# 主成分分析（PCA）原理总结

原创 石头 机器学习算法那些事 2019-03-01

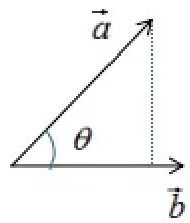
主成分分析（Principal components analysis，以下简称PCA）是最常用的降维方法之一，在数据压缩和消除冗余方面具有广泛的应用，本文由浅入深的对其降维原理进行了详细总结。

**目录**



## 1. 向量投影和矩阵投影的含义

如下图：

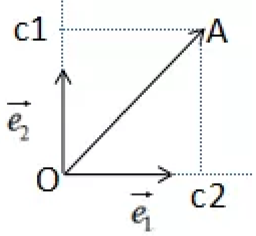



向量a在向量b的投影为：

$$\vec{a} * \cos\theta$$

其中，θ是向量间的夹角 。

向量a在向量b的投影表示向量a在向量b方向的信息，若θ=90°时，向量a与向量b正交，向量a无向量b信息，即向量间无冗余信息 。因此，向量最简单的表示方法是用基向量表示，如下图：




向量表示方法：

$$\overrightarrow{OA} = c1 * \overrightarrow{e1} + c2 * \overrightarrow{e2}$$

其中，c1是$\overrightarrow{OA}$在e1方向的投影，c2是$\overrightarrow{OA}$在e2方向的投影，e1和e2是基向量

我们用向量的表示方法扩展到矩阵，若矩阵$A_{n\times n}$的秩r(A)=n，$A=(a_1,a_2,...,a_n)$，其中ai（i=1,2,...,n）为n个维度的列向量，那么矩阵A的列向量表示为：

$$
\begin{aligned}
a_1 &= c_{11} * e_1 + c_{12} * e_2 + \cdots c_{1n} * e_n \\
a_2 &= c_{21} * e_1 + c_{22} * e_2 + \cdots c_{2n} * e_n \\
&\vdots \\
a_n &= c_{n1} * e_1 + c_{n2} * e_2 + \cdots c_{nn} * e_n
\end{aligned}
$$

其中，e1，e2，...，en为矩阵A的特征向量 。

若矩阵A是对称矩阵，那么特征向量为正交向量，我们对上式结合成矩阵的形式：

$$
A = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} * \begin{pmatrix} e_1^T \\ e_2^T \\ \vdots \\ e_n^T \end{pmatrix} \Rightarrow A(\mathrm{e}_1, e_2, ..., e_n) = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}
$$

$$
\Rightarrow (A\mathrm{e}_1, Ae_2, ..., Ae_n) = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}
$$

由上式可知，对称矩阵A在各特征向量的投影等于矩阵列向量展开后的系数，特征向量可理解为基向量。

### 2. 向量降维和矩阵降维含义

向量降维可以通过投影的方式实现，N维向量映射为M维向量转换为N维向量在M个基向量的投影，如N维向量 $\overrightarrow{OA} = (a_1, a_2, ..., a_n)^T$，M个基向量分别为 $\vec{e_1}, \vec{e_1}, ..., \vec{e_m}$，$\overrightarrow{OA}$ 在基向量的投影：

$$
\begin{cases}
a_1^{'} = \vec{e_1}^T * \overrightarrow{OA} \\
a_1^{'} = \vec{e_2}^T * \overrightarrow{OA} \\
\vdots \\
a_m^{'} = \vec{e_2}^T * \overrightarrow{OA}
\end{cases}
$$

