42 | PCA主成分分析（上）：如何利用协方差矩阵来降维？

2019-03-22 黄申

讲述：黄申
时长 11:09 大小 10.22M

你好，我是黄申。

在概率统计模块，我详细讲解了如何使用各种统计指标来进行特征的选择，降低用于监督式学习的特征之维度。接下来的几节，我会阐述两种针对数值型特征，更为通用的降维方法，它们是**主成分分析 PCA**（Principal Component Analysis）和**奇异值分解 SVD**（Singular Value Decomposition）。这两种方法是从矩阵分析的角度出发，找出数据分布之间的关系，从而达到降低维度的目的，因此并不需要监督式学习中样本标签和特征之间的关系。

PCA 分析法的主要步骤

我们先从主成分分析 PCA 开始看。

在解释这个方法之前，我先带你快速回顾一下什么是特征的降维。在机器学习领域中，我们要进行大量的特征工程，把物品的特征转换成计算机所能处理的各种数据。通常，我们增加物品的特征，就有可能提升机器学习的效果。可是，随着特征数量不断的增加，特征向量的维度也会不断上升。这不仅会加大机器学习的难度，还会影响最终的准确度。针对这种情形，我们需要过滤掉一些不重要的特征，或者是把某些相关的特征合并起来，最终达到在减少特征维度的同时，尽量保留原始数据所包含的信息。

了解了这些，我们再来看今天要讲解的 PCA 方法。它的主要步骤其实并不复杂，我一说你就能明白，但是为什么要这么做，你可能并不理解。咱们学习一个概念或者方法，不仅要知道它

是什么，还要明白是怎么来的，这样你就能知其然，知其所以然，明白背后的逻辑，达到灵活运用。因此，我先从它的运算步骤入手，给你讲清楚每一步，然后再解释方法背后的核心思想。

和线性回归的案例一样，我们使用一个矩阵来表示数据集。我们假设数据集中有 m 个样本、n 维特征，而这些特征都是数值型的，那么这个集合可以按照如下的方式来展示。

ID	特征1	特征2	特征3	...	特征n-1	特征n
1	1	3	-7	...	-10.5	-8.2
2	2	5	-14	...	2.7	4
...	...	...	...	...	...	...
m	-3	-7	2	...	55	13.6

那么这个样本集的矩阵形式就是这样的：

$$\begin{bmatrix} 1 & 3 & -7 & \cdots & -10.5 & -8.2 \\ 2 & 5 & -14 & \cdots & 2.7 & 4 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -3 & -7 & 2 & \cdots & 55 & 13.6 \end{bmatrix}$$

这个矩阵是 m×n 维的，其中每一行表示一个样本，而每一列表示一维特征。让我们把这个矩阵称作样本矩阵，现在，我们的问题是，能不能通过某种方法，找到一种变换，可以降低这个矩阵的列数，也就是特征的维数，并且尽可能的保留原始数据中有用的信息？

针对这个问题，PCA 分析法提出了一种可行的解决方案。它包括了下面这样几个主要的步骤：

1. 标准化样本矩阵中的原始数据；
2. 获取标准化数据的协方差矩阵；
3. 计算协方差矩阵的特征值和特征向量；
4. 依照特征值的大小，挑选主要的特征向量；
5. 生成新的特征。

下面，我们一步步来看。

1. 标准化原始数据

之前我们已经介绍过特征标准化，这里我们需要进行同样的处理，才能让每维特征的重要性具有可比性。为了便于你回顾，我把标准化的公式列在了这里。

$$x' = \frac{x-\mu}{\sigma}$$

其中 x 为原始值，u 为均值，σ 为标准差，x' 是变换后的值。需要注意的是，这里标准化的数据是针对同一种特征，也是在同一个特征维度之内。不同维度的特征不能放在一起进行标准化。

2. 获取协方差矩阵

首先，我们来看一下什么是协方差（Covariance），以及协方差矩阵。协方差是用于衡量两个变量的总体误差。假设两个变量分别是 x 和 y，而它们的采样数量都是 m，那么协方差的计算公式就是如下这种形式：

$$\operatorname{cov}(x,y) = \frac{\sum_{k=1}^{m}\left(x_k - \bar{x}\right)\left(y_k - \bar{y}\right)}{m-1}$$

