

创新创业学院实践报告

实践班名称： 人工智能实践班

课 程 ： 人工智能研究-上

题 目 ： 浅谈PCA的应用

院 系： 机械工程学院

班 级： 机创1801班

学生姓名： 陈可涵

学 号： 201862004

完成日期： 2019 年 11 月 26 日

大连理工大学创新创业学院

1. 概要

本学期主要学习了主成分分析(PCA—Principal Component Analysis)这种降维方式。本文将对PCA的主要知识进行一次回顾与梳理并且在Python中用Scikit-learnl对鸢尾花的数据进行一个简单的PCA操作来作为应用举例。

1. PCA(Principal Component Analysis)要点回顾
2. 引言

在机器学习领域主要有“监督学习”(Supervised Learning)、“无监督学习”(Unsupervised Learning)、“强化学习”(Reinforcement Learning)这三种方式。在监督学习中用来训练的所有数据都是有标签的，如果把原始数据比作一道数学题，那么他的标签就相当于是这道题目的标准答案，而我们要做的就是给模型大量的有标签的数据即“训练集”让该模型反复“训练”从而掌握数据与标签之间的映射关系，从而获得能够预测新的数据即“测试集”中的数据的能力。大一学年学习的线性回归就是一个典型例子。

在无监督学习中，数据是没有标签的，我们也不是为了做预测，而是为了能够获得数据的一些结构一些比较隐秘的关系，比如上学期学的K-means就是一个典型的例子。

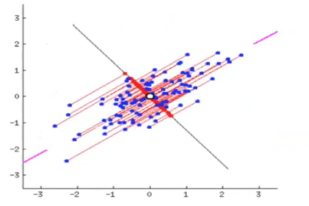
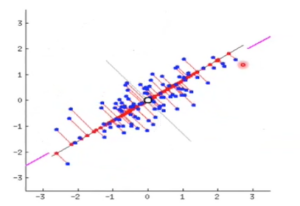
然而，随着样本的维数不断增加，模型训练的时间复杂度将会增大，训练会变得更加困难即“维数灾难”，为了解决这个问题我们需要一种降维手段，使得样本的位数能够减小并且还应当能够尽可能地保持数据的信息及其拓扑关系

1. 什么是PCA？

PCA，即主成分分析法就是一种无监督学习中的线性降维方式，是一种能够尽可能忠实的再现原始数据的所有信息的降维方法，换言之通过PCA降维能够较好的保留数据之间的拓扑关系。例如,下方有三张图，图(a)是一张《蒙娜丽莎的微笑》图中包含的信息量十分巨大,

1.  (b) (c)

比如它的颜色信息是RGB三通道图像，每一个像素点都是三个0-255的数；比如它的面部五官信息，背后的风景等等。降维要做的就是尽可能的减少一些信息但是又能让人们看出来这是《蒙娜丽莎的微笑》，如果按照(b)降维，显然失去了一些主要的信息，维度虽然降低了但是数据的信息也受到了很大的影响，但是如果按照(c)降维，这就正是我们想要的结果了。而所谓线性降维就是对于，寻找一线性变换得到。如下图所示，在二维空间中，线性降维就相当于是把一些离散的点投影到某根直线上，PCA做的工作就是寻找出一根合适的轴，使得数据点投影到该轴上以后相互之间的距离最大如(e)，而不是“糊成一团”如(d)

 (d) (e) zx

三、具体推导（二维空间向一维向量投影）

我们知道方差可以反映数据相对于均值的偏离程度，因此就希望投影得到的数据方差应该尽可能的大。

均值：

目标：

化简：

其中：即的协方差。而相当于对数据进行中心化操作，因此不妨记则有。

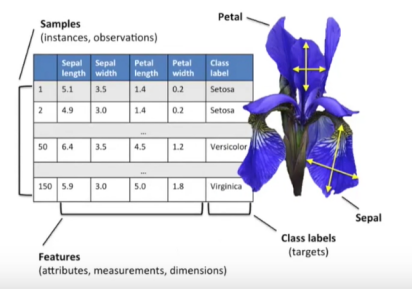
因为要使和尽可能的相似，满足是的正投影这一假设应有即。所以由拉格朗日乘数法可构造：

