PCA

2021-10-06

1. 背景知识

1. 协方差

X、Y为两个一维随机变量，它们的协方差为：

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])] \\ &= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{n-1}\end{aligned} \tag{35}$$

2. 协方差矩阵

对多维随机变量$\mathbf{X}$：

$$\mathbf{X}=\left(\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{array}\right)_{m\times n}=\left[\mathbf{X}_1,\mathbf{X}_2,\cdots,\mathbf{X}_n\right] \\ \text{其中，} \mathbf{X}_i=\left[x_{1i},x_{2i},\cdots,x_{mi}\right]\text{，为列向量} \tag{36}$$

计算其两两维度之间的协方差，各协方差组成了一个$n\times n$的矩阵，称为协方差矩阵，记为Σ。矩阵内的元素Σ_{ij}：

$$\Sigma_{ij}=\operatorname{cov}\left(X_i,X_j\right)=\mathrm{E}\left[\left(X_i-\mathrm{E}\left[X_i\right]\right)\left(X_j-\mathrm{E}\left[X_j\right]\right)\right] \tag{37}$$

样本协方差矩阵：

$$\begin{aligned}\Sigma &= \mathrm{E}\left[(\mathbf{X}-\mathrm{E}[\mathbf{X}])^{\mathrm{T}}(\mathbf{X}-\mathrm{E}[\mathbf{X}])\right] \\ &= \frac{1}{m-1}(\mathbf{X}-\mathrm{E}[\mathbf{X}])^{\mathrm{T}}(\mathbf{X}-\mathrm{E}[\mathbf{X}])\end{aligned} \tag{38}$$

若$\mathbf{X}$的列均值为0，则：

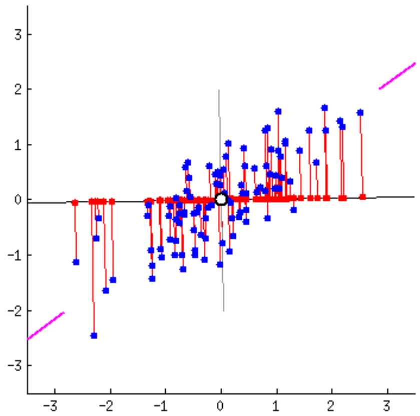
$$\begin{aligned}\Sigma &= \mathrm{E}\left[(\mathbf{X}-\mathrm{E}[\mathbf{X}])^{\mathrm{T}}(\mathbf{X}-\mathrm{E}[\mathbf{X}])\right] \\ &= \mathrm{E}\left[\mathbf{X}^{\mathrm{T}}\mathbf{X}\right] \\ &= \frac{1}{m-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\end{aligned} \tag{39}$$

2. PCA原理

2.1. 降维概述

不加思索地认为，对矩阵$\mathbf{X}$通过线性变换降维后，数据越分散，效果越好。如何度量数据的分散程度？各维度的方差越大，数据越分散。概括一下，降维后，各维度方差越大，降维效果越好。

举例：将二维数据降到一维，如果集中到了一点，那么数据就不具备可分性，自然可以认为降维效果不好。说明降维后数据越分散、越具有可分性，降维效果越好。在PCA中，选择"**方差**"作为衡量数据分散性的指标。



降维概述：给定原始矩阵$\mathbf{X}_{m\times n}$及线性变换矩阵$\mathbf{A}_{n\times k}$，通过$\mathbf{X}\mathbf{A}$线性变换，将矩阵$\mathbf{X}_{m\times n}$由n维降到k维，得到矩阵$\mathbf{Y}_{m\times k}$，降维后的维度k，又称为主成分个数。

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}_{m\times n} \tag{40}$$

2.2. 原理推导

假设：

1. 已经对矩阵$\mathbf{X}_{m\times n}$做了标准化，其均值$\mu = 0$。
2. 矩阵$\mathbf{A}_{n\times d}$为正交矩阵。

首先，考虑求第一主成分：

$$\begin{gathered}\mathbf{y}_1 = \mathbf{X}\mathbf{a}_1 \\ \text{其中，}\mathbf{a}_1 = \{a_{11}, a_{21}, \cdots, a_{n1}\}^{\mathrm{T}}\end{gathered} \tag{41}$$

