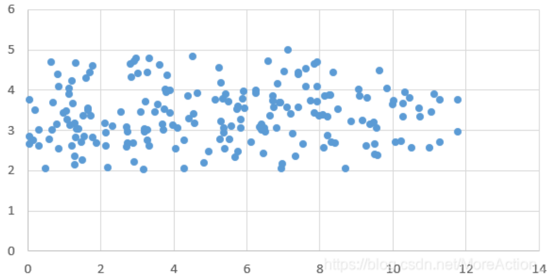
一文让你彻底搞懂主成成分分析PCA的原理及代码实现(超详细推导)

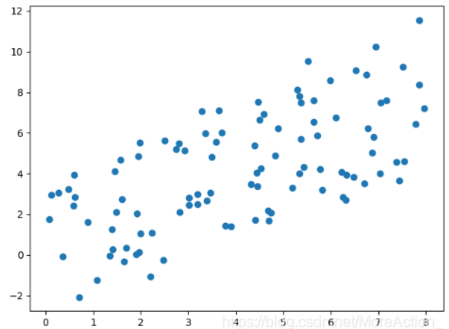
主成分分析（Principal components analysis）PCA是一个很重要的降维算法，可以用来降噪、消除冗余信息等，只要和数据打交道几乎是必学的。它
前置知识，我自己学的时候总是一知半解，后来才知道是这些前置知识基础没打牢固，为了彻底搞明白，我另外写了几篇文章，理清了其中用到的一
础不好的同学可以先过一下：
带你深入理解期望、方差、协方差的含义
一文读懂特征值分解EVD与奇异值分解SVD

引言

首先先举个例子来认识一下数据。
假设我们有一组二维数据(x,y)，它的分布如下：

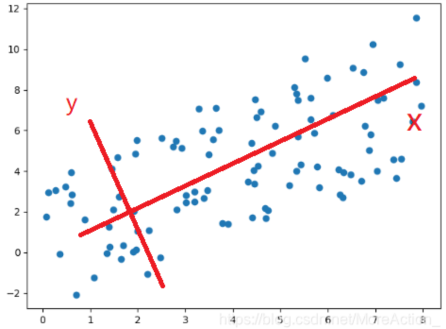


可以看到，数据在x轴上的变化大，而在y轴变化小，变化小意味着数据在这个特征上没有太大的差异，因此它包含的信息就比较少，那么我们就可以
重要的或者是噪音，从而可以直接将这个维度上的数据舍去，只用x轴上的数据来代替。
那么假如数据是这样分布的呢？



这个图我们就不太好看出到底是谁比较重要了，因为x和y变化都比较大，那么就不能降维了吗？非也，假如我们旋转一下坐标系

胤风 关注 61



新坐标系下数据的坐标值就是数据在坐标轴上的投影，这时候的情况就和上面那个例子一样了。

从这个例子也可以看到，数据本身的具体数值其实是不重要的，重要的是数据之间的关系，数据的整体分布。原来的数据是在E坐标系下，然后我们 标系来表示，本质上相当于对数据进行了一次正交变换（从数学公式看），在新的坐标系下，我们能更清楚的看到数据的特点，这为我们后续进一步 供了可能。

PCA其实做的就是这么一件事，求出了一个正交矩阵P，然后用这个矩阵对数据进行正交变换得到新的数据：

$$Y = PX$$

正交变换就相当于换了一个坐标系，所以其结果就是我们换了一个坐标系来观察数据，假如我们在这个坐标系下取前k个变化最大的轴上的数据，这就 维。

按照两阶段来理解PCA会容易得多，简单来说就是：第一阶段找了一个新的坐标系来表示数据，这个新的坐标系不是随便找的，是要求能最大限度的 上的数据变化大小，第二阶段在新坐标系下取前k个变化最大的轴上的数据，从而实现降维。

如果对正交变换不太了解，可以看下这篇文章：一文让你通俗易懂的理解正交变换和正交矩阵

PCA的原理

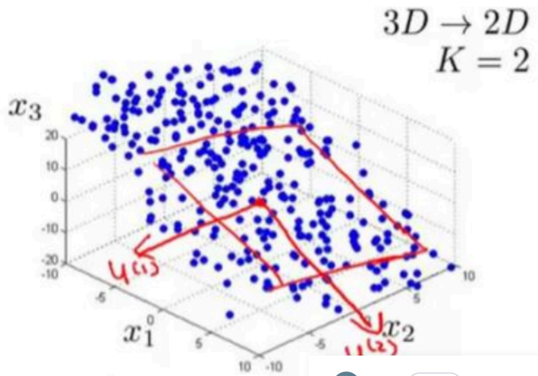
那么怎么才能找到这个新的坐标系呢？这就需要回到PCA所要解决的问题上。

PCA最本质的功能就是用来提取数据的主要信息，高维的数据存在维度灾难，而且许多变量之间可能存在相关性，也可能存在噪音，通过提取其中的 实现降维就是一种有效的解决手段。这里的"主要"指的是数据中包含的信息量较大的部分，那么一个很自然的问题就是：怎么来度量信息量的大小？

