PCA原理

PCA算法步骤

1. 首先求样本数据各维度的协方差;
2. 然后对这个协方差矩阵进行奇异值分解（SVD），通过SVD可以得到这个协方差矩阵的特征值对角矩阵和对应的特征向量矩阵;
3. 最后根据特征值的大小，从n个特征值中选取前k个较大特征值所对应的k个特征向量构成特征向量矩阵（转移矩阵P(n*k)），将样本数据（A(n,1)）转换到新的空间下（$A' = P^T A$）。

这样就实现了对数据的降维。

接下来我一步一步来解释其中的原理:

预备知识

空间变换

矩阵的乘法的几何意义就是空间变换。

$Ma = b$代表向量a经过M的变换后变成了向量b。

例如：矩阵$\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$将向量$\begin{bmatrix} x \\ y \end{bmatrix}$变为$\begin{bmatrix} x' \\ y' \end{bmatrix}$

$$\vec{b} = \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 3x + 1y \\ 1x + 2y \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = Ma \ , a = x\vec{i} + y\vec{j} \rightarrow b = x'\vec{u} + y'\vec{v}$$

所以从上式看出，矩阵的空间变换是一种映射。

如果M不是方阵，会使空间维度发生改变。例如一个3×2的矩阵可以将一个二维平面的向量映射为三维空间中的一个平面上的向量。而一个2×3的矩阵可以将一个三维空间的向量映射为二维平面上的一个向量，考虑整个三维空间就是将原空间做了一个投影。

基变换

在二维平面中，选取两个相互正交的单位基向量$i, j \in R^2$（$i; j$也叫一组标准正交基）；

同样在此平面中，选取另外两个基向量$u, v \in R^2$；

选取的u, v这一组基在以i, j为基向量的二维平面上可以表示为：$u = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, v = \begin{bmatrix} -2 \\ -1 \end{bmatrix}$。

$$(u, v) = (i, j)P \ , P = \begin{bmatrix} 3 & -2 \\ 1 & -1 \end{bmatrix}$$

P被称为由基i, j到基u, v的过渡矩阵。
所以一个向量在基i, j中的坐标为(x, y)，在基u, v中的坐标为(x', y')，那么可以根据过渡矩阵P相互转换:

$$\begin{pmatrix} x \\ y \end{pmatrix} = P \begin{pmatrix} x' \\ y' \end{pmatrix} \ , \begin{pmatrix} x' \\ y' \end{pmatrix} = P^{-1} \begin{pmatrix} x \\ y \end{pmatrix}$$

对于不同基下的空间变换来说：

如果对于标准正交基i, j来说，有一个空间变换矩阵A，可以将标准正交基i, j表示的向量$\vec{a}$映射至向量$\vec{b}$（$\vec{a} \rightarrow \vec{b}$）。

那么对于另一组基u, v来说，相同的变换矩阵M如何表示呢？也就是如何找到一个矩阵M，将一个用基u, v表示的向量$\overrightarrow{a'}$转换至$\overrightarrow{b'}$呢？

可以这样想，将一个用基u, v表示的向量$\overrightarrow{a'}$转换至标准正交基i, j下变成$\vec{a}$:
$\vec{a} = P\overrightarrow{a'}$
然后再使用A进行变换：
$\vec{b} = A\vec{a} = AP\overrightarrow{a'}$
然后将使用矩阵A变换后的用基i, j表示的向量转换至基u, v下：
$M\overrightarrow{a'} = \overrightarrow{b'} = P^{-1}\vec{b} = P^{-1}AP\overrightarrow{a'} \Rightarrow M = P^{-1}AP$
这说明，**对于形如$M = P^{-1}AP$的式子，它表示在不同基向量下同一个空间变换如何相互转化**。

特征值和特征向量

对n阶矩阵A，如果数λ和n维非零列向量x使关系式
$Ax = \lambda x$
成立，这样的数λ称为矩阵A的特征值，非零向量A的对应于特征值λ的特征向量。

上式可以这样解读，**对于变换A的特征向量x，A变换相当于只有伸缩变换没有旋转变换**

特征值分解

特征分解又称矩阵对角化，但不是所有的矩阵都可以对角化，而谱定理描述了什么样的矩阵可以被对角化，如实对称矩阵。

如果一个$n \times n$的变换矩阵A的特征向量能够构成这个向量空间的一组基（这个条件其实也就是可以对角化的条件之一：A具有n个各不相同的特征值，也就是有n个线性无关的特征向量），那么可以将其特征分解（矩阵对角化）得到：
$A = VMV^{-1}$
其中M是对角矩阵，对角线上的元素为A的特征值；而矩阵V的每一列都是M中每一个特征值所对应的特征向量。

这个式子在基变换的角度可以理解为:
$Ax = VMV^{-1}x$
特征值分解将标准正交基下的向量x，变为了在V下表示的向量，然后应用相同的变化M，然后再变换至标准正交基下。

