

**2020年春季学期  
计算学部《机器学习》课程**

**Lab 4实验报告**

|  |  |
| --- | --- |
| 姓名 | 王丁子睿 |
| 学号 | 1183710211 |
| 班号 | 1803104 |
| 电子邮件 | 1183710211@stu.hit.edu.cn |
| 手机号码 | 19845178018 |

**目录**

[1 问题描述 3](#_Toc55227136)

[2 数据生成 3](#_Toc55227137)

[3 问题求解 4](#_Toc55227138)

[3.1 PCA 4](#_Toc55227139)

[3.2 数据还原 4](#_Toc55227140)

[4 应用 5](#_Toc55227141)

# 问题描述

给定一个高维空间的点集，我们希望找到一种映射方式，使得可以用一个低维空间的点集来代表该高维空间的点集，同时损失的信息尽量少。

此外，我们可以用一定的方式，来还原原高维空间的点集。

# 数据生成

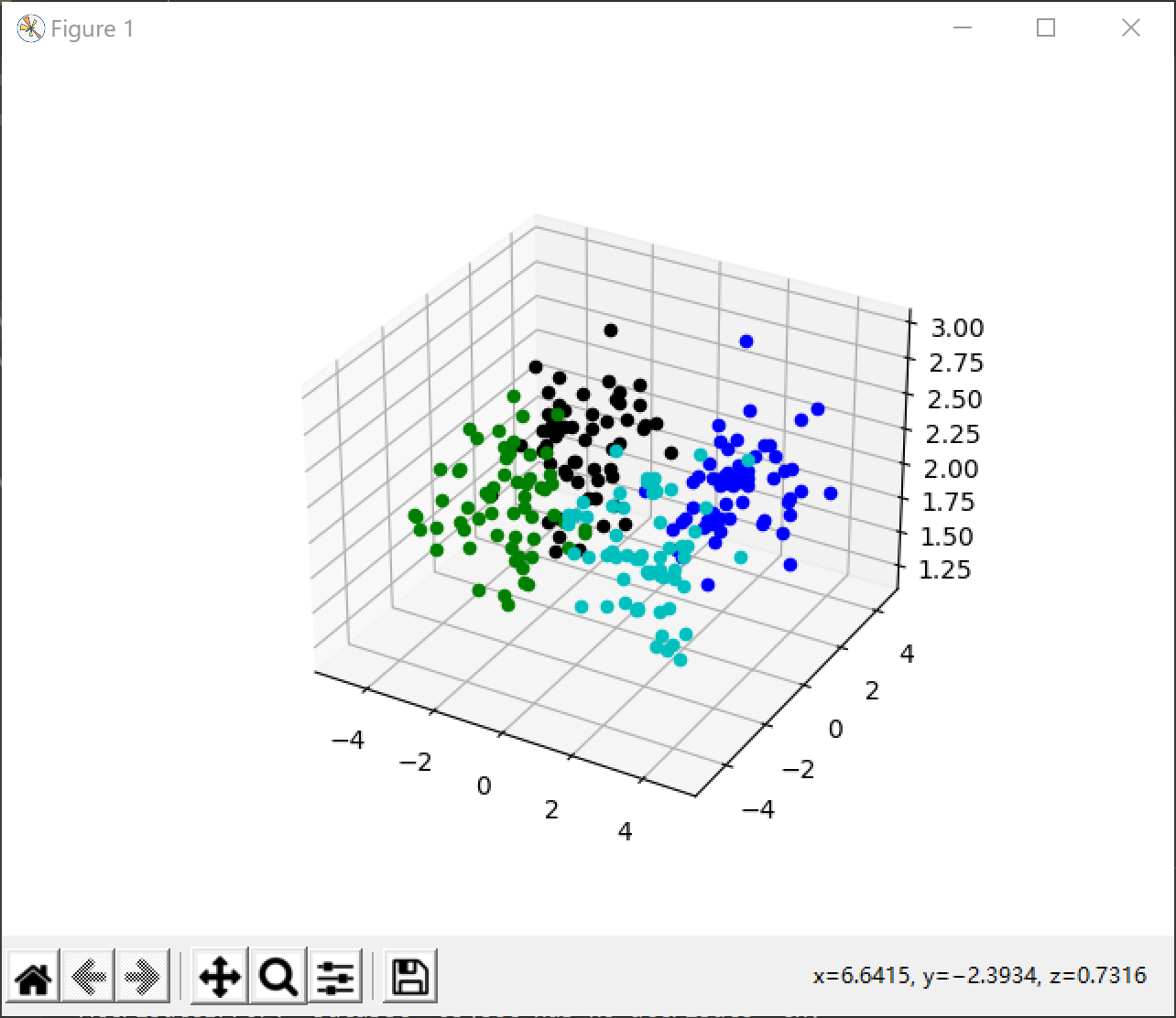
利用高斯分布来生成训练和测试数据。

在本次实验中，共利用四个三维高斯分布来生成数据，每个高斯分布的训练、验证、测试数据量之比为60:40:40。

为了能够实现降维的效果，生成数据集时第三维的方差较小

各高斯分布具体参数如下：

生成的数据集如下图所示：



下文中如不特殊说明，得到的结果均出自该数据集。

# 问题求解

## PCA

为了防止数值的大小对降维产生较大的影响，首先需要对数据进行归一化处理，即所有的数据减掉各维的平均值，然后除以各维的最大值减最小值，将数据压缩到[0, 1]的范围内。

然后我们计算得到数据集的协方差矩阵，考虑到特征向量可以一定程度地反应矩阵的性质，所以我们可以对协方差矩阵做奇异值分解，然后在按照奇异值由大到小的顺序排列的情况下，取前k个特征向量张成的空间作为结果空间，将各点投影到该空间中，从而起到降维的效果。