求导，化简可得：

所以

由此可以得出当等于数据的协方差矩阵的最大特征值对应的特征向量时可以使得投影过后的数据点方差最大，该特征向量也就是第一主成分。

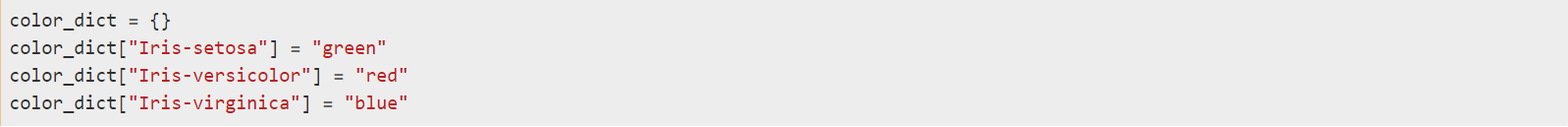
四、应用举例

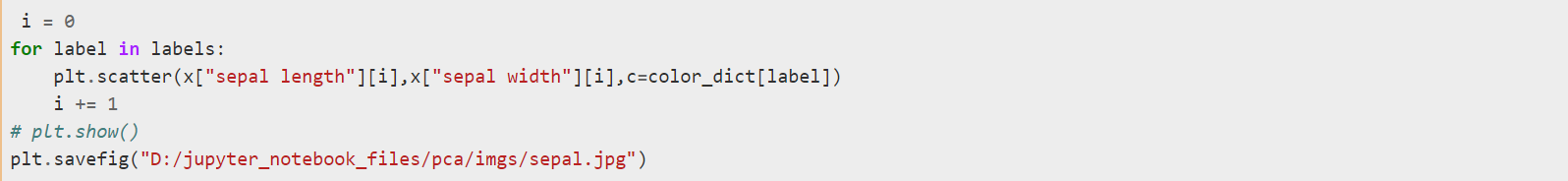
PCA的应用十分广泛，Google Scholar上可以查到诸如人脸识别、牛肉识别、基因数据聚类等应用。由于本人水平有限，难以对一些复杂的项目进行PCA降维操作，好在Python的scikit-learn库中自带Iris的数据并且有封装的PCA函数，非常适合用来实战。这是官方给出的Iris数据集，总共包含三种类别：Iris-setosa、

Iris-versicolor、Iris-virginica。每一种鸢尾花都有四个特征即：Sepal length、Sepal width、Petal length、Petal width。



可以在Python中直接下载获得并通过Pandas读取出相关信息：

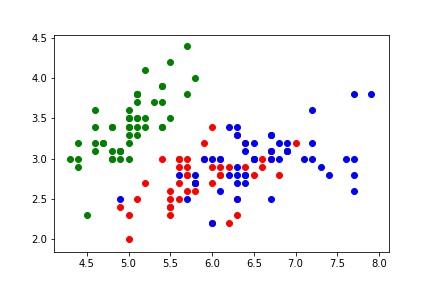
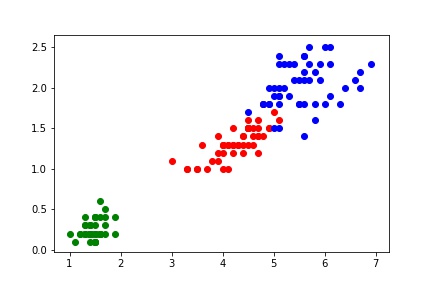