上面两个例子，我们是用数据在某个轴上的变化大小，也就分散程度来度量信息量的大小，这显然可以用数学上的方差来进行量化。

有了度量指标，接下来就是找到新的坐标系，那么就要逐个找到每个坐标轴，所以我们的目标就是先在整个数据空间中找到一个坐标轴（方向），使 个坐标轴上的投影（投影=坐标值）的方差达到最大，那么这个方向就是我们新的坐标系的第一个轴，然后再找一个方差次大的，而且与第一个轴垂直 （不限制垂直次大方向会和最大方向无限接近），作为新坐标系的第二个轴，依此类推，直到我们找出了K个，然后把原来的数据用这新的坐标系进行 就是进行投影），就得到了降维后的数据。

一个简单的示意图如下：



下面进行数学推导。

方差是每个元素与变量均值差的平方和的均值，一维数据的方差计算公式为：

$$\mathrm{Var}(a) = \frac{1}{m}\sum_{i=1}^{m}(a_i - \mu)^2$$

为了后续计算方便，我们先进行去中心化操作（使数据均值为0）。假设有m个n维数据，$X = [x_1, x_2, \ldots, x_m]$，其中的每个x是一个n维的列向量，

$$X = X - \frac{1}{m}\sum_{i=1}^{m} x_i$$

接下来，我们的目标就是在这个数据空间中找到一个方向，使得数据在这个方向上的投影的方差最大，假设这个方向为 w，$||w||_2 = 1$（单位向量） 数据在这个方向下的坐标值为：$w^T x_i$，于是有方差：

$$\begin{aligned} D(x) &= \frac{1}{m}\sum_{i=1}^{m}\left(w^T x_i\right)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(w^T x_i\right)\left(w^T x_i\right)^T \\ &= \frac{1}{m}\sum_{i=1}^{m} w^T x_i x_i^T w \\ &= w^T\left(\frac{1}{m}\sum_{i=1}^{m} x_i x_i^T\right) w \end{aligned}$$

其中，$\frac{1}{m}\sum_{i=1}^{m} x_i x_i^T$ 就是样本的协方差矩阵，令它为 C，那么我们的优化目标就是：

$$\begin{cases} \max\{w^T C w\} \\ s.t. w^T w = 1 \end{cases}$$

求解这种约束优化问题，用拉格朗日常数法最方便，构造函数：

$$L(w,\lambda) = w^T C w + \lambda\left(1 - w^T w\right)$$

然后对每个分量求导：

$$\begin{cases} \frac{\partial}{\partial w} f(w,\lambda) = 2Cw - 2\lambda w = 0 \\ \frac{\partial}{\partial w} f(w,\lambda) = w^T w - 1 = 0 \end{cases}$$

解得：

$$\begin{cases} Cw = \lambda w \\ w^T w = 1 \end{cases}$$

仔细看，这w不正是C的特征向量吗！代入目标函数中：

$$\max D(x) = \max\{w^T C w\} = \max\{w^T \lambda w\} = \max \lambda$$

于是要找的最大方差也就是协方差矩阵的最大特征值，而此时的方向就是最大特征值所对应的特征向量，那么次大方向自然就是第二大特征值对应的 依此类推，直到我们找出了K个，以这K个特征向量作为新的坐标系，然后将原始数据投影到这个坐标系下即可得到降维后的数据。

求解步骤

总结一下PCA算法的计算流程：

假设有m个n维数据，$X_{n\times m} = [x_1, x_2, \ldots, x_m]$，

1. 去中心化，$X = X - \frac{1}{m}\sum_{i=1}^{m} x_i$
2. 计算协方差矩阵，$C = \frac{1}{m}XX^T$
3. 对协方差矩阵进行特征值分解得到特征矩阵（按特征值从大到小以列排），取前k列组成矩阵$P_{n\times k}$，P就相当于是一个坐标系，P中的每一列就轴
4. 将原始数据投影到P坐标系下即得到降维后的数据，$Y_{k\times m} = P_{n\times k}^T X_{n\times m}$