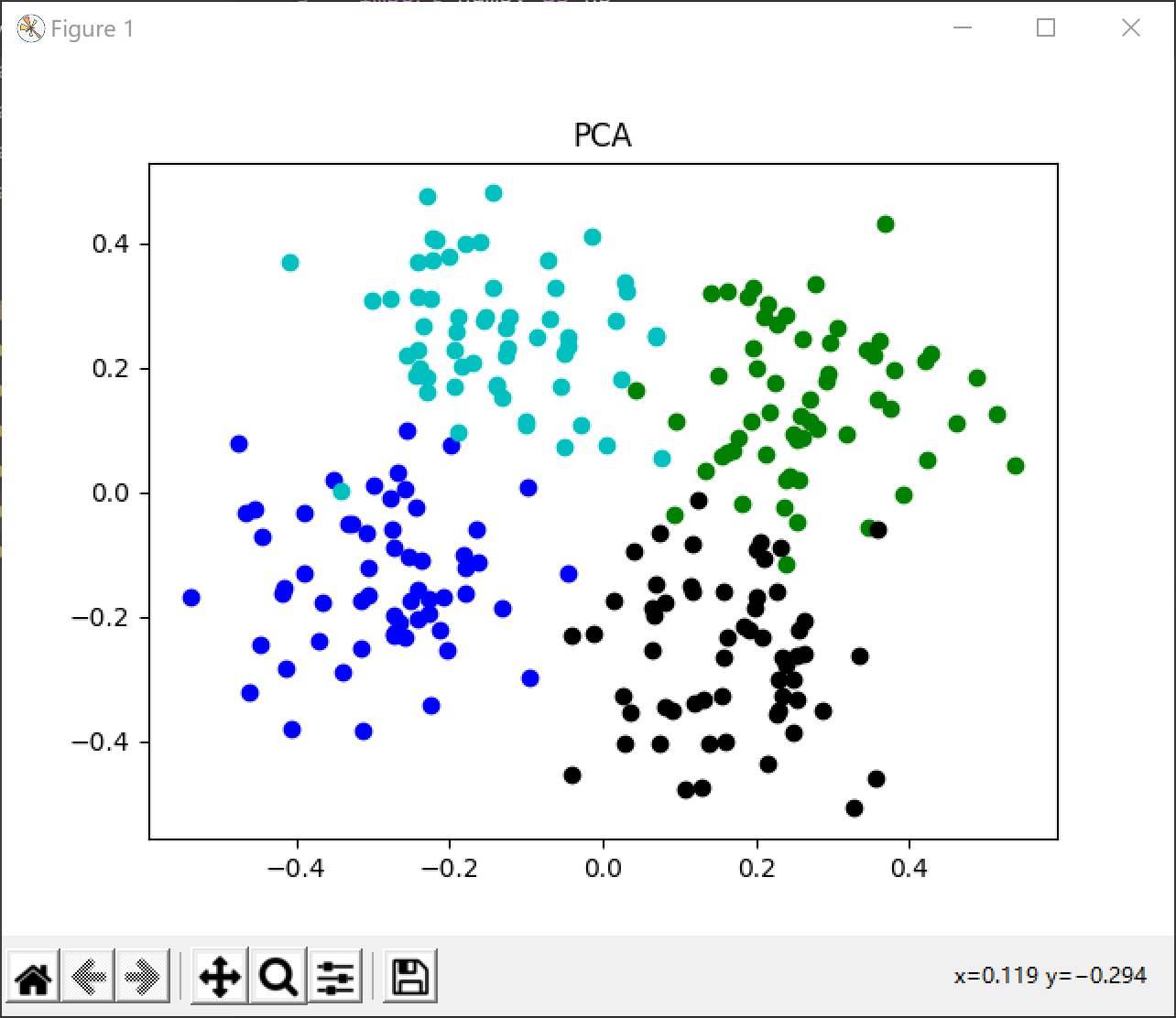
具体来说，样本协方差矩阵为

其中为样本点数，为点集向量矩阵。

对其做SVD，得到满足

其中从上到下由大到小排列，取的前列生成向量空间，将点集投影到这个空间中

即可得到最终结果，结果如下图所示

