PCA主成分分析学习

A tutorial on principal componet analysis这篇文章来自Google的研究员，他举了一个很形象的例子来说明PCA要解决的问题：假设目前有一个弹簧球沿着某条直线在运动，那么沿着该条直线的方向，我们可以给出一个非常简洁的运动方程，但是此时我们用来记录弹簧球运动的摄像头并不沿着这条直线的方向，与其同时，记录的位置信息还包括噪声，那么我们如何从这些数据中找到那个最简洁的运动方程呢，答案是通过PCA。PCA说白了很简单，就是对坐标基底进行线性变换， 换一个更好表示数据的方式。

$$PX = Y$$

很明显，这是一个优化问题，优化的目标是：1、重新表示X的最佳方式是什么？2、P作为基底，什么样的选择是好的？

对于运动方程这个二维的例子来讲，就是寻找一个方向使得信噪比（signal-to-noise ratio）最大：

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

噪声信号的方差应当较小，有意义信号的方差应当较大。

PCA就是一种降维方法，去掉数据中具有强烈相关性的部分，使得各类数据尽可能不相关，降维后再升维，重现误差最小，这就是优化目标。

$$A = \{a_1, a_2, \ldots, a_n\} \quad B = \{b_1, b_2, \ldots, b_n\}$$
$$\sigma^2_A = \frac{1}{n}\sum_i a_i^2 \quad \sigma^2_B = \frac{1}{n}\sum_i b_i^2$$
$$\sigma^2_{AB} = \frac{1}{n}\sum_i a_i b_i$$

但是在实际中σ^2_{AB}的系数为n-1，而不是n，这是由于均值和实际均值有差距的缘故。

可以将方差和协方差的计算表达为矩阵的形式：

$$\sigma^2_{ab} = \frac{1}{n} ab^T$$

对于矩阵形式的数据：

$$X = [x_1, \ldots, x_m]^T$$
$$C_X = \frac{1}{n} XX^T$$

C_X首先是一个对称方阵，对角线上的元素都是方差，其余位置都是协方差。协方差反映了数据中的冗余和噪声（即两个系数非常具有相关性，那么更适合用一个系数来表达，这也是降维的本质）。在对角项上，较大的数值表示较大的方差，同时也代表着这个测量值比较重要，在非对角项上，较大的数值代表较大的相关性，即冗余性质。

因此降维的目的就是：1、最小化冗余性；2、最大化信号。那么理想化的协方差矩阵Y应当是非对角项均是0，因此必须是个对角阵，而Y的对角项应当从大到小排列。

注意新的基底仍然是正交的。

让我们看看PCA做了哪些假设：

1、线性变换的基底（现在有的工作会讨论非线性变换的基底）；

2、大的方差具有重要的结构（这一点有的时候是不正确的）；

3、主成分分析后得到的基底仍然是正交的。

值得注意的是$C_Y = \frac{1}{n}YY^T$应当是一个对角阵

$$
\begin{aligned}
C_Y &= \frac{1}{n}YY^T \\
&= \frac{1}{n}(PX)(PX)^T \\
&= \frac{1}{n}PXX^TP^T \\
&= P(\frac{1}{n}XX^T)P^T \\
C_Y &= PC_XP^T
\end{aligned}
$$

这里C_X是一个对称矩阵，对称阵通过特征值分解，得到的特征向量是正交的，$A = EDE^T$，D是一个对角阵，而E各列均是特征向量。

$$
\begin{aligned}
C_Y &= PC_XP^T \\
&= P(E^TDE)P^T \\
&= P(P^TDP)P^T \\
&= D
\end{aligned}
$$

一个最优化问题，就被转换为一个矩阵分解的问题，因此进行PCA的步骤也是：

1、减去均值；

2、计算C_X的特征向量；

永远记住PCA的核心是基底的线性变化，无他。

性质：

a matrix is symmertric if and only if it is orthogonally diagonalizable.

这需要双向证明，先证明如果是可正交分解的，那么矩阵是对称的。

$$
\begin{aligned}
A &= EDE^T \\
A^T &= (EDE^T)^T = EDE^T = A
\end{aligned}
$$

反方向证明待续。

A symmertric matrix is diagonalized by a matrix of its orthonormal eigenvectors.

$$AE = ED$$
$$AE = [Ae_1 Ae_2 \dots Ae_n]$$
$$ED = [\lambda_1 e_1 \lambda_2 e_2 \dots \lambda_n e_n]$$
$$A = EDE^{-1}$$

而这正是特征向量和特征值的定义。

接着证明对称矩阵的性质：所有特征向量不仅线性无关，而且正交。

$$\begin{aligned}\lambda_1 e_1 \cdot e_2 &= (\lambda_1 e_1)^T e_2 \\ &= (Ae_1)^T e_2 \\ &= e_1^T A^T e_2 \\ &= e_1^T A e_2 \\ &= e_1^T (\lambda_2 e_2) \\ \lambda_1 e_1 \cdot e_2 &= \lambda_2 e_1 \cdot e_2\end{aligned}$$

