主成分分析PCA与奇异值分解SVD

0、

本文假设你有线性代数知识和概率论协方差矩阵知识

涉及变量全在实数域(复数域也适用，不过矩阵运算符会有变化，为了保证符号和概念不陌生所以限定在实数域)

1、PCA的解释方法

实际上有9到10种PCA的解释方法

(1)最大方差理论：选择一个子空间并在其上对数据进行投影并使得所有投影点之间尽可能稀疏，距离尽可能大，

(2)最小平方误差理论：选择一个子空间并在其上对数据进行投影并使得所有投影点到原始点的距离的平方和最小化

2、PCA数据预处理步骤

[1]预处理步骤

(1)求数据均值

(2)使数据均值为零

(3)求数据方差

(4)使数据方差为零

[2]预处理原因

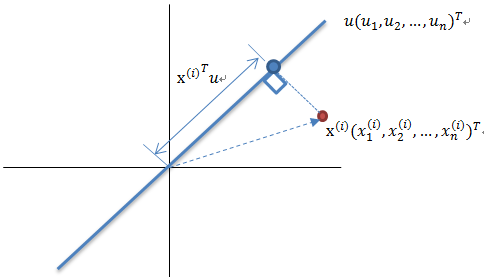
(1)数据均值为零，才可以保证对x拟合的轴u过零点；因为数据均值为0，可以直接使用x坐标当作本身的向量x，来计算x到u上的投影长度为 。

(2)数据方差为零，才能保证向量x各特征(元素)权重一致

3、最大方差理论解释

首先，考虑在一维空间 (k=1) 上的投影。我们可以使用 n 维向量u定义这个空间的方

向。为了方便(并且不失一般性)，我们假定选择一个单位向量，从而 (注意，我们只对u的方向感兴趣，而对 u本身的大小不感兴趣)。



所有向量x到轴u的投影长度平方和为：

注：计算投影长度的平方是因为，我们求投影长度绝对值之和，绝对值本身不便于计算，所以计算其平方的和

其中Σ是向量x的协方差矩阵，协方差矩阵是一个对称矩阵

注意到同时左乘u，

所以PCA实际上是一个特征值分解。

4、PCA算法原理

求出所有特征值并降序排列，最大特征值对应的特征向量被称为第一主成分。因此，我们只需要对协方差矩阵进行特征值分解，得到的前k大特征值对应的特征向量就是最佳的k维新特征，而且这k维新特征是正交的(对称矩阵不同特征值对应的特征向量必然正交)。得到前k个u以后，原始数据集x通过变换可以得到新的样本。

5、PCA的实际算法

在实际算法中我们使用的是奇异值分解而非特征值分解。

这是因为对于对称矩阵来说，奇异值等于特征值的绝对值；特征值大于零时，左奇异向量u=右奇异向量v=特征向量；特征值小于零时，左奇异向量u=负右奇异向量-v=特征向量

PCA仅仅使用了我们SVD的右奇异矩阵V，没有使用左奇异矩阵U。是因为左奇异矩阵U可以用于行数即样本数的压缩；右奇异矩阵V可以用于列数即特征维度的压缩，也就是我们的PCA降维。所以实际应用中只计算不用计算(注意)。

6、SVD 奇异值分解

奇异值分解（singular value decomposition）是线性代数中一种重要的矩阵分解。

其中k=min{m,n}, r=rank(A),k≥r,

A是m×n，U是m×m维的正交矩阵， V是n×n维的正交矩阵，Σ是m×n维的矩阵，除了主对角线上的元素以外全为0，主对角线上的每个元素都称为奇异值并由大到小排列。U由m个列向量构成，V由n个列向量构成。

[1]求解方法

假设A是m×n维的，U由m个列向量构成，V由n个列向量构成

根据性质

step1：首先求出和

step2：然后求出的特征值和特征向量以及的特征值和特征向量

step3：利用或者求出奇异值

step4：最后令，并相应调整U和V中列向量排序

7、SVD和PCA的关系

对矩阵A奇异值分解后，由性质，每个矩阵都可以表示为一系列秩为1的“小矩阵”之和，而奇异值则衡量了这些“小矩阵”对于矩阵A的权重。而且奇异值的减少特别的快，在很多情况下，前10%甚至1%的奇异值的和就占了全部的奇异值之和的99%以上的比例。也就是说，我们也可以用最大的k个的奇异值和对应的左右奇异向量来近似描述矩阵，故可以用于PCA降维。

