17.5.5（PCA详解）

主成分分析（PCA）原理详解 ：http://blog.jobbole.com/109015/

3. 数据降维

为了说明什么是数据的主成分，先从数据降维说起。数据降维是怎么回事儿？假设三维空间中有一系列点，这些点分布在一个过原点的斜面上，如果你用自然坐标系x,y,z这三个轴来表示这组数据的话，需要使用三个维度，而事实上，这些点的分布仅仅是在一个二维的平面上，那么，问题出在哪里？如果你再仔细想想，能不能把x,y,z坐标系旋转一下，使数据所在平面与x,y平面重合？这就对了！如果把旋转后的坐标系记为x′ ,y′ ,z′ ，那么这组数据的表示只用x′ 和y′ 两个维度表示即可！当然了，如果想恢复原来的表示方式，那就得把这两个坐标之间的变换矩阵存下来。这样就能把数据维度降下来了！但是，我们要看到这个过程的本质，如果把这些数据按行或者按列排成一个矩阵，那么这个矩阵的秩就是2！这些数据之间是有相关性的，这些数据构成的过原点的向量的最大线性无关组包含2个向量，这就是为什么一开始就假设平面过原点的原因！那么如果平面不过原点呢？这就是数据中心化的缘故！将坐标原点平移到数据中心，这样原本不相关的数据在这个新坐标系中就有相关性了！有趣的是，三点一定共面，也就是说三维空间中任意三点中心化后都是线性相关的，一般来讲n维空间中的n个点一定能在一个n-1维子空间中分析！

上一段文字中，认为把数据降维后并没有丢弃任何东西，因为这些数据在平面以外的第三个维度的分量都为0。现在，假设这些数据在z′ 轴有一个很小的抖动，那么我们仍然用上述的二维表示这些数据，理由是我们可以认为这两个轴的信息是数据的主成分，而这些信息对于我们的分析已经足够了，z′ 轴上的抖动很有可能是噪声，也就是说本来这组数据是有相关性的，噪声的引入，导致了数据不完全相关，但是，这些数据在z′ 轴上的分布与原点构成的夹角非常小，也就是说在z′ 轴上有很大的相关性，综合这些考虑，就可以认为数据在x′ ，y′ 轴上的投影构成了数据的主成分！

PCA的思想是将n维特征映射到k维上（k<n），这k维是全新的正交特征。这k维特征称为主成分，是重新构造出来的k维特征，而不是简单地从n维特征中去除其余n-k维特征。

协方差是用来表示两个随机变量之间关系的值。

二、为什么需要协方差

标准差和方差一般是用来描述一维数据的，但现实生活中我们常常会遇到含有多维数据的数据集，最简单的是大家上学时免不了要统计多个学科的考试成绩。面对这样的数据集，我们当然可以按照每一维独立的计算其方差，但是通常我们还想了解更多，比如，一个男孩子的猥琐程度跟他受女孩子的欢迎程度是否存在一些联系。协方差就是这样一种用来度量两个随机变量关系的统计量，我们可以仿照方差的定义：

$$\mathrm{var}(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{n-1}$$

来度量各个维度偏离其均值的程度，协方差可以这样来定义：

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

协方差的结果有什么意义呢？如果结果为正值，则说明两者是正相关的（从协方差可以引出"相关系数"的定义），也就是说一个人越猥琐越受女孩欢迎。如果结果为负值， 就说明两者是负相关，越猥琐女孩子越讨厌。如果为0，则两者之间没有关系，猥琐不猥琐和女孩子喜不喜欢之间没有关联，就是统计上说的"相互独立"。

在统计学与概率论中，协方差矩阵的每个元素是各个向量元素之间的协方差。是从标量随机变量到高维度随机向量的自然推广。引自《百度百科》

注：浅谈协方差矩阵，http://pinkyjie.com/2010/08/31/covariance/

协方差矩阵

SVD奇异值分解函数

奇异值：X是一个m*n的矩阵，X ‘*X是n*n的矩阵，是n个特征值的非负平方根，也就是X的奇异值。

返回值U 就是数据最小投射误差方向向量的矩阵，我们降到K维，就取前K列。

pca 算法就是求出怎么降维使原数据到投射平面的投影误差最小。（减少投射的平均均方误差）

PCA算法中怎么选择降到的最好维度。

svd函数求出来的三个值，第二个S 是一个n*n的矩阵，这个矩阵除了对角线上其余都为0 ，我们要使前K个对角线值的和 / 所有的对角线值的和 能大于等于0.99，就得出降维的最好维度K 。

其中的 S 是一个 n×n 的矩阵，只有对角线上有值，而其它单元都是 0，我们可以使用这个矩阵来计算平均均方误差与训练集方差的比例：

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}-x_{approx}^{(i)}\|^2}{\frac{1}{m}\sum_{i=1}^{m}\|x^{(i)}\|^2} = 1 - \frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{m} S_{ii}} \le 1\%$$

也就是：

$$\frac{\sum_{i=1}^{k} s_{ii}}{\sum_{i=1}^{n} s_{ii}} \ge 0.99$$

x 为 2 维，z 为 1 维，$z = U_{reduce}^{T} x$，相反的方程为：$x_{appox} = U_{reduce} \bullet z$ 。

利用PCA算法求出来的模型，进行预测的时候要把测试集进行PCA降维。

另一个常见的错误是，默认地将主要成分分析作为学习过程中的一部分，这虽然很多时候有效果，最好还是从所有原始特征开始，只在有必要的时候（算法运行太慢或者占用太多内存）才考虑采用主要成分分析。