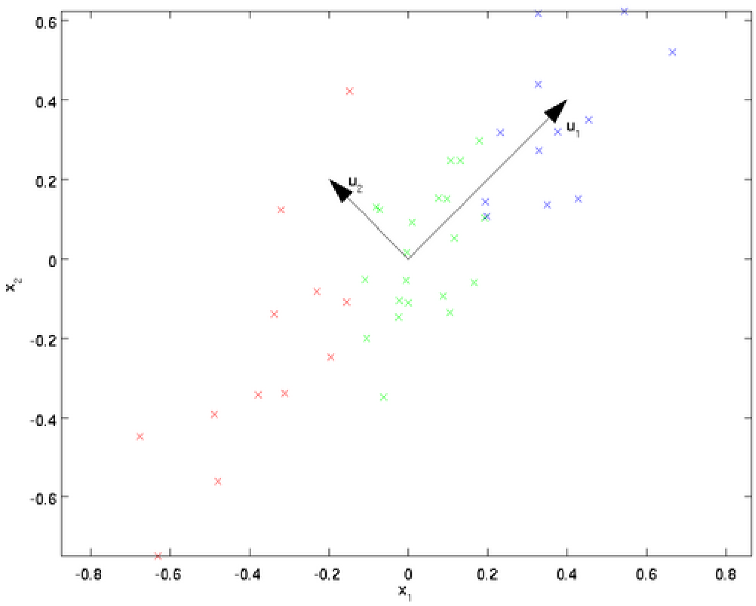
通过上式完成了降维，降维后的坐标为：

$$
(a_1^{'}, a_2^{'}, ..., a_m^{'})^T
$$

矩阵是由多个列向量组成的，因此矩阵降维思想与向量降维思想一样，只要求得矩阵在各基向量的投影即可，基向量可以理解为新的坐标系，投影就是降维后的坐标，那么问题来了，如何选择基向量？

## 3. 基向量选择算法

已知样本集的分布，如下图：




样本集共有两个特征x1和x2，现在对该样本数据从二维降到一维，图中列了两个基向量u1和u2，样本集在两个向量的投影表示了不同的降维方法，哪种方法好，需要有评判标准：（1）降维前后样本点的总距离足够近，即最小投影距离；（2）降维后的样本点（投影）尽可能的散开，即最大投影方差 。因此，根据上面两个评判标准可知选择基向量u1较好。

我们知道了基向量的选择标准，下面介绍基于这两个评判标准来推导基向量：

（1）基于最小投影距离

假设有n个n维数据 $\{x^{(1)},x^{(2)},...,x^{(n)}\}$ ，记为X。现在对该数据从n维降到m维，关键是找到m个基向量，假设基向量为{w1,w2,...,wm}，记为矩阵W，矩阵W的大小是n×m。

原始数据在基向量的投影： $z^{(i)}=(z_1^{(i)},z_2^{i},...,z_m^{i})^T$

投影坐标计算公式：

$$\begin{cases} z_1^{(i)}=w_1^T x^{(i)} \\ z_2^{(i)}=w_2^T x^{(i)} \\ \vdots \\ z_m^{(i)}=w_m^T x^{(i)} \end{cases}$$

根据投影坐标和基向量，得到该样本的映射点：

$$\overline{x^{(i)}} = z_1^{(i)} w_1 + z_2^{(i)} w_2 + \cdots + z_m^{(i)} w_m$$
$$\Rightarrow \overline{x^{(i)}} = \sum_{j=1}^{m} z_j^{(i)} w_j$$
$$\Rightarrow \overline{x^{(i)}} = W * z^{(i)}$$

最小化样本和映射点的总距离：

$$\min\left(\sum_{i=1}^{n} \| x^{(i)} - \overline{x^{(i)}} \|_2^2\right)$$

推导上式，得到最小值对应的基向量矩阵W，推导过程如下：

$$\min\left(\sum_{i=1}^{n} \| x^{(i)} - \overline{x^{(i)}} \|_2^2\right)$$
$$\sum_{i=1}^{n} \| x^{(i)} - \overline{x^{(i)}} \|_2^2 = \sum_{i=1}^{n} \| x^{(i)} - W * z^{(i)} \|_2^2 \qquad (1)$$
$$\Rightarrow \sum_{i=1}^{n} (x^{(i)})^T x^{(i)} - 2\sum_{i=1}^{n} (x^{(i)})^T (W * z^{(i)}) + \sum_{i=1}^{n} (W * z^{(i)})^T (W * z^{(i)}) \qquad (2)$$
$$\Rightarrow \sum_{i=1}^{n} (x^{(i)})^T x^{(i)} - 2\sum_{i=1}^{n} ((x^{(i)})^T W) * z^{(i)}) + \sum_{i=1}^{n} (W * z^{(i)})^T (W * z^{(i)}) \qquad (3)$$
$$\because (x^{(i)})^T W = (z^{(i)})^T$$

(3)式可得：

$$\Rightarrow \sum_{i=1}^{n} (x^{(i)})^T x^{(i)} - 2\sum_{i=1}^{n} (z^{(i)})^T * z^{(i)}) + \sum_{i=1}^{n} (W * z^{(i)})^T (W * z^{(i)}) \qquad (4)$$
$$\because (AB)^T = B^T A^T$$

