PCA主成分分析学习总结

zhuanlan.zhihu.com/p/32412043

鱼遇雨欲语与余微信公众号：Coggle数据科学

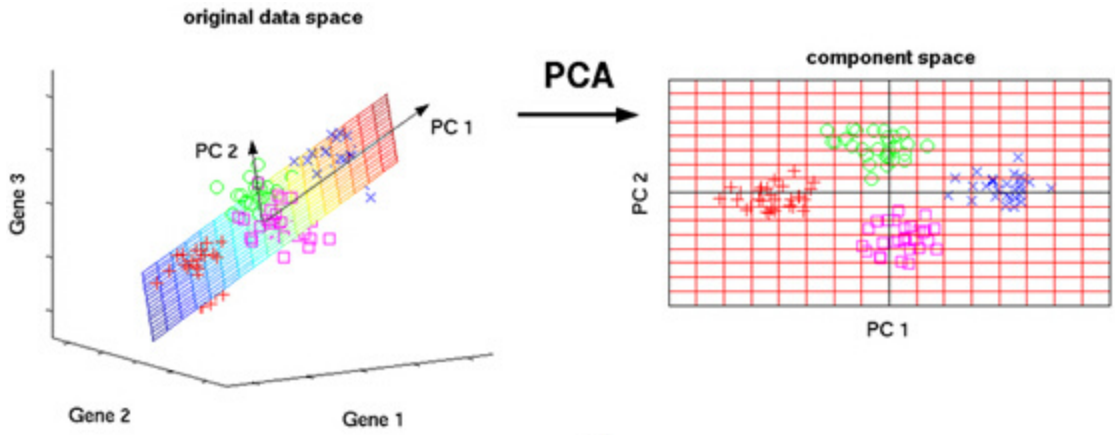


图 1

大概主成分分析（Principal components analysis，以下简称PCA）是最重要的降维方法之一。在数据压缩消除冗余和数据噪音消除等领域都有广泛的应用。一般我们提到降维最容易想到的算法就是PCA，下面我们就对PCA的原理做一个总结。

首先考虑一个问题：对于正交属性空间中的样本点，如何用一个超平面（直线的高维推广）对所有样本进行恰当的表达？

可以想到，若存在这样的超平面，那么它大概具有这样的性质：

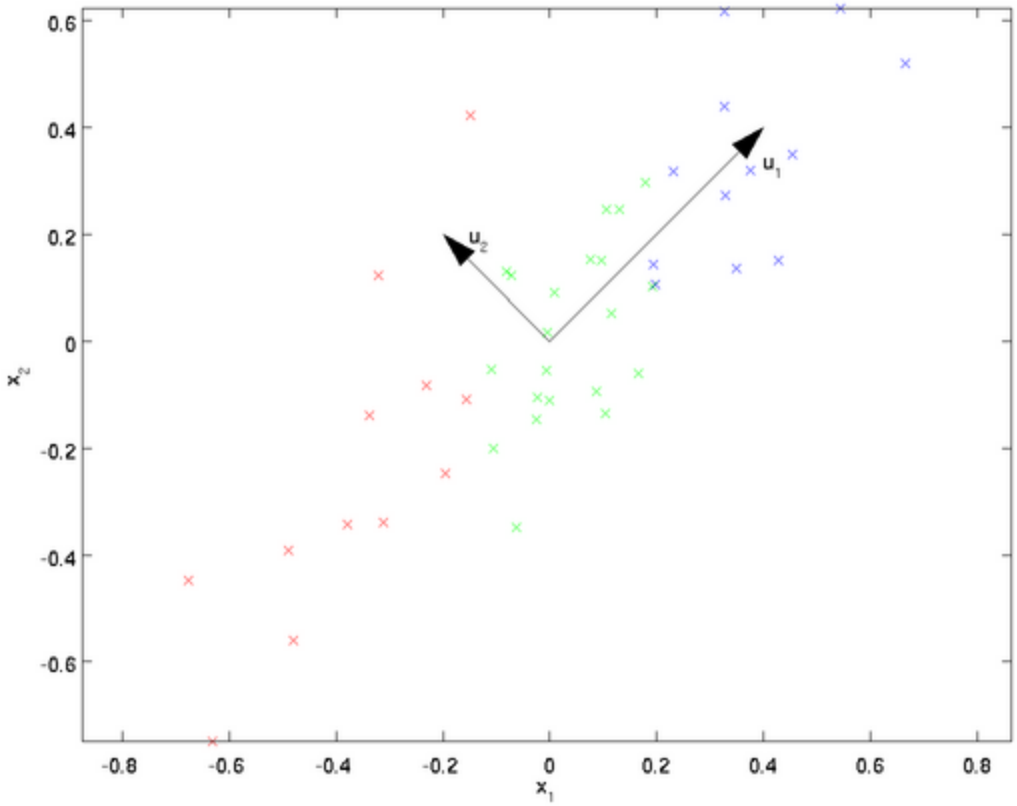
- 最近重构性：样本点到这个超平面的距离足够近
- 最大可分性：样本点在这个超平面上的投影能尽可能的分开

