第十章 降维与度量学习

10.0 本章线性代数基础知识

本部分内容参考于<线性代数(第五版)>以及"彬彬有礼的专栏", 博客地址: https://blog.csdn.net/jbb0523

10.0.1 符号说明

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**.

10.0.2 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**, 具体如:

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

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**, 具体如:

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

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**, 具体如:

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

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

$$=(\lambda_1x_{11}+\lambda_2x_{21}+\lambda_3x_{31},\lambda_1x_{12}+\lambda_2x_{22}+\lambda_3x_{32},\lambda_1x_{13}+\lambda_2x_{23}+\lambda_3x_{33})$$

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**, 具体如:

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

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

$$=(\lambda_1x_{11}+\lambda_2x_{12}+\lambda_3x_{13};\lambda_1x_{21}+\lambda_2x_{22}+\lambda_3x_{23};\lambda_1x_{31}+\lambda_2x_{32}+\lambda_3x_{33})$$

小结: **左乘**是对矩阵的**行操作**, 而**右乘**则是对矩阵的**列操作**.

> **注1: 什么是对角阵**
>
> 对角阵也即是对角矩阵, **它是一个主对角线之外的元素都为 0 的矩阵**. 对角线上的元素可以为 0 或其他值(不能全为 0). 令 $a_{i,j}$ 表示矩阵的元素, 则对角矩阵中满足
> $a_{i,j}=0 \text{ if } i \neq j \quad \forall i, j \in\{1,2, \ldots, n\}$.
>
> 常记为: $A=\operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$, 其中 $\lambda_1, \lambda_2, \cdots, \lambda_n$ 为主对角线上的元素.
>
> 同时**几个特殊的对角矩阵**:
>
> - 对角线上元素相等的对角矩阵称为**数量矩阵**;
> - 对角线上元素全为 1 的对角矩阵称为**单位矩阵**

10.0.3 矩阵的 F 范数与迹

1. **矩阵的 F 范数**

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

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

   其中, a_{ij} 为矩阵 $\mathbf{A}$ 第 i 行第 j 列元素, 即

   $$\mathbf{A}=\left[\begin{array}{cccccc}
a_{11} & a_{12} & \cdots & a_{1 j} & \cdots & a_{1 n} \\
a_{21} & a_{22} & \cdots & a_{2 j} & \cdots & a_{2 n} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
a_{i 1} & a_{i 2} & \cdots & a_{i j} & \cdots & a_{i n} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
a_{m 1} & a_{m 2} & \cdots & a_{m j} & \cdots & a_{m n}
\end{array}\right]$$

2. **列向量和行向量**表示的**矩阵的 F 范数**

   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 $\|\mathbf{A}\|_F^2=\sum_{j=1}^n\|\boldsymbol{\alpha}_j\|_2^2$; 且有 $\|\boldsymbol{\alpha}_j\|_2^2=\boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j$

   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 $\|\mathbf{A}\|_F^2=\sum_{i=1}^m\|\boldsymbol{\beta}_i\|_2^2$.且有 $\|\boldsymbol{\beta}_i\|_2^2=\boldsymbol{\beta}_i \boldsymbol{\beta}_i^T$

   > **注2:** 证明 $\|\mathbf{A}\|_F^2=\sum_{j=1}^n\|\boldsymbol{\alpha}_j\|_2^2$;
   >
   > 因为 $\|\boldsymbol{\alpha}_j\|_2^2=\sum_{i=1}^m|a_{ij}|^2$, 且 $\|\mathbf{A}\|_F^2=\sum_{i=1}^m \sum_{j=1}^n|a_{ij}|^2$, 同理, 可证 $\|\mathbf{A}\|_F^2=\sum_{i=1}^m\|\boldsymbol{\beta}_i\|_2^2$.

3. **矩阵的迹**

在线性代数中, 一个 $n \times n$ 矩阵 $\boldsymbol{A}$ 的**主对角线** (从左上方至右下方的对角线) 上**各个元素的总和**被称为**矩阵 $\boldsymbol{A}$ 的迹**(或迹数), 一般记作 $\operatorname{tr}(\mathbf{A})$

4. **方阵的特征值与迹之间的关系**

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

   (1) **矩阵 $\boldsymbol{A}$ 的特征值之和等于 $\boldsymbol{A}$ 的迹**, 也就是主对角线元素的总和. 即:

$$\lambda_1+\lambda_2+\cdots+\lambda_n=\operatorname{tr}(\mathbf{A})=a_{11}+a_{22}+\cdots+a_{mn}$$

   (2)**矩阵 $\boldsymbol{A}$ 的特征值的积等于 $\boldsymbol{A}$ 的行列式**, 即:

$$\lambda_1 \lambda_2 \cdots \lambda_n=|A|$$

5. 矩阵的**范数与迹、特征值**之间的关系

   若 λ_j 表示 n 阶方阵 $\mathbf{A}^{\top} \mathbf{A}$ 的第 j 个特征值, $\operatorname{tr}\left(\mathbf{A}^{\top} \mathbf{A}\right)$ 是 $\mathbf{A}^{\top} \mathbf{A}$ 的迹; λ_i 表示 m 阶方阵 $\mathbf{A} \mathbf{A}^{\top}$ 的第 i 个特征值, $\operatorname{tr}\left(\mathbf{A} \mathbf{A}^{\top}\right)$ 是 $\mathbf{A} \mathbf{A}^{\top}$ 的迹, 那么有:

