{i=1}^{m}\boldsymbol{z}_i^{\top}\boldsymbol{z}_j\\&=\operatorname{tr}(\mathbf{B})+mb_{jj}\end{aligned}$$

4、式(10.5)的推导

与式(10.4)类似：

$$\begin{aligned}\sum_{j=1}^{m} dist_{ij}^2 &= \sum_{j=1}^{m} \left(\|\boldsymbol{z}_i\|^2 + \|\boldsymbol{z}_j\|^2 - 2\boldsymbol{z}_i^\top \boldsymbol{z}_j\right) \\ &= \sum_{j=1}^{m} \|\boldsymbol{z}_i\|^2 + \sum_{j=1}^{m} \|\boldsymbol{z}_j\|^2 - 2\sum_{j=1}^{m} \boldsymbol{z}_i^\top \boldsymbol{z}_j \\ &= m b_{ii} + \operatorname{tr}(\mathbf{B})\end{aligned}$$

5、式(10.6)的推导

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 &= \sum_{i=1}^{m}\sum_{j=1}^{m} \left(\|\boldsymbol{z}_i\|^2 + \|\boldsymbol{z}_j\|^2 - 2\boldsymbol{z}_i^\top \boldsymbol{z}_j\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} \|\boldsymbol{z}_i\|^2 + \sum_{i=1}^{m}\sum_{j=1}^{m} \|\boldsymbol{z}_j\|^2 - 2\sum_{i=1}^{m}\sum_{j=1}^{m} \boldsymbol{z}_i^\top \boldsymbol{z}_j \\ &= 2m\ \operatorname{tr}(\mathbf{B})\end{aligned}$$

其中各子项推导如下：

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \|\boldsymbol{z}_i\|^2 = \sum_{i=1}^{m} \left(\|\boldsymbol{z}_i\|^2 \sum_{j=1}^{m} 1\right) = m\sum_{i=1}^{m} \|\boldsymbol{z}_i\|^2 = m\ \operatorname{tr}(\mathbf{B})$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \|\boldsymbol{z}_j\|^2 = \sum_{i=1}^{m} \operatorname{tr}(\mathbf{B}) = m\ \operatorname{tr}(\mathbf{B})$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \boldsymbol{z}_i^\top \boldsymbol{z}_j = 0$$

这里用到一个准则：与求和（积分）变量无关的项可以提到求和号（积分号）外面。

6、式(10.10)的推导

将式(10.5)代入式(10.7)：$dist_{i\cdot}^2 = \frac{1}{m}\left(\operatorname{tr}(\mathbf{B}) + m b_{ii}\right) = \frac{1}{m}\operatorname{tr}(\mathbf{B}) + b_{ii}$

将式(10.4)代入式(10.8)：$dist_{\cdot j}^2 = \frac{1}{m}\left(\operatorname{tr}(\mathbf{B}) + m b_{jj}\right) = \frac{1}{m}\operatorname{tr}(\mathbf{B}) + b_{jj}$

将式(10.6)代入式(10.9)：$dist_{\cdot\cdot}^2 = \frac{1}{m^2} 2m\ \operatorname{tr}(\mathbf{B}) = \frac{2}{m}\operatorname{tr}(\mathbf{B})$

将式(10.3)和以上结果代入：

$$\begin{aligned}&-\frac{1}{2}\left(dist_{ij}^2 - dist_{i\cdot}^2 - dist_{\cdot j}^2 + dist_{\cdot\cdot}^2\right) \\ &= -\frac{1}{2}\left((b_{ii} + b_{jj} - 2b_{ij}) - (\frac{1}{m}\operatorname{tr}(\mathbf{B}) + b_{ii}) - (\frac{1}{m}\operatorname{tr}(\mathbf{B}) + b_{jj}) + \frac{2}{m}\operatorname{tr}(\mathbf{B})\right) \\ &= b_{ij}\end{aligned}$$

在式(10.10)后紧跟着一句话：“由此即可通过降维前后保持不变的距离矩阵$\mathbf{D}$求取内积矩阵$\mathbf{B}$”，我们来解释一下这句话。

首先解释式(10.10)等号右侧的变量含义：$dist_{ij} = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|$表示降维后$\boldsymbol{z}_i$与$\boldsymbol{z}_j$的欧氏距离，注意这同时也应该是原始空间$\boldsymbol{x}_i$与$\boldsymbol{x}_j$的距离，因为降维的目标（也是约束条件）是“任意两个样本在d'维空间中的欧氏距离等于原始空间中的距离”，也就是说$dist_{ij}^2$是降维前

后的距离矩阵$\mathbf{D}$的元素$dist_{ij}$的平方；其次，式(10.10)等号左侧b_{ij}是降维后内积矩阵$\mathbf{B}$的元素，即$\mathbf{B}$的元素b_{ij}可以由距离矩阵$\mathbf{D}$来表达求取。

7、图 10.3 关于 MDS 算法的解释

首先要清楚此处降维算法要完成的任务：获得d维空间的样本集合$\mathbf{X} \in \mathbb{R}^{d\times m}$在$d'$维空间的表示$\mathbf{Z} \in \mathbb{R}^{d'\times m}$，并且保证距离矩阵$\mathbf{D} \in \mathbb{R}^{m\times m}$相同，其中$d' < d$，$m$为样本个数，距离矩阵即样本之间的欧氏距离。那么怎么由$\mathbf{X} \in \mathbb{R}^{d\times m}$得到$\mathbf{Z} \in \mathbb{R}^{d'\times m}$呢？

经过推导发现（式(10.3)~式(10.10)），在保证距离矩阵$\mathbf{D} \in \mathbb{R}^{m\times m}$相同的前提下，$d'$维空间的样本集合$\mathbf{Z} \in \mathbb{R}^{d'\times m}$的内积矩阵$\mathbf{B} = \mathbf{Z}^{\top}\mathbf{Z} \in \mathbb{R}^{m\times m}$可以由距离矩阵$\mathbf{D} \in \mathbb{R}^{m\times m}$得到（参见式(10.10)），此时只要对$\mathbf{B}$进行矩阵分解即可得到$\mathbf{Z}$；具体来说，对$\mathbf{B}$进行特征值分解可得$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\top}$，其中$\mathbf{V} \in \mathbb{R}^{m\times m}$为特征值向量矩阵，$\mathbf{\Lambda} \in \mathbb{R}^{m\times m}$为特征值构成的对角矩阵，接下来分类讨论：

(1) 当$d > m$时，即样本属性比样本个数还要多

此时，样本集合$\mathbf{X} \in \mathbb{R}^{d\times m}$的$d$维属性一定是线性相关的（即有冗余），因为矩阵$\mathbf{X}$的秩不会大于$m$（此处假设矩阵$\mathbf{X}$的秩恰好等于$m$），因此$\mathbf{\Lambda} \in \mathbb{R}^{m\times m}$主对角线有$m$个非零值，进而$\mathbf{B} = (\mathbf{V}\mathbf{\Lambda}^{1/2})(\mathbf{\Lambda}^{1/2}\mathbf{V}^{\top})$，得到的$\mathbf{Z} = \mathbf{\Lambda}^{1/2}\mathbf{V}^{\top} \in \mathbb{R}^{d'\times m}$实际将$d$维属性降成了$d' = m$维属性。

(2) 当$d < m$时，即样本个数比样本属性多

这是现实中最常见的一种情况。此时$\mathbf{\Lambda} \in \mathbb{R}^{m\times m}$至多有$d$个非零值（此处假设恰有$d$个非零值），因此$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\top}$可以写成$\mathbf{B} = \mathbf{V}_*\mathbf{\Lambda}_*{\mathbf{V}_*}^{\top}$，其中$\mathbf{\Lambda}_* \in \mathbb{R}^{d\times d}$为$d$个非零值特征值构成的特征值对角矩阵，$\mathbf{V}_* \in \mathbb{R}^{m\times d}$为$\mathbf{\Lambda}_* \in \mathbb{R}^{d\times d}$相应的特征值向量矩阵，进而$\mathbf{B} = (\mathbf{V}_*\mathbf{\Lambda}_*^{1/2})(\mathbf{\Lambda}_*^{1/2}\mathbf{V}_*^{\top})$，求得$\mathbf{Z} = \mathbf{\Lambda}_*^{1/2}\mathbf{V}_*^{\top} \in \mathbb{R}^{d\times m}$，此时属性没有冗余，因此按降维的规则（降维后距离矩阵不变）并不能实现有效降维。

由以上分析可以看出，降维后的维度d'实际为$\mathbf{B}$特征值分解后非零特征值的个数。

接下来我们通过 Matlab 简单直观的看一下：

第(1)种情况$(d > m)$：

```
>> m=4;d=8;rand('state',0);Z=rand(d,m),[V,D]=eig(Z'*Z)
Z =
    0.9501    0.8214    0.9355    0.1389
    0.2311    0.4447    0.9169    0.2028
    0.6068    0.6154    0.4103    0.1987
    0.4860    0.7919    0.8936    0.6038
    0.8913    0.9218    0.0579    0.2722
    0.7621    0.7382    0.3529    0.1988
    0.4565    0.1763    0.8132    0.0153
    0.0185    0.4057    0.0099    0.7468
V =
   -0.5613    0.5686    0.2420    0.5505
    0.6937   -0.0608    0.4070    0.5912
    0.0169   -0.1492   -0.8304    0.5365
   -0.4511   -0.8067    0.2935    0.2442
D =
```

```
    0.0318         0         0         0
         0    0.7322         0         0
         0         0    1.1085         0
         0         0         0    9.3487
```