如果Σ是实对称矩阵，实对称矩阵特征分解后得到的特征向量矩阵V可以正交化。而对于一个正交矩阵V，$V^{-1} = V^T$，所以对于实对称矩阵Σ:
$\Sigma = VMV^T$

奇异值分解 SVD

奇异值分解其实就是广义的特征值分解，它可以对任意的$m \times n$矩阵进行分解。

- 定义

假设M是一个$m \times n$阶矩阵，其中元素全部属于实数域或复数域。如此则存在一个分解使得
$M = U\Sigma V^*$。

其中U是$m \times m$阶酉矩阵；Σ是$m \times n$阶非负实数对角矩阵；而V^*，即V的共轭转置，是$n \times n$阶酉矩阵；Σ对角线上的元素即为M的奇异值。

协方差矩阵

对于m个样本，每一个样本都有n个不同的特征$X_1, X_2, \cdots X_n$，每一个特征都是一个随机变量。
那么这n组特征的协方差矩阵为：

$$\Sigma = \begin{bmatrix} cov(X_1, X_1) & cov(X_1, X_2) & \cdots & cov(X_1, X_n) \\ cov(X_2, X_1) & cov(X_2, X_2) & \cdots & cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_n, X_1) & cov(X_n, X_2) & \cdots & cov(X_n, X_n) \end{bmatrix}$$

显然Σ是一个实对角矩阵。

两个关键问题

为什么PCA要对协方差矩阵进行SVD?

协方差矩阵度量的是维度与维度之间的关系，而非样本与样本之间的。协方差矩阵的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。

借助该文章中的说法，方差描述的是各个维度的能量（这个表达有待斟酌，但是便于理解），协方差描述各个维度间的相关性。

在PCA中我们希望能够将n维的数据降维至k，降维并不是直接在原来的空间中舍弃n维数据中的一部分，而是对数据进行空间变换找到另一个空间，能够以更少的维度来尽可能描述同样的数据。

协方差矩阵Σ描述的是各个维度的相关性，因为要最小化描述数据的维度，那么就希望各个维度间的相关性尽可能的小。也就是说，我们要找到这样一个的过渡矩阵V，将数据转换到另一个空间中，让另一个空间中的协方差矩阵M中非对角线上的元素都基本为零。

这就转换成了一个**不同基下的空间变换的问题**，找到过渡矩阵V使$M = V^{-1}\Sigma V$

而这样的方法不就是矩阵对角化（特征值分解）吗?

所以对Σ进行特征值分解或者是奇异值分解，最后都能得到
$\Sigma = VMV^{-1} = VMV^T$

这样通过特征值分解，我们就找到了一个新的空间和将原始数据转移至这个新的空间的过渡矩阵V，在这个新的空间中是通过另外n个维度对数据进行描述的。

选取特征（奇异）值大的特征（奇异）向量?

通过上一步，我们已经得到了特征向量矩阵V和特征值对角矩阵M。最后一步就是选取前k个较大特征值所对应的k个特征向量构成$n \times k$的转移矩阵P。

这一步毋庸置疑，PCA确实是这样做的。我们知道M是另一空间下的协方差矩阵，对角线上是各个维度的方差，那么选择方差大的维度，真的能够尽可能保留原始数据的信息吗?

方差是用于描述数据的离散程度的，而信息量则是通过信息熵来衡量的，信息熵实际上度量的是随机变量的不确定度。这两者虽然有一定的联系但并不等价。随机变量的取值可以很不确定但并不是非常离散，方差并不和信息熵呈正相关，**所以，方差越大信息量越多这个说法是不正确的。**

那为什么主成分分析（PCA）的过程是寻找能使方差最大的方向以此保持最大的信息量？在这种方法中，为什么就可以认为找到使降维后的数据样本方差最大的基底就使损失的信息最小化?

不妨想一下PCA降维的目的，就是为了降噪。除去和结果关系不大的特征，保留最具相关性的特征。但是这些数据是以什么概率分布产生的，我们并不知道，这里的信息熵就没有什么太大意义了，不能开上帝视角找到最大信息熵的方向，PCA方法本来就是用来“揣测”和“创造”数据之间的规律。至于我们怎样区分出什么是噪声，什么是主成分，就是出于这种揣测的思路找到离散程度最高的方向，而离散程度低的

方向更有可能是由于噪声的干扰表现出同一性，或者反过来说就是因为太同一所以没什么分析价值。因此我们把注意力放在离散程度高的成分上，因为它的多样性可以帮助我们分析数据间潜在的关系。

参考文章

1. 空间变换及基变换：https://zhuanlan.zhihu.com/p/69069042
2. 特征值分解和奇异值分解：https://zhuanlan.zhihu.com/p/69069183
3. 方差和信息量：https://www.zhihu.com/question/36481348