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

> **注3: 证明上式**
>
> 证明 $\|\mathbf{A}\|_F^2=\operatorname{tr}\left(\mathbf{A}^{\top} \mathbf{A}\right)$:
>
> 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, 易知 $\operatorname{tr}(\mathbf{B})=\sum_{j=1}^n b_{jj}$, 且有
>
> $$\mathbf{B}=\mathbf{A}^{\top} \mathbf{A}=\left[\begin{array}{cccccc}a_{11} & a_{21} & \cdots & a_{i1} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{i2} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{1j} & a_{2j} & \cdots & a_{ij} & \cdots & a_{mj} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{in} & \cdots & a_{mn}\end{array}\right]\left[\begin{array}{cccccc}a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn}\end{array}\right]$$
>
> 由行列式的计算可知, b_{ij} 等于 $\mathbf{A}^{\top}$ 的第 j 行与 $\mathbf{A}$ 的第 j 列的内积, 也就是上面的红色元素对应积的和. 则
>
> $$\operatorname{tr}(\mathbf{B})=\sum_{j=1}^n b_{jj}=\sum_{j=1}^n\left(\sum_{i=1}^m\left|a_{ij}\right|^2\right)=\sum_{i=1}^m \sum_{j=1}^n\left|a_{ij}\right|^2=\|\mathbf{A}\|_F^2$$
>
> 因为 $\operatorname{tr}(\mathbf{B})=\operatorname{tr}\left(\mathbf{A}^{\top} \mathbf{A}\right)$, 所以 $\|\mathbf{A}\|_F^2=\operatorname{tr}\left(\mathbf{A}^{\top} \mathbf{A}\right)$,
>
> 同时由方阵与特征值的关系可得出结论
>
> $$\|\mathbf{A}\|_F^2=\operatorname{tr}\left(\mathbf{A}^{\top} \mathbf{A}\right)=\sum_{j=1}^n \lambda_j$$
>
> 同理, 可知: $\|\mathbf{A}\|_F^2=\operatorname{tr}\left(\mathbf{A} \mathbf{A}^{\top}\right)=\sum_{i=1}^m \lambda_i$, 得证.

10.1 k 近邻学习

10.1.1 k 近邻学习的概念

k **近邻** (k-Nearest Neighbor, 简称 kNN) 学习是一种常用的**监督学习**方法.

工作机制: 给定测试样本, 基于**某种距离度量**找出训练集中与其**最靠近的** k **个训练样本**, 然后基于这 k **个"邻居 "**的信息来进行**预测**.

如何基于 k 个"邻居"的信息进行预测:

- **在分类任务中:** 常用**"投票法"**, 即选择这 k 个样本中出现**最多的类别标记**作为**预测结果**.
- **在回归任务中:** 常用**"平均法"**, 即将这 k 个样本的**实值输出标记的平均值**作为**预测结果**.
- 还可以基于距离远近进行**加权平均**或**加权投票**, **距离越近**的样本**权重越大**.

10.1.2 懒惰学习和急切学习

k 近邻学习与前面的学习有一个很大的**不同之处**: k 近邻学习**没有显式的训练过程**.

- **懒惰学习**(lazy learning): 此类学习技术在**训练阶段仅仅是把样本保存起来**, 训练时间开销为零, **待收到测试样本后再进行处理**.
- **急切学习**(eager learning): 在**训练阶段就对样本进行学习处理**的方法, 称为急切学习.

10.1.3 参数 k 重要性

k **是一个重要参数**,

- 当 k **取不同值时, 分类结果会有显著不同**.
- 当**采用不同的距离计算方式,** 则找出的"近邻"可能有显著差别, 从而**导致分类结果有显著不同**.

图 10.1 给出了 k 近邻分类器的一个示意图.

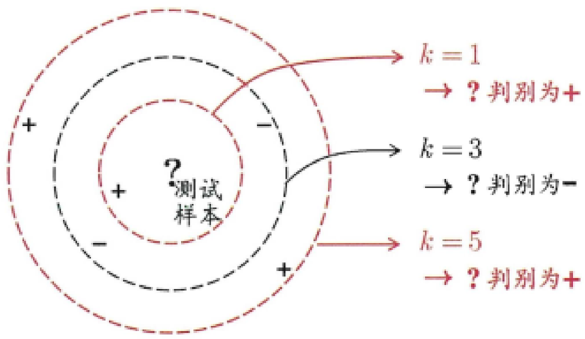


图 10.1 k 近邻分类器示意图. 虚线显示出等距线; 测试样本在 $k=1$ 或 $k=5$ 时被判别为正例, $k=3$ 时被判别为反例.

10.1.4 k 近邻的错误率

给定测试样本 $\boldsymbol{x}$, 若其最近邻样本为 $\boldsymbol{z}$, 则最近邻分类器出错的概率就是 $\boldsymbol{x}$ 与 $\boldsymbol{z}$ 类别标记不同的概率, 即

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

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**; 换言之, 对任意测试样本, 总能在任意近的范围内找到式 (10.1) 中的训练样本 δ . 令 $c^{*}=\arg\max_{c \in \mathcal{Y}} P(c|\boldsymbol{x})$ 表示贝叶斯最优分类器的结果, 有:

$$
\begin{aligned}
P(err) &= 1-\sum_{c \in \mathcal{Y}} P(c|\boldsymbol{x}) P(c|\boldsymbol{z}) \\
&\simeq 1-\sum_{c \in \mathcal{Y}} P^{2}(c|\boldsymbol{x}) \\
&\leqslant 1-P^{2}\left(c^{*}|\boldsymbol{x}\right) \\
&=\left(1+P\left(c^{*}|\boldsymbol{x}\right)\right)\left(1-P\left(c^{*}|\boldsymbol{x}\right)\right) \\
&\leqslant 2 \times\left(1-P\left(c^{*}|\boldsymbol{x}\right)\right)
\end{aligned} \tag{10.2}
$$

