



明德
知力
行

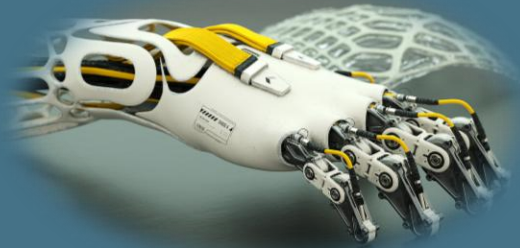
机器学习 第八讲

授课人：王闻博

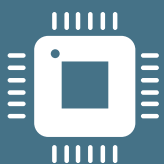
Email: wenbo_wang@kust.edu.cn

昆明理工大学 机电工程学院

2026年05月19日



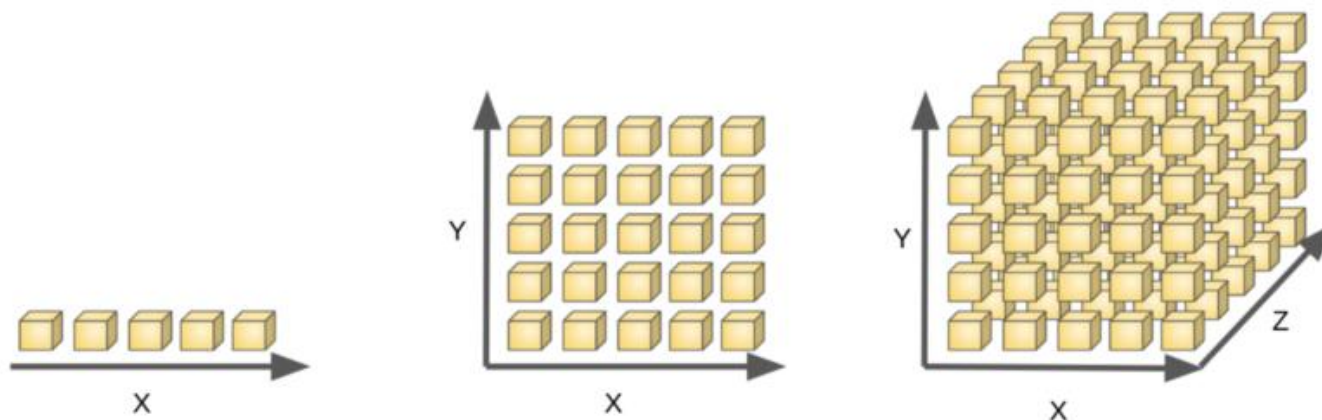
数据降维



1. 数据降维：概述
2. 数据降维的数学基础：奇异值分解 (SVD)
3. 主成分分析 (PCA)

• 维数灾难(Curse of Dimensionality)问题

- 在涉及到向量的计算的问题中，随着维数的增加，计算量呈指数倍增长的一种现象（见下图）；
- 维数灾难涉及数字分析、抽样、组合、机器学习、数据挖掘和数据库等诸多领域。



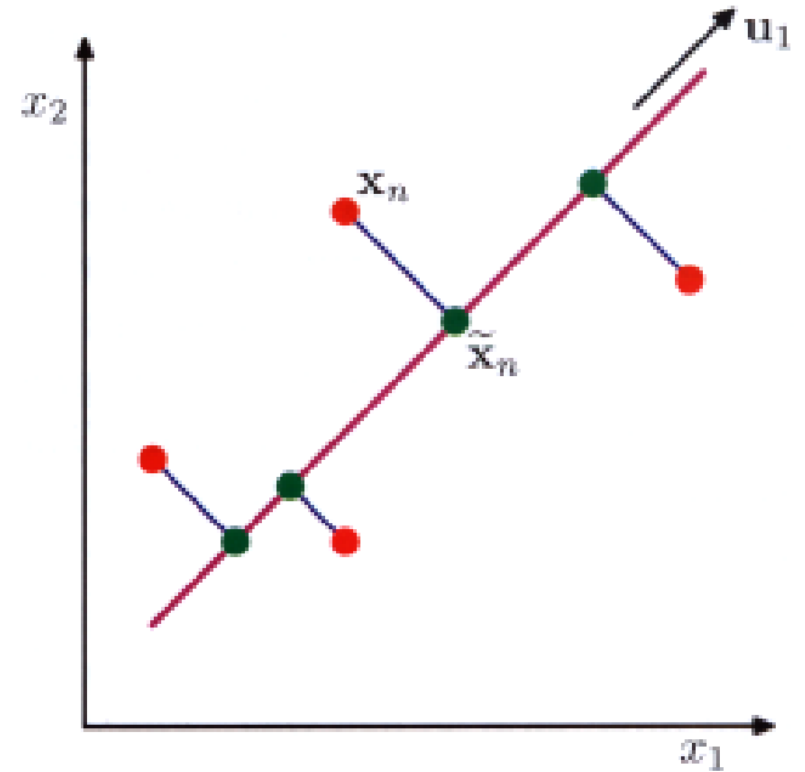
• 人工智能/机器学习任务重的维数灾难现象

- 在大数据驱动的机器学习模型汇总，训练集中的每条数据经常伴随着上千、甚至上万个特征；
- 随着特征数量的增多，计算量会变得很大。处理高维数据中所有的特征，不仅会让训练非常缓慢，还会极大增加搜寻最优解的困难；
- 机器学习模型的性能往往会随着特征维数的增加先上升后下降。

数据降维

• 降维 (Dimensionality Reduction)

