PCA

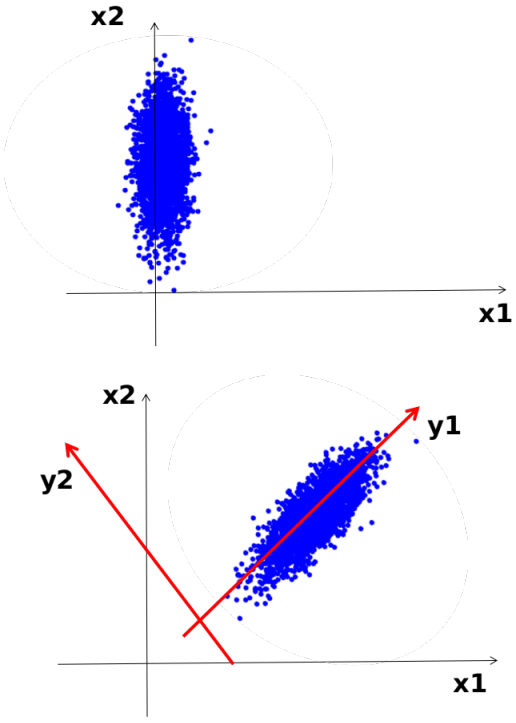
谢欣然

2021 年 6 月 4 日

1 原理

PCA（Principal Component Analysis）是一种常见的数据分析方式，常用于高维数据的降维，可用于提取数据的主要特征分量。

PCA 的数学推导可以从最大可分型和最近重构性两方面进行，前者的优化条件为划分后方差最大，后者的优化条件为点到划分平面距离最小，这里我将从最大可分性的角度进行证明。



1.1 基变换的矩阵表示

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_R \end{pmatrix} \begin{pmatrix} a_1 & a_2 & \cdots & a_M \end{pmatrix} = \begin{pmatrix} p_1a_1 & p_1a_2 & \cdots & p_1a_M \\ p_2a_1 & p_2a_2 & \cdots & p_2a_M \\ \vdots & \vdots & \ddots & \vdots \\ p_Ra_1 & p_Ra_2 & \cdots & p_Ra_M \end{pmatrix}$$

其中 P_i 是一个行向量，表示第 i 个基，a_j 是一个列向量，表示第 j 个原始数据记录。实际上也就是做了一个向量矩阵化的操作。两个矩阵相乘的意义是将右边矩阵中的每一列向量 a_j 变换到左边矩阵中以每一行行向量为基所表示的空间中去。也就是说一个矩阵可以表示一种线性变换。

1.2 最大可分性

上面我们讨论了选择不同的基可以对同样一组数据给出不同的表示，如果基的数量少于向量本身的维数，则可以达到降维的效果。

但是我们还没回答一个最关键的问题：如何选择基才是最优的。或者说，如果我们有一组 N 维向量，现在要将其降到 K 维（K 小于 N），那么我们应该如何选择 K 个基才能最大程度保留原有的信息？

一种直观的看法是：希望投影后的投影值尽可能分散，因为如果重叠就会有样本消失。当然这个也可以从熵的角度进行理解，熵越大所含信息越多。

上面的问题可以被形式化表述为：寻找一个一维基，使得所有数据变换为这个基上的坐标表示后，方差值最大。

在一维空间中我们可以用方差来表示数据的分散程度。而对于高维数据，我们用协方差进行约束，协方差可以表示两个变量的相关性。为了让两个变量尽可能表示更多的原始信息，我们希望它们之间不存在线性相关性，因为相关性意味着两个变量不是完全独立，必然存在重复表示的信息。

至此，我们得到了降维问题的优化目标：将一组 N 维向量降为 K 维，其目标是选择 K 个单位正交基，使得原始数据变换到这组基上后，各变量两两间协方差为 0，而变量方差则尽可能大（在正交的约束下，取最大的 K 个方差）。

1.3 协方差矩阵

我们看到，最终要达到的目的与变量内方差及变量间协方差有密切关系。因此我们希望能将两者统一表示，仔细观察发现，两者均可以表示为内积的形式，而内积又与矩阵相乘密切相关。于是我们有：

假设我们只有 a 和 b 两个变量，那么我们将它们按行组成矩阵 X：

$$X = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{pmatrix}$$

然后：

$$\frac{1}{m}XX^{\mathsf{T}} = \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m} a_i^2 & \frac{1}{m}\sum_{i=1}^{m} a_i b_i \\ \frac{1}{m}\sum_{i=1}^{m} a_i b_i & \frac{1}{m}\sum_{i=1}^{m} b_i^2 \end{pmatrix} = \begin{pmatrix} Cov(a,a) & Cov(a,b) \\ Cov(b,a) & Cov(b,b) \end{pmatrix}$$