(4)式可得：

$$\Rightarrow \sum_{i=1}^{n} (x^{(i)})^T x^{(i)} - 2\sum_{i=1}^{n} (z^{(i)})^T * z^{(i)}) + \sum_{i=1}^{n} (z^{(i)})^T W^T W * z^{(i)} \qquad (5)$$

$\because W^T W = E$，$E$为单位矩阵

(5)式可得：

$$\Rightarrow \sum_{i=1}^{n} (x^{(i)})^T x^{(i)} - 2\sum_{i=1}^{n} (z^{(i)})^T * z^{(i)}) + \sum_{i=1}^{n} (z^{(i)})^T z^{(i)} \qquad (6)$$
$$\Rightarrow \sum_{i=1}^{n} (x^{(i)})^T x^{(i)} - \sum_{i=1}^{n} (z^{(i)})^T z^{(i)} \qquad (7)$$

$\because (z^{(i)})^T z^{(i)} = \mathrm{tr}(z^{(i)} * (z^{(i)})^T)$，其中tr表示矩阵的迹

（7）式可得：

$$\Rightarrow \sum_{i=1}^{n}(x^{(i)})^T x^{(i)} - \sum_{i=1}^{n}\mathrm{tr}(z^{(i)} * (z^{(i)})^T) \quad (8)$$

$$\Rightarrow \sum_{i=1}^{n}(x^{(i)})^T x^{(i)} - \sum_{i=1}^{n}\mathrm{tr}(W^T \mathrm{x}^{(i)} * (W^T x^{(i)})^T) \quad (9)$$

$$\Rightarrow \sum_{i=1}^{n}(x^{(i)})^T x^{(i)} - \mathrm{tr}(\mathrm{W}^{\mathrm{T}} \sum_{i=1}^{n}(\mathrm{x}^{(i)} * (x^{(i)})^T) W) \quad (10)$$

$$\because \sum_{i=1}^{n}(\mathrm{x}^{(i)} * (x^{(i)})^T) = XX^T \ (11)$$

(11)式可得：

$$\Rightarrow \sum_{i=1}^{n}(x^{(i)})^T x^{(i)} - \mathrm{tr}(\mathrm{W^T XX^T W}) \quad (12)$$

$$\because \sum_{i=1}^{n}(x^{(i)})^T x^{(i)}\text{是常数}$$

$$\therefore (12)\text{式最小值等价于求}\mathrm{tr}(\mathrm{W^T XX^T W})\text{的最大值}$$

$$\text{令}\mathrm{J(W)} = \mathrm{tr}(\mathrm{W^T XX^T W}) \quad s.t. W^T W = I \quad (13)$$

利用拉格朗日函数，上式等价于：

$$\Rightarrow \mathrm{J(W)} = \mathrm{tr}(\mathrm{W^T XX^T W} + \lambda(\mathrm{W^T W - I})$$

$$\text{令}\mathrm{J}'(W) = 0\text{，得：}$$

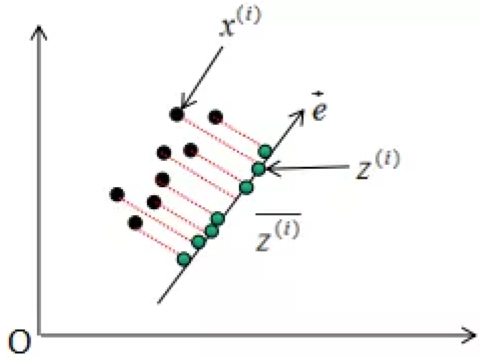
$$\mathrm{XX^T W} = -\lambda W \quad (14)$$