> **注1: 关于式 (10.2) 的推导**
>
> **1 "$\simeq$" 后边:**
>
> 因为是任意小正数 δ , 则 $f(x) \simeq f(x+\delta))$, 所以, $P(c|\boldsymbol{x}) \simeq P(c|\boldsymbol{z})$
>
> 2 **第一个 "$\leqslant$" 右边**:
>
> 因为 $P(c^{*}|\boldsymbol{x})$ 只是 $\sum_{c \in \mathcal{Y}} P^{2}(c|\boldsymbol{x})$ 的一部分, 因此 $P^{2}\left(c^{*}|\boldsymbol{x}\right) \leqslant \sum_{c \in \mathcal{Y}} P^{2}(c|\boldsymbol{x})$

可以得到这样一个**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**.

10.2 低维嵌入

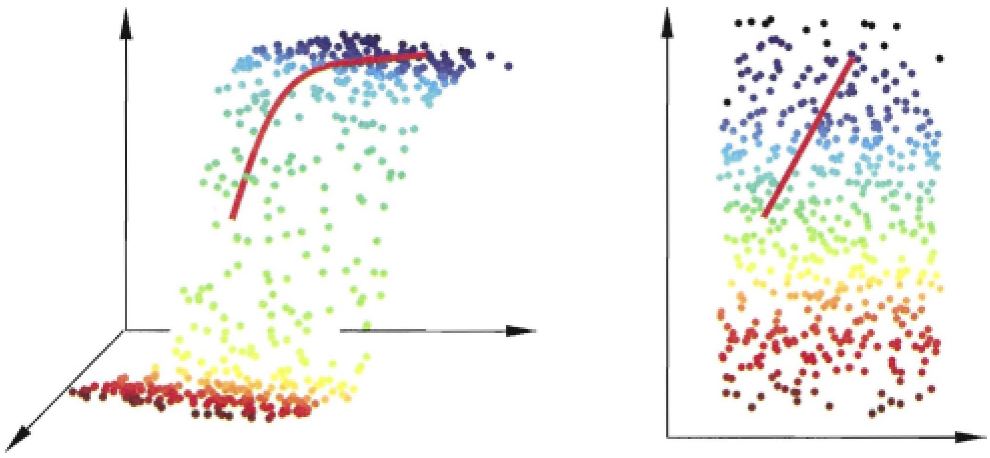
10.2.1 维数灾难的概念

在**高维**情形下出现的**数据样本稀疏**、**距离计算困难**等问题, 是所有机器学习方法共同面临的严重障碍, 被称为**"维数灾难"** (curse of dimensionality).

缓解维数灾难的一个重要途径是**降维** (dimension reduction), 亦称**"维数约简"** , 即通过某种**数学变换**将**原始高维属性空间**转变为一个**低维"子空间"** (subspace), 在这个子空间中**样本密度大幅提高**, **距离计算也变得更为容易** .

在很多时候, 人们观测或收集到的数据**样本虽是高维的**, 但**与学习任务密切相关**的也许仅是**某个低维分布**, 即**高维空间中的一个低维"嵌"** (embedding).

图 10.2 给出一个直观的列子.



(a) 三维空间中观察到的样本点　　(b) 二维空间中的曲面

图 10.2　低维嵌入示意图

若要求原始空间中样本之间的距离在低维空间中得以保持, 如图 10.2 所示, 即得到"**多维缩放**"(Multiple Dimensional Scaling, 简称 **MDS**) . MDS 是一种经典的降维方法.

10.2.2 MDS 的数学表示

1 MDS 的距离表示

假定 m 个样本在原始空间的距离矩阵为 $\mathbf{D} \in \mathbb{R}^{m\times m}$, 其第 i 行 j 列的元素 $dist_{ij}$ 为样本 $\boldsymbol{x}_i$ 到 $\boldsymbol{x}_j$ 的距离.

目标是获得样本在 d' 维空间的表示 $\mathbf{Z} \in \mathbb{R}^{d'\times m}, d' \leqslant d$, 且任意两个样本在 d' 维空间中的欧氏距离等于原始空间中的距离, 即 $\|z_i - z_j\| = dist_{ij}$.

令 $\mathbf{B} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z} \in \mathbb{R}^{m\times m}$, 其中 $\mathbf{B}$ 为降维后样本的内积矩阵, $b_{ij} = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j$, 有:

$$\begin{aligned} \mathrm{dist}_{ij}^2 &= \|z_i\|^2 + \|z_j\|^2 - 2\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j \\ &= b_{ii} + b_{jj} - 2b_{ij} \end{aligned} \tag{10.3}$$