基于最近重构性和最大可分性能分别得到主成分分析的两种等价推到，我们这里主要考虑最大可分性，并且一步一步推到出最终PCA。

1.PCA最大可分性的思想

PCA顾名思义，就是找出数据里最主要的方面，用数据里最主要的方面来代替原始数据。具体的，假如我们的数据集是 n 维的，共有 m 个数据 $(x_1, x_2, \ldots, x_m)$。我们希望将这 m 个数据的维度从 n 维降到 n' 维，希望这 m 个 n' 维的数据集尽可能的代表原始数据集。我们知道数据从 n 维降到 n' 维肯定会有损失，但是我们希望损失尽可能的小。那么如何让这 n' 维的数据尽可能表示原来的数据呢？

我们先看看最简单的情况，也就是 $n=2$ ， $n'=1$,也就是将数据从二维降维到一维。数据如下图。我们希望找到某一个维度方向，它可以代表这两个维度的数据。图中列了两个向量方向， u_1 和 u_2 ，那么哪个向量可以更好的代表原始数据集呢？



从直观上也可以看出， u_1 比 u_2 好，这就是我们所说的**最大可分性。**

2.基变换

一般来说，欲获得原始数据新的表示空间，最简单的是对原始数据进行线性变换（基变换）：

$$Y = PX$$

其中 Y 是样本在新空间的表达， P 是基向量， X 是原始样本。我们可知选择不同的基可以对一组数据给出不同的表示，同时当基的数量少于原始样本本身的维数则可达到降维的效果，矩阵表示如下：

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

其中， $p_i \in \mathbb{R}^{1*N}$ 是一个行向量，表示第 i 个基； ， $a_i \in \mathbb{R}^{N*1}$ 是一个列向量，表示第 j 个原始数据记录。特别要注意的是，这里 R 可以小于 N ，而 R 决定了变换后数据的维数。也就是说，我们可以将一