第(2)种情况：

```
>> m=8;d=4;rand('state',0);Z=rand(d,m),[V,D]=eig(Z'*Z)
Z =
    0.9501    0.8913    0.8214    0.9218    0.9355    0.0579    0.1389    0.2722
    0.2311    0.7621    0.4447    0.7382    0.9169    0.3529    0.2028    0.1988
    0.6068    0.4565    0.6154    0.1763    0.4103    0.8132    0.1987    0.0153
    0.4860    0.0185    0.7919    0.4057    0.8936    0.0099    0.6038    0.7468
V =
   -0.2550    0.2521   -0.2186    0.2536    0.7552    0.1857   -0.0879    0.3831
    0.3122   -0.3802   -0.4824    0.0014   -0.0782   -0.3129   -0.5370    0.3645
    0.3804   -0.3940    0.5004   -0.3146    0.2334    0.2928    0.1414    0.4366
   -0.0042    0.6127    0.2095   -0.4344   -0.0900   -0.4744   -0.0697    0.3925
   -0.3506   -0.1639    0.2356    0.5560   -0.3934   -0.1329    0.1932    0.5251
   -0.2434    0.1787   -0.0248   -0.1955   -0.3905    0.6820   -0.4741    0.1659
    0.6329    0.4024   -0.3041    0.2704   -0.2134    0.2667    0.3523    0.1762
   -0.3318   -0.2014   -0.5259   -0.4770   -0.0902    0.0535    0.5411    0.2036
D =
-0.0000         0         0         0         0         0         0         0
         0    0.0000         0         0         0         0         0         0
         0         0    0.0000         0         0         0         0         0
         0         0         0    0.0000         0         0         0         0
         0         0         0         0    0.2654         0         0         0
         0         0         0         0         0    0.5945         0         0
         0         0         0         0         0         0    0.9481         0
         0         0         0         0         0         0         0    9.4133
```

第(3)种情况：

```
>> m=8;d=4;rand('state',0);Z=rand(d,m);Z(end,:)=Z(1,:)+Z(2,:),[V,D]=eig(Z'*Z)
Z =
    0.9501    0.8913    0.8214    0.9218    0.9355    0.0579    0.1389    0.2722
    0.2311    0.7621    0.4447    0.7382    0.9169    0.3529    0.2028    0.1988
    0.6068    0.4565    0.6154    0.1763    0.4103    0.8132    0.1987    0.0153
    1.1813    1.6534    1.2661    1.6600    1.8524    0.4108    0.3417    0.4710
V =
   -0.0983    0.4100   -0.1981    0.1622    0.0342   -0.7519   -0.2524    0.3557
    0.1620   -0.4442   -0.7089    0.1474   -0.0977    0.1504    0.0604    0.4651
    0.0880   -0.5658    0.4825   -0.3761   -0.1446   -0.2745   -0.2481    0.3743
    0.3095    0.0818    0.4253    0.4895    0.3350    0.0963    0.3887    0.4522
   -0.5449    0.3320    0.0881   -0.2439   -0.3331    0.3444    0.1787    0.5139
   -0.0593    0.0797    0.0562    0.1897    0.2977    0.4329   -0.8078    0.1498
    0.7484    0.4306   -0.0361   -0.3295   -0.3171    0.1462   -0.1122    0.1007
```

```
    0.0023    0.0761   -0.1797   -0.6057    0.7457    0.0009    0.1514    0.1267
D =
-0.0000         0         0         0         0         0         0         0
      0   -0.0000         0         0         0         0         0         0
      0         0    0.0000         0         0         0         0         0
      0         0         0    0.0000         0         0         0         0
      0         0         0         0    0.0000         0         0         0
      0         0         0         0         0    0.2745         0         0
      0         0         0         0         0         0    0.7044         0
      0         0         0         0         0         0         0   19.9195
```

前两种情况即以上分析的两种情况；第三种情况中，发现 D 只有三个非零值，这是因为 Z 的第 4 维属性是第 1 维属性和第 2 维属性之和（Z(end,:)=Z(1,:)+Z(2,:)），即 Z 的 4 个属性只有三个是线性无关的。

通过以上分析和实验可以发现，若要严格遵守降维前后的距离矩阵不变的规则，则只能去除线性相关的属性，如第(1)种情况$d > m$时，若矩阵$\mathbf{X}$的秩恰好等于m，则d个属性中有$m-d$个属性线性相关；而第(2)种情况由于没有线性相关的属性，因此不能实现降维；第(3)种情况的第 4 维属性由第 1 维属性和第 2 维属性线性组合而成，因此可以实现降维。若属性线性无关时，即矩阵$\mathbf{X}$的行之间均线性无关，此时并不能实现有效的降维。“在现实应用中为了有效降维，往往仅需降维后的距离与原始空间的距离尽可能接近，而不必严格相等”，如何实现呢？此即图 10.3 中的第 4 步：“取d'个最大特征值所构成的对角矩阵…”，如此近似之后得内积矩阵不再严格等于根据式(10.10)计算出的矩阵$\mathbf{B}$，但很接近……

10.3 主成分分析

注意，作者在数次印刷中对本节符号进行修订，详见勘误修订，直接搜索页码即可，此处仅按个人推导需求定义符号，可能与不同印次书中符号不一致。

1、式(10.14)的推导

预备知识：

在一个坐标系中，任意向量等于其在各个坐标轴的坐标值乘以相应坐标轴单位向量之和。例如，在二维直角坐标系中，$\boldsymbol{x}$轴和$\boldsymbol{y}$轴的单位向量分别为$\boldsymbol{v}_1=(1;0)$和$\boldsymbol{v}_2=(0;1)$，向量$\boldsymbol{r}=(2;3)$可以表示为$\boldsymbol{r}=2\boldsymbol{v}_1+3\boldsymbol{v}_2$；其实$\boldsymbol{v}_1=(1;0)$和$\boldsymbol{v}_2=(0;1)$只是二维平面的一组标准正交基，但二维平面实际有无数标准正交基，如$\boldsymbol{v}_1'=(\frac{1}{\sqrt{2}};\frac{1}{\sqrt{2}})$和$\boldsymbol{v}_2'=(-\frac{1}{\sqrt{2}};\frac{1}{\sqrt{2}})$，此时向量$\boldsymbol{r}=\frac{5}{\sqrt{2}}\boldsymbol{v}_1'+\frac{1}{\sqrt{2}}\boldsymbol{v}_2'$，其中$\frac{5}{\sqrt{2}}=(\boldsymbol{v}_1')^\top\boldsymbol{r}, \frac{1}{\sqrt{2}}=(\boldsymbol{v}_2')^\top\boldsymbol{r}$，即新坐标系里的坐标。

下面开始推导：

对于d维空间$\mathbb{R}^{d\times 1}$来说，传统的坐标系为$\{\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k,\ldots,\boldsymbol{v}_d\}$，其中$\boldsymbol{v}_k$为除第$k$个元素为 1 其余元素均 0 的$d$维列向量；此时对于样本点$\boldsymbol{x}_i=(x_{i1};x_{i2};\ldots;x_{id})\in\mathbb{R}^{d\times 1}$来说亦可表示为$\boldsymbol{x}_i=x_{i1}\boldsymbol{v}_1+x_{i2}\boldsymbol{v}_2+\ldots+x_{id}\boldsymbol{v}_d$。

现假定投影变换后得到的新坐标系为$\{\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_k,\ldots,\boldsymbol{w}_d\}$（即一组新的标准正交基），则$\boldsymbol{x}_i$在新坐标系中的坐标为$\left(\boldsymbol{w}_1^\top\boldsymbol{x}_i;\boldsymbol{w}_2^\top\boldsymbol{x}_i;\ldots;\boldsymbol{w}_d^\top\boldsymbol{x}_i\right)$。若丢弃新坐标系中的部分坐标，即将维度降低到$d'<d$（不失一般性，假设丢掉的是后$d-d'$维坐标），并令

$$\mathbf{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'})\in\mathbb{R}^{d\times d'}$$

则$\boldsymbol{x}_i$在低维坐标系中的投影为

$$\begin{aligned}\boldsymbol{z}_i&=(z_{i1};z_{i2};\ldots;z_{id'})=\left(\boldsymbol{w}_1^\top\boldsymbol{x}_i;\boldsymbol{w}_2^\top\boldsymbol{x}_i;\ldots;\boldsymbol{w}_{d'}^\top\boldsymbol{x}_i\right)\\&=\mathbf{W}^\top\boldsymbol{x}_i\end{aligned}$$

若基于$\boldsymbol{z}_i$来重构$\boldsymbol{x}_i$，则会得到$\hat{\boldsymbol{x}}_i=\sum_{j=1}^{d'}z_{ij}\boldsymbol{w}_j=\mathbf{W}\boldsymbol{z}_i$（课本 P230 第 11 行）。

有了以上符号基础，接下来将式(10.14)化简成式(10.15)目标函数形式（可逐一核对各项维数以验证推导是否有误）：

$$\begin{aligned}\sum_{i=1}^{m}\left\|\sum_{j=1}^{d'}z_{ij}\boldsymbol{w}_j-\boldsymbol{x}_i\right\|_2^2&\overset{①}{=}\sum_{i=1}^{m}\|\mathbf{W}\boldsymbol{z}_i-\boldsymbol{x}_i\|_2^2\\&\overset{②}{=}\sum_{i=1}^{m}\left\|\mathbf{W}\mathbf{W}^\top\boldsymbol{x}_i-\boldsymbol{x}_i\right\|_2^2\\&\overset{③}{=}\sum_{i=1}^{m}\left(\mathbf{W}\mathbf{W}^\top\boldsymbol{x}_i-\boldsymbol{x}_i\right)^\top\left(\mathbf{W}\mathbf{W}^\top\boldsymbol{x}_i-\boldsymbol{x}_i\right)\\&\overset{④}{=}\sum_{i=1}^{m}\left(\boldsymbol{x}_i^\top\mathbf{W}\mathbf{W}^\top\mathbf{W}\mathbf{W}^\top\boldsymbol{x}_i-2\boldsymbol{x}_i^\top\mathbf{W}\mathbf{W}^\top\boldsymbol{x}_i+\boldsymbol{x}_i^\top\boldsymbol{x}_i\right)\\&\overset{⑤}{=}\sum_{i=1}^{m}\left(\boldsymbol{x}_i^\top\mathbf{W}\mathbf{W}^\top\boldsymbol{x}_i-2\boldsymbol{x}_i^\top\mathbf{W}\mathbf{W}^\top\boldsymbol{x}_i+\boldsymbol{x}_i^\top\boldsymbol{x}_i\right)\\&\overset{⑥}{=}\sum_{i=1}^{m}\left(-\boldsymbol{x}_i^\top\mathbf{W}\mathbf{W}^\top\boldsymbol{x}_i+\boldsymbol{x}_i^\top\boldsymbol{x}_i\right)\\&\overset{⑦}{=}\sum_{i=1}^{m}\left(-\left(\mathbf{W}^\top\boldsymbol{x}_i\right)^\top\left(\mathbf{W}^\top\boldsymbol{x}_i\right)+\boldsymbol{x}_i^\top\boldsymbol{x}_i\right)\\&\overset{⑧}{=}\sum_{i=1}^{m}\left(-\left\|\mathbf{W}^\top\boldsymbol{x}_i\right\|_2^2+\boldsymbol{x}_i^\top\boldsymbol{x}_i\right)\\&\overset{⑨}{\propto}-\sum_{i=1}^{m}\left\|\mathbf{W}^\top\boldsymbol{x}_i\right\|_2^2\end{aligned}$$