- 将训练数据中的样本(实例)从高维空间转换到低维空间;
- 该过程与信息论中有损压缩概念密切相关 (注意: **不存在完全无损的降维**) ;
- 在一般的降维方法中, 降维是通过对原始数据的线性变换实现的: 即将数据投影到维度更低的空间, 这种投影的建立总是由某种特定设计目标驱动
 - 设计目标举例 (见右图): 可以使用最小化平均投影损失, 如使用原始数据集和投影后数据的均方和距离 (Mean Square Distance) 。
 - 右图中, 使用M维向量 $\mathbf{u}_1$ 作为投影变换矩阵, 将n个数据 $\mathbf{x}_n$ 投影到 $\mathbf{u}_1^T \mathbf{x}_n$ 上, 即可建立带约束的 (如要求 $\mathbf{u}_1$ 是归一化向量: $1 - \mathbf{u}_1^T \mathbf{u}_1 = 0$) 均方和误差最小化问题, 解出 $\mathbf{u}_1$ 的值。



图来源: Bishop 《PRML》



数据降维和特征工程

• 总结：数据降维（或特征工程）的优点

- **缓解维度灾难**：通过压缩冗余（如线性相关特征）或与输出无因果关系特征，使数据分布更紧凑；
- **提升计算效率**：通过特征选择或特征提取，减少模型训练和推理的计算负担。
- **消除噪声与冗余特征**：高维数据中常包含无关特征（如传感器噪声）或高度相关的冗余特征（如像素间的相关性），干扰模型学习。
- **改善模型泛化能力**：高维数据易导致过拟合，尤其在样本量不足时（如基因数据通常有数万特征但仅数百样本）。
- **改善数据的可视化与可解释性**：帮助人类直观理解高维数据（如100维的文本嵌入向量）。

注意：数据降维不可避免地相对原始数据会丢失信息，且大部分降维技术需要经验法则确定降维算法的超参数。

示例：面向可视化的数据降维

- **t-分布随机邻域嵌入 (t-distributed Stochastic Neighbor Embedding, t-SNE)**

- t-SNE (TSNE) 将数据点之间的相似度转换为概率。原始空间中的相似度由高斯联合概率表示，嵌入空间的相似度由“学生t分布”表示。

- **算法步骤：**

- 1. 对每个高维点 $\mathbf{x}_i$ ，计算其与所有其他点 $\mathbf{x}_j$ 的条件概率 $p_{j|i}$ ，表示 $\mathbf{x}_i$ 作为 $\mathbf{x}_j$ 邻居的可能性：

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$$

- 2. 初始化低维空间（低维初始化）并定义相似性，随机生成低维点 $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ ，并使用自由度为1的t分布（柯西分布）计算低维点对的相似度：

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

示例：t-分布随机邻域嵌入（续）

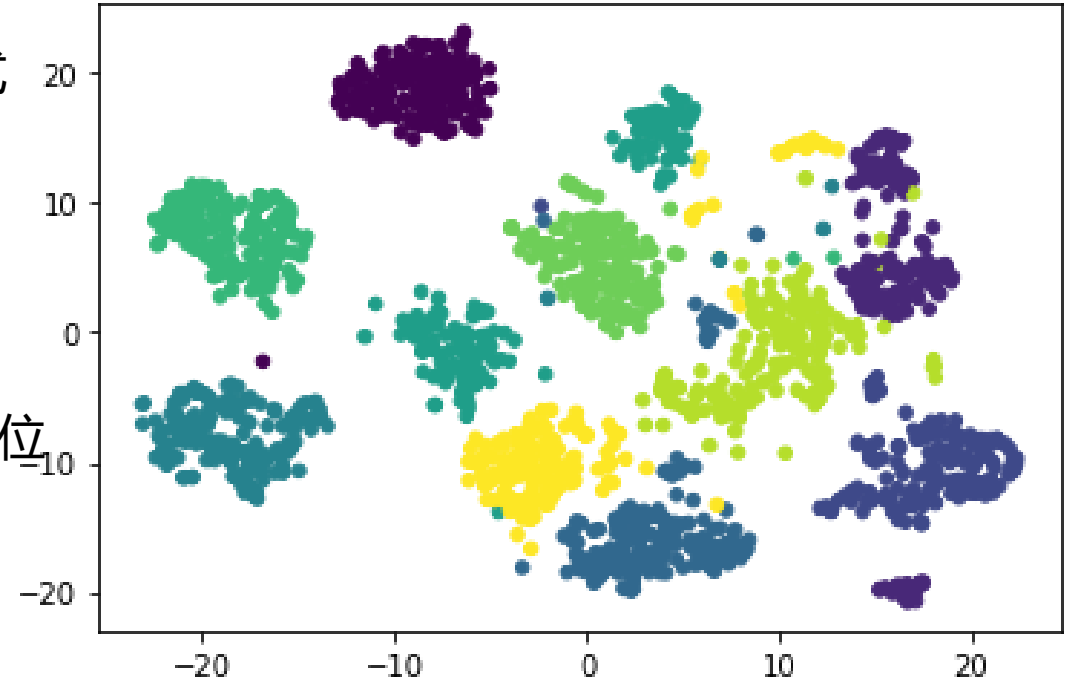
• （续）算法步骤

- 3. 优化低维嵌入：最小化高维和低维分布的KL散度，下式KL散度作为目标函数：

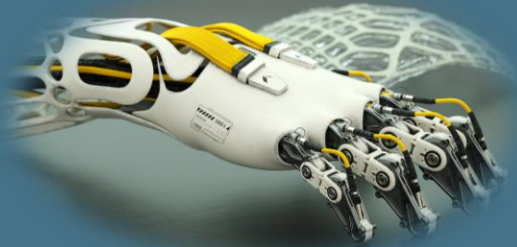
$$\text{KL}(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

- 3a. 梯度下降优化：通过计算梯度 $\frac{\partial \text{KL}}{\partial \mathbf{y}_i}$ ，迭代更新低维点位置。梯度计算如下：

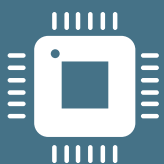
$$\frac{\partial \text{KL}}{\partial \mathbf{y}_i} = 4 \sum_{i \neq j} (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j)(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}$$