求$\mathbf{y}_1$的方差：

$$
\begin{aligned}
\operatorname{var}(\mathbf{y}_1) &= E\{(\mathbf{X}\mathbf{a_1}-\mu\mathbf{a_1})^{\mathrm{T}}(\mathbf{X}\mathbf{a_1}-\mu\mathbf{a_1})\} \\
&= \mathbf{a}_1^{\mathrm{T}}E\{(\mathbf{X}-\mu)^{\mathrm{T}}(\mathbf{X}-\mu)\}\mathbf{a}_1 \\
&= \mathbf{a}_1^{\mathrm{T}}E\{\mathbf{X}^{\mathrm{T}}\mathbf{X}\}\mathbf{a}_1 \\
&= \mathbf{a}_1^{\mathrm{T}}\Sigma\mathbf{a}_1
\end{aligned}
\tag{42}
$$

其中，Σ为矩阵$\mathbf{X}$的协方差矩阵

优化目标：

$$
\begin{aligned}
\max\ \ & \mathbf{a}_1^{\mathrm{T}}\Sigma\mathbf{a}_1 \\
s.t.\ \ & \mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 1
\end{aligned}
\tag{43}
$$

定义拉格朗日函数：

$$
L(\mathbf{a}_1,\lambda) = \mathbf{a}_1^{\mathrm{T}}\Sigma\mathbf{a}_1 - \lambda(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 - 1) \tag{44}
$$

求函数$L(\mathbf{a}_1,\lambda)$对$\mathbf{a}_1$的导数并令导数为0：

$$
\frac{\partial L}{\partial \mathbf{a}_1} = \Sigma\mathbf{a}_1 - \lambda\mathbf{a}_1 = 0 \tag{45}
$$

求得：

$$
\Sigma\mathbf{a}_1 = \lambda\mathbf{a}_1 \tag{46}
$$

所以，λ为协方差矩阵Σ的特征值，$\mathbf{a}_1$为对应的单位特征向量。

进一步，优化目标可化简为：

$$
\mathbf{a}_1^{\mathrm{T}}\Sigma\mathbf{a}_1 = \mathbf{a}_1^{\mathrm{T}}\lambda\mathbf{a}_1 = \lambda\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = \lambda \tag{47}
$$

假设$\mathbf{a}_1$是Σ的第一大特征值λ_1对应的单位特征向量，显然，$\mathbf{a}_1$和λ_1是以上问题的最优解。

然后，考虑求第二主成分：

$$
\begin{aligned}
\max\quad & \mathbf{a}_2^{\mathrm{T}}\Sigma\mathbf{a}_2 \\
\text{s.t.}\quad & \mathbf{a}_1^{\mathrm{T}}\Sigma\mathbf{a}_2 = 0 \\
& \mathbf{a}_2^{\mathrm{T}}\Sigma\mathbf{a}_1 = 0 \\
& \mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2 = 0 \\
& \mathbf{a}_2^{\mathrm{T}}\mathbf{a}_1 = 0 \\
& \mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 = 1
\end{aligned}
\tag{48}
$$

注意到：

$$
\mathbf{a}_1^{\mathrm{T}}\Sigma\mathbf{a}_2 = \mathbf{a}_2^{\mathrm{T}}\Sigma\mathbf{a}_1 = \mathbf{a}_2^{\mathrm{T}}\lambda_1\mathbf{a}_1 = \lambda_1\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_1 = \lambda_1\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2 = 0 \tag{49}
$$

说明，等式约束是等价的。

定义拉格朗日函数：

$$
L(\mathbf{a}_2,\lambda,\phi) = \mathbf{a}_2^{\mathrm{T}}\Sigma\mathbf{a}_2 - \lambda\left(\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 - 1\right) - \phi\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_1 \tag{50}
$$

求函数$L(\mathbf{a}_2,\lambda)$对$\mathbf{a}_2$的导数并令导数为0：

$$
\frac{\partial L}{\partial \mathbf{a}_2} = 2\Sigma\mathbf{a}_2 - 2\lambda\mathbf{a}_2 - \phi\mathbf{a}_1 = 0 \tag{51}
$$