上式从第三个等号到第四个等号：由于$\left(\mathbf{W}\mathbf{W}^\top\right)^\top=\left(\mathbf{W}^\top\right)^\top(\mathbf{W})^\top=\mathbf{W}\mathbf{W}^\top$，因此

$$\left(\mathbf{W}\mathbf{W}^\top\boldsymbol{x}_i\right)^\top=\boldsymbol{x}_i^\top\left(\mathbf{W}\mathbf{W}^\top\right)^\top=\boldsymbol{x}_i^\top\mathbf{W}\mathbf{W}^\top$$

代入即得第四个等号；从第四个等号到第五个等号：由于$\boldsymbol{w}_i^\top\boldsymbol{w}_j=0,(i\neq j),\|\boldsymbol{w}_i\|=1$，因此$\mathbf{W}^\top\mathbf{W}=\mathbf{I}\in\mathbb{R}^{d'\times d'}$，代入即得第五个等号。由于最终目标是寻找$\mathbf{W}$使目标函数(10.14)最小，而$\boldsymbol{x}_i^\top\boldsymbol{x}_i$与$\mathbf{W}$无关，因此在优化时可以去掉。令$\mathbf{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m)\in\mathbb{R}^{d\times m}$，即每列为一个样本，则式(10.14)可继续化简为（参见 10.0 中“矩阵的 F 范数与迹”）

$$
\begin{aligned}
-\sum_{i=1}^{m}\left\|\mathbf{W}^{\top}\boldsymbol{x}_i\right\|_2^2 &= -\left\|\mathbf{W}^{\top}\mathbf{X}\right\|_F^2 \\
&= -\mathrm{tr}\left(\left(\mathbf{W}^{\top}\mathbf{X}\right)\left(\mathbf{W}^{\top}\mathbf{X}\right)^{\top}\right) \\
&= -\mathrm{tr}\left(\mathbf{W}^{\top}\mathbf{X}\mathbf{X}^{\top}\mathbf{W}\right)
\end{aligned}
$$

这里$\mathbf{W}^{\top}\boldsymbol{x}_i=\boldsymbol{z}_i$，这里仅为得到式(10.15)的形式才最终保留$\mathbf{W}$和$\boldsymbol{x}_i$的；若令$\mathbf{Z}=(\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_m)\in\mathbb{R}^{d'\times m}$为低维坐标系中的样本集合，则$\mathbf{Z}=\mathbf{W}^{\top}\mathbf{X}$，即$\boldsymbol{z}_i$为矩阵$\mathbf{Z}$的第$i$列；而$\sum_{i=1}^{m}\left\|\mathbf{W}^{\top}\boldsymbol{x}_i\right\|_2^2=\sum_{i=1}^{m}\|\boldsymbol{z}_i\|_2^2$表示$\mathbf{Z}$所有列向量 2 范数的平方，也就是$\mathbf{Z}$所有元素的平方和，即为$\|\mathbf{Z}\|_F^2$，此即第一个等号的由来；而根据 10.0 中“矩阵的 F 范数与迹”中第(3)个结论，即对于矩阵$\mathbf{Z}$有$\|\mathbf{Z}\|_F^2=\mathrm{tr}\left(\mathbf{Z}^{\top}\mathbf{Z}\right)=\mathrm{tr}\left(\mathbf{Z}\mathbf{Z}^{\top}\right)$，其中$\mathrm{tr}(\cdot)$表示求矩阵的迹，即对角线元素之和，此即第二个等号的由来；第三个等号将转置化简即得。

到此即得式(10.15)的目标函数，约束条件$\mathbf{W}^{\top}\mathbf{W}=\mathbf{I}$已在推导中说明。

式(10.15)的目标函数式(10.14)结果略有差异，接下来推导$\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\top}=\mathbf{X}\mathbf{X}^{\top}$以弥补这个差异（这个结论推导一遍记下来好了）。

先化简$\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\top}$，首先：

$$
\boldsymbol{x}_i\boldsymbol{x}_i^{\top}=\begin{bmatrix}x_{i1}\\x_{i2}\\\vdots\\x_{id}\end{bmatrix}\begin{bmatrix}x_{i1} & x_{i2} & \cdots & x_{id}\end{bmatrix}=\begin{bmatrix}x_{i1}x_{i1} & x_{i1}x_{i2} & \cdots & x_{i1}x_{id}\\x_{i2}x_{i1} & x_{i2}x_{i2} & \cdots & x_{i2}x_{id}\\\vdots & \vdots & \ddots & \vdots\\x_{id}x_{i1} & x_{id}x_{i2} & \cdots & x_{id}x_{id}\end{bmatrix}_{d\times d}
$$

整体代入求和号$\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\top}$，得

$$
\begin{aligned}
\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^{\top} &= \sum_{i=1}^{m}\begin{bmatrix}x_{i1}x_{i1} & x_{i1}x_{i2} & \cdots & x_{i1}x_{id}\\x_{i2}x_{i1} & x_{i2}x_{i2} & \cdots & x_{i2}x_{id}\\\vdots & \vdots & \ddots & \vdots\\x_{id}x_{i1} & x_{id}x_{i2} & \cdots & x_{id}x_{id}\end{bmatrix}_{d\times d}\\
&= \begin{bmatrix}\sum_{i=1}^{m}x_{i1}x_{i1} & \sum_{i=1}^{m}x_{i1}x_{i2} & \cdots & \sum_{i=1}^{m}x_{i1}x_{id}\\\sum_{i=1}^{m}x_{i2}x_{i1} & \sum_{i=1}^{m}x_{i2}x_{i2} & \cdots & \sum_{i=1}^{m}x_{i2}x_{id}\\\vdots & \vdots & \ddots & \vdots\\\sum_{i=1}^{m}x_{id}x_{i1} & \sum_{i=1}^{m}x_{id}x_{i2} & \cdots & \sum_{i=1}^{m}x_{id}x_{id}\end{bmatrix}_{d\times d}
\end{aligned}
$$

再化简$\mathbf{X}\mathbf{X}^{\top}\in\mathbb{R}^{d\times d}$：

$$
\mathbf{X}\mathbf{X}^{\top}=\begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_d\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_1^{\top}\\\boldsymbol{x}_2^{\top}\\\vdots\\\boldsymbol{x}_d^{\top}\end{bmatrix}
$$

将列向量$\boldsymbol{x}_i=(x_{i1};x_{i2};\ldots;x_{id})\in\mathbb{R}^{d\times 1}$代入：

$$\mathbf{X}\mathbf{X}^\top = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1d} & x_{2d} & \cdots & x_{md} \end{bmatrix}_{d\times m} \bullet \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} \end{bmatrix}_{m\times d}$$
$$= \begin{bmatrix} \sum_{i=1}^m x_{i1}x_{i1} & \sum_{i=1}^m x_{i1}x_{i2} & \cdots & \sum_{i=1}^m x_{i1}x_{id} \\ \sum_{i=1}^m x_{i2}x_{i1} & \sum_{i=1}^m x_{i2}x_{i2} & \cdots & \sum_{i=1}^m x_{i2}x_{id} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^m x_{id}x_{i1} & \sum_{i=1}^m x_{id}x_{i2} & \cdots & \sum_{i=1}^m x_{id}x_{id} \end{bmatrix}_{d\times d}$$

综合$\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^\top$和$\mathbf{X}\mathbf{X}^\top$的化简结果，即$\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^\top = \mathbf{X}\mathbf{X}^\top$（协方差矩阵）。

根据刚刚推导得到的结论，式(10.14)最后的结果即可化为式(10.15)的目标函数：

$$\operatorname{tr}\left(\mathbf{W}^\top\left(\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)\mathbf{W}\right) = \operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right)$$

式(10.15)描述的优化问题的求解详见式(10.17)最后的解释。

2、式(10.16)的解释

先说什么是方差：

对于包含n个样本的一组数据$X = \{x_1, x_2, \ldots, x_n\}$来说，均值$M$为

$$M = \frac{x_1 + x_2 + \ldots + x_n}{n} = \sum_{i=1}^n x_i$$

则方差σ_X^2公式为

$$\sigma^2 = \frac{(x_1 - M)^2 + (x_2 - M)^2 + \ldots + (x_n - M)^2}{n}$$
$$= \frac{1}{n}\sum_{i=1}^n (x_i - M)^2$$

方差衡量了该组数据偏离均值的程度；样本越分散，其方差越大。

再说什么是协方差：

若还有包含n个样本的另一组数据$X' = \{x'_1, x'_2, \ldots, x'_n\}$，均值为$M'$，则下式

$$\sigma_{XX'}^2 = \frac{(x_1 - M)(x'_1 - M') + (x_2 - M)(x'_2 - M') + \ldots + (x_n - M)(x'_n - M')}{n}$$
$$= \frac{1}{n}\sum_{i=1}^n (x_i - M)(x'_i - M')$$