对MNIST数据集进行t-SNE降维可视化操作



数据降维



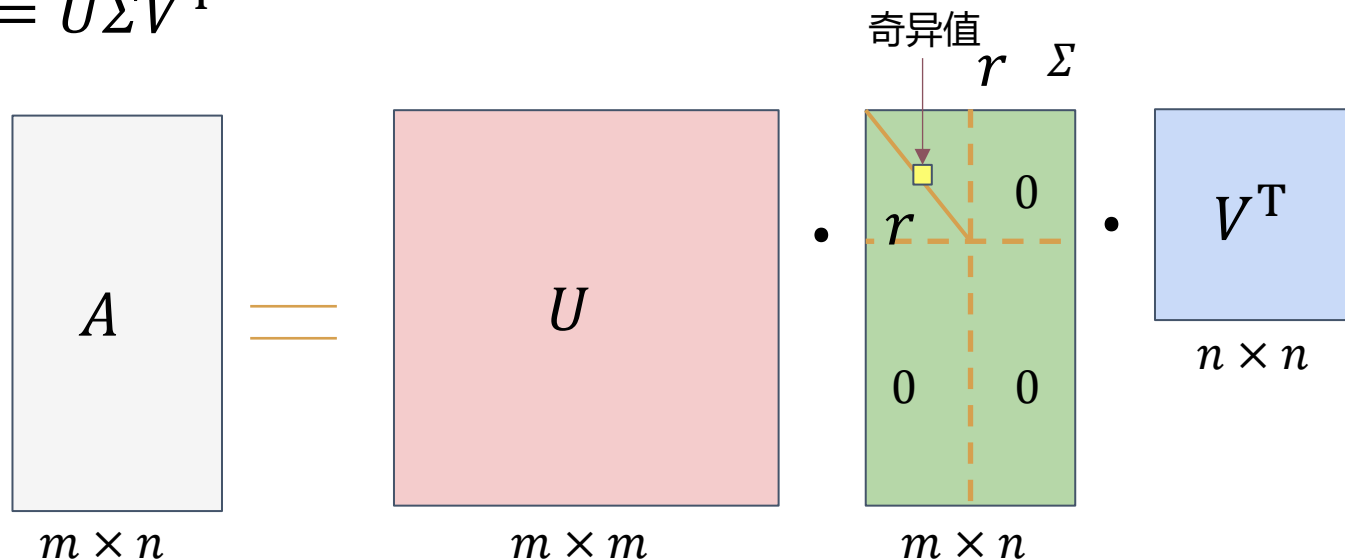
1. 数据降维：概述
2. 数据降维的数学基础：奇异值分解 (SVD)
3. 主成分分析 (PCA)

奇异值分解 (Singular Value Decomposition)

• 奇异值分解——SVD可以将一个任意维度的矩阵 A 分解为三个矩阵的乘积：

- 正交矩阵 (orthogonal matrix) : U
- 对角矩阵 (diagonal matrix) : Σ
- 正交矩阵的转置: V^T 。

• 分解形式: $A = U\Sigma V^T$





奇异值分解 (SVD, 续)

• 符号定义

$$A = U\Sigma V^T = u_1\sigma_1v_1^T + \cdots + u_r\sigma_rv_r^T$$

- U 是一个 $m \times m$ 的方阵，每个特征向量 u_i 叫做 A 的左奇异向量，所有左奇异向量都是正交的。
- Σ 是一个 $m \times n$ 的矩阵，除了主对角线上的元素以外全为 0，主对角线上的每个元素都称为奇异值 σ 。
- V 是一个 $n \times n$ 的方阵，每个特征向量 v_i 叫做 A 的右奇异向量，所有右奇异向量都是正交的。
- U 和 V 都是酉矩阵，即满足： $U^T U = I, V^T V = I$ 。
- r 为矩阵 A 的秩 (rank) 。

由于 U 、 V 都是酉矩阵，本页顶部公式的矩阵表达形式可写为 $AV = U\Sigma$ ；
向量表达形式可写为 $Av_i = \sigma_i u_i$ 。

SVD求解

- **(求解目标) 构造两组奇异值向量u和v, 使得: $A = U\Sigma V^T$:**

- 为此, 构造对称矩阵 $A^T A$ 和 AA^T :

$$AA^T = (U\Sigma V^T)(V\Sigma^T U^T) = U\Sigma^T U^T \quad (a)$$

$$A^T A = (V\Sigma^T U^T)(U\Sigma V^T) = V\Sigma^T V^T \quad (b)$$

- 注意: 上述二式都产生了对称矩阵, 通常 $A^T A \neq AA^T$ 。

- **U矩阵求解**

- AA^T 为 $m \times m$ 的方阵, 由上述公式 (a), 可以进行特征分解, 求左奇异值向量U:

$$(AA^T)u_i = \lambda_i u_i$$

- 其中, λ_i 对应降序排列的 $\sigma_1^2 \geq \sigma_2^2 \geq \dots, \geq \sigma_r^2 \geq 0$ (回忆: r为矩阵A的秩), λ_i 是方阵 AA^T 的非零特征值。



SVD求解 (续)

- 同理， U 矩阵求解

$$A^T A = (V \Sigma^T U^T)(U \Sigma V^T) = V \Sigma^T V^T \quad (b)$$

- $A^T A$ 为 $n \times n$ 的方阵，由公式 (b)，可以进行特征分解，求右奇异值向量 V ：

$$(A^T A)v_i = \lambda_i v_i$$

- 其中， λ_i 对应 $\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_r^2 \geq 0$ （回忆： r 为矩阵 A 的秩），是方阵 $A^T A$ 的非零特征值。

