主成分分析

方佳艳

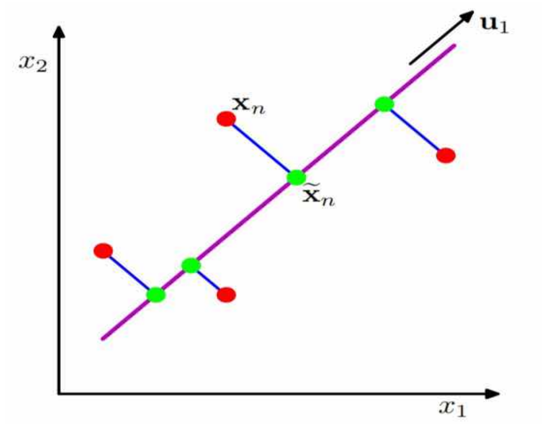
1. 概述

主成分分析（Principle Component Analysis）,或者称为 PCA ,是一种被广泛使用的技术，应用的领域包括维度降低、有损数据压缩、特征抽取、数据可视化。它也被称为 Karhunen-Loève 变换。

通常可以分别采用几何和代数的语言来描述主成分分析。

从几何的观点来看，有两种经常使用的 PCA 的定义，它们会给出同样的算法。PCA 可以被定义为数据在低维线性空间上的正交投影，这个线性空间被称为主子空间（principal subspace），使得投影数据的方差被最大化。等价地，它也可以被定义为使得平均投影代价最小的线性投影。平均投影代价是指数据点和它们的投影之间的平均平方距离。如下图所示，主成分分析的目标是寻找一个低维空间，被称为主子平面，用紫色的线表示，使得数据点（红点）在子空间上的正交投影能够最大化投影点（绿点）的方差。PCA 的另一个定义基于的是投影误差的平方和的最小值，用蓝线表示。

从代数的观点来看，PCA 相当于是对数据矩阵的一种低秩近似（low rank approximation），这里需要使用 SVD 分解（singular value decomposition）来求解最优低秩近似矩阵。



2. 几何的观点

2.1. 几何上的定义

令 $P=\{P_1,P_2,...,P_n\}\subseteq R^d$ 是 d 维实内积空间中的 n 个数据点组成的集合，PCA 的目标是在空间中找到一个最优 k 维子空间（超平面）来近似这 n 个数据点在空间中的分布，即，用 $\{\pi(P_1),\pi(P_2),...,\pi(P_n)\}$ 来近似 $\{P_1,P_2,...,P_n\}$，其中 $\pi(P_i)$ 为数据点 P_i 在 k 维子空间上的正交投影。

设目标函数为$\sum_{i=1}^{n}\left\|P_i-\pi(P_i)\right\|^2$，这里我们采用$L_2-$范数来度量近似距离误差。PCA 的任务是最小化该目标函数值。

2.2. 最简单情况：$k=0$

当$k=0$时，即用一个点来近似所有数据点的分布。于是有：

$$\pi(P_1)=\pi(P_2)=\cdots=\pi(P_n)=q$$

目标函数可以写成为：

$$\min\sum_{i=1}^{n}\left\|P_i-q\right\|^2$$

证明：最优解$q=\frac{1}{n}\sum_{i=1}^{n}P_i$

反证法：令$\mu=\frac{1}{n}\sum_{i=1}^{n}P_i$，假设$q\neq\mu$，

$$\begin{aligned}\sum_{i=1}^{n}\left\|P_i-q\right\|^2&=\sum_{i=1}^{n}\left\|P_i-\mu+\mu-q\right\|^2\\&=\sum_{i=1}^{n}\left(\left\|P_i-\mu\right\|^2+2\langle P_i-\mu,\mu-q\rangle+\left\|\mu-q\right\|^2\right)\\&=\sum_{i=1}^{n}\left\|P_i-\mu\right\|^2+2\left\langle\sum_{i=1}^{n}P_i-\mu,\mu-q\right\rangle+n\left\|\mu-q\right\|^2\\&=\sum_{i=1}^{n}\left\|P_i-\mu\right\|^2+\ n\left\|\mu-q\right\|^2\end{aligned}$$

于是，$\min\sum_{i=1}^{n}\left\|P_i-q\right\|^2\Leftrightarrow\min n\left\|\mu-q\right\|^2$，即，$q=\mu$，矛盾。

从而，当$k=0$时，最优近似点即为所有数据点的均值μ。

2.3. 用$k\geq1$维的子空间近似数据分布

由 2.2.部分可知，当只用一个点来近似所有数据时，这个最优的近似点即为所有数据点的均值。

问题：当$k\geq1$时，最优k维超平面F是否一定包含了全体数据的均值μ？

同样采用反证法：

假设$\mu\notin F$，此时可以将F平移到经过μ的位置，得到一个新的超平面F'，并且

$$F^{'} = F + \mu - \pi(\mu)$$

idea：设法将$k \geq 1$的情形规约到$k = 0$的情形。

考虑 F 的正交补空间，在有限维实内积空间V中，对于任一子空间F，都存在它的一个正交补空间$F^{\perp}$，使得$V = F \oplus F^{\perp}$，于是F和它的正交补$F^{\perp}$中的标准正交基就张成了整个空间V。