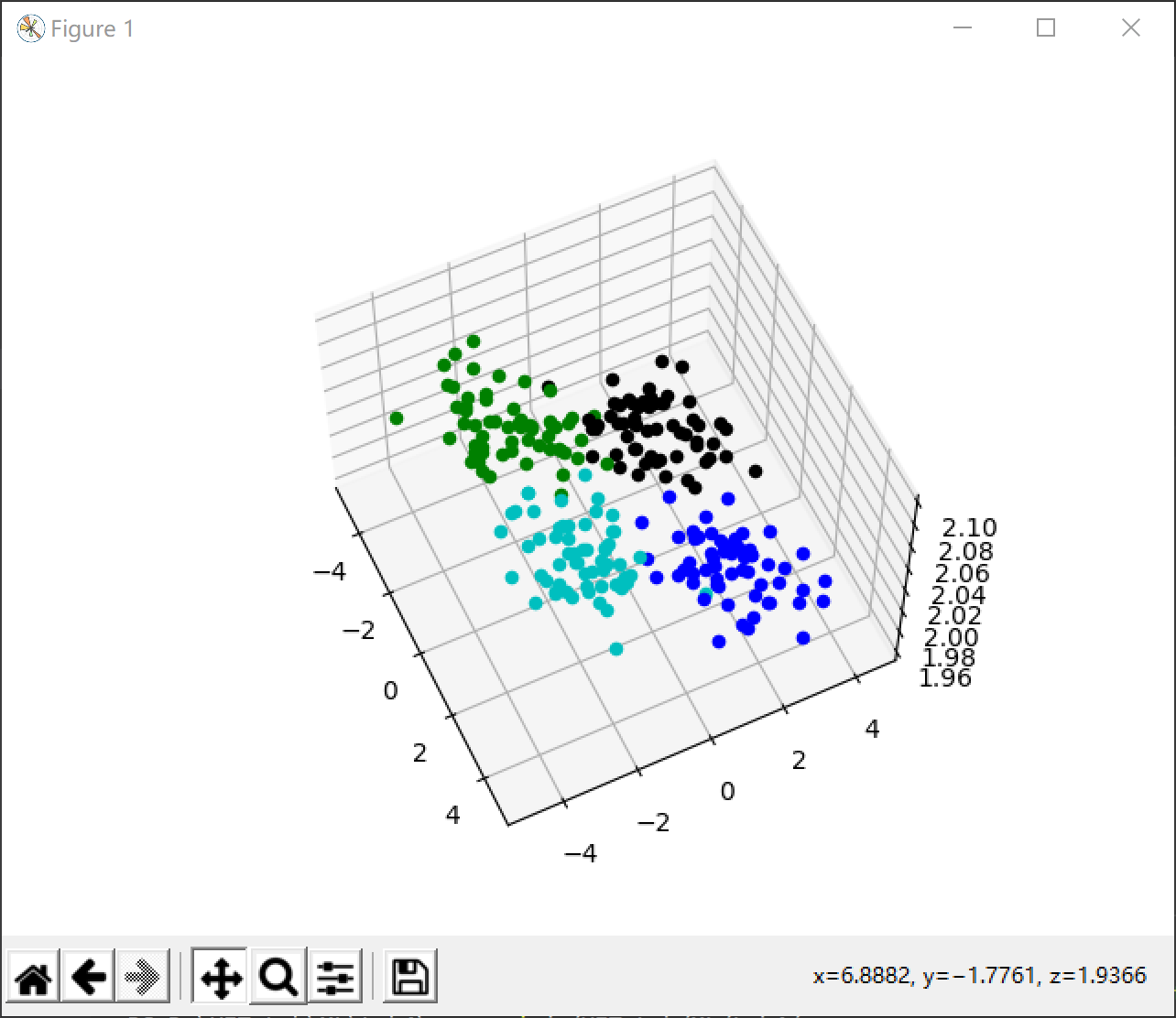


可以发现，执行过PCA的数据依然有较好的区分度。

## 数据还原

对于数据还原，我们执行上述操作的逆变换，即

对上述数据做逆变换，结果如图



可以看到起到了较好的还原效果，但方差较小的第三维的还原效果欠佳。

# 应用