8、奇异值和特征值的关系

[1]数学解释

(1)奇异值分解适用与矩阵，而特征值分解只适用于方阵

矩阵的非零奇异值是或的非零特征值的算术平方根，即

(2)对称方阵的奇异值和特征值

对于对称矩阵来说，奇异值等于特征值的绝对值；特征值大于零时，左奇异向量u=右奇异向量v=特征向量；特征值小于零时，左奇异向量u=负右奇异向量-v=特征向量

证明(unfinished)：假设A为n阶对称矩阵，

[2]几何解释

线性变换y=Ax可以分解为旋转、缩放、投影这三种基本线性变换。

(1)特征值分解其实是对旋转、缩放两种效应的归并(有投影效应的矩阵不是方阵，没有特征值)。特征向量由得到，它表示如果一个向量v处于A的特征向量方向，那么Av对v的线性变换作用只是一个缩放。也就是说，求特征向量和特征值的过程，我们找到了这样一组基，在这组基下，矩阵的作用效果仅仅是存粹的缩放。

(2)特征值分解和奇异值分解都是给一个矩阵(线性变换)找一组特殊的基，特征值分解找到了特征向量这组基，在这组基下该线性变换只有缩放效果。而奇异值分解则是找到另一组基，这组基下线性变换的旋转、缩放、投影三种功能独立地展示出来了。

9、PCA特性

PCA本质上是将方差最大的方向(即数据最离散/稀疏的方向)作为主要特征，并且在各个正交方向上将数据“离相关”，也就是让它们在不同正交方向上没有相关性。

[1]优点：

(1)PCA也存在一些限制，例如它可以很好的解除线性相关，但是对于高阶相关性就没有办法了，对于存在高阶相关性的数据，可以考虑Kernel PCA，通过Kernel函数将非线性相关转为线性相关

(2)最后需要说明的是，PCA是一种无参数技术，也就是说面对同样的数据，如果不考虑清洗，谁来做结果都一样，没有主观参数的介入，所以PCA便于通用实现，但是本身无法个性化的优化。

[2]缺点：

(1)如果用户对观测对象有一定的先验知识，掌握了数据的一些特征，却无法通过参数化等方法对处理过程进行干预，可能会得不到预期的效果，效率也不高。

(2)贡献率小的主成分往往可能含有对样本差异的重要信息，PCA不考虑任何与结果变量y有关的信息，因此可能会丢失非常重要的特征。所以PCA不能用来减少过拟合

(3)特征值矩阵的正交向量空间是否唯一有待讨论。

(4)在非高斯分布的情况下，PCA方法得出的主元可能并不是最优的，此时在寻找主元时不能将方差作为衡量重要性的标准。

(5)另外一个问题就是可能出现重复的特征向量。应用PCA时，有的时候特征向量是可以在其所在的子空间中自由旋转的。这种情况在存在重复的或者接近的特征值时就会出现，将其视为一组表示数据的基础向量坐标轴，但是有些时候这些基础向量可以随意旋转，而细微的数值变化，可能引起特征向量的巨大变化，所以很多时候一次只考虑一个特征向量是没用的、很危险的。但是由前k个特征向量确定的子空间大致是相同的。

10、应用

Visualization

Compression

Learning

Anomaly detection

Matching or find better distance calculations

11、参考文章

《矩阵分析与应用(第二版)》

see.stanford.edu/materials/aimlcs229/cs229-notes10.pdf

blog.csdn.net/u013719780/article/details/78352262

www.mathworks.com/content/dam/mathworks/mathworks-dot-com/moler/eigs.pdf 10.5节

www.zhihu.com/question/49959130

www.zhihu.com/question/19666954

www.zhihu.com/question/22237507/answer/53804902