# 主成分分析法

# 1 概述

主成分分析方法（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。PCA的工作就是从原始的空间中顺序地找一组相互正交的坐标轴，新的坐标轴的选择与数据本身是密切相关的。其中，第一个新坐标轴选择是原始数据中方差最大的方向，第二个新坐标轴选取是与第一个坐标轴正交的平面中使得方差最大的，第三个轴是与第1,2个轴正交的平面中方差最大的。依次类推，可以得到n个这样的坐标轴。通过这种方式获得的新的坐标轴，大部分方差都包含在前面k个坐标轴中，后面的坐标轴所含的方差几乎为0。于是，可以忽略余下的坐标轴，只保留前面k个含有绝大部分方差的坐标轴。事实上，这相当于只保留包含绝大部分方差的维度特征，而忽略包含方差几乎为0的特征维度，实现对数据特征的降维处理。

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。由于得到协方差矩阵的特征值特征向量有两种方法：特征值分解协方差矩阵、奇异值分解协方差矩阵，所以PCA算法有两种实现方法：基于特征值分解协方差矩阵实现PCA算法、基于SVD分解协方差矩阵实现PCA算法。

# 2 两种实现步骤

## 2.1基于特征值分解协方差矩阵

输入：数据集，其中n代表样本数量，m代表指标个数，需要将样本数量降到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。
3. 用特征值分解方法求协方差矩阵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。

### 2.1.1 特征值分解矩阵

(1) 特征值与特征向量

如果一个向量v是矩阵A的特征向量，将一定可以表示成下面的形式：

其中，λ是特征向量v对应的特征值，一个矩阵的一组特征向量是一组正交向量。

(2) 特征值分解矩阵

对于矩阵A，有一组特征向量v，将这组向量进行正交化单位化，就能得到一组正交单位向量。特征值分解，就是将矩阵A分解为如下式：

其中，Q是矩阵A的特征向量组成的矩阵，则是一个对角阵，对角线上的元素就是特征值。

## 2.2 基于SVD分解协方差矩阵实现PCA算法

输入：数据集，其中n代表样本数量，m代表指标个数，需要将样本数量降到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。

（3）通过SVD计算协方差矩阵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列向量组成特征向量矩阵。

（5）将数据转换到k个特征向量构建的新空间中。

### 2.2.1 SVD分解矩阵原理

奇异值分解是一个能适用于任意矩阵的一种分解的方法，对于任意矩阵A总是存在一个奇异值分解：

假设A是一个m\*n的矩阵，那么得到的U是一个m\*m的方阵，U里面的正交向量被称为左奇异向量。是一个m\*n的矩阵，除了对角线其它元素都为0，对角线上的元素称为奇异值。是v的转置矩阵，是一个n\*n的矩阵，它里面的正交向量被称为右奇异值向量。而且一般来讲，我们会将Σ上的值按从大到小的顺序排列。

SVD分解矩阵A的步骤：

求的特征值和特征向量，用单位化的特征向量构成 U。

(2) 求的特征值和特征向量，用单位化的特征向量构成 V。

(3) 将或者的特征值求平方根，然后构成 Σ。

# 3 案例（基于特征值求解）

以X为例，我们用PCA方法将这两行数据降到一行。

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：

3)求协方差矩阵的特征值与特征向量：

求解后的特征值为：，；

特征向量为：，。

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

，

矩阵P为：

用P的第一行乘以数据矩阵X，就得到了降维后的表示：