Homework 02 PCA

王晓捷 (11521053)

May 18, 2016

1 引言

PCA(Principal Components Analysis)，即主成分分析，是一种现代数据分析中常用的分析、简化数据集的方法。主成分分析经常用于减少数据集的维数，同时保持数据集中的对方差贡献最大的特征[1]。

主成分分析有卡尔·皮尔逊在1901年发明，用于分析数据及建立数理模型。其主要思想是通过移动坐标轴使得原先的n维特征映射到k维上（$k < n$）,这k维是全新的正交特征，称为“主元”。其方法主要是通过对协方差矩阵进行特征分解，以得出数据的主成分与它们的权值。PCA 是最简单的以特征量分析多元统计分布的方法。PCA 通过解释哪一个方向上的数据值对方差的影响最大的方式，提供了一种降低数据维度的有效办法。这种方法数据信息损失量将会是最小的。

2 方法概述

在PCA中，我们队数据的坐标进行了旋转，该旋转的过程取决于数据的本身。第一条坐标轴旋转到覆盖数据的最大方差的位置。在选择了覆盖数据最大差异性的坐标轴之后，接下来选择第二条坐标轴，加入该坐标轴与第一条坐标轴垂直，也就是正交，它就是覆盖数据次大差异性的坐标轴，依次类推。在线性代数的学习中，我们知道特征值分析能够通过数据的一般格式来揭示数据的“真实”结构。因此，通过数据集的协方差矩阵及其特征值分析，可以求得这些主成分的值。一旦得到了协方差矩阵的特征向量，就可以保留最大的N个值。这些特征向量也给出了N个最重要特征的真是结构。可以通过将数据乘上这N个特征向量而将它们转换到新的空间。

其伪代码大致如下[2]：

- 去除平均值
- 计算协方差矩阵
- 计算协方差军阵的特征值和特征向量
- 将特征值从大到小排序
- 保留最大的N个特征向量
- 将数据转换到上述N个特征向量构架的新空间中

3 实验结果

- 数据集：all digit ’3’ in Optical Recognition of Handwritten Digits Data Set
- 原始维度：64维
- PCA降维：2维

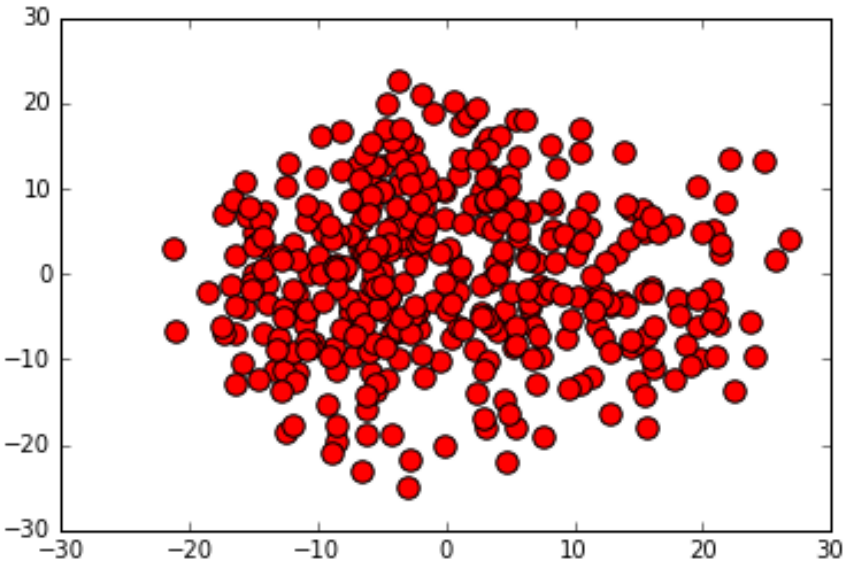

图 1: PCA降维后效果图

4 小结与讨论

由第3部分中的实验结果可以看到利用PCA降维后的数据表示。降维使数据变得更易使用，并且在一定程度上能够去除数据中的噪声，使得其他机器学习的任务更加精确。

参考文献

[1] Andrew Ng. Cs229 lecture notes. CS229 Lecture notes, 1(1):1–3, 2000.

[2] Jonathon Shlens. A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100, 2014.