我们可以看到这个矩阵对角线上的分别是两个变量的方差，而其它元素是 a 和 b 的协方差。两者被统一到了一个矩阵里。

我们很容易被推广到一般情况：

设我们有 m 个 n 维数据记录，将其排列成矩阵 $X_{n,m}$，设 $C = \frac{1}{m}XX^T$，则 C 是一个对称矩阵，其对角线分别对应各个变量的方差，而第 i 行 j 列和 j 行 i 列元素相同，表示 i 和 j 两个变量的协方差。

1.4 矩阵对角化

根据我们的优化条件，我们需要将除对角线外的其它元素化为 0，并且在对角线上将元素按大小从上到下排列（变量方差尽可能大）。

设原始数据矩阵 X 对应的协方差矩阵为 C，而 P 是一组基按行组成的矩阵，设 Y=PX，则 Y 为 X 对 P 做基变换后的数据。设 Y 的协方差矩阵为 D，我们推导一下 D 与 C 的关系：

$$
\begin{aligned}
D &= \frac{1}{m} Y Y^T \\
&= \frac{1}{m} (PX)(PX)^T \\
&= \frac{1}{m} PXX^T P^T \\
&= P(\frac{1}{m} XX^T) P^T \\
&= PCP^T
\end{aligned}
$$

我们要找的 P 是能让原始协方差矩阵对角化的 P。换句话说，优化目标变成了寻找一个矩阵 P，满足 PCP^T 是一个对角矩阵，并且对角元素按从大到小依次排列，那么 P 的前 K 行就是要寻找的基，用 P 的前 K 行组成的矩阵乘以 X 就使得 X 从 N 维降到了 K 维并满足上述优化条件。

由上文知道，协方差矩阵 C 是一个是对称矩阵，在线性代数中实对称矩阵有一系列非常好的性质：

1. 实对称矩阵不同特征值对应的特征向量必然正交。2. 设特征向量 λ 重数为 r，则必然存在 r 个线性无关的特征向量对应于 λ，因此可以将这 r 个特征向量单位正交化。

由上面两条可知，一个 n 行 n 列的实对称矩阵一定可以找到 n 个单位正交特征向量，设这 n 个特征向量为 $e_1, e_2, \cdots e_n$，我们将其按列组成矩阵：$E = (e_1, e_2, \cdots e_n)$ 。

则对协方差矩阵 C 有如下结论：

$$
E^T C E = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}
$$

其中 Λ 为对角矩阵，其对角元素为各特征向量对应的特征值（可能有重复）。

到这里，我们发现我们已经找到了需要的矩阵: $P = E^T$。

P 是协方差矩阵的特征向量单位化后按行排列出的矩阵，其中每一行都是 C 的一个特征向量。如果设 P 按照 Λ 中特征值的从大到小，将特征向量从上到下排列，则用 P 的前 K 行组成的矩阵乘以原始数据矩阵 X，就得到了我们需要的降维后的数据矩阵 Y。

1.5 证明方差最大化

我们知道样本点 x_i 在基 w 下的坐标为：$x_i^T w$ ，于是我们有方差 (我们令均值为 0)：

$$\begin{aligned} D(x) &= \frac{1}{m}\sum_{i=1}^{m}(x_i^T w)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}(x_i^T w)^T(x_i^T w) \\ &= \frac{1}{m}\sum_{i=1}^{m} w^T x_i x_i^T w \\ &= w^T(\frac{1}{m}\sum_{i=1}^{m} x_i x_i^T)w \end{aligned}$$

我们看到 $\frac{1}{m}\sum_{i=1}^{m} x_i x_i^T$ 就是原样本的协方差，我们另这个矩阵为 Λ ，于是我们有：

$$\begin{cases} max\{w^T \Lambda w\} \\ s.t.\, w^T w = 1 \end{cases}$$

然后构造拉格朗日函数：

$$L(w) = w^T \Lambda w + \lambda(1 - w^T w)$$

对 w 求导：

$$\Lambda w = \lambda w$$

此时我们的方差为：

$$D(x) = w^T \Lambda w = \lambda w^T w = \lambda$$

于是我们发现，x 投影后的方差就是协方差矩阵的特征值。我们要找到最大方差也就是协方差矩阵最大的特征值，最佳投影方向就是最大特征值所对应的特征向量，次佳就是第二大特征值对应的特征向量，以此类推。

至此我们完成了基于最大可分性的 PCA 数学证明

1.6 求解步骤

设有 m 条 n 维数据。

1. 将原始数据按列组成 n 行 m 列矩阵 X；

2. 将 X 的每一行进行零均值化，即减去这一行的均值；

3. 求出协方差矩阵 $C = \frac{1}{m} X_X^T$；

4. 求出协方差矩阵的特征值及对应的特征向量；

5. 将特征向量按对应特征值大小从上到下按行排列成矩阵，取前 k 行组成矩阵 P；

6.$Y = PX$ 即为降维到 k 维后的数据。