- **SVD要求** $Av_i = \sigma_i u_i$

- 这个式子将每个右奇异向量 v_i 关联到左奇异向量 u_i ；
- 选择 $A^T A$ 的单位正交特征向量 $v_1, v_2, \dots, v_r$ ，然后选择 $\sigma_i = \sqrt{\lambda_i}$ ，**对特征值进行降序排列**；
- 此时，只需要先计算一次特征分解，确定右奇异向量 v 后，即可确定 u ： $u_i = \frac{Av_i}{\sigma_i}$ 。

注意：通过两次特征分解，由方程 $Av_i = \sigma_i u_i$ ，也可解得 $\sigma_i = Av_i / u_i$ 。



SVD计算案例

例1：设矩阵 A 为： $A = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$ 则 A 的秩 $r = 2$ 。

$$A^T A = \begin{bmatrix} 3 & 4 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} = \begin{bmatrix} 25 & 20 \\ 20 & 25 \end{bmatrix}$$

$$A A^T = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 3 & 4 \\ 0 & 5 \end{bmatrix} = \begin{bmatrix} 9 & 12 \\ 12 & 41 \end{bmatrix}$$

两者都有相同的迹，为50。



SVD计算案例 (续)

例1：设矩阵A为： $A = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$

由前页计算，求得 $A^T A$ 的特征值和特征向量：

$$\begin{vmatrix} 25 - \lambda & 20 \\ 20 & 25 - \lambda \end{vmatrix} = (25 - \lambda)^2 - 400 = (\lambda - 45)(\lambda - 5) = 0$$

求解得到特征值： $\lambda_1 = 45, \lambda_2 = 5$

由 $\sigma_i = \sqrt{\lambda_i}$ ，可以得到奇异值为： $\sigma_1 = \sqrt{45}, \sigma_2 = \sqrt{5}$



SVD计算案例 (续)

接着求出 AA^T 的特征值和单位正交特征向量:

同理, 求得: $\lambda_1 = 45, \lambda_2 = 5$

$$v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \quad v_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

利用 $Av_i = \sigma_i u_i, i = 1, 2$, 可求右奇异向量 u :

$$Av_1 = \begin{bmatrix} \frac{3}{\sqrt{2}} \\ 9 \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \sqrt{45} \begin{bmatrix} \frac{1}{\sqrt{10}} \\ 3 \\ \frac{1}{\sqrt{10}} \end{bmatrix} = \sigma_1 u_1, \quad Av_2 = \begin{bmatrix} -\frac{3}{\sqrt{2}} \\ 1 \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \sqrt{5} \begin{bmatrix} -\frac{3}{\sqrt{10}} \\ 1 \\ \frac{1}{\sqrt{10}} \end{bmatrix} = \sigma_2 u_2$$



SVD计算案例 (续)

最终得到 A 的奇异值分解为:

$$U = \begin{bmatrix} \frac{1}{\sqrt{10}} & -\frac{3}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sqrt{45} & 0 \\ 0 & \sqrt{5} \end{bmatrix}, \quad V = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$
$$A = U\Sigma V^T = \begin{bmatrix} \frac{1}{\sqrt{10}} & -\frac{3}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{bmatrix} \begin{bmatrix} \sqrt{45} & 0 \\ 0 & \sqrt{5} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$$

SVD在数据科学中的重要结论

- SVD将矩阵分解为秩为1的部分矩阵的和：

$$A = U\Sigma V^T = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T$$

- 由于奇异值按降序排列，这些部分矩阵按其重要性依次出现。
- 第一部分 $\sigma_1 u_1 v_1^T$ 是最近似A的秩为1的矩阵；不仅如此，前k部分之和是秩为k的任意矩阵中能够找到的对A的最佳近似（见：Eckart-Young定理）。
- 换言之，用最大的 k 个的奇异值和对应的左右奇异向量来近似描述矩阵A，可得：

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T \approx U_{m \times k} \Sigma_{k \times k} V_{k \times n}^T = A_k = \sigma_1 u_1 v_1^T + \cdots + \sigma_k u_k v_k^T$$

Eckart-Young定理：对所有秩为k的矩阵B，有 $\|A - A_k\| \leq \|A - B\|$ ，其中 $\|\cdot\|$ 为矩阵的范数（如Frobenius范数）。（注意：此定理对应的最优近似性由奇异值降序排列保证。）

（补充知识）Frobenius范数用奇异值定义：

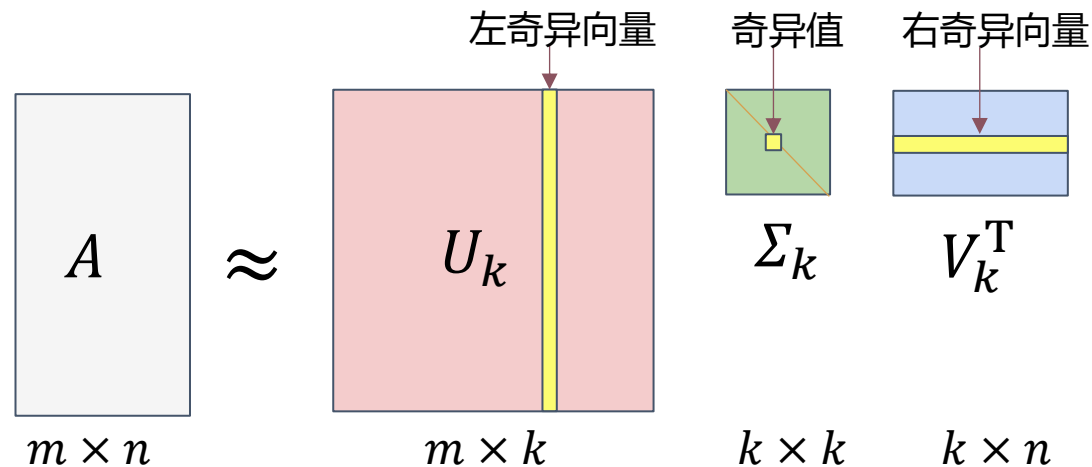
$$\|A\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$$