其中 x_k 表示变量 x 的第 k 个采样数据，$\bar{x}$ 表示这 k 个采样的平均值。而当两个变量是相同时，协方差就变成了方差。

那么，这里的协方差矩阵又是什么呢？我们刚刚提到了样本矩阵，假设 $X_{,1}$ 表示样本矩阵 X 的第 1 列，$X_{,2}$ 表示样本矩阵 X 的第 2 列，依次类推。而 $cov(X_{,1}, X_{,1})$ 表示第 1 列向量和自己的协方差，而 $cov(X_{,1}, X_{,2})$ 表示第 1 列向量和第 2 列向量之间的协方差。结合之前协方差的定义，我们可以得知：

$$\operatorname{cov}(X_{,i}, X_{,j}) = \frac{\sum_{k=1}^{m}\left(x_{k,i} - \overline{X_{,i}}\right)\left(x_{k,j} - \overline{X_{,j}}\right)}{m-1}$$

其中，$x_{k,i}$ 表示矩阵中第 k 行，第 i 列的元素。 $\bar{X_{,i}}$ 表示第 i 列的平均值。

有了这些符号表示，我们就可以生成下面这种协方差矩阵。

$$COV = \begin{bmatrix} \text{cov}(X_{,1}, X_{,1}) & \text{cov}(X_{,1}, X_{,2}) & \cdots & \text{cov}(X_{,1}, X_{,n-1}) & \text{cov}(X_{,1}, X_{,n}) \\ \text{cov}(X_{,2}, X_{,1}) & \text{cov}(X_{,2}, X_{,2}) & \cdots & \text{cov}(X_{,2}, X_{,n-1}) & \text{cov}(X_{,2}, X_{,n}) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \text{cov}(X_{,n}, X_{,1}) & \text{cov}(X_{,n}, X_{,2}) & \cdots & \text{cov}(X_{,n}, X_{,n-1}) & \text{cov}(X_{,n}, X_{,n}) \end{bmatrix}$$

从协方差的定义可以看出，$cov(X_{,i}, X_{,j}) = cov(X_{,j}, X_{,i})$，所以 COV 是个对称矩阵。另外，我们刚刚提到，对于 $cov(X_{,i}, X_{,j})$，如果 $i = j$，那么 $cov(X_{,i}, X_{,j})$ 也就是 $X_{,j}$ 这组数的方差。所以这个对称矩阵的主对角线上的值就是各维特征的方差。

3. 计算协方差矩阵的特征值和特征向量

需要注意的是，这里所说的矩阵的特征向量，和机器学习中的特征向量（Feature Vector）完全是两回事。矩阵的特征值和特征向量是线性代数中两个非常重要的概念。对于一个矩阵 X，如果能找到向量 v 和标量 λ，使得下面这个式子成立。

$$Xv = \lambda v$$

那么，我们就说 v 是矩阵 X 的特征向量，而 λ 是矩阵 X 的特征值。矩阵的特征向量和特征值可能不止一个。说到这里，你可能会好奇，特征向量和特征值表示什么意思呢？我们为什么要关心这两个概念呢？简单的来说，我们可以把向量 v 左乘一个矩阵 X 看做对 v 进行旋转或拉伸，而这种旋转和拉伸都是由于左乘矩阵 X 后，所产生的“运动”所导致的。特征向量 v 表示了矩阵 X 运动的方向，特征值 λ 表示了运动的幅度，这两者结合就能描述左乘矩阵 X 所带来的效果，因此被看作矩阵的“特征”。在 PCA 中的主成分，就是指特征向量，而对应的特征值的大小，就表示这个特征向量或者说主成分的重要程度。特征值越大，重要程度越高，我们要优先现在这个主成分，并利用这个主成分对原始数据进行变换。

如果你还是有些困惑，我会在下面一节，讲解更多的细节。现在，让我们先来看看给定一个矩阵，如何计算它的特征值和特征向量，并完成 PCA 分析的剩余步骤。我在下面列出了计算特征值的推导过程：

$$\begin{aligned} &Xv = \lambda v \\ &Xv - \lambda v = 0 \\ &Xv - \lambda Iv = 0 \\ &(X - \lambda I)v = 0 \end{aligned}$$