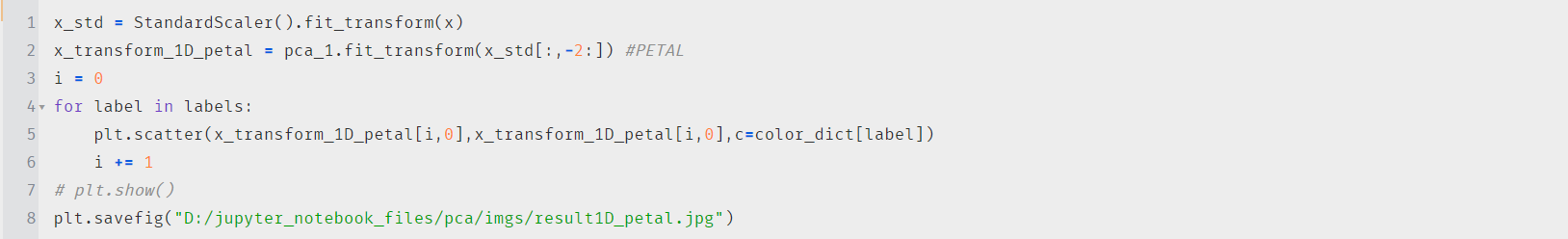
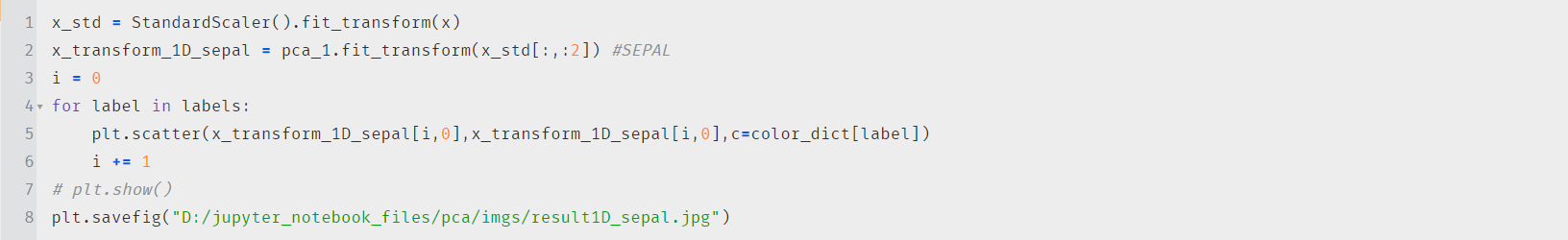
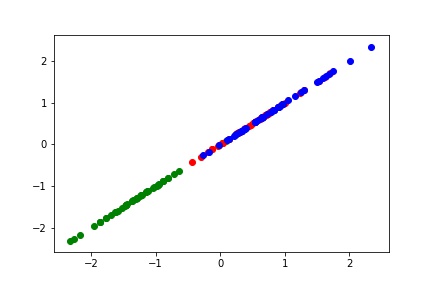
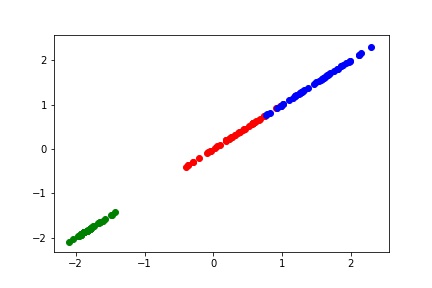