SVD用于数据压缩

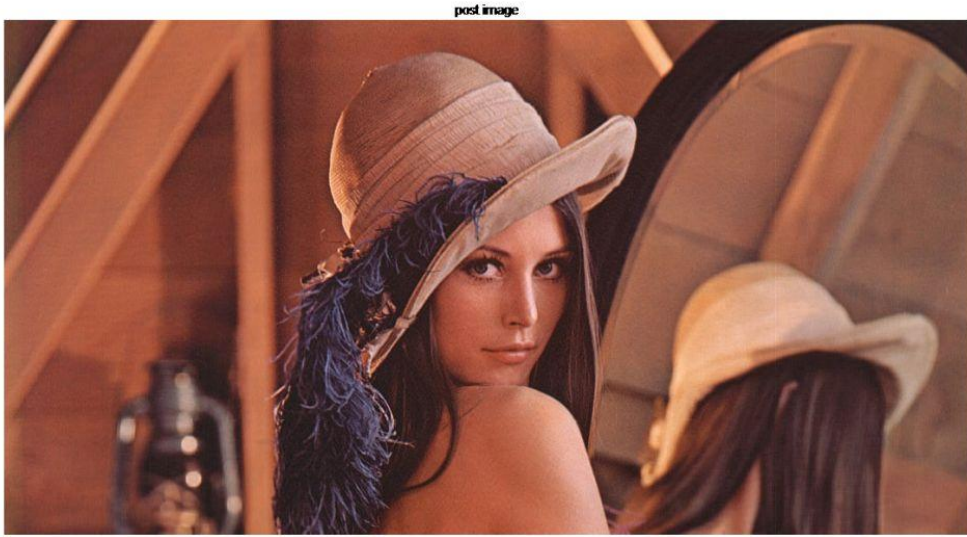
- 由 k 个最重要部分 A_k (对应最大的 k 个奇异值) 对矩阵 A 进行近似:

$$A_k = \sigma_1 u_1 v_1^T + \cdots + \sigma_k u_k v_k^T$$

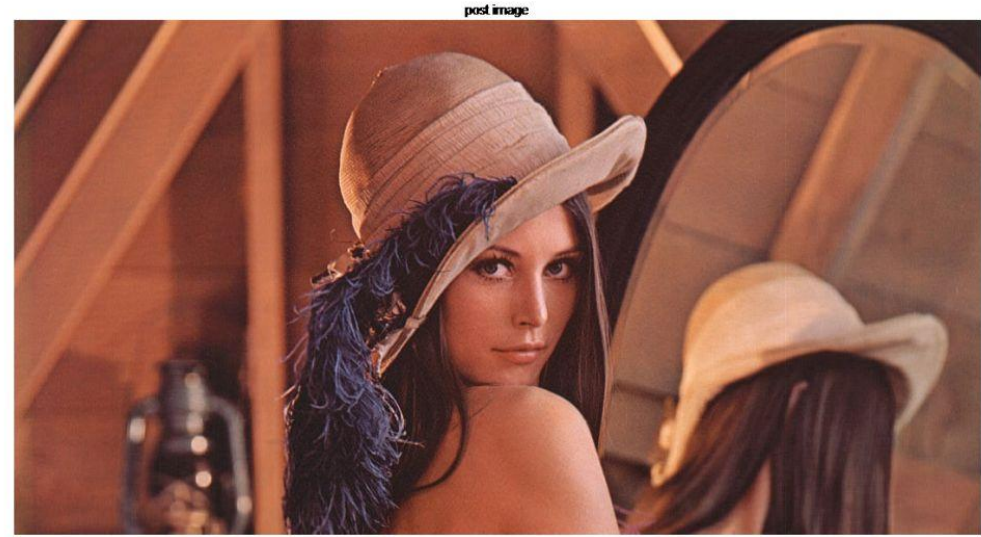
- 其中 k 要比 n 小很多, 即一个大的矩阵 A 可以用三个较小的矩阵 $U_{m \times k}, \Sigma_{k \times k}, V_{k \times n}^T$ 来表示。
- (回忆上页) 上式写成矩阵形式有: $A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T \approx U_{m \times k} \Sigma_{k \times k} V_{k \times n}^T$ 。
- 换言之, 用最大的 k 个的奇异值和对应的左右奇异向量来近似描述矩阵 A , 可得:
- 如图所示, 矩阵 A 的维度对比构建 A_k 矩阵所需的三个较小矩阵的维度有:



应用SVD进行奇异值分解的案例



原始图像



处理后的图像

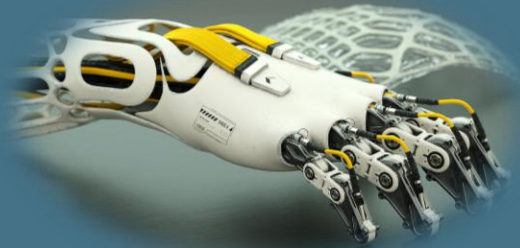
$$m = 575, \quad n = 1081, \quad k = 150$$

$$\text{原始维度} A = 575 \times 1081 \times 3 = 1864725$$

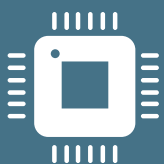
$$\text{则原始图像经过压缩后的维度: } 3 \times (575 \times 150 + 150 \times 150 + 1081 \times 150) = 812700$$

设 $k = 150$ ，则经过SVD分解后的矩阵及维度：

$$U_{m \times k} = 575 \times 150, \quad \Sigma_{k \times k} = 150 \times 150, \quad V_{k \times n}^T = 1081 \times 150$$



数据降维



1. 数据降维：概述
2. 数据降维的数学基础：奇异值分解 (SVD)
3. 主成分分析 (PCA)

主成分分析 (Principal Component Analysis, PCA)



