PCA (Principal Component Analysis)

PCA(Principal Component Analysis)，PCA的作用主要是降低数据集的维度，然后挑选出主要的特征。

PCA的主要思想是移动坐标轴，**找到方差最大的方向上的特征值**

在用统计分析方法研究这个多变量的课题时，变量个数太多就会增加课题的复杂性。人们自然希望变量个数较少而得到的信息较多。在很多情形，变量之间是有一定的相关关系的，当两个变量之间有一定相关关系时，可以解释为这两个变量反映此课题的信息有一定的重叠。主成分分析是对于原先提出的所有变量，建立尽可能少的新变量，使得这些新变量是两两不相关的，而且这些新变量在反映课题的信息方面尽可能保持原有的信息。

设法将原来变量重新组合成一组新的互相无关的几个综合变量，同时根据实际需要从中可以取出几个较少的综合变量尽可能多地反映原来变量的信息的统计方法叫做主成分分析或称主分量分析，也是数学上处理降维的一种方法。

主成分分析主要步骤如下：

1. 指标数据标准化；
2. 指标之间的相关性判定；
3. 计算特征值与特征向量
4. 计算主成分贡献率及累计贡献率
5. 计算主成分载荷

基本步骤：

**（1）首先去除平均值，计算数据集的协方差矩阵**

**（2）计算协方差矩阵的特征值和特征向量**

**（3）保留最重要的n个特征：【1.将特征值从大到小排序；2.保留最上面的N个特征向量，3.强数据转换到上述N各特征向量构成的新空间中】**

定义是变量向量减去均值向量，然后乘以变量向量减去均值向量的***[转置](http://zhidao.baidu.com/search?word=%E8%BD%AC%E7%BD%AE&fr=qb_search_exp&ie=utf8" \t "_blank)***再求均值。例如x是变量，μ是均值，***[协方差矩阵](http://zhidao.baidu.com/search?word=%E5%8D%8F%E6%96%B9%E5%B7%AE%E7%9F%A9%E9%98%B5&fr=qb_search_exp&ie=utf8" \t "_blank)***等于E[(x-μ)(x-μ)^t]，***[物理意义](http://zhidao.baidu.com/search?word=%E7%89%A9%E7%90%86%E6%84%8F%E4%B9%89&fr=qb_search_exp&ie=utf8" \t "_blank)***是这样的，例如x=（x1,x2,...,xi）那么***[协方差矩阵](http://zhidao.baidu.com/search?word=%E5%8D%8F%E6%96%B9%E5%B7%AE%E7%9F%A9%E9%98%B5&fr=qb_search_exp&ie=utf8" \t "_blank)***的第m行n列的数为xm与xn的***[协方差](http://zhidao.baidu.com/search?word=%E5%8D%8F%E6%96%B9%E5%B7%AE&fr=qb_search_exp&ie=utf8" \t "_blank)***，若m=n，则是xn的***[方差](http://zhidao.baidu.com/search?word=%E6%96%B9%E5%B7%AE&fr=qb_search_exp&ie=utf8" \t "_blank)***。如果x的元素之间是独立的，那么***[协方差矩阵](http://zhidao.baidu.com/search?word=%E5%8D%8F%E6%96%B9%E5%B7%AE%E7%9F%A9%E9%98%B5&fr=qb_search_exp&ie=utf8" \t "_blank)***只有对角线是有值，因为x独立的话对于m≠n的情况xm与xn的[***协方差***](http://zhidao.baidu.com/search?word=%E5%8D%8F%E6%96%B9%E5%B7%AE&fr=qb_search_exp&ie=utf8)为0。另外[***协方差***](http://zhidao.baidu.com/search?word=%E5%8D%8F%E6%96%B9%E5%B7%AE&fr=qb_search_exp&ie=utf8)矩阵是对称的。