> **注4: 关于 10.3 的相关推导**
>
> - **关于 $b_{ij} = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j$**
>
> 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 $\boldsymbol{z}_i = (z_{1i}; z_{2i}; \ldots; z_{d'i}) \in \mathbb{R}^{d'\times 1}$; 降维后的内积矩阵 $\mathbf{B} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z} \in \mathbb{R}^{m\times m}$, 其中矩阵 $\mathbf{B}$ 的第 i 行第 j 列元素 b_{ij} ,具体的矩阵表示如下:
>
> $$\mathbf{B} = \mathbf{Z}^{\top}\mathbf{Z} = \begin{bmatrix} z_{11} & z_{21} & \cdots & z_{i1} & \cdots & z_{m1} \\ z_{12} & z_{22} & \cdots & z_{i2} & \cdots & z_{m2} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{1j} & z_{2j} & \cdots & z_{ij} & \cdots & z_{mj} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{1d'} & z_{2d'} & \cdots & z_{id'} & \cdots & z_{md'} \end{bmatrix} \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1j} & \cdots & z_{1d'} \\ z_{21} & z_{22} & \cdots & z_{2j} & \cdots & z_{2d'} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{i1} & z_{i2} & \cdots & z_{ij} & \cdots & z_{id'} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mj} & \cdots & z_{md'} \end{bmatrix}$$
>
> **有几个特别的:**(其中, 红色的就是 b_{jj})

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

- **关于 (10.3) 推导:**

因为 $\boldsymbol{z}_i-\boldsymbol{z}_j$ 是列向量, 则 $\|\boldsymbol{z}_i-\boldsymbol{z}_j\|^2=(\boldsymbol{z}_i-\boldsymbol{z}_j)^\top(\boldsymbol{z}_i-\boldsymbol{z}_j)$, 则可得:

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

小注: 因为 $\boldsymbol{z}_j^\top\boldsymbol{z}_i$ 和 $\boldsymbol{z}_j^\top\boldsymbol{z}_j$ 都是一个值, 是一个标量, 因此 $\boldsymbol{z}_j^\top\boldsymbol{z}_i=\boldsymbol{z}_j^\top\boldsymbol{z}_j$.

同时, 为了便于讨论, 令降维后的样本 $\mathbf{Z}$ 被中心化, 即 $\sum_{i=1}^m\boldsymbol{z}_i=\mathbf{0}$. 那么, 矩阵 $\mathbf{B}$ 的行与列之和均为零, 即 $\sum_{i=1}^m b_{ij}=\sum_{j=1}^m b_{ij}=0$.

注5: $\sum_{i=1}^m b_{ij}=\sum_{j=1}^m b_{ij}=0$ 的证明

$$\begin{aligned}\textstyle\sum_{i=1}^m b_{ij}=\sum_{i=1}^m\boldsymbol{z}_j^\top\boldsymbol{z}_i=\boldsymbol{z}_j^\top\sum_{i=1}^m\boldsymbol{z}_i=\boldsymbol{z}_j^\top\cdot\mathbf{0}_{d'\times1}=0\\\textstyle\sum_{j=1}^m b_{ij}=\sum_{j=1}^m\boldsymbol{z}_i^\top\boldsymbol{z}_j=\boldsymbol{z}_i^\top\sum_{j=1}^m\boldsymbol{z}_j=\boldsymbol{z}_i^\top\cdot\mathbf{0}_{d'\times1}=0\end{aligned}$$

易得:

$$\sum_{i=1}^m dist_{ij}^2=\operatorname{tr}(\mathbf{B})+mb_{jj}\tag{10.4}$$

$$\sum_{j=1}^m dist_{ij}^2=\operatorname{tr}(\mathbf{B})+mb_{ii}\tag{10.5}$$

$$\sum_{i=1}^m\sum_{j=1}^m dist_{ij}^2=2m\operatorname{tr}(\mathbf{B})\tag{10.6}$$

注6: 关于 (10.4), (10.5) 和 (10.6) 的证明 (根据 (10.3))

(两个证明, 一个是根据 (10.3) 第一个等式, 一个根据第二个等式)

根据第一个等式 $\mathrm{dist}_{ij}^2=\|z_i\|^2+\|z_j\|^2-2\boldsymbol{z}_i^\top\boldsymbol{z}_j$ 来进行证明.

式 (10.4) 证明:

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

又根据定义可知:

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

且:

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

带入可得:

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

(10.5)类似可得

(10.6) 推导:

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

其中各项的求解如下:

$$
\begin{aligned}
&\textstyle\sum_{i=1}^{m}\sum_{j=1}^{m}\|\boldsymbol{z}_i\|^2=\sum_{i=1}^{m}\left(\|\boldsymbol{z}_i\|^2\sum_{j=1}^{m}1\right)=m\sum_{i=1}^{m}\|\boldsymbol{z}_i\|^2=m\operatorname{tr}(\mathbf{B})\\
&\textstyle\sum_{i=1}^{m}\sum_{j=1}^{m}\|\boldsymbol{z}_j\|^2=\sum_{i=1}^{m}\operatorname{tr}(\mathbf{B})=m\operatorname{tr}(\mathbf{B})\\
&\textstyle\sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{z}_i^\top\boldsymbol{z}_j=0
\end{aligned}
$$

即可得证.