其它细节

(1)求导公式

拉格朗日函数求极值那块需要进行矩阵求导，具体推导就不介绍了，直接给出用到的两个公式：

$$\frac{\partial x^T a}{\partial x} = \frac{\partial a^T x}{\partial x} = a$$

$$\frac{\partial x^T A x}{\partial x} = Ax + A^T x$$

如果矩阵A是对称的：

$$Ax + A^T x = 2Ax$$

这两个公式比较常用，在最小二乘法中也用到，最好直接记下来。

(2)数据还原（重建）

PCA可以对高维数据进行降维以达到压缩数据的目的，比如图像处理领域就经常用到PCA作图像压缩，有压缩就有还原，但是因为PCA降维是有损失压缩后的数据没有保持原来数据的全部信息，所以根据压缩数据无法还原回原来的高维数据，但是还原的数据可以看作原来数据的一种近似。

最后经计算得到变换矩阵$P_{n\times k}$，那么降维的数据就是：$Y_{k\times m} = P_{n\times k}^T X_{n\times m}$，根据$Y$和$P$，我们可以还原得到$X$的近似：

$$\widetilde{X}_{n\times m} = P_{n\times k} Y_{k\times m}$$

(3)数据占比

我们会好奇降维后的数据到底保持了原来数据的多少信息，而PCA中数据的方差就代表着信息，从推导的结果中可以看到C的特征值λ就是方差，我们个，所以信息占比就是：

$$r = \frac{\sum_{j=1}^{k}\lambda_j}{\sum_{j=1}^{n}\lambda_j}$$

PCA算法只有一个参数，就是降维数据的维度k，有时候我们会要求降维后的数据信息占比多少以上，就可以根据这个公式计算出k

(4)PCA与SVD的关系

PCA与SVD其实并没有什么直接联系，只不过是PCA可以用SVD来实现，比较方便。
我们都知道SVD是一种用于任意矩阵分解的方法：

$$A = U\Sigma V^T$$

其中，U的列向量就是 AA^T 的特征向量，而PCA中我们要计算的就是$C = \frac{1}{m}XX^T$的特征向量，那么我们令$A = C$就可以直接算出特征向量了，方便。

(5)PCA的性质

1. 降噪：PCA舍弃了那些方差较小的数据，这些数据可能是噪音；
2. 去除冗余信息：PCA各主成分之间正交， 可以去除原始数据中具有较强线性关系的feature
3. 可解释性差：降维的数据是对原始数据的重组，各个特征维度的含义不具有原始样本特征的解释性

代码实现

这里介绍两种，一种是采用SVD手动降维，一种是直接调PCA api

SVD手动降维


```
import numpy as np

# 3 x 5 , 3维特征，5组数据
data = np.array([[-1,-1,0,2,1],[2,0,0,-1,-1],[2,0,1,1,0]], dtype=np.float)
n, m = data.shape
k = 2
data = data - data.mean(axis=1, keepdims=True)
# 协方差矩阵
C = np.dot(data, data.T) / m
# u的每一列是特征向量
u, d, v = np.linalg.svd(C)
```

调用sklearn中的PCA api

注意，sklearn中的输入数据是以行排列的，每一行是一个数据！

```
import numpy as np
from sklearn.decomposition import PCA

# 3 x 5 , 3维特征，5组数据
data = np.array([[-1,-1,0,2,1],[2,0,0,-1,-1],[2,0,1,1,0]], dtype=np.float)
k = 2

pca = PCA(n_components=k)
# 注意转置
new_data = pca.fit_transform(data.T)
# 降维后的数据各个维度特征 所占信息比例
```

PCA常用参数介绍：

n_components： 这个参数类型有int型，float型，string型，默认为None，常用的设置有两种：

- 若0<n_components<1，则n_components的值为主成分方差的阈值； 通过设置该变量，即可调整主成分数量K；
- 若n_components≥1，则降维后的特征数为n_components；

whiten：参数为bool型，是否对降维后的数据的每个特征进行归一化，默认是False。

其它常用方法：
fit(X,y=None)：用训练数据X训练模型，由于PCA是无监督降维，因此y=None。
transform(X,y=None)：训练好模型后，对输入数据X进行降维。
fit_transform(X)：用训练数据X训练模型，并对X进行降维。相当于先用fit(X)，再用transform(X)。
inverse_transform(X)：将降维后的数据还原成原始数据的近似。(PCA的重建)

PCA对象常用属性：
components：array, shape (n_components, n_features)，降维后各主成分方向，并按照各主成分的方差值大小排序。
explained_variance：array, shape (n_components,)，降维后各主成分的方差值，方差值越大，越主要。
explained_variance_ratio：array, shape (n_components,)，降维后的各主成分的方差值占总方差值的比例，比例越大，则越主要。

reference
https://www.cnblogs.com/XDU-Lakers/p/11612094.html
https://blog.csdn.net/xxdragon126/article/details/90748254
https://zhuanlan.zhihu.com/p/77151308
https://blog.csdn.net/zhongkelee/article/details/44064401
https://blog.csdn.net/hustqb/article/details/78394058
https://www.cnblogs.com/pinard/p/6239403.html

如果对你有帮助，请点个赞让我知道:-D