教程 8 : PCA

提交日期：2020.7.25

提交人：詹紫琦

目录

[1. 题目 3](#_Toc28019)

[1.1 主成成分分析（PCA） 3](#_Toc4508)

[1.2 运行环境 3](#_Toc18615)

[2. 算法阐述或实验步骤说明 3](#_Toc29011)

[2.1变量说明 3](#_Toc9127)

[2.2 PCA算法阐述及步骤说明 3](#_Toc1629)

[3. 实验结果与截图 5](#_Toc13972)

[3.1 PCA数据可视化应用实验结果 5](#_Toc3225)

[3.2 PCA图片压缩实验结果 6](#_Toc18307)

[4. 总结 7](#_Toc19719)

[4.1 总结 7](#_Toc4642)

[5. 参考文献 7](#_Toc6286)

# 题目

## 1.1 主成成分分析（PCA）

PCA是一种降维技术，通过分析数据中的主成成分，来保留数据中最具有特征意义维度的数据从而达到降低维度的目的。PCA这种技术一般用于数据可视化中，例如当训练数据特征数大于三个维度时，无法将其可视化，这里就可以通过降低数据维度，同时也保持一定的分布特征来可视化数据。同时PCA还应用于图片压缩存储当中，能将图片对应的多维矩阵进行降维存储从而减少存储空间。PCA在计算中，主要应用到一些矩阵变换的知识，例如归一化，协方差矩阵，以及矩阵的奇异值分解等基础变换。本次实验就介绍了PCA技术应用下的高维数据可视化以及图片压缩还原等场景。

## 1.2 运行环境

系统：Ubuntu20.04，python3.7，Anaconda集成工具Jupyter编写。

# 算法阐述或实验步骤说明

## 2.1变量说明

，每一个数据中包含的n维特征。数据集中总共包含m项数据。

代表所包含数据的特征向量，用于求解协方差矩阵。

代表协方差矩阵。

分别代表奇异值中分解的左奇异矩阵，奇异值矩阵，以及右奇异矩阵

代表降维所需的降维矩阵。

是压缩重现后的矩阵，与原数据矩阵的范数相同

## 2.2 PCA算法阐述及步骤说明

在进行主成成分分析之前，需要对数据进行归一化处理，归一化处理就是将数据的每一列特征的均值归一化为0，公式如2-1：

 （2-1）

进行均值归一化之后开始求解协方差矩阵，由于数据为n维特征，所以协方差矩阵是一个矩阵。计算协方差矩阵的公式如2-2：

 （2-2）

计算出协方差矩阵后就需要编写SVD函数了，SVD函数即为奇异值分解函数，输入为协方差矩阵，输出为三个矩阵，分别为，得到之后，就可以根据的前k列来决定需要降低到的维度。的计算步骤如下：

由公式2-3可得出公式2-4，2-5：

 （2-3）

 （2-4）

同理可得到公式2-5：

 （2-5）

可知，的列由单位化过的特征向量构成，的列由单位化过的特征向量构成，的对角元素来源于或者是特征值的平方根由大到小排序。由此取的前k列构成，通过原数据矩阵与降维矩阵相乘得到新的降维后的数据，至此PCA结束。

在进行PCA的过程中一般会计算降维前后的矩阵误差大小，可以通过SVD分解得到的矩阵S计算出精确程度，如公式2-6：

 （2-6）

一般我们会把精确度设置为99%。

# 实验结果与截图

## 3.1 PCA数据可视化应用实验结果

这里我们使用的数据集是鸢尾花数据集，该数据集包含的数据维度为四，无法用一般的数据可视化展示出来，进行初步分析，这里我们进行PCA降维，首先将四维数据降低至三维，得到三维可视化数据如图3-1：

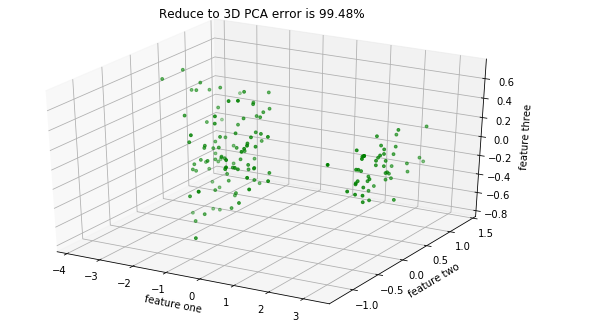


图3-1 鸢尾花数据集PCA降维至3维数据可视化

降维后计算PCA前后的误差可以得出是99.48%，也就是说在进行还原数据时，还原的数据与原始数据的相似程度是99.48%。大于0.99，说明鸢尾花数据集从四维降至三维后对特征的分布影响不大。

降至三维后，我们继续进行PCA降维，先降至二维，得到的数据如图3-2：

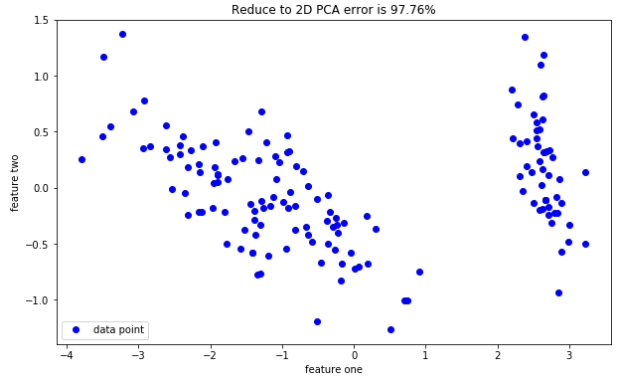


图3-2 鸢尾花数据集PCA降维至2维数据可视化

降维后我们可以发现，准确度只有97.76%，这个误差已经比较大了，还原后的数据与原数据的相似程度低于99%，在这个评判标准上将四维的鸢尾花数据集降维至2维得到的数据不能满足要求，所以在进行该数据集上的PCA降维数据可视化时，降至三维然后进行数据可视化能达到较好的效果。

## 3.2 PCA图片压缩实验结果

读取图片，将其转换为灰度图像，减去其RGB类型。根据得到的图像矩阵进行降维处理，进行PCA后，找到误差能满足99%的相似程度的k值，并找到最小的k值，k值便为能降维至的最低维。由于原图矩阵为1458x1458矩阵，找到满足条件的最小k值为1302，然后进行降维存储，最后进行还原，得到的原图与还原图的对比如图3-3：

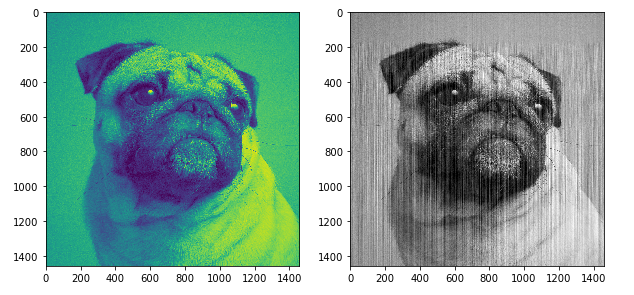


图3-3 1458x1458通过PCA降至1458x1302后还原图与原始图对比

# 总结

## 4.1 总结

PCA是一种主成成分分析技术，通过主分量分析技术，旨在利用降维的思想，把多指标转化为少数几个综合指标，是一种简化数据集的技术，是一种通过矩阵完成的线性变换技术，降维实质上就是投影。通过从高维投影到低维，最后进行误差分析，寻找满足误差的最大投影程度。

# 参考文献

[1] 周志华等.机器学习（西瓜书）