根据 (10.3)第二个等式证明, 即 $\mathrm{dist}_{ij}^2==b_{ii}+b_{jj}-2b_{ij}$

式 (10.4) 证明:

$$
\begin{aligned}
\sum_{i=1}^{m} dist_{ij}^2 &= \sum_{i=1}^{m}(b_{ii}+b_{jj}-2b_{ij})\\
&=\sum_{i=1}^{m}b_{ii}+\sum_{i=1}^{m}b_{jj}-2\sum_{i=1}^{m}b_{ij}\\
&=\operatorname{tr}(\boldsymbol{B})+mb_{jj}-0\\
&=\operatorname{tr}(\boldsymbol{B})+mb_{jj}
\end{aligned}
$$

同理可证式 (10.5)

> 式 (10.6) 证明:
>
> $$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 &= \sum_{i=1}^{m}\sum_{j=1}^{m}(b_{ii}+b_{jj}-2b_{ij}) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} b_{ii} + \sum_{i=1}^{m}\sum_{j=1}^{m} b_{jj} - 2\sum_{i=1}^{m}\sum_{j=1}^{m} b_{ij} \\ &= \sum_{i=1}^{m} b_{ii}\cdot\sum_{j=1}^{m}\cdot 1 + \sum_{j=1}^{m} b_{jj}\cdot\sum_{i=1}^{m}\cdot 1 + \sum_{i=1}^{m}\sum_{j=1}^{m} b_{ij} \\ &= m\,\mathrm{tr}(\boldsymbol{B}) + m\,\mathrm{tr}(\boldsymbol{B}) - 2\sum_{i=1}^{m}\sum_{j=1}^{m} b_{ij} \\ &= 2m\,\mathrm{tr}(\boldsymbol{B})\end{aligned}$$
>
> 注意 $\sum_{i=1}^{m}\sum_{j=1}^{m} b_{ij} = 0$
>
> $$\sum_{i=1}^{m}\sum_{j=1}^{m} b_{ij} = \sum_{j=1}^{m}\cdot\sum_{j=1}^{m} b_{ij} = 0$$

其中, $\mathrm{tr}(\cdot)$ 表示矩阵的迹 (trace), $\mathrm{tr}(\mathbf{B}) = \sum_{i=1}^{m}\|\boldsymbol{z}_i\|^2$, 令

$$\mathrm{dist}_{i\cdot}^2 = \frac{1}{m}\sum_{j=1}^{m} dist_{ij}^2 \tag{10.7}$$

$$dist_{\cdot j}^2 = \frac{1}{m}\sum_{i=1}^{m} dist_{ij}^2 \tag{10.8}$$

$$dist_{\cdot\cdot}^2 = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2 \tag{10.9}$$

根据式 (10.3) 和式 (10.4)~(10.9), 易得:

$$b_{ij} = -\frac{1}{2}\left(dist_{ij}^2 - dist_{i\cdot}^2 - dist_{\cdot j}^2 + dist_{\cdot\cdot}^2\right) \tag{10.10}$$

由此即可通过降维前后不变的距离矩阵 $\mathbf{D}$ 求取内积矩阵 $\mathbf{B}$.

2 MDS 的特征值求解及完整算法描述

对矩阵 $\mathbf{B}$ 做特征值分解 (eigenvalue decomposition), $\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$ (注: $\mathbf{B}$ 是对称矩阵), 其中, $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ 为特征值构成的对角矩阵, $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_d$, $\mathbf{V}$ 为特征向量矩阵. 假定其中有 d^* 个非零特征值, 它们构成对角矩阵 $\mathbf{\Lambda}_* = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{d^*})$, 令 $\mathbf{V}^*$ 表示相应的特征向量矩阵, 则 $\mathbf{Z}$ 可表达为:

$$\mathbf{Z} = \mathbf{\Lambda}_*^{1/2}\mathbf{V}_*^{\mathrm{T}} \in \mathbb{R}^{d^*\times m} \tag{10.11}$$

> 注7: 式 (10.11) 目前不知道如何来的, 暂放, 标记.

而在现实中, 为了有效降维, 往往仅需降维后的距离与原始空间中的距离尽可能接近, 而不必严格相等. 此时可取 $d' \ll d$ 个最大特征值构成对角矩阵 $\tilde{\mathbf{\Lambda}} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{d'})$, 令 $\tilde{\mathbf{\Lambda}}$ 表示相应的特征向量矩阵, 则 $\mathbf{Z}$ 可表达为:

$$\mathbf{Z} = \tilde{\mathbf{\Lambda}}^{1/2}\tilde{\mathbf{V}}^{\mathrm{T}} \in \mathbb{R}^{d'\times m} \tag{10.12}$$

图 10.3 给出了 MDS 算法的描述.

输入: 距离矩阵 $\mathbf{D} \in \mathbb{R}^{m\times m}$, 其元素 $dist_{ij}$ 为样本 $\boldsymbol{x}_i$ 到 $\boldsymbol{x}_j$ 的距离;
低维空间维数 d'.

过程:

1: 根据式(10.7)~(10.9)计算 $dist_{i\cdot}^2$, $dist_{\cdot j}^2$, $dist_{\cdot\cdot}^2$;
2: 根据式(10.10)计算矩阵 $\mathbf{B}$;
3: 对矩阵 $\mathbf{B}$ 做特征值分解;
4: 取 $\tilde{\mathbf{\Lambda}}$ 为 d' 个最大特征值所构成的对角矩阵, $\tilde{\mathbf{V}}$ 为相应的特征向量矩阵.

输出: 矩阵 $\tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}^{1/2} \in \mathbb{R}^{m\times d'}$, 每行是一个样本的低维坐标

图 10.3 MDS 算法

10.2.3 高维空间的线性变换

一般来说, 欲获得低维子空间, 最简单的是对**原始高维空间进行线性变换**. 给定 d 维空间中的样本 $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) \in \mathbb{R}^{d\times m}$, 变换之后得到 $d' \leqslant d$ 维空间中的样本:

$$\mathbf{Z} = \mathbf{W}^{\mathrm{T}}\mathbf{X} \tag{10.13}$$

其中, $\mathbf{W} \in \mathbb{R}^{d\times d'}$ 是变换矩阵, $\mathbf{Z} \in \mathbb{R}^{d'\times m}$ 是样本在新空间中的表示.