获得了鸢尾花的数据，我们不妨看一下在使用PCA 之前它的四种特征是一个什么样的分布：

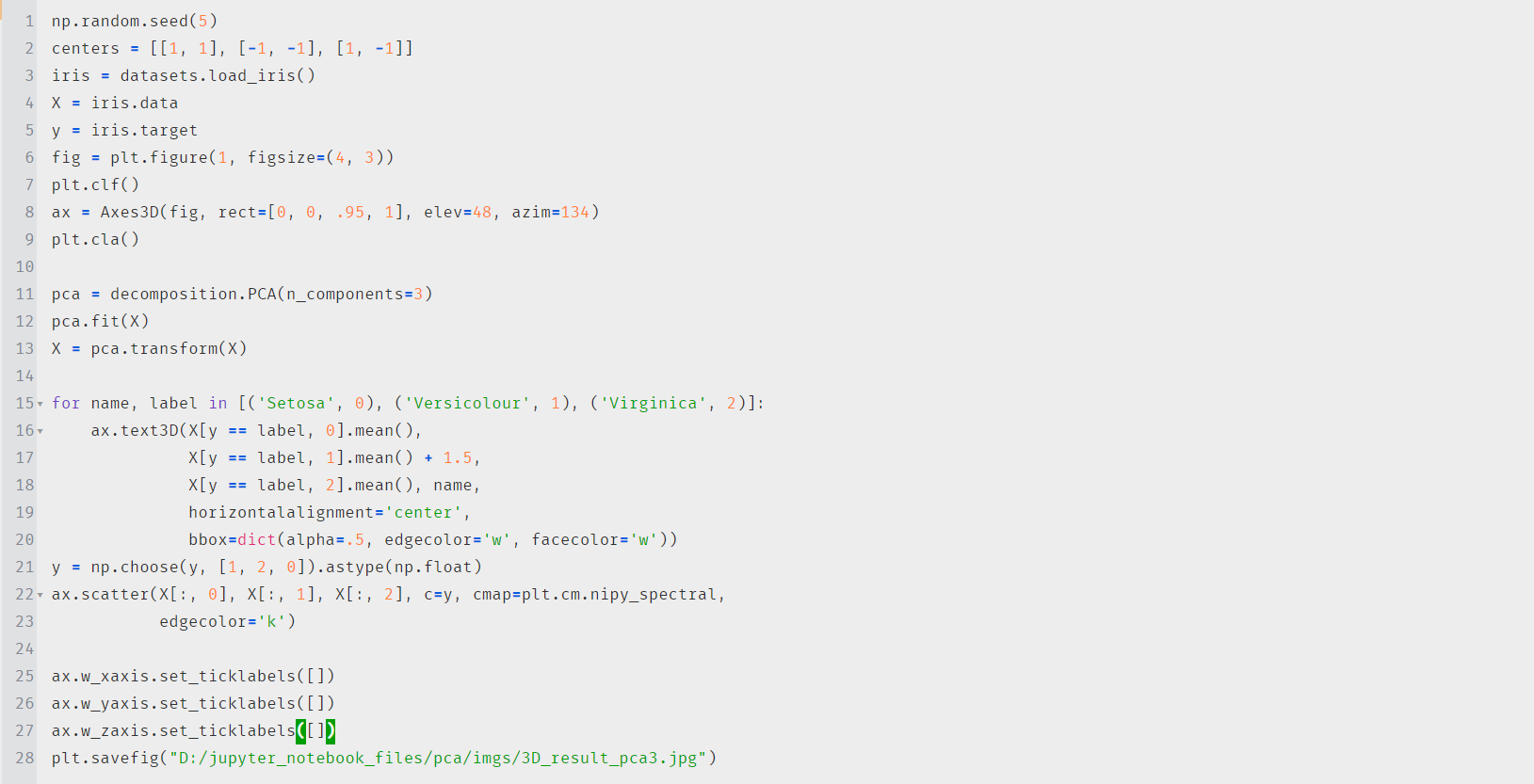
首先，生活常识告诉我们，Sepal length和Sepal width都是关于sepal的信息，计算机或许会将它们看作孤立的数据，但是我们不妨先将他们看作一种特征，因此他的维度就是2这样就可以在二维平面内将其可视化，如下图，横坐标代表length，纵坐标代表width；绿色代表Iris-setosa、红色代表Iris-versicolor、蓝色代表Iris-virginica。其实对于这两幅图我们可以很容的看出来进行了PCA之后这些点会被投影到那一条直线上，下面就来验证一下：