$$p_i \in \{p_1, p_2, \ldots, p_R\}$$

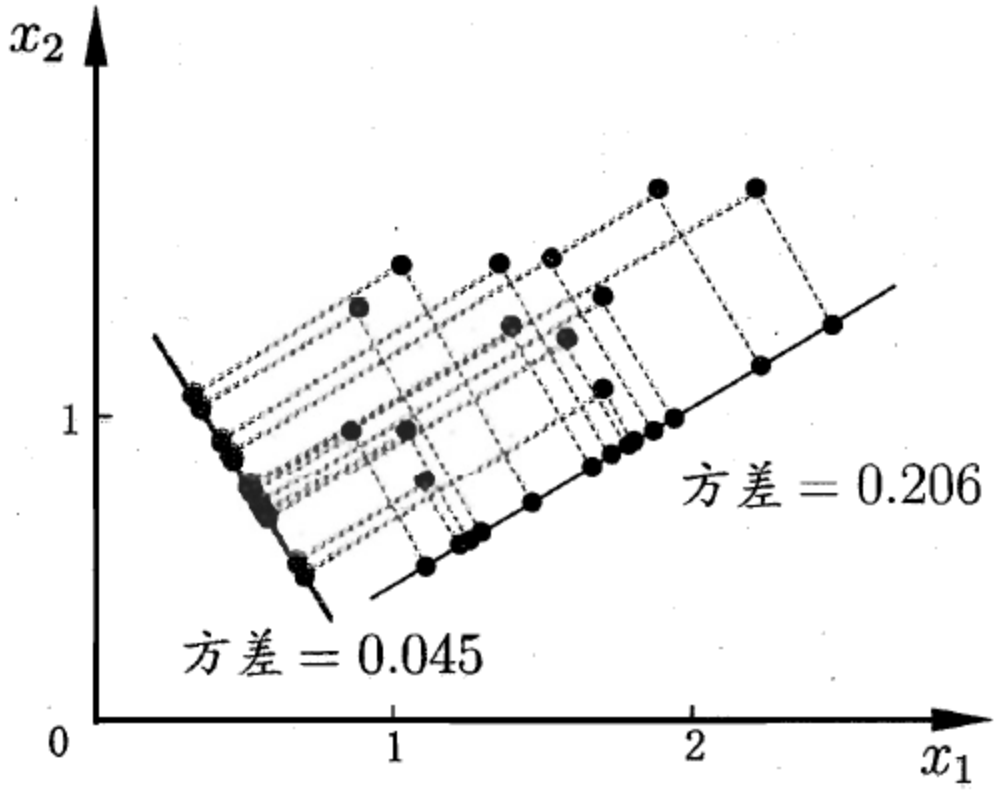
$$a_j \in \{a_1, a_2, \ldots a_M\}$$

个 N 维数据变换到更低维度的空间中去，变换后的维度取决于基的数量。**从原本** $X \in \mathbb{R}^{N*M}$ **降维到** $Y \in \mathbb{R}^{R*M}$ **。**因此这种矩阵相乘的表示也可以表示降维变换。

最后，上述分析同时给矩阵相乘找到了一种物理解释：两个矩阵相乘的意义是将右边矩阵中的每一列列向量变换到左边矩阵中每一行行向量为基所表示的空间中去。更抽象的说，一个矩阵可以表示一种线性变换。很多同学在学线性代数时对矩阵相乘的方法感到奇怪，但是如果明白了矩阵相乘的物理意义，其合理性就一目了然了。

3.方差

那么考虑，如何选择一个方向或 者基才是最优的？观察下图



周志华《机器学习》插图

我们将所有的点分别向两条直线做投影，基于前面PCA最大可分思想，我们要找的方向是降维后损失最小，可以理解为投影后的数据尽可能的分开，那么这种分散程度可以用数学上的方差来表示，方差越大数据越分散。方差公式如下：

对数据进行了中心化后（可以方便后面的操作）：

$$Var(a) = \frac{1}{m}\sum_{i=1}^{m}(a_i - \mu)^2$$

现在我们已经知道了以下几点：

- 对原始样本进行（线性变换）基变换可以对原始样本给出不同的表示
- 基的维度小于数据的维度可以起到降维的效果

- 对基变换后新的样本求其方差，选取使其方差最大的基

那么在下面我们来考虑一个新的问题

$$Var(a) = \frac{1}{m}\sum_{i=1}^{m} {a_i}^2$$

上面我们导出了优化目标，但是这个目标似乎不能直接作为操作指南（或者说算法），因为它只说要什么，但根本没有说怎么做。所以我们要继续在数学上研究计算方案。

4.协方差

从二维降到一维可以使用方差最大来选出能使基变换后数据分散最大的方向（基），但如果遇到高维的变换，当完成第一个方向（基）选择后，第二个投影方向应该与第一个“几乎重合在一起”，这显然是没有用的，因此要有其它的约束条件。我们希望两个字段尽可能表示更多的信息，使其不存在相关性。