10.3 主成分分析

10.3.1 主成分分析的基本概念

主成分分析 (Principal Component Analysis, 简称 PCA) 是最常用的一种**降维方法**.

对于**正交属性空间中的样本点**, 如何用一个**超平面**对所有样本进行恰当的表达?

若存在这样的超平面, 它大概具有这样的性质:

- **最近重构性**: 样本点到这个超平面的距离都**足够近**;
- **最大可分性**: 样本点在这个超平面上的**投影尽可能分开**.

基于最近重构性和最大可分性, 能分别得到主成分分析的**两种等价推导**.

10.3.2 基于最近重构性推导主成分分析

1 基本假定和条件

首先, 假定数据样本进行了中心化, 也即 $\sum_i \boldsymbol{x}_i = \mathbf{0}$;

同时, 再假定投影变换后得到的新坐标系为 $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d\}$, 其中, $\boldsymbol{w}_i$ 是标准正交基向量, 也即是满足 $\|\boldsymbol{w}_i\|_2 = 1, \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{w}_j = 0 (i \neq j)$

若丢弃新坐标系中的部分坐标, 即将纬度降低到 $d' < d$, 则样本点 $\boldsymbol{x}_i$ 在低维坐标系中的投影是 $\boldsymbol{z}_i = (z_{i1}; z_{i2}; \ldots; z_{id'})$, 其中 $z_{ij} = \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{x}_i$ 是 $\boldsymbol{x}_i$ 在低维坐标系下第 j 维坐标. 若基于 $\boldsymbol{z}_i$ 来重构 $\boldsymbol{x}_i$, 则会得到 $\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d'} z_{ij} \boldsymbol{w}_j$.

注8: 关于1, 2, 3 的解释和推导

对于 1 : 理解中心化的概念

因为数据进行了中心化, 易知: $\sum_i \boldsymbol{x}_i = \mathbf{0}$

对于 2 : 理解标准正交基的概念

- 先理解 **基** 的概念, 根据<线性代数>(第五版) $p141$ 关于基的定义:

定义 2 在线性空间 V 中,如果存在 n 个元素 $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n$,满足:

(i) $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n$ 线性无关;

(ii) V 中任一元素 $\boldsymbol{\alpha}$ 总可由 $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n$ 线性表示,

那么, $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n$ 就称为线性空间 V 的一个基, n 称为线性空间 V 的维数. 只含一个零元素的线性空间没有基, 规定它的维数为 0.

- 再理解**正交**的概念: <线性代数>(第五版) $p112$ 关于正交的概念:

当两个向量的**内积**为 0 时(即 $[\boldsymbol{x}, \boldsymbol{y}] = 0$) , 称向量 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 正交. 同时注意到 $[\boldsymbol{x}, \boldsymbol{y}] = \boldsymbol{x}^T \boldsymbol{y}$, 其中 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 都是列向量.

- 最后**规范正交基**的概念: <线性代数>(第五版) $p113$ 关于规范正交基(也叫标准正交基)的定义:

定义 3 设 n 维向量 $\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_r$ 是向量空间 $V (V \subset \mathbb{R}^n)$ 的一个基, 如果 $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_r$ 两两正交, 且都是单位向量, 则称 $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_r$ 是 V 的一个规范正交基.

对于3: 理解坐标变换和基于 $\boldsymbol{z}_i$ 来重构 $\boldsymbol{x}_i$

- 先理解坐标的表示, 根据<线性代数>(第五版) $P113$ 关于标准正交基中的坐标计算公式

  若 $e_1, \cdots, e_r$ 是 V 的一个规范正交基, 那么 V 中任一向量 $\boldsymbol{a}$ 应能由 $e_1, \cdots, e_r$ 线性表示, 设表示式为

$$\boldsymbol{a} = \lambda_1 \boldsymbol{e}_1 + \lambda_2 \boldsymbol{e}_2 + \cdots + \lambda_r \boldsymbol{e}_r$$

  为求其中的系数 $\lambda_i (i = 1, \cdots, r)$, 可用 $\boldsymbol{e}_i^T$ 左乘上式, 有

$$\boldsymbol{e}_i^{\mathrm{T}} \boldsymbol{a} = \lambda_i \boldsymbol{e}_i^{\mathrm{T}} \boldsymbol{e}_i = \lambda_i$$

  即

$$\lambda_i = \boldsymbol{e}_i^{\mathrm{T}} \boldsymbol{a} = [\boldsymbol{a}, \boldsymbol{e}_i]$$

  这就是向量在规范正交基中的坐标的计算公式. 利用这个公式能方便地求得向量的坐标.