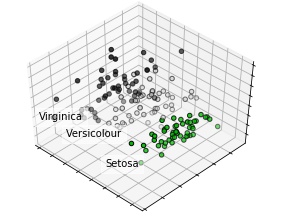
Petal Sepal

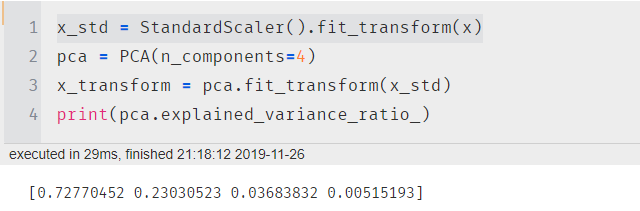


首先将数据标准化，使其尽可能分布在[-1,1]之内，再使用PCA，得到结果如作图（左一为Petal、左二为Sepal）和预计的结果几乎一致。

接下来，不妨再以计算机的角度思考一下，如果把这忽略length与width的关系，那么Iris的数据就是四维的，这样我们是难以进行可视化的，但是如果我们能将这个数据降到三维，也就是利用PCA选出3个主成分，把四维的数据投影到一个三维的空间坐标系中就可以利用可视化的一些手段来观察这些数据了。

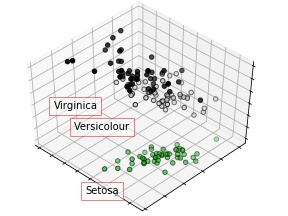






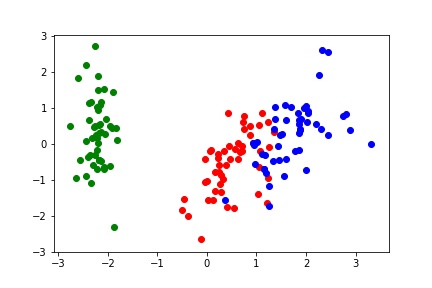
运行结果如图，不难发现原本难以想象的四维空间经过PCA降维，映射到了三维空间并且体现出了一定的层次结构。Python中可以利用explain\_variance\_ratio命令获得每一个成分的方差占总体的百分比，换言之能告诉我们4个成分中到底哪三个才是“主成分”，显然，在本例中，主成分是Sepal length、Sepal width、Petal length，那么，不妨看一下在降维之前这三者在空间是怎么分布的。

得到结果如下图：



显然，经过PCA操作后，原本混在一起的Iris-versicolor和Iris-virginica被分开了。

那么，如果只选择两个特征作为主成分呢？那就应当是把四维空间的点映射到一个平面上，如下图：



综上所述：

本例利用PCA对鸢尾花的数据进行了分析，所谓主成分分析在此应用中可以理解成从Sepal length、Sepal width、Petal length、Petal width这四个信息中筛选出最能用来判断鸢尾花种类的成分，从而起到降维作用。至于在具体应用过程中降到几维还需根据具体情况而定。

1. 总结

本例最后将四维的数据降到了二维，其实可以发现效果并不是很好，红点与蓝点之间的决策边界并不是很明显。我想在成这个问题的主要原因可能是PCA采用的是线性降维，对于数据重叠的处理能力不是很好，即PCA不能很好地处理弯曲状的数据，如果那PCA来处理“瑞士卷”想必结果会更糟糕。在Google Scholar上有一些“Kernel PCA”的文章，提供了一种利用核函数进行非线性降维的方法。这种核方法是将训练样本进行非线性函数变换，在变换后的特征空间中进行主成分分析。例如对于一些难以直接利用PCA进行分析的二维数据，可以先将其升维，在高维空间里找到更合适的平面去投影这些数据，即在高维空间使用PCA。当然还有拉普拉斯特征映射等方法，此处不再展开。