数学上用协方差表示其相关性：

当 $Cov(a,b) = 0$ 时，表示两个字段完全独立，这也是我们的优化目标。

$$Cov(a,b) = \frac{1}{m}\sum_{i=1}^{m} a_i b_i$$

5.协方差矩阵

我们想达到的目标与字段内方差及字段间协方差有密切关系，假如只有 a 、 b 两个字段，那么我们将它们按行组成矩阵 X ,表示如下：

然后我们用 X 乘以 X 的转置，并乘上系数 $\frac{1}{m}$ ：

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

可见，协方差矩阵是一个对称的矩阵，而且对角线是各个维度的方差，而其它元素是 a 和 b 的协方差，然后会发现两者被统一到了一个矩阵的。

$$\frac{1}{m}XX^{\top} = \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m} a_i^2 & \frac{1}{m}\sum_{i=1}^{m} a_i b_i \\ \frac{1}{m}\sum_{i=1}^{m} a_i b_i & \frac{1}{m}\sum_{i=1}^{m} b_i^2 \end{pmatrix}$$

6.协方差矩阵对角化

我们的目标是使 ，根据上述推倒，可以看出我们的优化目标 $C = \frac{1}{m}XX^{\top}$

等价于协方差矩阵对角化。即除对角线外的其它元素（如 $\frac{1}{m}\sum_{i=1}^{m} a_i b_i$ ）化

$$\frac{1}{m}\sum_{i=1}^{m} a_i b_i = 0$$

为0，并且在对角线上将元素按大小从上到下排列，这样我们就达到了优化目的。这样说可能还不是很明晰，我们进一步看下原矩阵与基变换后矩阵协方差矩阵的关系：

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：

$$D=\frac{1}{m}YY^{\top}$$

$$=PCP^{\top}$$

$$=\frac{1}{m}(PX)(PX)^{\top}$$

$$=\frac{1}{m}PXX^{\top}P^{\top}$$

$$=P(\frac{1}{m}XX^{\top})P^{\top}$$

$$=P\begin{pmatrix}\frac{1}{m}\sum_{i=1}^{m}a_i^2 & \frac{1}{m}\sum_{i=1}^{m}a_ib_i\\ \frac{1}{m}\sum_{i=1}^{m}a_ib_i & \frac{1}{m}\sum_{i=1}^{m}b_i^2\end{pmatrix}P^{\top}$$

可见，我们要找的 P 不是别的，而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 $PCP^{\top}$ 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。

我们希望的是投影后的方差最大化，于是我们的优化目标可以写为：

利用拉格朗日函数可以得到:

对 P 求导有 ，整理下即为：

于是，只需对协方差矩阵 C 进行特征分解，对求得的特征值进行排序，再对 取前 K 列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。

$$\max_{P}\ tr(PCP^{\top})$$

$$s.t.\ \ PP^{\top}=I$$

$$J(P)=tr(PCP^{\top})+\lambda(PP^{\top}-I)$$

$$CP^{\top}+\lambda P^{\top}=0$$

$$CP^{\top}=\ (-\lambda)\ P^{\top}$$

$$P^{\top}=\ (P_1,P_2,...,P_R)$$

7.PCA算法流程

从上面两节我们可以看出，求样本 x_i 的 n' 维的主成分其实就是求样本集的协方差矩阵 $\frac{1}{m}XX^{\top}$ 的前 n' 个特征值对应特征向量矩阵 P ，然后对于每个样本 x_i ,做如下变换 $y_i=Px_i$ ，即达到降维的PCA目的。

下面我们看看具体的算法流程：

输入： n 维样本集 ，要降维到的维数 n' .

$$X=(x_1,x_2,...,x_m)$$

输出：降维后的样本集 Y

1.对所有的样本进行中心化

$$x_i=x_i-\frac{1}{m}\sum_{j=1}^{m}x_j$$

2.计算样本的协方差矩阵 $C=\frac{1}{m}XX^{\mathsf{T}}$

3.求出协方差矩阵的特征值及对应的特征向量

4.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

5.Y=PX即为降维到k维后的数据

注意：

有时候，我们不指定降维后的 n' 的值，而是换种方式，指定一个降维到的主成分比重阈值 t 。这个阈值t在 $(0,1]$ 之间。假如我们的 n 个特征值为 ,则n'可以通过下式得到:

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$$

$$\frac{\sum_{i=1}^{n'}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \geq t$$