  再看书中的投影是 $\boldsymbol{z}_i = (z_{i1}; z_{i2}; \ldots; z_{id'})$, $\boldsymbol{z}_i$ 也就是坐标向量, $(z_{i1}; z_{i2}; \ldots; z_{id'})$ 也就是各个坐标系数, 那么 $\boldsymbol{z}_i = (z_{i1}; z_{i2}; \ldots; z_{id'}) = \left(\boldsymbol{w}_1^\top \boldsymbol{x}_i; \boldsymbol{w}_2^\top \boldsymbol{x}_i; \ldots; \boldsymbol{w}_{d'}^\top \boldsymbol{x}_i\right)$. 易知, $z_{ij} = \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{x}_i$

- 再来看坐标表示的过程

对于 d 维空间 $\mathbb{R}^{d\times 1}$ 来说, 传统的坐标系为 $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k, \ldots, \boldsymbol{v}_d\}$ (标准正交基), 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 : $\boldsymbol{x}_i = x_{i1}\boldsymbol{v}_1 + x_{i2}\boldsymbol{v}_2 + \ldots + x_{id}\boldsymbol{v}_d$,

现在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), 那么根据前面的坐标表示, 我们可以得到 $\boldsymbol{x}_i$ 在新坐标系中的坐标为 $(\boldsymbol{w}_1^\top \boldsymbol{x}_i; \boldsymbol{w}_2^\top \boldsymbol{x}_i; \ldots; \boldsymbol{w}_d^\top \boldsymbol{x}_i)$.

若丢弃新坐标系中的部分坐标, 即将维度降低到 $d' < d$, 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:

$$\begin{aligned}\boldsymbol{z}_i &= (z_{i1}; z_{i2}; \ldots; z_{id'}) = \left(\boldsymbol{w}_1^\top \boldsymbol{x}_i; \boldsymbol{w}_2^\top \boldsymbol{x}_i; \ldots; \boldsymbol{w}_{d'}^\top \boldsymbol{x}_i\right) \\ &= \mathbf{W}^\top \boldsymbol{x}_i \in \mathbb{R}^{d'\times 1}\end{aligned}$$

同时也易知, $z_{ij} = \boldsymbol{w}_j^\mathrm{T} \boldsymbol{x}_i$.

- 最后来看基于 $\boldsymbol{z}_i$ 来重构 $\boldsymbol{x}_i$, 得到 $\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j$

  此处, 卡了我非常久的时间, 我之前老是从式子本身来推导, 发现如何都推不出来. 后来发现, 自己想多了,可能也是自己脑子不太灵光. 现在, 特别记录在此, 防止以后有人看到这里, 会有一样疑惑.

  其实, 从坐标表示的定义就可以直接得到.

  看前面的坐标表示过程

$$\boldsymbol{x}_i = x_{i1}\boldsymbol{v}_1 + x_{i2}\boldsymbol{v}_2 + \ldots + x_{id}\boldsymbol{v}_d$$

  也即是

$$\boldsymbol{x}_i = \sum_{j=1}^{d} \boldsymbol{x}_{ij}\boldsymbol{v}_j$$

  当坐标系从 $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k, \ldots, \boldsymbol{v}_d\}$ 换成 $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d\}$, 丢弃部分坐标, 即 $d' < d$, 那么就是

$$\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d'} \boldsymbol{z}_{ij}\boldsymbol{w}_j = \mathbf{W} z_i$$

2 样本点 $\boldsymbol{x}_i$ 与投影重构的样本点 $\hat{\boldsymbol{x}}_i$ 之间的距离

考虑整个训练集, 样本点 $\boldsymbol{x}_i$ 与投影重构的样本点 $\hat{\boldsymbol{x}}_i$ 之间的距离为:

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

> 注9: 上式的推导过程
>
> **第四个等式到第五个等式:**
>
> 由于 $\boldsymbol{w}_i^\top\boldsymbol{w}_j=0,(i\neq j),\|\boldsymbol{w}_i\|=1$, 且 $\mathbf{W}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'})\in\mathbb{R}^{d\times d'}$, 那么 $\mathbf{W}^\top\mathbf{W}=\mathbf{I}\in\mathbb{R}^{d'\times d'}$. 即得.
>
> **第八个等式到最后一个式子:**
>
> 由于是寻找 $\mathbf{W}$ 使得目标函数最小, 而 $\boldsymbol{x}_i^\top\boldsymbol{x}_i$ 与 $\mathbf{W}$ 无关, 因此, 优化时可以略去.

同时, 令 $\mathbf{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m)\in\mathbb{R}^{d\times m}$, 再根据前面的预备知识中关于矩阵范数中的知识, 即 $\|\mathbf{A}\|_F^2=\sum_{j=1}^{n}\|\boldsymbol{\alpha}_j\|_2^2$. 那么上面的式子可以继续化简为:

$$
\begin{aligned}
-\sum_{i=1}^{m}\left\|\mathbf{W}^\top\boldsymbol{x}_i\right\|_2^2 &= -\left\|\mathbf{W}^\top\mathbf{X}\right\|_F^2 \\
&= -\operatorname{tr}\left(\left(\mathbf{W}^\top\mathbf{X}\right)\left(\mathbf{W}^\top\mathbf{X}\right)^\top\right) \\
&= -\operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right)
\end{aligned}
\tag{10.14}
$$

> **注10: (10.14) 的推导**
>
> **第一个等式:**
>
> 根据 $\|\mathbf{A}\|_F^2=\sum_{j=1}^{n}\|\boldsymbol{\alpha}_j\|_2^2$ 即可推得:
>
> $\sum_{i=1}^{m}\|\mathbf{W}^\top\boldsymbol{x}_i\|_2^2=\|\mathbf{W}^\top\mathbf{X}\|_F^2$
>
> **第二个等式:**

$\|\mathbf{A}\|_F^2 = \operatorname{tr}(\mathbf{A}^\top\mathbf{A})$

即可得

那么根据最近重构性, 就可以得到最终的优化目标和约束条件, 即:

$$\begin{aligned} &\min_{\mathbf{W}} -\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right) \\ &\text{s.t. } \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I} \end{aligned} \tag{10.15}$$

这就是主成分分析的优化目标.