称为两组数据的协方差。$\sigma_{XX'}^2$能说明第一组数据$x_1, x_2, \ldots, x_n$和第二组数据$x'_1, x'_2, \ldots, x'_n$的变化情况。具体来说，如果两组数据总是同时大于或小于自己的均值，则$(x_i - M)(x'_i - M') > 0$，此时$\sigma_{XX'}^2 > 0$；如果两组数据总是一个大于(或小于)自己的均值而别一个小于(或大于)自己的均值，则$(x_i - M)(x'_i - M') < 0$，此时$\sigma_{XX'}^2 < 0$；如果两组数据与自己的均值的大小关系无规律，则$(x_i - M)(x'_i - M')$的正负号随机变化，其平均数$\sigma_{XX'}^2$则会趋近于 0。引用百度百科协方差词条原话：“从直观上来看，协方差表示的是两个变量总体误差的期望。如果两个变量的变化趋势一致，也就是说如果其中一个大于自身的期望值时另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值；如果两个变量的变化趋势相反，即其中一个变量大于自身的期望值时另外一个却小于自身的期望值，

那么两个变量之间的协方差就是负值。如果两个变量是统计独立的，那么二者之间的协方差就是 0，但是，反过来并不成立。协方差为 0 的两个随机变量称为是不相关的。”

最后说什么是协方差矩阵：

结合本书中的符号：

$$\mathbf{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m)=\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1d} & x_{2d} & \cdots & x_{md} \end{bmatrix}_{d\times m}$$

矩阵$\mathbf{X}$每一行表示一维特征，每一列表示该数据集的一个样本；而本节开始已假定数据样本进行了中心化，即$\sum_{i=1}^{m}\boldsymbol{x}_i=\mathbf{0}\in\mathbb{R}^{d\times 1}$（中心化过程可通过$\mathbf{X}(\mathbf{I}-\frac{1}{m}\mathbf{1}\mathbf{1}^{\top})$实现，其中$\mathbf{I}\in\mathbb{R}^{m\times m}$为单位阵，$\mathbf{1}\in\mathbb{R}^{m\times 1}$为全 1 列向量，参见习题 10.3），即上式矩阵的每一行平均值等于零（其实就是分别对所有$\boldsymbol{x}_i$的每一维坐标进行中心化，而不是分别对单个样本$\boldsymbol{x}_i$中心化）。对于包含d个特征的特征空间（或称d维特征空间）来说，每一维特征可以看成是一个随机变量，而$\mathbf{X}$中包含m个样本，也就是说每个随机变量有m个数据，根据前面$\mathbf{X}\mathbf{X}^{\top}$的矩阵表达形式：

$$\frac{1}{m}\mathbf{X}\mathbf{X}^{\top}=\frac{1}{m}\begin{bmatrix} \sum_{i=1}^{m}x_{i1}x_{i1} & \sum_{i=1}^{m}x_{i1}x_{i2} & \cdots & \sum_{i=1}^{m}x_{i1}x_{id} \\ \sum_{i=1}^{m}x_{i2}x_{i1} & \sum_{i=1}^{m}x_{i2}x_{i2} & \cdots & \sum_{i=1}^{m}x_{i2}x_{id} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{m}x_{id}x_{i1} & \sum_{i=1}^{m}x_{id}x_{i2} & \cdots & \sum_{i=1}^{m}x_{id}x_{id} \end{bmatrix}_{d\times d}$$

根据前面的结果知道$\frac{1}{m}\mathbf{X}\mathbf{X}^{\top}$的第$i$行第$j$列的元素表示$\mathbf{X}$中第$i$行和$\mathbf{X}^{\top}$第$j$列（即$\mathbf{X}$中第$j$行）的方差$(i=j)$或协方差$(i\neq j)$。**注意：协方差矩阵对角线元素为各行的方差**。

接下来正式解释式(10.16)：

对于$\mathbf{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m)\in\mathbb{R}^{d\times m}$，将其投影为$\mathbf{Z}=(\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_m)\in\mathbb{R}^{d'\times m}$，其中$\mathbf{Z}=\mathbf{W}^{\top}\mathbf{X}$，变换矩阵$\mathbf{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'}\}\in\mathbb{R}^{d\times d'}$为一组新的标准正交基。从最大可分性出发，我们希望在新空间的每一维坐标轴上样本都尽可能分散（即每维特征尽可能分散，也就是$\mathbf{Z}$各行方差最大；参见图 10.4 所示，原空间只有两维坐标，现考虑降至一维，希望在新坐标系下样本尽可能分散，图中画出了一种映射后的坐标系，显然橘红色坐标方向样本更分散，方差更大），即寻找$\mathbf{W}\in\mathbb{R}^{d\times d'}$使协方差矩阵$\frac{1}{m}\mathbf{Z}\mathbf{Z}^{\top}$对角线元素之和（矩阵的迹）最大（即使$\mathbf{Z}$各行方差之和最大），由于$\mathbf{Z}=\mathbf{W}^{\top}\mathbf{X}$，而常系数$\frac{1}{m}$在最大化时并不发生影响，求矩阵对角线元素之和即为矩阵的迹，综上即得式(10.16)。

另外，中心化后$\mathbf{X}$的各行均值为零，变换后$\mathbf{Z}=\mathbf{W}^{\top}\mathbf{X}$的各行均值仍为零，这是因为$\mathbf{Z}$的第$i$行$(1\leqslant i\leqslant d')$为$\{\boldsymbol{w}_i^{\top}\boldsymbol{x}_1,\boldsymbol{w}_i^{\top}\boldsymbol{x}_2,\ldots,\boldsymbol{w}_i^{\top}\boldsymbol{x}_m\}$，该行之和$\boldsymbol{w}_i^{\top}\sum_{j=1}^{m}\boldsymbol{x}_j=\boldsymbol{w}_i^{\top}\mathbf{0}=0$。

最后，有关方差的公式，有人认为应该除以样本数量m，有人认为应该除以样本数量减 1 即$m-1$。简单来说，根据总体样本集求方差就除以总体样本数量，而根据抽样样本集求方差就除以抽样样本集数量减 1；总体样本集是真正想调查的对象集合，而抽样样本集是从

总体样本集中被选出来的部分样本组成的集合，用来估计总体样本集的方差；一般来说，总体样本集是不可得的，我们拿到的都是抽样样本集。CSDN 博主 hearthougan 有篇博客《彻底理解样本方差为何除以 n-1》通过严格的数学推导证明了样本方差应该除以 n-1 才会得到总体样本的无偏估计，若除以 n 则得到的是有偏估计。

式(10.16)描述的优化问题的求解详见式(10.17)最后的解释。

3、式(10.17)的推导

注意若要对式(10.16)使用拉格朗日乘子法应先将最大化问题转为式(10.15)最小化问题。

对式(10.15)使用拉格朗日乘子法，写出拉格朗日函数（此处与书中符号有差别）：

$$L(\mathbf{W},\mathbf{\Lambda})=-\mathrm{tr}\left(\mathbf{W}^{\top}\mathbf{X}\mathbf{X}^{\top}\mathbf{W}\right)+\left(\mathbf{W}^{\top}\mathbf{W}-\mathbf{I}\right)\mathbf{\Lambda}$$

其中

$$\mathbf{\Lambda}=\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_{d'}\end{bmatrix}\in\mathbb{R}^{d'\times d'},\ \mathbf{I}=\begin{bmatrix}1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}\in\mathbb{R}^{d'\times d'}$$

对$\mathbf{W}\in\mathbb{R}^{d\times d'}$求导：

$$\begin{aligned}\frac{\partial L(\mathbf{W},\mathbf{\Lambda})}{\partial\mathbf{W}}&=-\frac{\partial\mathrm{tr}\left(\mathbf{W}^{\top}\mathbf{X}\mathbf{X}^{\top}\mathbf{W}\right)}{\partial\mathbf{W}}+\frac{\partial\left(\mathbf{W}^{\top}\mathbf{W}-\mathbf{I}\right)}{\partial\mathbf{W}}\mathbf{\Lambda}\\&=-\mathbf{X}\mathbf{X}^{\top}\mathbf{W}-(\mathbf{X}\mathbf{X}^{\top})^{\top}\mathbf{W}+2\mathbf{W}\mathbf{\Lambda}\\&=-2\mathbf{X}\mathbf{X}^{\top}\mathbf{W}+2\mathbf{W}\mathbf{\Lambda}\end{aligned}$$

有关矩阵的迹求导可以搜索《The Matrix Cookbook(Version: November 15, 2012)》（式 108）：
http://202.119.24.249/cache/13/03/www2.imm.dtu.dk/e5ee3ed66202b93eaa50e31d5cf2fb3a/imm3274.pdf

令偏导$\frac{\partial L(\mathbf{W},\mathbf{\Lambda})}{\partial\mathbf{W}}=0$，得

$$\mathbf{X}\mathbf{X}^{\top}\mathbf{W}=\mathbf{W}\mathbf{\Lambda}$$

此即第 13 次印刷及之前的式(10.17)的正确形式；将此式拆成d'个式子

$$\mathbf{X}\mathbf{X}^{\top}\boldsymbol{w}_i=\lambda_i\boldsymbol{w}_i,\ 1\leqslant i\leqslant d'$$

此即第 14 次印刷修订后的式(10.17)。注意，第 13 次印刷及之前的式(10.17)是有误的，等号右侧不应该是$\lambda\mathbf{W}$，而应该是对$\mathbf{W}$的每列乘不同的拉格朗日乘子，即$\mathbf{W}\mathbf{\Lambda}$。

这个形式是不是似曾相似？对，就是求矩阵$\mathbf{X}\mathbf{X}^{\top}\in\mathbb{R}^{d\times d}$特征值的形式！

4、根据式(10.17)求解式(10.16)