其中 I 是单位矩阵。对于上面推导中的最后一步，我们需要计算矩阵的行列式。

$$\left|(X-\lambda I)\right|=\left\|\begin{bmatrix} x_{1,1}-\lambda & x_{1,2} & x_{1,3} & \cdots & x_{1,n-1} & x_{1,n} \\ x_{2,1} & x_{2,2}-\lambda & x_{2,3} & \cdots & x_{2,n-1} & x_{2,n} \\ x_{3,1} & x_{3,2} & x_{3,3}-\lambda & \cdots & x_{3,n-1} & x_{3,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n-1,1} & x_{n-1,2} & x_{n-1,3} & \cdots & x_{n-1,n-1}-\lambda & x_{n-1,n} \\ x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,n-1} & x_{n,n}-\lambda \end{bmatrix}\right\|=0$$

$$(x_{1,1}-\lambda)(x_{2,2}-\lambda)\ldots(x_{n,n}-\lambda)+x_{1,2}x_{2,3}\ldots x_{n-1,n}x_{n,1}+\ldots)-(x_{n,1}x_{n-1,2}\ldots x_{2,n-1}a$$

最后，通过解这个方程式，我们就能求得各种λ的解，而这些解就是特征值。计算完特征值，我们可以把不同的λ值代入 $\lambda E-A$，来获取特征向量。

$$(\lambda I-X)\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$$

4. 挑选主要的特征向量，转换原始数据

假设我们获得了 k 个特征值和对应的特征向量，那么我们就有：

$$Xv_1=\lambda_1 v_1$$
$$Xv_2=\lambda_2 v_2$$
$$\ldots$$
$$Xv_k=\lambda_k v_k$$

按照所对应的λ数值的大小，对这 k 组的 v 排序。排名靠前的 v 就是最重要的特征向量。

假设我们只取前 k1 个最重要的特征，那么我们使用这 k1 个特征向量，组成一个 n×k1 维的矩阵 D。

把包含原始数据的 m×n 维矩阵 X 左乘矩阵 D，就能重新获得一个 m×k1 维的矩阵，达到了降维的目的。

有的时候，我们无法确定 k1 取多少合适。一种常见的做法是，看前 k1 个特征值的和占所有特征值总和的百分比。假设一共有 10 个特征值，总和是 100，最大的特征值是 80，那么第一大特征值占整个特征值之和的 80%，我们认为它能表示 80% 的信息量，还不够多。那我们就继续看第二大的特征值，它是 15，前两个特征值之和有 95，占比达到了 95%，如果我们认为足够了，那么就可以只选前两大特征值，把原始数据的特征维度从 10 维降到 2 维。

小结

这一节，我首先简要地重温了为什么有时候需要进行特征的降维和基于分类标签的特征选择。随后，我引出了和特征选择不同的另一种方法，基于矩阵操作的 PCA 主成分分析。这种方法的几个主要步骤包括，标准化原始数据、获得不同特征的协方差矩阵、计算协方差矩阵的特征值和特征向量、选择最重要的主成分，以及通过所选择的主成分来转换原始的数据集。

要理解 PCA 分析法是有一定难度的，主要是因为两点原因：第一，计算的步骤有些复杂。第二，这个方法的核心思路有些抽象。这两点可能会让刚刚接触 PCA 的学习者，感到无从下手。

为了帮助你更好的理解，下一节，我会使用一个示例的矩阵进行详细的推算，并用两种 Python 代码进行结果的验证。除此之外，我还会分析几个要点，包括 PCA 为什么使用协方差矩阵？这个矩阵的特征值和特征向量又表示什么？为什么特征值最大的主成分涵盖最多的信息量？明白了这些，你就能深入理解为什么 PCA 分析法要有这些步骤，以及每一步都代表什么含义。