8.PCA算法总结

这里对PCA算法做一个总结。作为一个非监督学习的降维方法，它只需要特征值分解，就可以对数据进行压缩，去噪。因此在实际场景应用很广泛。为了克服PCA的一些缺点，出现了很多PCA的变种，比如为解决非线性降维的KPCA，还有解决内存限制的增量PCA方法Incremental PCA，以及解决稀疏数据降维的PCA方法Sparse PCA等。

PCA算法的主要优点有：

- 仅仅需要以方差衡量信息量，不受数据集以外的因素影响。
- 各主成分之间正交，可消除原始数据成分间的相互影响的因素。
- 计算方法简单，主要运算是特征值分解，易于实现。

PCA算法的主要缺点有：

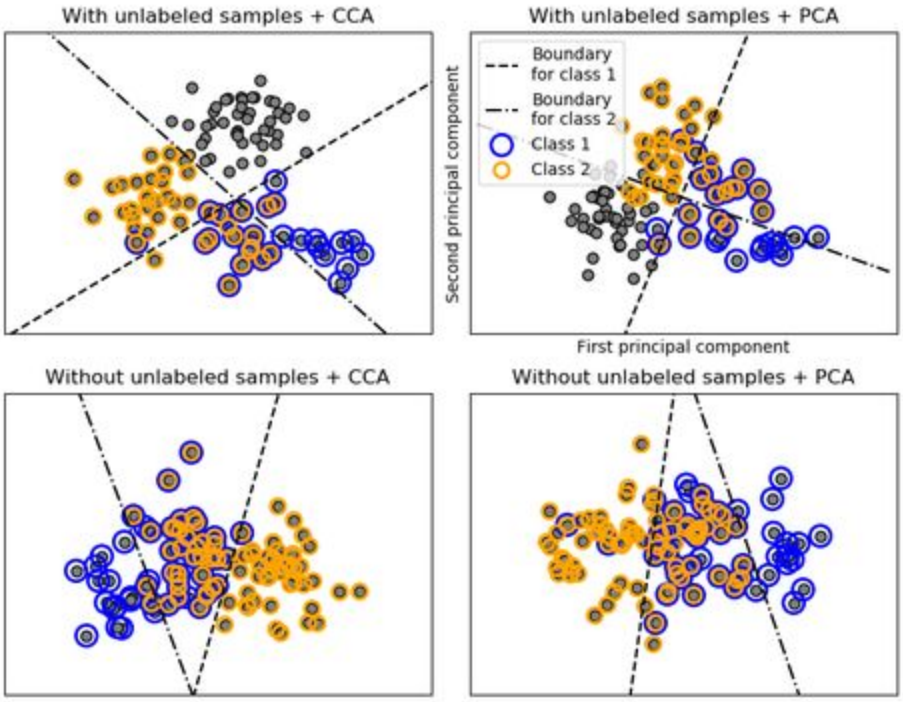
- 主成分各个特征维度的含义具有一定的模糊性，不如原始样本特征的解释性强。
- 方差小的非主成分也可能含有对样本差异的重要信息，因降维丢弃可能对后续数据处理有影响。

继续阅读

在PCA降维过程中，当进行协方差矩阵上求解特征值时，如果面对维度高达 $10000 * 10000$ ，可想而知耗费的计算量程平方级增长。面对这样一个难点，从而引出奇异值分解(SVD)，利用SVD不仅可以解出PCA的解，而且无需大的计算量。

Betten：奇异值分解(SVD)原理111 赞同 · 13 评论文章

PCA（主成分分析）和LDA（线性判别分析）有很多的相似点，其本质是要将初始样本映射到维度更低的样本空间中，但是PCA和LDA的映射目标不一样：PCA是为了让映射后的样本具有最大的发散性；而LDA是为了让映射后的样本有最好的分类性能。所以说PCA是一种无监督的降维方法，而LDA是一种有监督的降维方法。



Betten：LDA线性判别分析196 赞同 · 29 评论文章

参考资料

Pattern Recognition and Machine Learning

《机器学习》

主成分分析（Principal components analysis）-最大方差解释

Betten：机器学习面试干货精讲882 赞同 · 45 评论文章