同时注意一点, $\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} = \mathbf{X}\mathbf{X}^\top$ 是协方差矩阵.

> 注11: **本笔记中的式 (10.14) 和书本上的式 (10.14) 有些许差别**
>
> 先给出结论: $\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^\top = \mathbf{X}\mathbf{X}^\top$
>
> 具体证明过程略, 主要是各自展开左右两式子即可得证.

10.3.3 基于最大可分性推导主成分分析

从最大可分性出发, 能得到主成分分析的另一种解释.

样本点 $\boldsymbol{x}_i$ 在新空间中超平面上的投影是 $\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i$. 若所有样本点的投影能尽可能分开, 则应该使投影后样本点的方差最大化, 如图 10.4 所示:

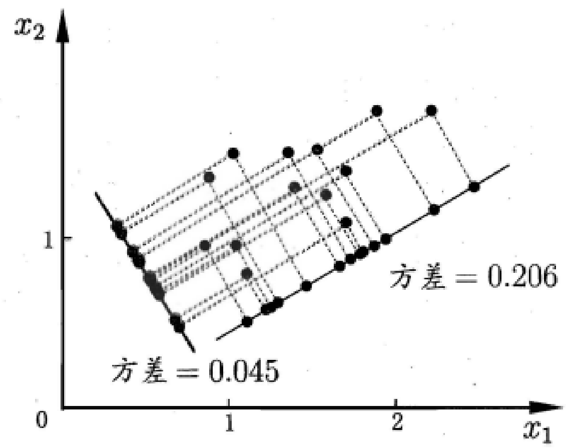


图 10.4 使所有样本的投影尽可能分开(如图中红线所示), 则需最大化投影点的方差

投影后样本点的方差是 $\sum_i \mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\mathbf{W}$, 于是优化目标可以写成:

$$\begin{aligned} &\max_{\mathbf{W}} \operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right) \\ &\text{s.t. } \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I} \end{aligned} \tag{10.16}$$

> ---
>
> **注12: 关于式 (10.16) 的推导**
>
> 先考虑协方差矩阵 $\mathbf{X}\mathbf{X}^\top$:

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

那么我们可以知道 $\frac{1}{m}\mathbf{X}\mathbf{X}^\top$ 的第 i 行第 j 列的元素表示 $\mathbf{X}$ 中第 i 行和 $\mathbf{X}^\top$ 中第 j 列(其实也就是 $\mathbf{X}$ 中第 j 行)的方差(当 $i=j$)或协方差(当 $i\neq j$) . 同时, 我们可以看到, 协方差矩阵的对角线元素为隔行的方差.

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 $\mathbf{Z}=\mathbf{W}^\top\mathbf{X}$, 其中 $\mathbf{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_{d'}\}\in\mathbb{R}^{d\times d'}$ 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**各行方差最大**)

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 (即使各行方差之和最大) . 同时, $\mathbf{Z}=\mathbf{W}^\top\mathbf{X}$, 且 $\frac{1}{m}$ 为常数, 不影响优化过程. **求矩阵对角线元素之和即为矩阵的迹.** 即可得:

$$\max_{\mathbf{W}}\operatorname{tr}\left(\mathbf{W}^\mathrm{T}\mathbf{X}\mathbf{X}^\mathrm{T}\mathbf{W}\right)$$

显然, 式 (10.16) 与式 (10.15) 是等价的.

对式 (10.15) 或式 (10.16) 使用拉格朗日乘子法可得:

$$\mathbf{X}\mathbf{X}^\top\mathbf{W}=\mathbf{W}\boldsymbol{\Lambda} \tag{10.17}$$

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

还可以进一步将此式拆成 d' 个子式子:

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

注13: 关于式 (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:

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

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

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

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

关于向量的求导前面已经说过了, 关于迹的求导, 目前还不清楚, 结果参考开头博客.

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

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

还可以进一步将此式拆成 d' 个子式子:

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

于是, 只需对协方差矩阵 $\mathbf{X}\mathbf{X}^\mathrm{T}$ 进行特征值分解, 将求得的特征值排序: $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_d$, 再取前 d' 个特征值对应的特征向量构成 $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}) \in \mathbb{R}^{d\times d'}$.这就是主成分分析的解. PCA 算法描述如图 10.5 所示:

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
低维空间维数 d'.

过程:

1: 对所有样本进行中心化: $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{m}\sum_{i=1}^m \boldsymbol{x}_i$;
2: 计算样本的协方差矩阵 $\mathbf{X}\mathbf{X}^\mathrm{T}$;
3: 对协方差矩阵 $\mathbf{X}\mathbf{X}^\mathrm{T}$ 做特征值分解;
4: 取最大的 d' 个特征值所对应的特征向量 $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}$.

输出: 投影矩阵 $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$.

图 10.5 PCA 算法

注14: 对式 (10.17) 的另外的进一步解释

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

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

这里, 我们使用了约束条件 $\mathbf{W}^\top \mathbf{W} = \mathbf{I}$.

上面的式子的左边与式 (10.16) 的优化目标对应的矩阵相同, 而右边 $\mathbf{\Lambda} \in \mathbb{R}^{d'\times d'}$ 是由 $\mathbf{X}\mathbf{X}^\mathrm{T}$ 的 d' 个特征值组成的对角阵, 两边同时取矩阵的迹, 得

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

左边的优化目标相当于最大化 $\sum_{i=1}^{d'} \lambda_i$.