- 主成分分析 (PCA) : 对数据集 $\{X_1, \dots, X_N\}$, 其中的一个样本 X_n 的维度为 D 。通过将数据集投影到较小的维度空间 $M < D$ (此空间称为主空间, *Principal Subspace*) , 同时最大化投影数据的方差。
- PCA的朴素原理
 - 将样本数据向量 x 从某个坐标系映射到另一个坐标系, 这并未改变向量自身, 而只改变了该向量的定量表示形式。
 - 若希望将给定数据集中的样本数据由 D 降至 M 维 ($D > M$) , 则可依条件选择 M 个线性无关的 D 维向量 $w_1, w_2, \dots, w_M$ 作为 M 维空间的一组基 (因为线性无关) , 将样本集中的 D 维向量通过线性变换, 映射到以 $w_1, w_2, \dots, w_M$ 个基向量张成的 M 维空间中, 由此实现对 D 维数据集中样本向量的降维。

PCA计算的矩阵形式

- 数据矩阵 $\mathbf{A}_0$ 是由 N 个样本列向量排成的矩阵，每个样本是 D 维的列向量

- **步骤1**：求样本矩阵每行（每个分量）的平均值，写成向量表达形式有

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$$

对每个样本进行标准化（即每个样本的每行都减去其对应的样本分量平均值），得到中心化矩阵 $\mathbf{A} = \mathbf{A}_0 - \mathbf{1}_N \bar{\mathbf{x}}^T$ 。此时 $\mathbf{A}$ 的 n 个列是空间 $\mathbb{R}^D$ 中的 n 个点，它们的和为0（变换后样本均值为0）。

- **步骤2**：求样本方差/协方差矩阵（ $D \times D$ 维矩阵）：

$$S = \frac{1}{N-1} \sum_{k=1}^N (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T$$

上式中协方差矩阵可表示为：

$$S = \frac{1}{N-1} \mathbf{A} \mathbf{A}^T$$

方差：每行到均值的距离的平方和， $\mathbf{A} \mathbf{A}^T$ 的对角元；
协方差： $\mathbf{A}$ 的第 i 行和第 j 行的点积， $\mathbf{A} \mathbf{A}^T$ 的非对角元；
因子使用 $N-1$ ：因为有一个自由度已用于样本矩阵中心化（均值置0）。

PCA计算的矩阵形式 (续)

- **步骤3：构造标准正交基** $\{w_1, w_2, \dots, w_M\}$

- 3.1 用特征分解方法计算协方差矩阵 S 的特征值和**正交特征向量**;
- 3.2 对特征值从大到小排序, 选择其中最大的 M 个。然后**将其对应的** D 个特征向量 (**称为主成分**) 分别作为**行向量**组成特征向量矩阵 W 。
- 3.3 将数据转换到 M 个特征向量构建的新空间中, 即 $Y = WX$ 。

- **PCA算法和SVD的关系:** 由 $S = \frac{1}{N-1} \mathbf{A} \mathbf{A}^T$ 回忆: $\mathbf{A} \mathbf{A}^T = (\mathbf{U} \mathbf{\Sigma} \mathbf{V}^T)(\mathbf{V} \mathbf{\Sigma}^T \mathbf{U}^T) = \mathbf{U} \mathbf{\Sigma}^T \mathbf{U}^T$

- S 的特征向量是中心化矩阵 A 的左奇异向量;
- 给出 A 中的数据, 计算 S 的特征值分解并不是一个明智的决定。对大矩阵而言, 对 A 直接进行SVD更快、更准确;

- 注意: 当 A 中各样本幅值不一时, 需要引进标准化处理: $A_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, i = 1, 2, \dots, N, j = 1, 2, \dots, D$
 分量标准差

