PCA Study Notes

Zhengyang Chen{chenzhengyang117@gmail.com}

January 19, 2019

Abstract

在这里我对自己在学习PCA过程中认为重要的点做以记录,同时也为了防止自己忘记.

Contents

1 为什么要进行PCA

在我们在对数据进行分析的时候,往往会面临数据的维度很高的问题,这会给我们后面的处理计算带来很大的问题。而PCA的出现就是为了解决这一问题,PCA可以有效的实现对数据的降维和除噪。那么PCA的降维和除噪为什么是有意义的呢?

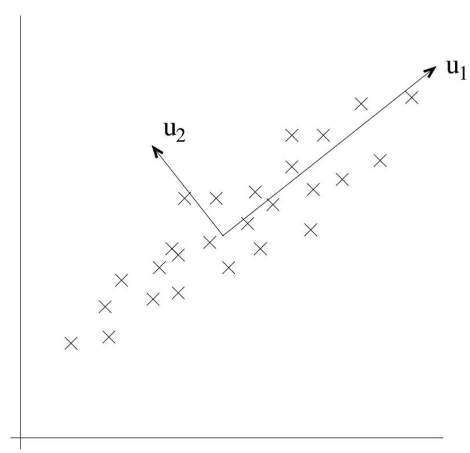
假如我们有数据$x^{(i)}; i = 1, ...m$, $x^{(i)} \in \mathcal{R}^n$, x_j表示的是x的第j个特征。我们假设这个数据表示的是学校同学的成绩信息。 x_i和 x_j分别表示的是学生的物理和数学成绩。我们知道，一般来说一个人的数学成绩好，那么他的物理成绩一般也会很好。也就是说这两个特征是相关的。也许我们只需用数学或者物理成绩中的一维就可以表示一个学生在这两门课中的学习情况了。而PCA便可以发现这种数据中的相关性，并可以实现特征之间的去相关，最后通过留下来比较重要的特征实现降维。

此外，噪声一般和我们的数据是不相关的，所以PCA很容易的将噪声分离出来。

2 PCA的数学推导

2.1 结果推导

那PCA究竟在做什么呢？我们以一个二维的图为例子：

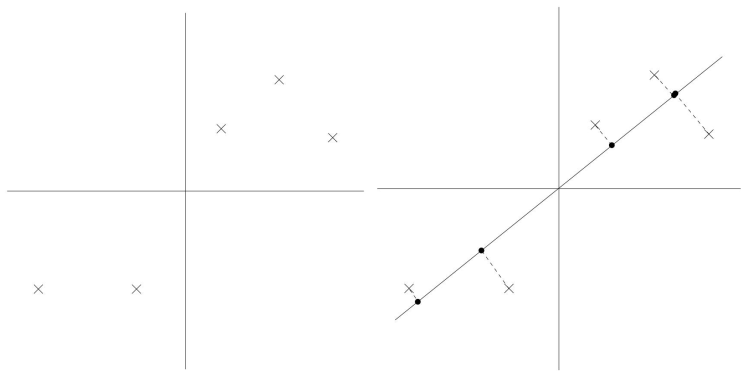


(a) 二维数据

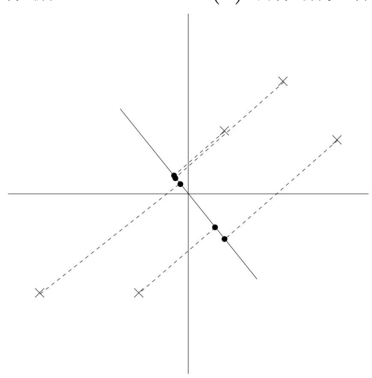
我们假设横轴为x，纵轴为y。可以看出两个坐标轴是有着很强的相关性的，即随着x的增大y也在增大。这样的话，也许我们用一个一维的数据就可以很好的表示这个二维的数据了。

如果要对一个二维的数据做PCA，将数据降到一维，那这件事是怎样做到的呢?