因此，当基向量是$XX^T$的特征向量时，具有最小投影距离。

所以我们选择$XX^T$的特征向量作为投影的基向量 。

### (2) 基于最大投影方差

我们希望降维后的样本点尽可能分散，方差可以表示这种分散程度。




如上图所示，$x^{(i)}$表示原始数据，$z^{(i)}$表示投影数据，$\overline{z^{(i)}}$表示投影数据的平均值。所以最大化投影方差表示为：

$$\max(\sum_{i=1}^{n}(z^{(i)}-\overline{z^{(i)}})^2)$$

下面推导上式，得到相应的基向量矩阵W，推导过程如下：

$$\sum_{i=1}^{n}(z^{(i)}-\overline{z^{(i)}})^2 \quad (1)$$

$\because$ 原始数据进行去中心化处理

$$\therefore \overline{x^{(i)}}=0$$

$$\therefore \overline{z^{(i)}}=W^T\overline{x^{(i)}}=0$$

(1)式可得：

$$\Rightarrow \sum_{i=1}^{n}(z^{(i)})^2$$

$$\Rightarrow \sum_{i=1}^{n}tr(z^{(i)}*z^{(i)T}) \quad (2)$$

$$\Rightarrow \sum_{i=1}^{n}\mathrm{tr}(W^Tx^{(i)}*(W^Tx^{(i)})^T)$$

$$\Rightarrow \sum_{i=1}^{n}\mathrm{tr}(W^Tx^{(i)}x^{(i)T}W) \quad (3)$$

$$\because \sum_{\mathrm{i}=1}^{n}x^{(i)}x^{(i)T}=XX^T$$

(3)式可得：

$$\Rightarrow \sum_{i=1}^{n}\mathrm{tr}(W^TXX^TW) \quad s.t. \quad W^TW=E \quad (4)$$

我们发现(4)式与上一节的(13)式是相同的。

因此，基向量矩阵W满足下式：

$$XX^TW=-\lambda W$$

小结：降维通过样本数据投影到基向量实现的，基向量的个数等于降维的个数，基向量是通过上式求解的。

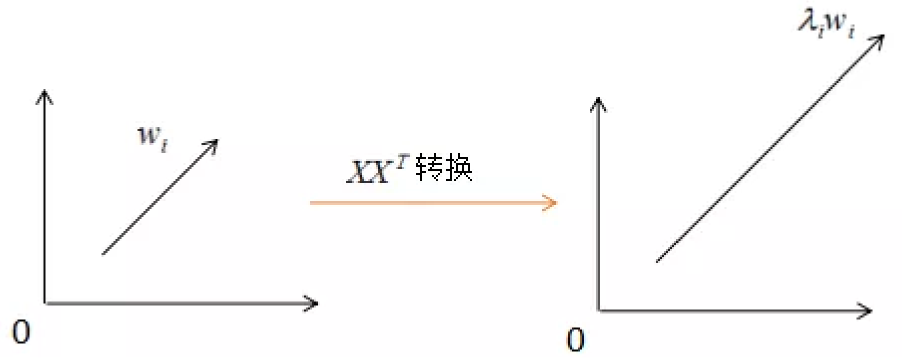
## 4. 基向量个数的确定

我们知道怎么求解基向量，但是我们事先确定了基向量的个数，如上节的m个基向量，那么怎么根据样本数据自动的选择基向量的个数了？在回答这一问题前，简单阐述下特征向量和特征值的意义。

假设向量wi，λi分别为$XX^T$的特征向量和特征值，表达式如下：

$$XX^Tw_i=\lambda_iw_i$$

对应的图：




由上图可知，$XX^T$没有改变特征向量wi的方向，只在wi的方向上伸缩或压缩了λi倍。特征值代表了$XX^T$在该特征向量的信息分量。特征值越大，包含矩阵$XX^T$的信息分量亦越大。因此，我们可以用λi去选择基向量个数。我们设定一个阈值threshold，该阈值表示降维后的数据保留原始数据的信息量，假设降维后的特征个数为m，降维前的特征个数为n，m应满足下面条件：

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \geq threshold$$

因此，通过上式可以求得基向量的个数m，即取前m个最大特征值对应的基向量 。

投影的基向量：

$$W=(w_1,w_2,...,w_m)$$

投影的数据集：

$$z^{(i)}=W^T x^{(i)}$$

## 5. 中心化的作用

我们在计算协方差矩阵$XX^T$的特征向量前，需要对样本数据进行中心化，中心化的算法如下：

$$x^{(i)}=x^{(i)}-\frac{1}{n}\sum_{i=1}^{n}x^{(i)}$$

中心化数据各特征的平均值为0，计算过程如下：

对上式求平均：

$$\overline{x^{(i)}}=\frac{1}{n}\sum_{i=1}^{n}(x^{(i)}-\frac{1}{n}\sum_{i=1}^{n}x^{(i)})$$

