**深度学习实验报告（一）**

**一、PCA简介**

主成分分析的主要目的是希望用较少的变量去解释原来资料中的大部分变异，将我们手中许多相关性很高的变量转化成彼此相互独立或不相关的变量，从而达到降维的目的。在原始数据“预处理”阶段通常要先对它们采用PCA的方法进行降维。本质上讲，PCA就是将高维的数据通过线性变换投影到低维空间上去，但并非随意投影，而是需要遵循一个规则：希望降维后的数据不能失真，也就是说被PCA降掉的那些维度只能是噪声或是冗余的数据。

**二、PCA基本原理及方法步骤**

**1、自编码器角度**

考虑一个数据降维任务，目标在于保留原始数据的重要特征的同时降低数据的维度。尤其是在传统PCA方法中，目标是把数据通过一个线性函数嵌入到一个低维度的空间中，使得数据方差最大化。也就是说，我们想要找到线性函数和使得和之间的距离最小化。通常，我们令和，这里，，而且假设函数不存在截距项。事实上，PCA本质上就是要极小化平方损失函数：

事实上，PCA是一种比较特殊的自编码网络，它也被称为不完全自编码器。

**2、特征向量角度**

假设现有p个n维随机变量作为样本，将样本写成矩阵形式，有



令，求解，其中，且有限制式，。

是所对应的特征向量，是*X*的协差阵S的第k个特征值，且。由定义，可得：



其中。若使得，因此有

基于上述样本PCA的原理，总结样本PCA的具体步骤，如下：

STEP 1 将初始数据进行标准化，可得到处理后的矩阵*Y*。

STEP 2 对新得到的矩阵*Y*求协差阵∑。

STEP 3 得出∑的特征根与特征向量。由得特征向量，且，因此可得到。

STEP 4 根据累计方差贡献比率确定主要成分。由，前k个主成分的累计方差贡献率:。通常的，，此时取前k个成分，有

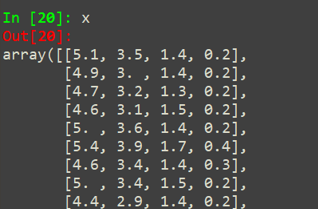
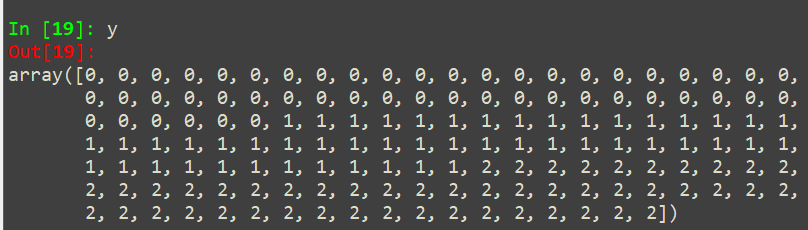


**三、基于python的实验分析**

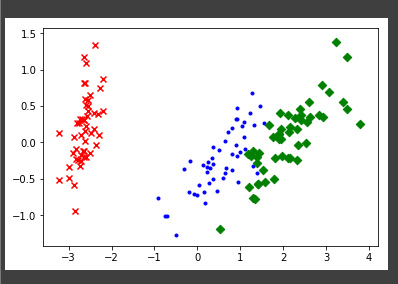
1、实验目标：通过Python的sklearn库来实现鸢尾花数据进行降维。

2、实验过程

首先导入相关库(matplotlib库用于可视化，sklearn库中自带PCA算法包以及可操作的数据集)。查看鸢尾花数据集，该数据集样本量为150,具有4个特征和3种label（鸢尾花的三种颜色）



之后利用Python中PCA算法包进行降维将数据属性压缩成两个，并可视化得到下图：



从上图中可以清楚地看到数据被压缩到二维平面中，在二维平面上可以较好展示各类数据的聚集分布情况。

**四、代码**



