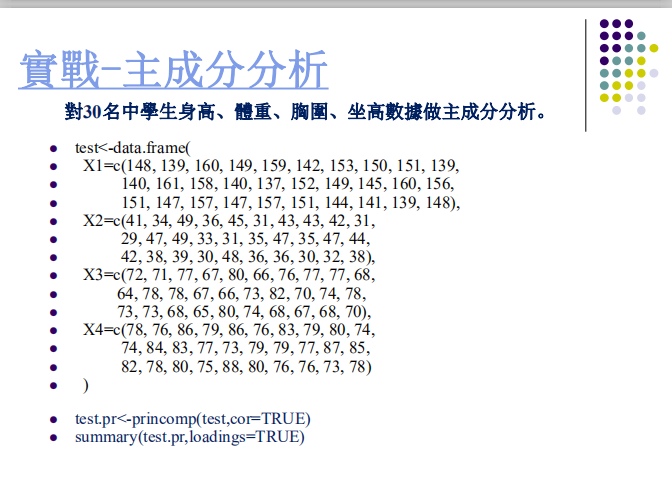
1、题目要求

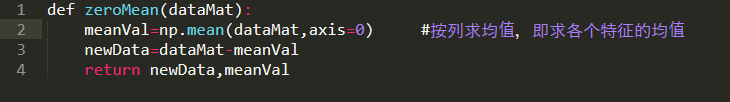


解析：

1. PCA算法实现【语言：python,函数库：numpy】

①零均值化

假如原始数据集为矩阵dataMat,dataMat中每一行代表着一个样本（对应本题c中的每一个数据），每一列代表同一个特征（对应本题的身高、体重等）（所以每一行就是一个特征向量，一个特征向量就是一个样本）。零均值化就是求每一列的平均值，然后该列上的所有数都减去这个均值。也就是说，这里零均值化是对每一个特征而言的。



函数中用numpy中的mean方法来求均值，axis=0表示按列求均值。

该函数返回两个变量，newData是零均值化后的数据，meanVal是每个特征的均值，是给后面重构数据用的。

②求协方差矩阵



numpy中的cov函数用于求协方差矩阵，参数rowvar很重要！若rowvar=0，说明传入的数据一行代表一个样本，若非0，说明传入的数据一列代表一个样本。因为newData每一行代表一个样本，所以将rowvar设置为0。

covMat即所求的协方差矩阵。

③求特征值、特征矩阵

调用numpy中的线性代数模块linalg中的eig函数，可以直接由covMat求得特征值和特征向量：



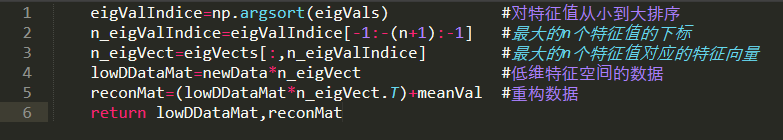
eigVals存放特征值，行向量。

eigVects存放特征向量，每一列代表一个特征向量。

特征值和特征向量是一一对应的

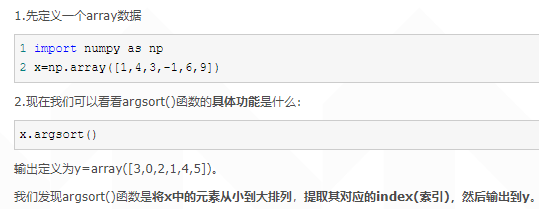
④保留主要的成分[即保留值比较大的前n个特征]

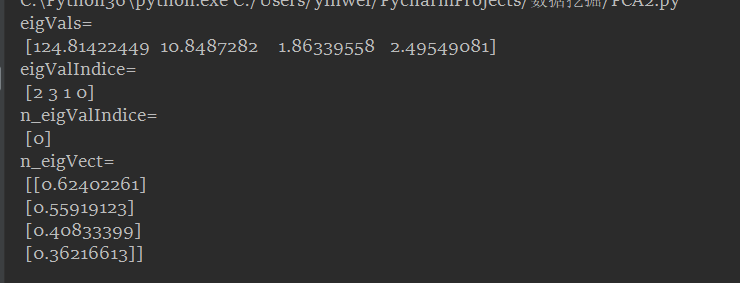
第三步得到了特征值向量eigVals，假设里面有m个特征值，我们可以对其排序，排在前面的n个特征值所对应的特征向量就是我们要保留的，它们组成了新的特征空间的一组基n\_eigVect。将零均值化后的数据乘以n\_eigVect就可以得到降维后的数据。代码如下