$$\Rightarrow \overline{x^{(i)}}=\frac{1}{n}\sum_{i=1}^{n}x^{(i)}-\frac{1}{n}\sum_{i=1}^{n}x^{(i)}$$

$$\Rightarrow \overline{x^{(i)}}=0$$

中心化的目的是简化算法，我们重新回顾下协方差矩阵，以说明中心化的作用 。

$X=(x^{(1)},x^{(2)},...,x^{(n)})$，X表示共有n个样本数。

每个样本包含n个特征，即：

$$x^{(i)}=(x_1^{(i)},x_2^{(i)},\ldots,x_n^{(i)})$$

展开$XX^T$：

$$XX^T=(x^{(1)},x^{(2)},\ldots,x^{(n)})\begin{pmatrix}x^{(1)T}\\x^{(2)T}\\\vdots\\x^{(n)T}\end{pmatrix}\qquad(1)$$

$$\Rightarrow XX^T=x^{(1)}x^{(1)T}+x^{(2)}x^{(2)T}+\ldots+x^{(n)}x^{(n)T}\qquad(2)$$

为了阅读方便，我们只考虑两个特征的协方差矩阵：

$$\mathrm{cov}(x_1,x_2)=E[x_1-E(x_1)(x_2-E(x_2))]$$

$$\Rightarrow \mathrm{cov}(x_1,x_2)=\frac{1}{n}\sum_{i=1}^{n}[(x_1^{(i)}-\overline{x_1})(x_2^{(i)}-\overline{x_2})]$$

$$\because \text{数据经过中心化处理}, \therefore \overline{x_1}=0,\overline{x_2}=0$$

$$\Rightarrow \mathrm{cov}(x_1,x_2)=\frac{1}{n}\sum_{i=1}^{n}x_1^{(i)}x_2^{(i)}\qquad(3)$$

由(3)式推导(2)式得：

$$XX^T=\begin{pmatrix}\mathrm{cov}(x_1,x_1),\mathrm{cov}(x_1,x_2),\ldots,\mathrm{cov}(x_1,x_2)\\\mathrm{cov}(x_1,x_2),\mathrm{cov}(x_1,x_2),\ldots,\mathrm{cov}(x_1,x_2)\\\vdots\qquad\qquad\vdots\qquad\qquad\vdots\\\mathrm{cov}(x_1,x_2),\mathrm{cov}(x_1,x_2),\ldots,\mathrm{cov}(x_1,x_2)\end{pmatrix}$$

所以$XX^T$是样本数据的协方差矩阵，但是，切记必须事先对数据进行中心化处理 。

## 6. PCA算法流程

1）样本数据中心化。

2）计算样本的协方差矩阵$XX^T$。

3）求协方差矩阵$XX^T$的特征值和特征向量，并对该向量进行标准化（基向量）。

3）根据设定的阈值，求满足以下条件的降维数m。

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{n}\lambda_i}\geq threshold$$

4）取前m个最大特征值对应的向量，记为W。

$$W=(w_1,w_2,\ldots,w_m)$$

5）对样本集的每一个样本 $x^{(i)}$，映射为新的样本 $z^{(i)}$。

$$z^{(i)} = W^T x^{(i)}$$

6）得到映射后的样本集D'。

$$D' = (z^{(1)}, z^{(2)}, \ldots, z^{(m)})$$

## 7. 核主成分分析（KPCA）介绍

因为 $XX^T$ 可以用样本数据内积表示：

$$XX^T = \sum_{i=1}^{n} x^{(i)} x^{(i)T}$$

由核函数定义可知，可通过核函数将数据映射成高维数据，并对该高维数据进行降维：

$$\sum_{i=1}^{n} \phi(x^{(i)}) \phi(x^{(i)})^T W = \lambda W$$

KPCA一般用在数据不是线性的，无法直接进行PCA降维，需要通过核函数映射成高维数据，再进行PCA降维 。

## 8. PCA算法总结

PCA是一种非监督学习的降维算法，只需要计算样本数据的协方差矩阵就能实现降维的目的，其算法较易实现，但是降维后特征的可解释性较弱，且通过降维后信息会丢失一些，可能对后续的处理有重要影响。

参考

https://www.cnblogs.com/pinard/p/6239403.html#undefined

A Singularly Valuable Decompostion: The SVD of a Matrix

推荐阅读

XGBoost算法原理小结

LightGBM算法原理小结

深入浅出核函数