因此对称矩阵的不同特征值对应的特征向量是正交的，即$e_1 \cdot e_2$

因此正交阵的性质为$E^T = E^{-1}$

因此$A = EDE^T$

For any arbitrary m*n matrix X, the symmetric matrix $X^T X$ has a set of orthonormal eigenvectors $\{v_1, v_2, \dots, v_n\}$ and a set of associated eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$. The set of vectors $\{Xv_1, Xv_2, \dots, Xv_n\}$ then form an orthogonal basis, where each vector Xv_i is of length $\sqrt{\lambda_i}$.

$$\begin{aligned}(Xv_i) \cdot (Xv_j) &= (Xv_i)^T (Xv_j) \\ &= v_i^T X^T X v_j \\ &= v_i^T (\lambda_j v_j) \\ &= \lambda_j v_i \cdot v_j\end{aligned}$$

因为v是正交的，因此：

$$(Xv_i) \cdot (Xv_j) = \begin{cases} \lambda_j & i = j \\ 0 & i \neq j \end{cases}$$

因此得证。

这个证明引出了一个很棒的结果，我们将所有的特征值放在对角线组成一个主对角阵，由于V和U只包含r（矩阵的秩）个特征向量，我们再补齐m和n个正交向量基底，补充的基底对应特征值为0，所以事实上是无用的，就可以得到SVD分解：

$$Xv_i = \sigma_i u_i$$
$$V = [v_1, v_2, \dots, v_m]$$
$$U = [u_1, u_2, \dots, u_n]$$
$$XV = U\Sigma$$
$$X = U\Sigma V^T$$

我们要注意，v就是X^TX的特征向量，假设我们定义Y：

$$Y = \frac{1}{\sqrt{n}}X^T$$

$$\begin{aligned} Y^TY &= (\frac{1}{\sqrt{n}}X^T)^T(\frac{1}{\sqrt{n}}X^T) \\ &= \frac{1}{n}XX^T \\ Y^TY &= C_X \end{aligned}$$

也就是说对Y做SVD分解，我们得到的是C_X的特征向量，而这正是我们在PCA中所需要的向量！因此SVD和PCA紧密地联系在了一起，SVD是对任何矩阵做分解，而PCA是针对数据X计算得到的对称矩阵X^TX做分解。

- Organize data as an m*n matrix, where m is the number of measurement types and n is the number of samples.
- Subtract off the mean for each measurement type.
- Calculate the SVD or the eigenvectors of the covariace.

最后PCA可以用来做降维，它是一种非参数化的方法，得到的结果是线性变化基底以及各基底相互正交。

它在各变量线性相关的情况下很有效，但是在非线性和有用的数据非正交的情况下无效。

非线性的情况，我们可以使用流型学习和kernel PCA解决。

非正交的情况，我们可以要求分解的结果必须统计无关（ICA），这样却使得计算量更大。

PCA可以消灭数据的冗余性质，那么SVD的实际意义又是什么呢？

首先SVD可以进行数据压缩，数据去噪。保留奇异值较大对应的特征向量部分，去掉奇异值较小对应的特征向量部分。

奇异值往往对应着矩阵中隐藏的重要信息，而且重要性和奇异值大小正相关。每个矩阵A都可以表示为一系列rank为1的小矩阵之和，而奇异值则衡量了这些小矩阵对于A的权重。

那么奇异值的几何意义又是什么？

$$Mv_1 = \sigma_1 u_1, Mv_2 = \sigma_2 u_2$$

v_1和$v2$本身就是正交的，在经过M变化后，生成的u_1和u_2仍然是正交的，而奇异值代表这个向量的长度。

假设矩阵A的奇异值分解为：

$$A = \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix}$$

假设：

$$x = \xi_1 v_1 + \xi_2 v_2$$
$$y = Ax = A[v_1 \quad v_2][\begin{matrix}\xi_1 \\ \xi_2\end{matrix}] = 3\xi_1 u_1 + \xi_2 u_2$$
$$\eta_1 = 3\xi_1, \eta_2 = \xi_2$$

如果x是在单位圆$\xi_1^2 + \xi_2^2 = 1$，那么y就会在椭圆$\frac{\eta_1^2}{3^2} + \frac{\eta_2^2}{1^2} = 1$上，这表明矩阵A将单位圆变成了椭圆，而椭圆的半轴长正好是对应的奇异值。

推广到一般情况：一个矩阵A将单位球变换为超椭球面，那么矩阵A的每一个奇异值恰好就是超椭球的每条半轴长度。