注意式(10.16)中$\mathbf{W}\in\mathbb{R}^{d\times d'}$，只有$d'$列，而式(10.17)可以得到$d$列，如何根据式(10.17)求解式(10.16)呢？

对$\mathbf{X}\mathbf{X}^{\top}\mathbf{W}=\mathbf{W}\mathbf{\Lambda}$两边同乘$\mathbf{W}^{\top}$，得

$$\mathbf{W}^{\top}\mathbf{X}\mathbf{X}^{\top}\mathbf{W}=\mathbf{W}^{\top}\mathbf{W}\mathbf{\Lambda}=\mathbf{\Lambda}$$

注意使用了约束条件$\mathbf{W}^{\top}\mathbf{W}=\mathbf{I}$；上式左边与式(10.16)的优化目标对应矩阵相同，而右边$\mathbf{\Lambda}\in\mathbb{R}^{d'\times d'}$是由$\mathbf{X}\mathbf{X}^{\top}$的$d'$个特征值组成的对角阵，两边同时取矩阵的迹，得

$$\mathrm{tr}(\mathbf{W}^{\top}\mathbf{X}\mathbf{X}^{\top}\mathbf{W})=\mathrm{tr}(\mathbf{\Lambda})=\sum_{i=1}^{d'}\lambda_i$$

左边即式(10.16)的优化目标，最大化优化目标相当于最大化$\sum_{i=1}^{d'}\lambda_i$。但$\mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{d\times d}$共有$d$个特征值，因此当然是取出最大的前$d'$个特征值，而$\mathbf{W}$即特征值对应的标准化特征向量组成的矩阵。

在 Malab 中可以直接使用[V,D] = eig(X*X')即可，其中 V 每列是一个特征向量，D 为特征值组成的对角阵，默认按从小到大排列，与 V 相对应，因此取 V 的后d'列组成$\mathbf{W}$即可。

特别注意，图 10.5 只是得到了投影矩阵$\mathbf{W}$，而降维后的样本为$\mathbf{Z}=\mathbf{W}^\top\mathbf{X}$。

10.4 核化线性降维

本节符号在第 14 次印刷中进行了修订，详见勘误修订，直接搜索页码即可。

另外略有一点混乱的是，在上一节中用$\boldsymbol{z}_i$表示$\boldsymbol{x}_i$降维后的像，而本节用$\boldsymbol{z}_i$表示$\boldsymbol{x}_i$在高维特征空间中的像。

本节推导实际上有一个前提，以式(10.19)为例（式(10.21)仅将$\boldsymbol{z}_i$换为$\phi(\boldsymbol{x}_i)$而已），那就是$\boldsymbol{z}_i$已经中心化（计算方差要用样本减去均值，式(10.19)是均值为零时特殊形式，详见式(10.16)的解释），但$\boldsymbol{z}_i=\phi(\boldsymbol{x}_i)$是$\boldsymbol{x}_i$高维特征空间中的像，即使$\boldsymbol{x}_i$已进行中心化，但$\boldsymbol{z}_i$却不一定是中心化的，此时本节推导均不再成立。推广工作详见 KPCA 原始文献的附录 A。

1、式(10.19)的解释

首先，类似于式(10.14)的推导后半部分内容可知$\sum_{i=1}^{m}\boldsymbol{z}_i\boldsymbol{z}_i^\top=\mathbf{Z}\mathbf{Z}^\top$，其中$\mathbf{Z}$的每一列为一个样本，设高维空间的维度为$h$，则$\mathbf{Z}\in\mathbb{R}^{h\times m}$，其中$m$为数据集样本数量。

其次，式(10.19)中的$\mathbf{W}$为从高维空间降至低维（维度为d）后的正交基，在第 14 次印刷中加入表述$\mathbf{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_d)$，其中$\mathbf{W}\in\mathbb{R}^{h\times d}$，降维过程为$\mathbf{X}=\mathbf{W}^\top\mathbf{Z}$。

最后，式(10.19)类似于式(10.17)，是为了求解降维投影矩阵$\mathbf{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_d)$。但问题在于$\mathbf{Z}\mathbf{Z}^\top\in\mathbb{R}^{h\times h}$，当维度$h$很大时（注意本节为核化线性降维，第六章核方法中高斯核会把样本映射至无穷维），此时根本无法求解$\mathbf{Z}\mathbf{Z}^\top$的特征值和特征向量。因此才有了后面的式(10.20)。

第 14 次印刷及之后印次，式(10.19)为$\left(\sum_{i=1}^{m}\boldsymbol{z}_i\boldsymbol{z}_i^\top\right)\boldsymbol{w}_j=\lambda_j\boldsymbol{w}_j$，而在之前的印次中表达有误，实际应该为$\left(\sum_{i=1}^{m}\boldsymbol{z}_i\boldsymbol{z}_i^\top\right)\mathbf{W}=\mathbf{W}\mathbf{\Lambda}$，类似于式(10.17)。而这两种表达本质相同，$\lambda_j\boldsymbol{w}_j$为$\mathbf{W}\mathbf{\Lambda}$的第$j$列，仅此而已。

2、式(10.20)的解释

本节为核化线性降维，而式(10.19)是在维度为h的高维空间运算，式(10.20)变形（咋一看似乎有点无厘头）的目的是为了避免直接在高维空间运算，即想办法能够使用第 6 章的式(6.22)的核技巧，也就是后面的式(10.24)。

第 14 次印刷及之后印次该式没问题，之前的式(10.20)应该是：

$$\begin{aligned}\mathbf{W}&=\left(\sum_{i=1}^{m}\boldsymbol{z}_i\boldsymbol{z}_i^\top\right)\mathbf{W}\mathbf{\Lambda}^{-1}=\sum_{i=1}^{m}\left(\boldsymbol{z}_i\left(\boldsymbol{z}_i^\top\mathbf{W}\mathbf{\Lambda}^{-1}\right)\right)\\&=\sum_{i=1}^{m}\left(\boldsymbol{z}_i\boldsymbol{\alpha}_i\right)\end{aligned}$$

其中$\boldsymbol{\alpha}_i=\boldsymbol{z}_i^\top\mathbf{W}\mathbf{\Lambda}^{-1}\in\mathbb{R}^{1\times d}$，$\boldsymbol{z}_i^\top\in\mathbb{R}^{1\times h}$，$\mathbf{W}\in\mathbb{R}^{h\times d}$，$\mathbf{\Lambda}\in\mathbb{R}^{d\times d}$为对角阵。这个结果

看似等号右侧也包含$\mathbf{W}$，但将此式代入式(10.19)后经化简可避免在高维空间的运算，而将目标转化为求低维空间的$\boldsymbol{\alpha}_i \in \mathbb{R}^{1\times d}$，详见式(10.24)的推导。

3、式(10.21)的解释

该式即为将式(10.19)中的$\boldsymbol{z}_i$换为$\phi(\boldsymbol{x}_i)$的结果。

4、式(10.22)的解释

该式即为将式(10.20)中的$\boldsymbol{z}_i$换为$\phi(\boldsymbol{x}_i)$的结果。

5、式(10.24)的推导

我们先以第 14 次印刷及之后的版本来推导。

鉴于式(10.19)(10.20)和式(10.21)(10.22)的区别仅在于$\boldsymbol{z}_i$和$\phi(\boldsymbol{x}_i)$的不同表达形式，此处以式(10.19)(10.20)来推导。

式(10.19)写为矩阵形式为$\mathbf{Z}\mathbf{Z}^{\top}\boldsymbol{w}_j = \lambda_j \boldsymbol{w}_j$；

式(10.20)写为矩阵形式为$\boldsymbol{w}_j = \mathbf{Z}\boldsymbol{\alpha}^j$，其中$\boldsymbol{\alpha}^j = (\alpha_1^j; \alpha_2^j; \ldots; \alpha_m^j) \in \mathbb{R}^{m\times 1}$；

将式(10.20)代入式(10.19)得，$\mathbf{Z}\mathbf{Z}^{\top}\mathbf{Z}\boldsymbol{\alpha}^j = \lambda_j \mathbf{Z}\boldsymbol{\alpha}^j$；

两边同时左乘$\mathbf{Z}^{\top}$得，$\mathbf{Z}^{\top}\mathbf{Z}\mathbf{Z}^{\top}\mathbf{Z}\boldsymbol{\alpha}^j = \lambda_j \mathbf{Z}^{\top}\mathbf{Z}\boldsymbol{\alpha}^j$；

注意，若$\boldsymbol{z}_i = \phi(\boldsymbol{x}_i)$，式(10.23)的核函数对应的核矩阵为$\mathbf{K} = \mathbf{Z}^{\top}\mathbf{Z}$；

因此$\mathbf{Z}^{\top}\mathbf{Z}\mathbf{Z}^{\top}\mathbf{Z}\boldsymbol{\alpha}^j = \lambda_j \mathbf{Z}^{\top}\mathbf{Z}\boldsymbol{\alpha}^j$变为$\mathbf{K}\mathbf{K}\boldsymbol{\alpha}^j = \lambda_j \mathbf{K}\boldsymbol{\alpha}^j$，此式即为本节开头提到的 KPCA 的原始参考文献中的式(11)。相比为式(10.24)，等号两边多余$\mathbf{K}$，若$\mathbf{K}$可逆，则两边左乘其逆矩阵即可消去，但核矩阵$\mathbf{K}$一般仅为半正定矩阵，不一定可逆。KPCA 的原始参考文献中此时提到“As $\mathbf{K}$ is symmetric, it has a set of Eigenvectors which spans the whole space”，因此就得到式(10.24)，即原始参考文献中的式(12)，具体原因有待琢磨。

若改为第 13 次印刷及之前的版本，则：

式(10.19)写为矩阵形式为$\mathbf{Z}\mathbf{Z}^{\top}\mathbf{W} = \mathbf{W}\boldsymbol{\Lambda}$；