将上式左乘以$\mathbf{a}_1^{\mathrm{T}}$得：

$$
2\mathbf{a}_1^{\mathrm{T}}\Sigma\mathbf{a}_2 - 2\lambda\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2 - \phi\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 0 \tag{52}
$$

此式前两项为0，且$\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 1$，所以，$\phi = 0$。根据式(12)可得：

$$\Sigma\mathbf{a}_2 - \lambda\mathbf{a}_2 = 0 \tag{53}$$

所以，λ是协方差矩阵Σ的特征值，$\mathbf{a}_2$为对应的单位特征向量。于是，目标函数：

$$\mathbf{a}_2^{\mathrm{T}}\Sigma\mathbf{a}_2 = \mathbf{a}_2^{\mathrm{T}}\lambda\mathbf{a}_2 = \lambda\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 = \lambda \tag{54}$$

假设$\mathbf{a}_2$是Σ的第二大特征值λ_2对应的单位特征向量，显然，$\mathbf{a}_2$和λ_2是以上问题的最优解。

一般地，$\mathbf{A}_{n\times d}$的第k主成分是$\mathbf{y}_k = \mathbf{X}\mathbf{a}_k$，并且$\mathrm{var}(\mathbf{y}_k) = \lambda_k$，这里$\lambda_k$为$\Sigma$第$k$大的特征值并且$\mathbf{a}_k$为对应的单位特征向量。

综上所述，为将$\mathbf{X}_{m\times n}$进行降维，只需要求其协方差矩阵的特征向量即可。

3. PCA求解

3.1. 协方差矩阵的特征值分解法

1. 数据规范化处理

   将$\mathbf{X}_{m\times n}$做规范化处理，处理后，每一列的均值为0。

$$\mathbf{X}_{m\times n} = \frac{\mathbf{X}_{m\times n} - \text{每列均值}}{\text{每列标准差}} \tag{55}$$

2. 计算协方差矩阵

$$\Sigma = \frac{1}{m-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} \tag{56}$$

3. 求协方差矩阵的特征值和特征向量

   特征值(从大到小排列)：$\lambda_1, \lambda_2, \cdots, \lambda_k$

   特征向量：$\mathbf{A}_{n\times k} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k], \mathbf{a}_i = [a_{1i}, a_{2i}, \cdots, a_{ni}]^{\mathrm{T}}$

4. 求k个主成分

$$\mathbf{Y}_{m\times k} = \mathbf{X}_{m\times n}\mathbf{A}_{n\times k} \tag{57}$$

5. 计算k个主成分的方差贡献率

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \tag{58}$$

3.2. 数据矩阵的奇异值分解算法

给定矩阵$\mathbf{X}_{m\times n}$，并且该矩阵每列均值为0。构造一个新矩阵：

$$\mathbf{X}' = \frac{1}{\sqrt{m-1}}\mathbf{X} \tag{59}$$

那么：

$$\begin{aligned}\mathbf{X}'^{\mathrm{T}}\mathbf{X}' &= \left(\frac{1}{\sqrt{m-1}}\mathbf{X}\right)^{\mathrm{T}}\left(\frac{1}{\sqrt{m-1}}\mathbf{X}\right) \\ &= \frac{1}{m-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\end{aligned} \tag{60}$$

$$
\begin{aligned}
&= \frac{}{m-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} \\
&= \mathbf{S}
\end{aligned}
$$

不难得知，$\mathbf{S}$为$\mathbf{X}$的协方差矩阵。

根据SVD分解定理：

$$
\begin{gathered}
\mathbf{X}' = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} \\
\mathbf{S} = \mathbf{X}'^{\mathrm{T}}\mathbf{X}' = \mathbf{V}\left(\Sigma^{T}\Sigma\right)\mathbf{V}^{\mathrm{T}}
\end{gathered} \tag{61}
$$