代码中有几点要说明一下，首先argsort对特征值是从小到大排序的，那么最大的n个特征值就排在后面，所以eigValIndice[-1:-(n+1):-1]就取出这个n个特征值对应的下标。【python里面，list[a:b:c]代表从下标a开始到b，步长为c。】（相当于list[::-1]）

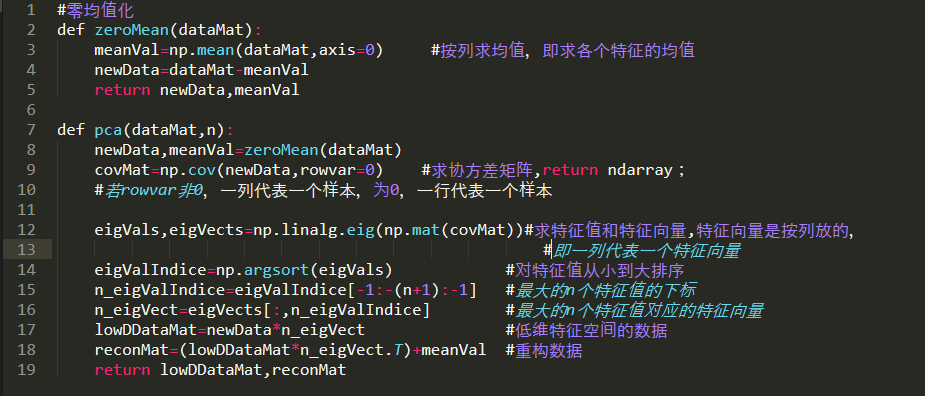
所以其实np.argsort()是对特征值对应的下标进行排序。

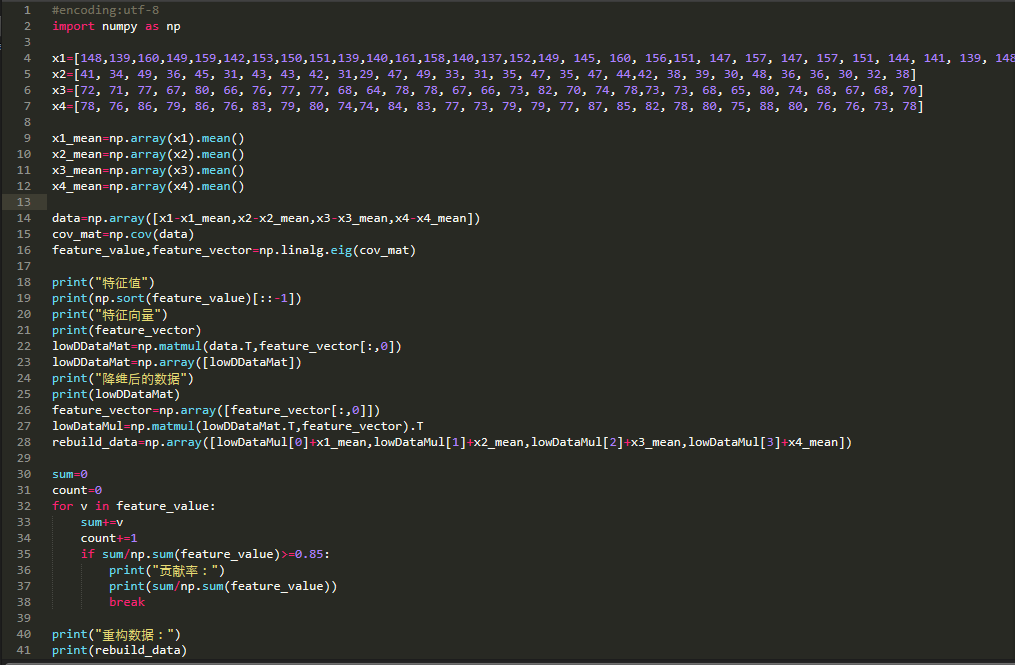




reconMat是重构的数据，他是降维数据乘以n\_eigVect的转置矩阵，再加上均值meanVal得到的。

OK，这四步下来就可以从高维的数据dataMat得到低维的数据lowDDataMat，另外，程序也返回了重构数据reconMat，有些时候reconMat课便于数据分析。





特征降维形象店说就好比，矩阵中有两行是线性相关的，通过特征降维，只会保留其中一行；你要给一个三维盆栽拍照片，你可以从上下左右45度等角度拍二维照片，有的角度得到的信息多，更能表现原物体特征，但有的角度得到的信息就少。

利用PCA进行特征选择时，首先要把原来的特征空间做了映射，使得新的映射后的特征空间数据彼此正交，这样做就尽可能保留下具备区分性的低维数据特征。