式(10.20)写为矩阵形式为$\mathbf{W} = \mathbf{Z}\mathbf{A}$，其中$\mathbf{A} = (\boldsymbol{\alpha}_1; \boldsymbol{\alpha}_2; \ldots; \boldsymbol{\alpha}_m) \in \mathbb{R}^{m\times d}$；

将式(10.20)代入式(10.19)得，$\mathbf{Z}\mathbf{Z}^{\top}\mathbf{Z}\mathbf{A} = \mathbf{Z}\mathbf{A}\boldsymbol{\Lambda}$；

两边同时左乘$\mathbf{Z}^{\top}$得，$\mathbf{Z}^{\top}\mathbf{Z}\mathbf{Z}^{\top}\mathbf{Z}\mathbf{A} = \mathbf{Z}^{\top}\mathbf{Z}\mathbf{A}\boldsymbol{\Lambda}$，即$\mathbf{K}\mathbf{K}\mathbf{A} = \mathbf{K}\mathbf{A}\boldsymbol{\Lambda}$，进而根据原始文献中相同的原因得到式(10.24)，正确形式应为$\mathbf{K}\mathbf{A} = \mathbf{A}\boldsymbol{\Lambda}$。

实际上，两个版次之间的差别仅在于较新版是较旧版某一列的表达形式而已。

由式(10.24)可知，求解$\boldsymbol{\alpha}^j$（或$\mathbf{A}$）只需对$m \times m$的核矩阵$\mathbf{K}$进行特征值分解即可，而不必在高维特征空间进行运算，这正是核技巧的作用。

6、式(10.25)的解释

式(10.25)仅需将第 14 次印刷中式(10.22)的$\boldsymbol{w}_j$表达式转置后代入即可。

该式的意义在于，求解新样本$\boldsymbol{x} \in \mathbb{R}^{d\times 1}$映射至高维空间$\phi(\boldsymbol{x}) \in \mathbb{R}^{h\times 1}$后再降至低维空间的像$\boldsymbol{y} \in \mathbb{R}^{d'\times 1}$时（即$\mathbb{R}^{h\times 1} \to \mathbb{R}^{d'\times 1}$，此时与样本原始特征$\mathbb{R}^{d\times 1}$没关系），可以避免在高维空间$\mathbb{R}^{h\times 1}$的运算。但是由于此处没有类似第 6 章支持向量的概念，可以发现式(10.25)计算时需要对所有样本求和，因此它的计算开销比较大。

注意，此处书中符号使用略有混乱，因为在式(10.19)中$\boldsymbol{z}_i$表示$\boldsymbol{x}_i$在高维特征空间中的像，而此处又用z_j表示新样本$\boldsymbol{x}$映射为$\phi(\boldsymbol{x})$后再降维至$\mathbb{R}^{d'\times 1}$空间时的第j维坐标。

10.5 流形学习

不要被“流形学习”的名字所欺骗，本节开篇就明确说了，它是一类借鉴了拓扑流形概念的降维方法而已，因此称为“流形学习”。10.2 节 MDS 算法的降维准则是要求原始空间中样本之间的距离在低维空间中得以保持，10.3 节 PCA 算法的降维准则是要求低维子空间对样本具有最大可分性，因为它们都是基于线性变换来进行降维的方法（参见式(10.13)），故称为线性降维方法。

1、等度量映射(Isomap)的解释

如图 10.8 所示，Isomap 算法与 10.2 节的 MDS 算法的区别仅在于距离矩阵$\mathbf{D} \in \mathbb{R}^{m\times m}$的计算方法不同。在 10.2 节的 MDS 算法中，距离矩阵$\mathbf{D} \in \mathbb{R}^{m\times m}$即为普通的样本之间欧氏距离；而本节的 Isomap 算法中，距离矩阵$\mathbf{D} \in \mathbb{R}^{m\times m}$由图 10.8 的 Step1~Step5 生成，即遵循流形假设。当然，对新样本降维时也有不同，这在图 10.8 下的一段话中已阐明。

另外解释一下测地线距离，欧氏距离即两点之间的直线距离，而测地线距离是实际中可以到达的路径，如图 10.7(a)中黑线（欧氏距离）和红线（测地线距离）。

2、式(10.28)的推导

推导过程参考了刘建平 Pinard 的博客：https://www.cnblogs.com/pinard/p/6266408.html 。

式(10.28)是式(10.27)的解，书中省略了较长的推导，整个推导过程如下：

由约束条件$\sum_{j\in Q_i} w_{ij} = 1$，则$\boldsymbol{x}_i = \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_i$，代入式(10.27)的 2 范数求和项：

$$\begin{aligned}\left\|\boldsymbol{x}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_j\right\|_2^2 &= \left\|\sum_{j\in Q_i} w_{ij}\boldsymbol{x}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_j\right\|_2^2 \\ &= \left\|\sum_{j\in Q_i} w_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right\|_2^2\end{aligned}$$

设K为 LLE 算法所取的近邻个数，令

$$\begin{aligned}\boldsymbol{w}_i &= (w_{i\kappa_1}; w_{i\kappa_2}; \ldots; w_{i\kappa_K}) \in \mathbb{R}^{K\times 1} \\ \mathbf{D}_i &= (\boldsymbol{x}_i - \boldsymbol{x}_{\kappa_1}, \boldsymbol{x}_i - \boldsymbol{x}_{\kappa_2}, \ldots, \boldsymbol{x}_i - \boldsymbol{x}_{\kappa_K}) \in \mathbb{R}^{d\times K}\end{aligned}$$

其中$\boldsymbol{w}_i$元素下标$_{\kappa_j}$表示K近邻中第j个样本的下标$(1 \leqslant j \leqslant K)$。即$\boldsymbol{w}_i$是$K \times 1$的列向量，$\mathbf{D}_i$是$d \times K$的矩阵，$d$表示样本$\boldsymbol{x}_i$原始特征空间的维度。再令$\mathbf{C}_i = \mathbf{D}_i^\top \mathbf{D}_i$，则

$$\begin{aligned}\left\|\boldsymbol{x}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_j\right\|_2^2 &= \left\|\sum_{j\in Q_i} w_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right\|_2^2 = \|\mathbf{D}_i\boldsymbol{w}_i\|_2^2 \\ &= (\mathbf{D}_i\boldsymbol{w}_i)^\top \mathbf{D}_i\boldsymbol{w}_i = \boldsymbol{w}_i^\top \mathbf{D}_i^\top \mathbf{D}_i\boldsymbol{w}_i \\ &= \boldsymbol{w}_i^\top \mathbf{C}_i\boldsymbol{w}_i\end{aligned}$$

令$\mathbf{I}_K$为元素均为 1 的$K \times 1$列向量，则$\sum_{j\in Q_i} w_{ij} = \boldsymbol{w}_i^\top \mathbf{I}_K$。此时式(10.27)重写为

$$\min_{\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_m} \sum_{i=1}^{m} \boldsymbol{w}_i^{\top}\mathbf{C}_i\boldsymbol{w}_i$$
$$\text{s.t. } \boldsymbol{w}_i^{\top}\mathbf{I}_K = 1, i = 1, 2, \ldots, m$$

其实可以发现，式(10.27)实际上可以拆为m个独立的优化问题：

$$\min_{\boldsymbol{w}_i} \boldsymbol{w}_i^{\top}\mathbf{C}_i\boldsymbol{w}_i \ , \ \text{s.t. } \boldsymbol{w}_i^{\top}\mathbf{I}_K = 1$$

应用拉格朗日乘子法，拉格朗日函数为：

$$L(\boldsymbol{w}_i, \lambda) = \boldsymbol{w}_i^{\top}\mathbf{C}_i\boldsymbol{w}_i + \lambda(\boldsymbol{w}_i^{\top}\mathbf{I}_K - 1)$$

对$\boldsymbol{w}_i$和λ分别求偏导，并令导数等于 0，得：

$$\frac{\partial L(\boldsymbol{w}_i, \lambda)}{\partial \boldsymbol{w}_i} = 2\mathbf{C}_i\boldsymbol{w}_i + \lambda\mathbf{I}_K = 0$$
$$\frac{\partial L(\boldsymbol{w}_i, \lambda)}{\partial \lambda} = \boldsymbol{w}_i^{\top}\mathbf{I}_K - 1 = 0$$

由第一个式子可得$\boldsymbol{w}_i = -\frac{\lambda}{2}\mathbf{C}_i^{-1}\mathbf{I}_K$，将该结果代入第二个式子，得

$$\left(-\frac{\lambda}{2}\mathbf{C}_i^{-1}\mathbf{I}_K\right)^{\top}\mathbf{I}_K - 1 = 0 \Rightarrow -\frac{\lambda}{2}\mathbf{I}_K^{\top}\left(\mathbf{C}_i^{-1}\right)^{\top}\mathbf{I}_K = 1$$

由于$\mathbf{C}_i = \mathbf{D}_i^{\top}\mathbf{D}_i$是对称矩阵，所以$\left(\mathbf{C}_i^{-1}\right)^{\top} = \mathbf{C}_i^{-1}$，因此$-\frac{\lambda}{2} = \frac{1}{\mathbf{I}_K^{\top}\mathbf{C}_i^{-1}\mathbf{I}_K}$，代入$\boldsymbol{w}_i$的表示式，即得

$$\boldsymbol{w}_i = \frac{\mathbf{C}_i^{-1}\mathbf{I}_K}{\mathbf{I}_K^{\top}\mathbf{C}_i^{-1}\mathbf{I}_K}$$