记α为空间F中的任一向量在的正交补空间$F^{\perp}$上的投影。则有：

$$\sum_{i=1}^{n} \left\| P_i - \pi(P_i) \right\|^2 = \sum_{i=1}^{n} \left\| \alpha(P_i) - \alpha(F) \right\|^2$$

由于$\alpha(F)$是一个点，于是，我们通过在F的正交补上的投影操作，使得目标函数退化成了$k = 0$的情形。

因为$\alpha(F') = \frac{1}{n}\sum_{i=1}^{n} \alpha(P_i)$，由$k = 0$时的讨论可知，$\alpha(F) = \alpha(F')$。于是，$F = F'$，

从而，$\mu \in F$。证毕。

2.4. 确定最优k维超平面的位置

由 2.2.和 2.3.的讨论可知，最优k维超平面一定包含所有数据点的均值μ。

为方便起见，假设μ是原点，确定最优k维超平面只需要找到k个标准正交基$t_1, \cdots, t_k$即可，$F = span\{t_1, \cdots, t_k\}$。

2.4.1. 最简单的情况（k=1）

当$k = 1$时，也就是用一条直线t_1去近似所有数据的分布。这里假设t_1是单位向量。

由勾股定理可知：

$$\min \sum_{i=1}^{n} \left\| P_i - \pi(P_i) \right\|^2 = \sum_{i=1}^{n} \left\| P_i \right\|^2 - \sum_{i=1}^{n} \left(\langle P_i, t_1 \rangle \right)^2$$

于是，$\min \sum_{i=1}^{n} \left\| P_i - \pi(P_i) \right\|^2 \Leftrightarrow \max_{\|t_1\|=1} \sum_{i=1}^{n} \left(\langle P_i, t_1 \rangle \right)^2$

设数据矩阵$A = \begin{pmatrix} P_1^T \\ P_2^T \\ \vdots \\ P_n^T \end{pmatrix}$，则有：

$$\max_{\|t_1\|=1} \sum_{i=1}^{n} \left(\langle P_i, t_1 \rangle \right)^2 = \max_{\|t_1\|=1} \left\| A t_1 \right\|_F^2$$

对矩阵A作 SVD 分解，有$A = U \cdot S \cdot V^T$，其中，U和V都是正交矩阵，S为对角矩阵。

于是，$At_1 = A \cdot V \cdot V^T \cdot t_1 = U \cdot S \cdot V^T \cdot t_1$。

由于V是欧氏空间中的正交变换，因此它是保距同构的。而实内积空间中的任一保距同构都保持向量长度不变，于是，$V^T t_1$仍然是一个单位向量。

$U = [u_1, u_2, \cdots, u_n]$，$V = [v_1, v_2, \cdots, v_d]$，$S = diag\{\sigma_1, \cdots, \sigma_d\}, d \le n$

$U \cdot S = [\sigma_1 \mu_1, \sigma_2 \mu_2, \cdots, \sigma_d \mu_d, 0, \cdots, 0]$

先考虑$d = 2$时的情形：

$U \cdot S = [\sigma_1 \mu_1, \sigma_2 \mu_2, 0, \cdots, 0]$，$V^T t_1 = [\lambda_1, \lambda_2]^T$，其中，$\lambda_1{}^2 + \lambda_2{}^2 = 1$

于是，$U \cdot S \cdot V^T \cdot t_1 = \lambda_1 \sigma_1 \mu_1 + \lambda_2 \sigma_2 \mu_2$，令$x = \lambda_1 \sigma_1 \mu_1$，$y = \lambda_2 \sigma_2 \mu_2$，

由$\lambda_1{}^2 + \lambda_2{}^2 = 1$得，$\left(\frac{x}{\sigma_1}\right)^2 + \left(\frac{y}{\sigma_2}\right)^2 = 1$，其中，$\sigma_1 \ge \sigma_2$

由 $\max\limits_{\|t_1\|=1} \sum_{i=1}^{n} \left(\langle P_i, t_1 \rangle\right)^2 = \max\limits_{\|t_1\|=1} \|At_1\|_F^2 = \max\limits_{\|t_1\|=1} U \cdot S \cdot V^T \cdot t_1 = \max\limits_{\|t_1\|=1} \lambda_1 \sigma_1 \mu_1 + \lambda_2 \sigma_2 \mu_2$

得：可$x = \pm\sigma_1$，$y = 0 \Rightarrow V^T t_1 = (\pm 1, 0) \Rightarrow t_1 = \pm v_1$