为了将下面三个小图(a)中的数据映射为一维，我们希望找到一个新的基（坐标轴），在这个基下原数据的信息不会损失太多。换句话说，也就是在新



(a) 原数据　(b) 将数据映射到某个坐标轴上



(c) 将数据映射到另一个坐标轴上

的基下原数据保留了很多的信息。在PCA中，我们用**方差**来度量这个信息的相对大小。

那现在我们的目标就是，找到一个新的基 v ,将原数据映射到这个基之后方差最大。通过线性代数的知识我们很容易求得 x_i 映射到 基 v 之后的坐标 v^Tx_i

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(v^Tx_i-0)^2 = \frac{1}{N}\sum_{i=1}^{N}(v^Tx_i)(v^Tx_i)^T \\ &= \frac{1}{N}\sum_{i=1}^{N}v^Tx_i{x_i}^Tv \\ &= v^T\left(\frac{1}{N}\sum_{i=1}^{N}x_i{x_i}^T\right)v = v^TCv\end{aligned}$$

上面的 C 对应的就是数据 x 对应的方差矩阵。这里我们加上 v 的模为1的限制：

$$v = argmax_{v \in R^d, ||v||=1} v^T C v$$

鉴于是带等式约束的优化问题，遂采用拉格朗日乘数法，写出拉格朗日乘数式如下：

$$f(x, \lambda) = v^T C v - \lambda(vv^T - 1)$$

上式对 v 和 λ 求导，并令导数等于0 ：

$$\frac{\partial f}{\partial v} = 2Cv - 2\lambda v = 0 \Rightarrow Cv = \lambda v$$
$$\frac{\partial f}{\partial \lambda} = v^T v = 1$$

根据上面的式子我们有：

$$v^T C v = v^T \lambda v = \lambda v^T v = \lambda$$

最终我们要求的那个基便是协方差矩阵 C 的最大特征值对应的特征向量。

2.2 特征向量的计算

2.2.1 什么是特征向量

对于一个 mxm 的矩阵S，如果有满足以下条件的 v ，使得

$$Sv = \lambda v$$

那么我们就称v是矩阵S的一个特征向量， λ 是这个特征向量对应的特征值。

2.2.2 一个矩阵的特征值有多少个

为了求解特征值，我们需要求解以下的方程：

$$Sv = \lambda v \Longleftrightarrow (S - \lambda I)v = 0 \Longleftrightarrow |S - \lambda I| = 0$$

特征值的个数也就是方程解的个数，也就是矩阵的**S的秩**。

2.2.3 特征向量的性质

定理（矩阵对角化定理）：存在以下特征值分解

$$S = U \Lambda U^{-1}$$

U的每个列向量就是矩阵S的特征向量，Λ是一个对角矩阵，Λ对角线上的第n个值对应着U的第n个列向量对应的特征值。

从2.1中我们知道，如果对向量 x 做 $v^T x$ 变换，那么向量x的协方差矩阵 C 将变为 $v^T C v$.另外,如果U的各列向量互相垂直且模为1.那么$U^T = U^{-1}$.

也就是说如果S是向量 x 的协方差矩阵，我们以协方差矩阵的特征向量作为x的变换矩阵 ，那么变换之后向量的协方差矩阵就是 Λ,也就意味着变换之后各维度之间是互不相关的，这也是PCA为什么可以**去相关**的原因。

3 PCA降维

在实际应用中，我们的数据往往是高维的（大于2维），我们也不可能将这些数据压缩到一维，这样我们损失的信息太多。在这种情况下，我们可以保留协方差矩阵 C 前 k 大的特征值 λ 。这样我们就可以将高维数据映射为 k 维数据。

我们知道协方差矩阵 C 是对称的，对一个对称矩阵而言，它的特征向量是彼此正交的。也就是说我们新的映射空间中的所有的基都是正交的。

由线性代数的知识知道，向量 x 在基 v 之下的坐标为 $v^T x$ 。 现在我们如果想把 N 维的数据 x 映射到 K 维，也就是我们想得到 x 在每个 $v_j, j = 1, ..., K$下的坐标。那么映射的方式就是：

$$y^{(i)} = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_K^T \end{bmatrix} x^{(i)} \in R^n, y^{(i)} \in R^k$$