这就是式(10.28)的矩阵形式，分母$\mathbf{I}_K^{\top}\mathbf{C}_i^{-1}\mathbf{I}_K$即式(10.28)的$\sum_{l,s\in Q_i} C_{ls}^{-1}$，是一个常数；分子$\mathbf{C}_i^{-1}\mathbf{I}_K$为$K \times 1$的列向量，$K$个元素分别对应式(10.28)$w_{ij}$分子$(1 \leqslant j \leqslant K)$。需要特别注意的是，在式(10.28)的$w_{ij}$表达式中，$C_{jk}^{-1}$表示$\mathbf{C}_i$的逆矩阵$\mathbf{C}_i^{-1}$的第$j$行第$k$列的元素，而不是$\mathbf{C}_i$第$j$行第$k$列的元素的倒数。另外，式(10.28)解出的$w_{ij}$仅对应$\boldsymbol{x}_i$的$K$个近邻，即$j \in Q_i$；对于$j \notin Q_i$的$w_{ij}$统一设置为零，即图 10.10 的第 4 步。

除了上述推导之外，再提供另外一种推导：

采用与上述推导相同的符号表示，目标函数可以变形为：

$$\left\|\boldsymbol{x}_i - \sum_{j\in Q_i} w_{ij}\boldsymbol{x}_j\right\|_2^2 = \|\boldsymbol{x}_i - (w_{i\kappa_1}\boldsymbol{x}_{\kappa_1} + w_{i\kappa_2}\boldsymbol{x}_{\kappa_2} + w_{i\kappa_K}\boldsymbol{x}_{\kappa_K})\|_2^2$$
$$= \left\|\begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{bmatrix} - \begin{bmatrix} x_{\kappa_1 1}w_{i\kappa_1} + x_{\kappa_2 1}w_{i\kappa_2} + \ldots + x_{\kappa_K 1}w_{i\kappa_K} \\ x_{\kappa_1 2}w_{i\kappa_1} + x_{\kappa_2 2}w_{i\kappa_2} + \ldots + x_{\kappa_K 2}w_{i\kappa_K} \\ \vdots \\ x_{\kappa_1 d}w_{i\kappa_1} + x_{\kappa_2 d}w_{i\kappa_2} + \ldots + x_{\kappa_K d}w_{i\kappa_K} \end{bmatrix}\right\|_2^2$$

可以看出，最小化目标函数相当于求解如下方程组的最小二乘解：

$$
\begin{aligned}
&x_{\kappa_1 1}w_{i\kappa_1}+x_{\kappa_2 1}w_{i\kappa_2}+\ldots+x_{\kappa_K 1}w_{i\kappa_K}=x_{i1}\\
&x_{\kappa_1 2}w_{i\kappa_1}+x_{\kappa_2 2}w_{i\kappa_2}+\ldots+x_{\kappa_K 2}w_{i\kappa_K}=x_{i2}\\
&\cdots\\
&x_{\kappa_1 d}w_{i\kappa_1}+x_{\kappa_2 d}w_{i\kappa_2}+\ldots+x_{\kappa_K d}w_{i\kappa_K}=x_{id}
\end{aligned}
$$

即$\mathbf{X}_K\boldsymbol{w}_i=\boldsymbol{x}_i$，其中$\mathbf{X}_K=[\boldsymbol{x}_{\kappa_1},\boldsymbol{x}_{\kappa_2},\ldots,\boldsymbol{x}_{\kappa_K}]$，$\boldsymbol{w}_i=(w_{i\kappa_1};w_{i\kappa_2};\ldots;w_{i\kappa_K})$，易知该方程组的最小二乘解为$\boldsymbol{w}_i=(\mathbf{X}_K^\top\mathbf{X}_K)^{-1}\mathbf{X}_K^\top\boldsymbol{x}_i$，一般由此得到的$\boldsymbol{w}_i$已经满足约束条件，不放心的话可以自行验证一下$\boldsymbol{w}_i$元素之和是否为 1，或者自行进行规范化。需要注意的是，根据$\mathbf{X}_K$的特征，以上方程组可能有无数组解、唯一解或无解（此时即需要求最小二乘解），具体参见《压缩感知中的数学知识：线性方程组的解》(https://blog.csdn.net/jbb0523/article/details/41577721)。

可以做个简单实验验证，该最小二乘解与前面的推导结果数值上是相同的。

3、式(10.31)的推导

以下推导需要使用 10.0 预备知识中的第 3 条：矩阵的 F 范数与迹。

观察式(10.29)，求和号内实际是一个列向量的 2 范数平方，令$\boldsymbol{v}_i=\boldsymbol{z}_i-\sum_{j\in Q_i}w_{ij}\boldsymbol{z}_j$，$\boldsymbol{v}_i$的维度与$\boldsymbol{z}_i$相同，$\boldsymbol{v}_i\in\mathbb{R}^{d'\times 1}$，则式(10.29)可重写为

$$
\begin{aligned}
&\min_{\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_m}\sum_{i=1}^{m}\|\boldsymbol{v}_i\|_2^2\\
&\quad\text{s.t. }\boldsymbol{v}_i=\boldsymbol{z}_i-\sum_{j\in Q_i}w_{ij}\boldsymbol{z}_j,i=1,2,\ldots,m
\end{aligned}
$$

令$\mathbf{Z}=(\boldsymbol{z}_1,\boldsymbol{z}_2,\ldots,\boldsymbol{z}_i,\ldots,\boldsymbol{z}_m)\in\mathbb{R}^{d'\times m}$，$\mathbf{I}_i=(0;0;\ldots;1;\ldots;0)\in\mathbb{R}^{m\times 1}$，即$\mathbf{I}_i$为$m\times 1$的列向量，除第$i$个元素等于 1 之外其余元素均为零，则

$$\boldsymbol{z}_i=\mathbf{Z}\mathbf{I}_i$$

令$(\mathbf{W})_{ij}=w_{ij}$（P237 页第 1 行），即$\mathbf{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_i,\ldots,\boldsymbol{w}_m)^\top\in\mathbb{R}^{m\times m}$，也就是说$\mathbf{W}$的第$i$行的转置（没错，就是第$i$行）对应第$i$个样本$\boldsymbol{x}_i$的由式(10.28)解出来的系数$\boldsymbol{w}_i$（这里符号之所以别扭是因为$w_{ij}$已用来表示列向量$\boldsymbol{w}_i$的第$j$个元素，但为了与习惯保持一致即$w_{ij}$表示$\mathbf{W}$的第$i$行第$j$列元素，只能忍忍，此处暂时别扭着），即

$$
\mathbf{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_i,\ldots,\boldsymbol{w}_m)^\top=\begin{bmatrix}
w_{11} & w_{21} & \cdots & w_{i1} & \cdots & w_{m1}\\
w_{12} & w_{22} & \cdots & w_{i2} & \cdots & w_{m2}\\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\
w_{1j} & w_{2j} & \cdots & w_{ij} & \cdots & w_{mj}\\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots\\
w_{1m} & w_{2m} & \cdots & w_{im} & \cdots & w_{mm}
\end{bmatrix}^\top
$$

对于$\boldsymbol{w}_i\in\mathbb{R}^{m\times 1}$来说，只有$\boldsymbol{x}_i$的$K$个近邻样本对应的下标对应的$w_{ij}\neq 0,j\in Q_i$，且它们的和等于 1，则

$$\sum_{j\in Q_i}w_{ij}\boldsymbol{z}_j=\mathbf{Z}\boldsymbol{w}_i$$

因此

$$\boldsymbol{v}_i=\boldsymbol{z}_i-\sum_{j\in Q_i}w_{ij}\boldsymbol{z}_j=\mathbf{Z}\mathbf{I}_i-\mathbf{Z}\boldsymbol{w}_i=\mathbf{Z}(\mathbf{I}_i-\boldsymbol{w}_i)$$

令$\mathbf{V}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_i,\ldots,\boldsymbol{v}_m)\in\mathbb{R}^{d'\times m}$，$\mathbf{I}=(\mathbf{I}_1,\mathbf{I}_2,\ldots,\mathbf{I}_i,\ldots,\mathbf{I}_m)\in\mathbb{R}^{m\times m}$，则

$$\mathbf{V}=\mathbf{Z}(\mathbf{I}-\mathbf{W}^\top)=\mathbf{Z}(\mathbf{I}^\top-\mathbf{W}^\top)=\mathbf{Z}(\mathbf{I}-\mathbf{W})^\top$$

根据前面的预备知识，并将上式$\mathbf{V}$和式(10.30)代入，得式(10.31)目标函数：

$$\begin{aligned}\sum_{i=1}^{m}\|\boldsymbol{v}_i\|_2^2&=\|\mathbf{V}\|_F^2\\&=\mathrm{tr}\left(\mathbf{V}\mathbf{V}^\top\right)\\&=\mathrm{tr}\left(\left(\mathbf{Z}(\mathbf{I}-\mathbf{W})^\top\right)\left(\mathbf{Z}(\mathbf{I}-\mathbf{W})^\top\right)^\top\right)\\&=\mathrm{tr}\left(\mathbf{Z}(\mathbf{I}-\mathbf{W})^\top(\mathbf{I}-\mathbf{W})\mathbf{Z}^\top\right)\\&=\mathrm{tr}\left(\mathbf{Z}\mathbf{M}\mathbf{Z}^\top\right)\end{aligned}$$

接下来求解式(10.31)。

参考式(10.17)的推导，应用拉格朗日乘子法，先写出拉格朗日函数

$$L(\mathbf{Z},\mathbf{\Lambda})=\mathrm{tr}\left(\mathbf{Z}\mathbf{M}\mathbf{Z}^\top\right)+\left(\mathbf{Z}\mathbf{Z}^\top-\mathbf{I}\right)\mathbf{\Lambda}$$

令$\mathbf{P}=\mathbf{Z}^\top$（否则有点别扭），则拉格朗日函数变为

$$L(\mathbf{P},\mathbf{\Lambda})=\mathrm{tr}\left(\mathbf{P}^\top\mathbf{M}\mathbf{P}\right)+\left(\mathbf{P}^\top\mathbf{P}-\mathbf{I}\right)\mathbf{\Lambda}$$