于是，我们得到了，当$d = 2$时，最优近似直线的方向即为数据矩阵A作 SVD 分解后的正交矩阵V的列向量v_1的方向（正向或反向）。

同理可得，当$d \ge 2$时，$t_1 = \pm v_1$

2.4.2. 用$k \ge 1$维的超平面近似数据分布

根据 2.4.1.部分的讨论，可以猜测：$F = span\{v_1, v_2, \cdots, v_k\}$

证明：（采用反证法）

假设$v_1 \notin F$，设$F = span\{t_1, t_2, \cdots, t_k\}$

令$t_1 = \frac{\pi(v_1)}{\|\pi(v_1)\|}$

于是我们可以构造一个新的超平面$F' = span\{v_1, t_2, \cdots, t_k\}$。$F'$实际是$F$以$span\{t_2, \cdots, t_k\}$为轴作旋转得到的一个经过$v_1$的超平面。

类比$k = 1$的情形，有：

$$\sum_{i=1}^{n} \|P_i - \pi(P_i)\|^2 = \sum_{i=1}^{n} \|P_i\|^2 - \sum_{i=1}^{n} \left(\left(\langle P_i, t_1 \rangle\right)^2 + \cdots + \left(\langle P_i, t_k \rangle\right)^2\right)$$

于是，$\min\sum_{i=1}^{n}\|P_i-\pi(P_i)\|^2 \Leftrightarrow \max\sum_{i=1}^{n}\left((\langle P_i,t_1\rangle)^2+(\langle P_i,t_2\rangle)^2+\cdots+(\langle P_i,t_k\rangle)^2\right)$

在 F'上有：

$$\sum_{i=1}^{n}\left((\langle P_i,v_1\rangle)^2+(\langle P_i,t_2\rangle)^2+\cdots+(\langle P_i,t_k\rangle)^2\right)\text{，记}$$

$$\text{记}\alpha_F=\sum_{i=1}^{n}\left((\langle P_i,t_1\rangle)^2+(\langle P_i,t_2\rangle)^2+\cdots+(\langle P_i,t_k\rangle)^2\right)$$

$$\alpha_{F'}=\sum_{i=1}^{n}\left((\langle P_i,v_1\rangle)^2+(\langle P_i,t_2\rangle)^2+\cdots+(\langle P_i,t_k\rangle)^2\right)$$

$\Rightarrow \alpha_{F'}\geq\alpha_F$

$\Rightarrow F'$是最优的超平面

$\Rightarrow v_1$在最优的k维超平面上

以此类推：$\{v_1,\cdots,v_k\}\subseteq$最优的$k$维超平面$F$上。

由于$v_1,\cdots,v_k$是一组互相正交的单位向量，因此构成了F的一个基。从而，最优的k维超平面F处在由$v_1,\cdots,v_k$张成的一个椭球上。于是，我们最终确定了最优超平面的位置。

3. 代数的观点

3.1. 代数上的定义

令A为数据矩阵，PCA 的目标是希望找到一个秩为k的矩阵A_k，使得矩阵$A-A_k$的 Frobenious 范数$\|A-A_k\|_F$最小。（$\|X\|_F=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d}{x_{ij}}^2}$）

$$A=\begin{pmatrix}P_1^T\\P_2^T\\\vdots\\P_n^T\end{pmatrix}\text{，}A_k=\begin{pmatrix}\pi(P_1)^T\\\pi(P_2)^T\\\vdots\\\pi(P_n)^T\end{pmatrix}$$

由于$v_1,\cdots,v_k$是最优超平面F上的一个标准正交基，因此

$$\pi(P_i)=\langle P_i,v_1\rangle v_1+\cdots+\langle P_i,v_k\rangle v_k$$

$$\Rightarrow \pi(P_i)^T = P_i^T \cdot [v_1, \cdots, v_k] \begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \end{bmatrix}$$

$$\Rightarrow A_k = A \cdot [v_1, \cdots, v_k] \begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \end{bmatrix}$$

将 A 作 SVD 分解：

$A = (U)_{n\times n}(S)_{d\times d}(V^T)_{d\times d}$

4. 几何与代数定义上的等价性

几何上：

$$\min \sum_{i=1}^{n} \left\| P_i - \pi(P_i) \right\|^2$$

代数上：

$$A = \begin{pmatrix} P_1^T \\ P_2^T \\ \vdots \\ P_n^T \end{pmatrix}, \quad A_k = \begin{pmatrix} \pi(P_1)^T \\ \pi(P_2)^T \\ \vdots \\ \pi(P_n)^T \end{pmatrix}$$

$$\min \left\| A - A_k \right\|_F = \sum_{i=1}^{n} \left\| P_i - \pi(P_i) \right\|^2$$

5. PCA 的缺点

1）复杂度较高：对数据矩阵 A 作 SVD 分解时的计算复杂度是：并不是线性复杂度。
2）需要近似计算特征向量，数值精度和稳定性会受到影响。
3）有些数据不适合 PCA。
4）对于大规模的数据集需要对数据作多次读取（比如用 QR 分解），因此不太适应于分布式计算和处理流数据问题。