所以，$\mathbf{V}$为协方差矩阵$\mathbf{S}$的特征向量，同时也是$\mathbf{X}'$的右奇异矩阵。

至此，将求解矩阵$\mathbf{X}_{m\times n}$协方差矩阵的特征向量，转为求解矩阵$\mathbf{X}'$的右奇异矩阵。

基于SVD分解的PCA降维求解过程：

1. 构造新矩阵

$$
\mathbf{X}' = \frac{1}{\sqrt{m-1}}\mathbf{X} \tag{62}
$$

2. 对矩阵$\mathbf{X}'$进行截断奇异值分解

$$
\begin{gathered}
\mathbf{X}' \approx \mathbf{U}_k\Sigma_k\mathbf{V}_k^{\mathrm{T}} \\
\text{其中，}\mathbf{U}_k\text{是}m\times k\text{矩阵，}\Sigma_k\text{是}k\text{阶对角矩阵，}\mathbf{V}_k\text{是}n\times k\text{矩阵}
\end{gathered} \tag{63}
$$

矩阵$\mathbf{V}_k$是k个样本主成分的线性变换矩阵。

3. 降维：n维 $\to$ k维

$$
\mathbf{Y}_{m\times k} = \mathbf{X}\mathbf{V}_k \tag{64}
$$

3.3. 示例

问题：给定矩阵$\mathbf{X}$，从2维降到1维。

$$
\mathbf{X} = \begin{vmatrix} -1, & -2 \\ -1, & 0 \\ 0, & 0 \\ 2, & 1 \\ 0, & 1 \end{vmatrix}_{5\times 2} \tag{65}
$$

降维步骤：

1. 数据标准化处理：减均值，除以标准差

$$
\mathbf{X} = \begin{vmatrix} -0.91, & -1.83 \\ -0.91, & 0 \\ 0, & 0 \\ 1.83, & 0.91 \\ 0, & 0.91 \end{vmatrix}_{5\times 2} \tag{66}
$$

2. 求协方差矩阵

$$
\mathbf{S} = \begin{vmatrix} 1.25, & 0.83 \\ 0.83, & 1.25 \end{vmatrix}_{5\times 2} \tag{67}
$$

3. 求协方差矩阵的特征值和特征向量

特征值：$[2.08, 0.42]$

特征向量：$[[0.71, 0.71], [-0.71, 0.71]]$

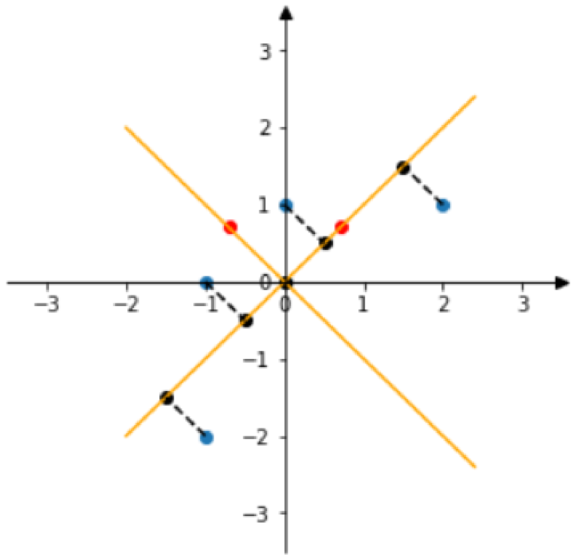
4. 求降维后的数据（第一主成分）

$$\mathbf{X}_{new} = \begin{vmatrix} -2.12 \\ -0.71 \\ 0 \\ 2.12 \\ 0.71 \end{vmatrix}_{5\times 2} \tag{68}$$

5. 求第一主成分的方差贡献

$2.08/(2.08+0.42) = 0.83$

降维结果可视化：



4. 参考链接

1. [PCA与SVD应用中负号问题 - 深海里的猫 - 博客园 (cnblogs.com)]
2. [主成分分析（Principle Component Analysis) - swolf的博客 (mrswolf.github.io)]
3. [使用numpy来理解PCA和SVD - 知乎 (zhihu.com)]
4. [CodingLabs - PCA的数学原理]