求导并令导数等于 0：

$$\begin{aligned}\frac{\partial L(\mathbf{P},\mathbf{\Lambda})}{\partial\mathbf{P}}&=\frac{\partial\mathrm{tr}\left(\mathbf{P}^\top\mathbf{M}\mathbf{P}\right)}{\partial\mathbf{P}}+\frac{\partial\left(\mathbf{P}^\top\mathbf{P}-\mathbf{I}\right)}{\partial\mathbf{P}}\mathbf{\Lambda}\\&=2\mathbf{M}\mathbf{P}-2\mathbf{P}\mathbf{\Lambda}=\mathbf{0}\end{aligned}$$

有关矩阵的迹求导可以搜索《The Matrix Cookbook(Version: November 15, 2012)》（式 108）：
http://202.119.24.249/cache/13/03/www2.imm.dtu.dk/e5ee3ed66202b93eaa50e31d5cf2fb3a/imm3274.pdf

解得$\mathbf{M}\mathbf{P}=\mathbf{P}\mathbf{\Lambda}$，因此$\mathbf{P}\in\mathbb{R}^{m\times d'}$和$\mathbf{\Lambda}\in\mathbb{R}^{d'\times d'}$分别为$\mathbf{M}\in\mathbb{R}^{m\times m}$的部分特征向量矩阵和特征值对角阵；然后两边再同时左乘$\mathbf{P}^\top$并取矩阵的迹，注意$\mathbf{P}^\top\mathbf{P}=\mathbf{I}\in\mathbb{R}^{d'\times d'}$，得

$$\mathrm{tr}(\mathbf{P}^\top\mathbf{M}\mathbf{P})=\mathrm{tr}(\mathbf{P}^\top\mathbf{P}\mathbf{\Lambda})=\mathrm{tr}(\mathbf{\Lambda})$$

因此，$\mathbf{P}=\mathbf{Z}^\top$是由$\mathbf{M}\in\mathbb{R}^{m\times m}$最小的$d'$个特征值对应的特征向量组成的矩阵。

10.6 度量学习

回忆 10.5.1 节的 Isomap 算法相比与 10.2 节的 MDS 算法的区别在于距离矩阵的计算方法不同，Isomap 算法在计算样本间距离时使用的（近似）测地线距离，而 MDS 算法使用的是欧氏距离，也就是说二者的距离度量不同。

1、式(10.34)的解释

为了推导方便，令$\boldsymbol{u}=(u_1;u_2;\ldots;u_d)=\boldsymbol{x}_i-\boldsymbol{x}_j\in\mathbb{R}^{d\times 1}$，其中$u_k=x_{ik}-x_{jk}$，则式(10.34)重写为$\boldsymbol{u}^\top\mathbf{M}\boldsymbol{u}=\|\boldsymbol{u}\|_{\mathbf{M}}^2$，其中$\mathbf{M}\in\mathbb{R}^{d\times d}$，具体到元素级别的表达：

$$
\begin{aligned}
\boldsymbol{u}^{\top}\mathbf{M}\boldsymbol{u} &= \begin{bmatrix} u_1 & u_2 & \dots & u_d \end{bmatrix} \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1d} \\ m_{21} & m_{22} & \dots & m_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ m_{d1} & m_{d2} & \dots & m_{dd} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{bmatrix} \\
&= \begin{bmatrix} u_1 & u_2 & \dots & u_d \end{bmatrix} \begin{bmatrix} u_1 m_{11} + u_2 m_{12} + \ldots + u_d m_{1d} \\ u_1 m_{21} + u_2 m_{22} + \ldots + u_d m_{2d} \\ \vdots \\ u_1 m_{d1} + u_2 m_{d2} + \ldots + u_d m_{dd} \end{bmatrix} \\
&= u_1 u_1 m_{11} + u_1 u_2 m_{12} + \ldots + u_1 u_d m_{1d} \\
&\quad + u_2 u_1 m_{21} + u_2 u_2 m_{22} + \ldots + u_2 u_d m_{2d} \\
&\quad \ldots \\
&\quad + u_d u_1 m_{d1} + u_d u_2 m_{d2} + \ldots + u_d u_d m_{dd}
\end{aligned}
$$

注意，对应到本式符号，式(10.33)的结果即为红色部分，即

$$u_1 u_1 m_{11} + u_2 u_2 m_{22} + \ldots + u_d u_d m_{dd}$$

而式(10.32)的结果则要更进一步，去除红色部分中的权重$m_{ii}(1 \leqslant i \leqslant d)$部分，即

$$u_1 u_1 + u_2 u_2 + \ldots + u_d u_d$$

对比以上三个结果，即式(10.32)的**平方欧氏距离**，式(10.33)的**加权平方欧氏距离**，式(10.34)的**马氏距离**，可以细细体会度量矩阵究竟带来了什么~

因此，所谓“度量学习”，即将系统中的平方欧氏距离换为式(10.34)的马氏距离，通过优化某个目标函数，得到最恰当的度量矩阵$\mathbf{M}$（新的距离度量计算方法）的过程。书中在式(10.34)~(10.38)介绍的 NCA 即为一个具体的例子，可以从中品味“度量学习”的本质。

对于度量矩阵$\mathbf{M}$要求半正定，文中提到必有正交基$\mathbf{P}$使得$\mathbf{M}$能写为$\mathbf{M} = \mathbf{P}\mathbf{P}^{\top}$，此时马氏距离$\boldsymbol{u}^{\top}\mathbf{M}\boldsymbol{u} = \boldsymbol{u}^{\top}\mathbf{P}\mathbf{P}^{\top}\boldsymbol{u} = \left\|\mathbf{P}^{\top}\boldsymbol{u}\right\|_2^2$。

2、式(10.35)的解释

这就是一种定义而已，没什么别的意思。传统近邻分类器使用多数投票法，有投票权的样本为$\boldsymbol{x}_i$最近的 K 个近邻，即 KNN；但也可以将投票范围扩大到整个样本集，但每个样本的投票权重不一样，距离$\boldsymbol{x}_i$越近的样本投票权重越大，例如可取为第 5 章式(5.19)当 $\beta_i = 1$ 时的高斯径向基函数$\exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2\right)$。从式中可以看出，若$\boldsymbol{x}_j$与$\boldsymbol{x}_i$重合，则投票权重为 1，距离越大该值越小。式(10.35)的分母是对所有投票值规一化至$[0, 1]$范围，使之为概率。

可能会有疑问：式(10.35)分母求和变量l是否应该包含$\boldsymbol{x}_i$的下标即$l = i$？其实无所谓，进一步说其实是否进行规一化也无所谓，熟悉 KNN 的话就知道，在预测时是比较各类投票数的相对大小，各类样本对$\boldsymbol{x}_i$的投票权重的分母在式(10.35)中相同，因此不影响相对大小。

注意啊，这里有计算投票权重时用到了距离度量，所以可以进一步将其换为马氏距离，通过优化某个目标（如式(10.38)）得到最优的度量矩阵$\mathbf{M}$。

3、式(10.36)的解释

先简单解释留一法(LOO)，KNN 是选出样本$\boldsymbol{x}_i$的在样本集中最近的 K 个近邻，而现在将范围扩大，使用样本集中的所有样本进行投票，每个样本的投票权重为式(10.35)，将各类样本的投票权重分别求和，注意$\boldsymbol{x}_i$自己的类别肯定与自己相同（现在是训练阶段，还没到

对未见样本的预测阶段，训练集样本的类别信息均已知），但自己不能为自己投票吧，所以要将自己除外，即留一法。

假设训练集共有N个类别，Ω_n表示第n类样本的下标集合$(1 \leqslant n \leqslant N)$，对于样本$\boldsymbol{x}_i$来说，可以分别计算$N$个概率：

$$p_n^{\boldsymbol{x}_i} = \sum_{j \in \Omega_n} p_{ij}, 1 \leqslant n \leqslant N$$

注意，若样本$\boldsymbol{x}_i$的类别为n_*，则在根据上式计算$p_{n_*}^{\boldsymbol{x}_i}$时要将$\boldsymbol{x}_i$的下标去除，即刚刚解释的留一法（自己不能为自己投票）。$p_{n_*}^{\boldsymbol{x}_i}$即为训练集将样本$\boldsymbol{x}_i$预测为第$n_*$类的概率，若$p_{n_*}^{\boldsymbol{x}_i}$在所有的$p_n^{\boldsymbol{x}_i}(1 \leqslant n \leqslant N)$中最大，则预测正确，反之预测错误。

其中$p_{n_*}^{\boldsymbol{x}_i}$即为式(10.36)。

4、式(10.37)的解释

换为刚才式(10.36)的符号，式(10.37)即为$\sum_{i=1}^{m} p_{n_*}^{\boldsymbol{x}_i}$，也就是所有训练样本被训练集预测正确的概率之和。我们当然希望这个概率和最大，但若采用平方欧氏距离时，对于某个训练集来说这个概率和是固定的；但若采用了马氏距离，这个概率和与度量矩阵$\mathbf{M}$有关。

5、式(10.38)的解释

刚才式(10.37)中提到希望寻找一个度量矩阵$\mathbf{M}$使训练样本被训练集预测正确的概率之和最大，即$\max_{\mathbf{M}} \sum_{i=1}^{m} p_{n_*}^{\boldsymbol{x}_i}$，但优化问题习惯是最小化，所以改为$\min_{\mathbf{M}} -\sum_{i=1}^{m} p_{n_*}^{\boldsymbol{x}_i}$即可，而式(10.38)目标函数中的常数 1 并不影响优化结果，有没有无所谓的。

式(10.38)中有关将$\mathbf{M} = \mathbf{P}\mathbf{P}^{\top}$代入的形式参见前面式(10.34)的解释最后一段。

6、式(10.39)的解释

式(10.39)是本节第二个“度量学习”的具体例子。优化目标函数是要求必连约束集合$\mathcal{M}$中的样本对之间的距离之和尽可能的小，而约束条件则是要求勿连约束集合$\mathcal{C}$中的样本对之间的距离之和大于 1。

这里的“1”应该类似于第 6 章 SVM 中间隔大于“1”，纯属感觉，没有推导。