思考题

给定这样一个矩阵：

$$\begin{bmatrix} 1 & 3 & -7 \\ 2 & 5 & -14 \\ -3 & -7 & 2 \end{bmatrix}$$

假设这个矩阵的每一列表示一个特征的维度，每一行表示一个样本。请完成

1. 按照列（也就是同一个特征维度）进行标准化。
2. 生成这个矩阵的协方差矩阵。

欢迎留言和我分享，也欢迎你在留言区写下今天的学习笔记。你可以点击“请朋友读”，把今天的内容分享给你的好友，和他一起精进。

上一篇　41 | 线性回归（下）：如何使用最小二乘法进行效果验证？

下一篇　43 | PCA主成分分析（下）：为什么要计算协方差矩阵的特征值和特征向量？

精选留言 (4)

写留言

qinggeouy...
2019-03-31

1

markdown 语法支持不是很好

(1) 标准化原始数据

$$x' = \frac{x - \mu}{\sigma}$$

第一列

均值 $\mu_1 = 0$，方差 ${\sigma_1}^2 = [(1-0)^2 + (2-0)^2 + (-3-0)^2]/3 = 14/3$

...

第二列

均值 $\mu_2 = 1/3$，方差 ${\sigma_2}^2 = [(3-1/3)^2 + (5-1/3)^2 + (-7-1/3)^2]/3 = 248/9$

第三列

均值 $\mu_3 = -19/3$，方差 ${\sigma_3}^2 = [(-7+19/3)^2 + (-14+19/3)^2 + (2+19/3)^2]/3 = 386/9$

则，

$$\mathbf{X}' = \begin{vmatrix} 0.46291005 & 0.50800051 & -0.10179732 \\ 0.9258201 & 0.88900089 & -1.17066918 \\ -1.38873015 & -1.3970014 & 1.2724665 \end{vmatrix}$$

(2)协方差矩阵

$$\sum^{m} (x_{i\cdot} - \bar{X}_{\cdot})(x_{i\cdot} - \bar{X}_{\cdot})$$

$$\mathbf{cov(X_{,i}, X_{,j})} = \frac{\sum_{k=1}(x_{k,i} - \bar{X}_{,i})(x_{k,j} - \bar{X}_{,j})}{m-1}$$

$$\mathbf{X}'.mean(asix = 0) = [0, 0, -7.401486830834377e-17]$$

$$\mathbf{cov(X_{,i}, X_{,j})} = \frac{(\mathbf{X}'[:, \mathbf{i-1}] - \mathbf{X}'[:, \mathbf{i-1}].mean()).transpose().dot(\mathbf{X}'[:, \mathbf{j}}{m-1}$$

协方差矩阵(对角线上是各维特征的方差)：

$$\mathbf{COV} = \begin{vmatrix} \mathbf{cov(X_{,1}, X_{,1})} & \mathbf{cov(X_{,1}, X_{,2})} & \mathbf{cov(X_{,1}, X_{,3})} \\ \mathbf{cov(X_{,2}, X_{,1})} & \mathbf{cov(X_{,2}, X_{,2})} & \mathbf{cov(X_{,2}, X_{,3})} \\ \mathbf{cov(X_{,3}, X_{,1})} & \mathbf{cov(X_{,3}, X_{,2})} & \mathbf{cov(X_{,3}, X_{,3})} \end{vmatrix} = \begin{vmatrix} 1.5 \\ 1.4991357 \\ -1.44903232 \end{vmatrix}$$

展开

Joe 1
2019-03-22

一直有个问题为什么协方差是除以m-1，而不是m。方差，均方根等公式也是除m-1。好奇怪。

作者回复: 这是个很好的问题，涉及的内容比较多，我可以放在后面答疑来解释

余泽锋
2019-04-11

import numpy as np
import pandas as pd
from sklearn.preprocessing import scale
array = np.array([[1, 3, -7], [2, 5, -14], [-3, -7, 2]])
array = scale(array)...
展开

---

yaya
2019-03-22

所以上只是讲解pca的步骤吗？非常赞同要明白他是为什么被提出的，怎么来的观点，但是pca如果只是记步骤很容易忘记，觉得还是从如何建模，然后推导而来更有印象。

作者回复: 非常同意，我会在下一篇解释为什么PCA要这么做