2.0.PCA(Principal Component Analysis)

1.PCA的思想

PCA的主要思想是将n维特征映射到k维上，这k维是全新的**正交特征**也被称为主成分，**是在原有n维特征的基础上重新构造出来的k维特征。**

1.1PCA做法

通过计算数据矩阵的协方差矩阵，然后得到**协方差矩阵的特征值特征向量**，选择**特征值最大**(即方差最大)的k个特征所对应的特征向量组成的矩阵。

这样就可以将数据矩阵转换到新的空间当中，实现数据特征的降维。

1.2做法解释

PCA的目标：

1）使得保留下来的 维度间 的相关性尽可能小。

2）使得保留下来的维度 含有尽可能多的原始信息（方差大）

所以，要知道各 维度间 的相关性 以及 各维度上的方差。那这个时候就想到了协方差矩阵。

协方差矩阵的含义：

主对角线上的元素是各个维度上的方差。

其他元素是两两维度间的协方差（即相关性）。

两者结合：

PCA目标的第一条反应到协方差矩阵中就是，使协方差矩阵中的非对角元素基本为0。

现在的目标是：Y= PX，

找到一个P，使Y的协方差矩阵（YY^T)变成一个对角矩阵。

$$\begin{aligned} D &= \frac{1}{m} Y Y^{\top} \\ &= \frac{1}{m}(PX)(PX)^{\top} \\ &= \frac{1}{m} P X X^{\top} P^{\top} \end{aligned}$$

$$
\begin{aligned}
&= P(\frac{1}{m}XX^\intercal)P^\intercal \\
&= PCP^\intercal
\end{aligned}
$$

我们要找的P不是别的，**而是能让原始协方差矩阵（XX^T）对角化的P。**所以后面就是对X的协方差矩阵分解即可得到最优的P，完成降维。

2.协方差矩阵

样本均值:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{N} x_i$$

样本方差:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

样本X和样本Y的协方差:

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

由上面的公式，我们可以得到以下结论：

(1) **方差的计算公式是针对一维特征**，即针对同一特征不同样本的取值来进行计算得到；

而协方差则必须要求至少满足二维特征；**方差是协方差的特殊情况**。**Cov(X,X)就是X的方差。**

(2) 方差和协方差的除数是n-1,这是为了得到方差和协方差的无偏估计。

协方差为正时，说明X和Y是正相关关系；协方差为负时，说明X和Y是负相关关系；协方差为0时，说明X和Y是相互独立。

当样本是n维数据时，它们的协方差实际上是协方差矩阵(对称方阵)。例如，对于3维数据(x,y,z)，计算它的协方差就是：

$$Cov(X,Y,Z)=\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

3.PCA算法两种实现方法

3.1基于特征值分解协方差矩阵实现PCA算法

3.1.1过程（记忆这个）

输入：数据集 $X=\{x_1,x_2,x_3,\ldots,x_n\}$，需要降到k维。

1) 去平均值(即去中心化)，即每一位特征减去各自的平均值。、

2) 计算协方差矩阵 $\frac{1}{n}XX^T$,注：这里除或不除样本数量n或n-1,其实对求出的特征向量没有影响。

3) 用特征值分解方法求协方差矩阵 $\frac{1}{n}XX^T$ 的特征值与特征向量。

4) 对特征值从大到小排序，选择其中最大的k个。然后将其对应的k个特征向量分别作为行向量组成特征向量矩阵P。

5) 将数据转换到k个特征向量构建的新空间中，即Y=PX。

3.1.2实例

以X为例，我们用PCA方法将这两行数据降到一行。

$$X=\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

解：

1) 因为X矩阵的每行已经是零均值，所以不需要去平均值。

2) 求协方差矩阵：

$$C = \frac{1}{5}\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{6}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} \end{pmatrix}$$

3)求协方差矩阵的特征值与特征向量。

求解后的特征值为：

$$\lambda_1 = 2, \quad \lambda_2 = \frac{2}{5}$$

对应的特征向量为：

$$c_1\begin{pmatrix} 1 \\ 1 \end{pmatrix}, c_2\begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

其中对应的特征向量分别是一个通解，C1和C2可以取任意实数。那么标准化后的特征向量为：

$$\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

4)矩阵P为：

$$P = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

5)最后我们用P的第一行乘以数据矩阵X，就得到了降维后的表示：

$$Y = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{3}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \frac{3}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

结果如图1所示：

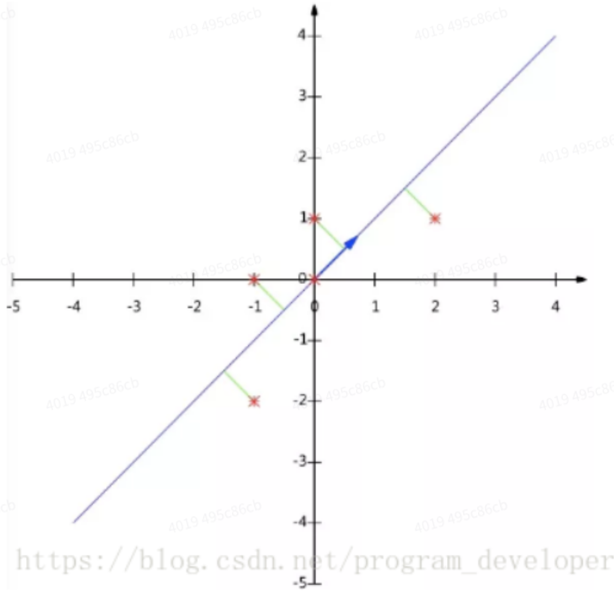

图1：数据矩阵X降维投影结果

4.选择降维后的维度K(主成分的个数)

如何选择主成分个数K呢？先来定义两个概念：

- average squared projection error：$\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}-x_{approx}^{(i)}\right\|^2$，其中$x_{approx}^{(i)}$为映射值。
- total variation in the data：$\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}\right\|^2$

选择不同的K值，然后用下面的式子不断计算，选取能够满足下列式子条件的最小K值即可。

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}-x_{approx}^{(i)}\right\|^2}{\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}\right\|^2}\leq t$$

其中t值可以由自己定，比如t值取0.01，则代表了该PCA算法保留了99%的主要信息。当你觉得误差需要更小，你可以把t值设置的更小。