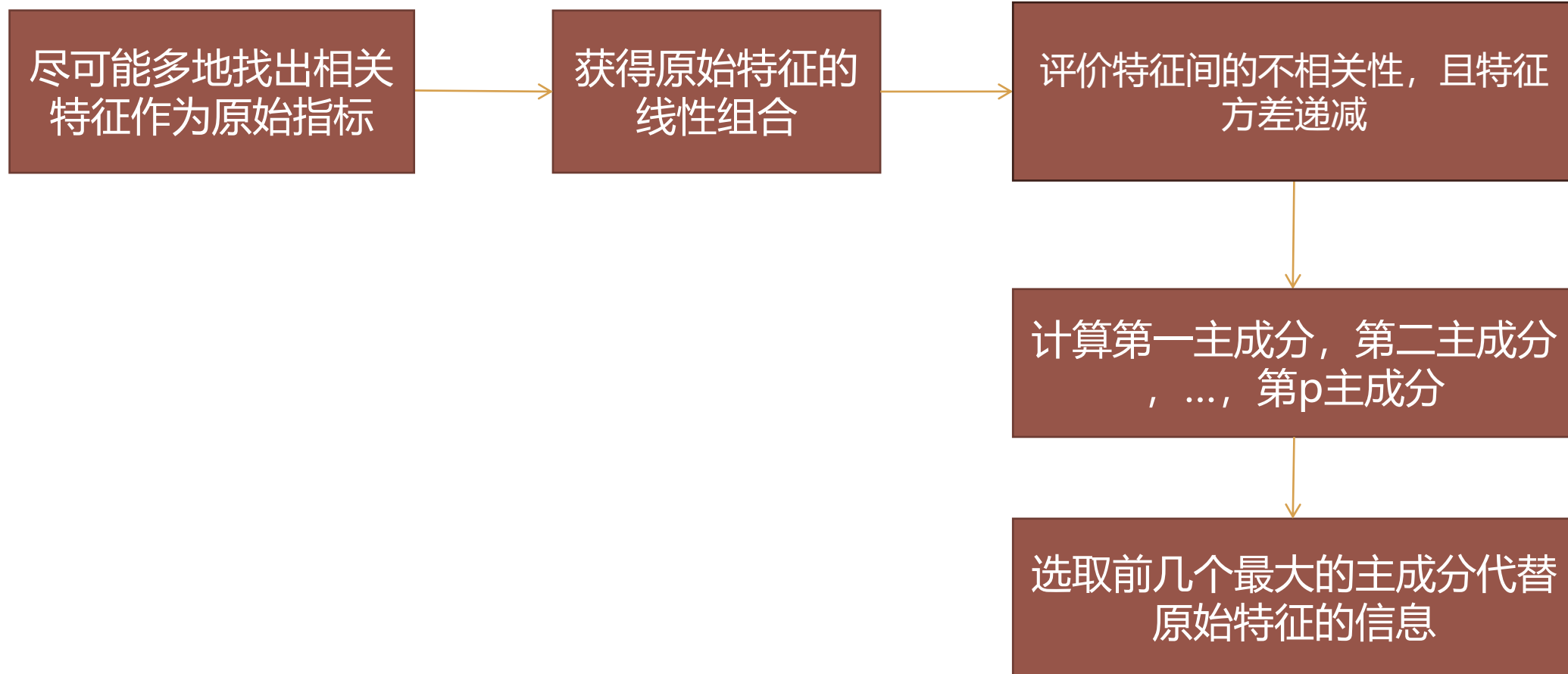
前文注释：协方差矩阵的标量形式

- 将标准化后数据组成新的数据矩阵 $\tilde{A} = (Z_1, Z_2, \dots, Z_N)$ 并构造其协方差矩阵 S

$$S = \frac{1}{N-1} A^T A = \begin{pmatrix} \frac{1}{N-1} \sum_{j=1}^N z_{1j}^2 & \frac{1}{N-1} \sum_{j=1}^N z_{1j} z_{2j} & \cdots & \frac{1}{N-1} \sum_{j=1}^N z_{1j} z_{Nj} \\ \frac{1}{N-1} \sum_{j=1}^N z_{1j} z_{2j} & \frac{1}{N-1} \sum_{j=1}^N z_{2j}^2 & \cdots & \frac{1}{N-1} \sum_{j=1}^N z_{2j} z_{Nj} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{N-1} \sum_{j=1}^N z_{1j} z_{Nj} & \frac{1}{N-1} \sum_{j=1}^N z_{2j} z_{Nj} & \cdots & \frac{1}{N-1} \sum_{j=1}^N z_{Nj}^2 \end{pmatrix}$$

此处，用 z_i 表示对中心化矩阵 A 中的列向量样本按特征分量（行）进行归一化操作后获得的列向量。

小结：主成分分析算法流程



PCA原理分析

- 讨论：为何数据协方差矩阵的特征值分解能够满足PCA构建要求？

- 标准正交基的构建要求**：标准化样本数据 $\tilde{A} = (Z_1, Z_2, \dots, Z_N)$ 变换到以 $\{w_1, w_2, \dots, w_M\}$ 为坐标系的 M 维线性空间中，能够得到最大的数据方差。
- 令矩阵 $X' = (X'_1, X'_2, \dots, X'_N)$ 表示数据集 $\tilde{A}$ 在以 $w_1, w_2, \dots, w_M$ 为基向量的 k 维空间中样本数据，则有线性变换：

$$X' = (w_1, w_2, \dots, w_M)^T \tilde{A} = \begin{pmatrix} w_1 \\ w_2 \\ \dots \\ w_M \end{pmatrix} \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1D} \\ z_{21} & z_{22} & \dots & z_{2D} \\ \dots & \dots & \dots & \dots \\ z_{N1} & z_{N2} & \dots & z_{ND} \end{pmatrix}^T$$

上式可写为矩阵形式： $X' = W\tilde{A}$ 。

PCA原理分析 (续)

- (续) 讨论：为何数据协方差矩阵的特征值分解能够满足PCA构建要求？
 - **标准正交基的构建要求**：标准化样本数据 $\tilde{A} = (Z_1, Z_2, \dots, Z_N)$ 变换到以 $\{w_1, w_2, \dots, w_M\}$ 为坐标系的 M 维线性空间中，能够得到最大的数据方差。

- 由 $X' = (w_1, w_2, \dots, w_M)^T \tilde{A} = \begin{pmatrix} w_1 \\ w_2 \\ \dots \\ w_M \end{pmatrix} \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1D} \\ z_{21} & z_{22} & \dots & z_{2D} \\ \dots & \dots & \dots & \dots \\ z_{N1} & z_{N2} & \dots & z_{ND} \end{pmatrix}^T$

- 以 X' 第一个特征属性（上式 X' 矩阵的第一行）的方差作为目标函数，最大化特征分量方差，可以构建如下最大化目标函数（因为样本已中心化，均值为零）：

$$J(w_1) = \frac{1}{N} \sum_{i=1}^N (Z_i^T w_1)^2 = \frac{1}{N} \sum_{i=1}^N (Z_i^T w_1)^T (Z_i^T w_1) = \frac{1}{N} w_1^T \left(\sum_{i=1}^N Z_i Z_i^T \right) w_1 = \frac{1}{N} w_1^T \tilde{A} \tilde{A}^T w_1$$

$$\sum_{i=1}^n Z_i Z_i^T = \tilde{A} \tilde{A}^T$$

PCA原理分析 (续)

- (续) 讨论：为何数据协方差矩阵的特征值分解能够满足PCA构建要求？

- 变换后数据集的第一个特征分量的方差最大化问题，其目标函数：

$$J(\mathbf{w}_1) = \frac{1}{N} \sum_{i=1}^N (Z_i^T \mathbf{w}_1)^2 = \frac{1}{N} \sum_{i=1}^N (Z_i^T \mathbf{w}_1)^T (Z_i^T \mathbf{w}_1) = \frac{1}{N} \mathbf{w}_1^T \left(\sum_{i=1}^N Z_i Z_i^T \right) \mathbf{w}_1 = \frac{1}{N} \mathbf{w}_1^T \tilde{\mathbf{A}} \tilde{\mathbf{A}}^T \mathbf{w}_1$$

- 由于 $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$ 为标准正交基，故需满足约束条件 $\mathbf{w}_1^T \mathbf{w}_1 = 1$ ，因此，有如下带约束的最大化问题：

$$\max_{\mathbf{w}_1} \mathbf{w}_1^T \tilde{\mathbf{A}} \tilde{\mathbf{A}}^T \mathbf{w}_1 \quad \text{s.t.} \quad \mathbf{w}_1^T \mathbf{w}_1 = 1$$

- 因为协方差矩阵 $\mathbf{S} = \frac{1}{N-1} \tilde{\mathbf{A}} \tilde{\mathbf{A}}^T$ ，故可由上面的优化问题构造Lagrange函数（对应无约束优化问题）如下，其中 α 是拉格朗日乘子：

$$\max_{\mathbf{w}_1} \mathbf{w}_1^T \mathbf{S} \mathbf{w}_1 - \alpha (\mathbf{w}_1^T \mathbf{w}_1 - 1)$$



PCA原理分析（续）

- (续) 讨论：为何数据协方差矩阵的特征值分解能够满足PCA构建要求？

- （接上页底）解由下式表达的无约束最大化问题：

$$\max_{\mathbf{w}_1} \mathbf{w}_1^T S \mathbf{w}_1 - \alpha (\mathbf{w}_1^T \mathbf{w}_1 - 1)$$

即令上述优化问题的目标函数对 $\mathbf{w}_1$ 的偏导数为0，则有 $2S\mathbf{w}_1 - 2\alpha\mathbf{w}_1=0$ ，得到 $S\mathbf{w}_1 = \alpha\mathbf{w}_1$ 。

- $\mathbf{w}_1$ 是协方差矩阵 S 的一个特征向量， α 是与该特性向量对应的特征根。由于 $\mathbf{w}_1$ 是单位向量，有

$$\mathbf{w}_1^T S \mathbf{w}_1 = \alpha \mathbf{w}_1^T \mathbf{w}_1 = \alpha$$

因此，要使得 $\mathbf{w}_1^T Z Z^T \mathbf{w}_1$ 取得最大化，即使得 $\mathbf{w}_1^T S \mathbf{w}_1$ 取得最大化， $\mathbf{w}_1$ 即为协方差矩阵 S 的最大特征根 λ_1 所对应的特征向量，可由此获得样本数据 $\tilde{\mathbf{A}}$ 的第一个主成分

$$X'_1 = \mathbf{w}_1^T \tilde{\mathbf{A}}$$

PCA原理分析 (续)

- (续) 类似地, 可以通过数据线性映射后的第二个主成分 X'_2 对应的基向量 w_2 的构造, 拓展一维压缩后的样本空间维度

- 第二个主成分 X'_2 的构造, 可通过求解如下条件优化问题获得 (第三项是基向量正交约束) :

$$\max_{w_2} w_2^T S w_2 - \alpha (w_2^T w_2 - 1) - \beta (w_2^T w_1 - 0) \text{ s.t. } w_2^T w_2 = 1$$

- 令上述优化问题的目标函数对 w_2 的偏导数为0, 有

$$2S w_2 - 2\alpha w_2 - \beta w_1 = 0 \quad (a)$$

用 w_1^T 左乘上式, 得:

$$2w_1^T S w_2 - 2\alpha w_1^T w_2 - \beta w_1^T w_1 = 0 \quad (b)$$

$w_1^T w_2 = 0$ 且 $w_1^T S w_2$ 作为标量等于其转置 $w_2^T S w_1$ 。

- 注意到 w_1 为协方差矩阵 S 的以 λ_1 为特征根的特征向量, 故有 $w_1^T S w_2 = w_2^T S w_1 = \beta w_2^T w_1 = 0$, 即由 (b) 式可得, $\beta w_1^T w_1 = 0$ 。
- 由此, (a) 式可简化为 $S w_2 = \alpha w_2$, 即需要满足矩阵 S 除 λ_1 之外的最大特征值 $\lambda_2 = \alpha$ 所对应的特征向量。同理, 可求得第三、第四、……直到第 M 个基向量和其对应的主成分。

PCA的计算案例

例2: $X = \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$

用PCA的方法将这组二维数据降到一维。

因为这个矩阵的每行已经是零均值，所以我们可以直接求协方差矩阵：

$$\Sigma = \frac{1}{5} \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{6}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} \end{pmatrix}$$



PCA的计算案例 (续)

然后，求 Σ 的特征值和特征向量：

$$|A - \lambda I| = \begin{vmatrix} \frac{6}{5} - \lambda & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} - \lambda \end{vmatrix} = (\frac{6}{5} - \lambda)^2 - \frac{16}{25} = (\lambda - 2)(\lambda - 2/5) = 0$$

求解，得到特征值： $\lambda_1 = 2, \lambda_2 = 2/5$

其对应的特征向量分别是： $\sigma_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \sigma_2 \begin{pmatrix} -1 \\ 1 \end{pmatrix}$



PCA的计算案例 (续)

由于对应的特征向量分别是一个通解, σ_1 和 σ_2 可取任意实数。
标准化后的特征向量为:

$$\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}, \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

因此变换矩阵 P 是:

$$P = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

可以验证协方差矩阵 Σ 的对角化:

$$P\Sigma P^T = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 6/5 & 4/5 \\ 4/5 & 6/5 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2/5 \end{pmatrix}$$

用 P 的第一行乘以数据矩阵, 就得到降维后的数据表示:

$$Y = \left(\frac{1}{\sqrt{2}} \quad 1/\sqrt{2}\right) \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} -3/\sqrt{2} & -1/\sqrt{2} & 0 & 3/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

PCA的计算案例 (续)



降维后的投影结果如图：

