《系统工程导论》第五次作业
主成分分析

2017011010 杜澍淦 自71

题目1（10 points）

使用PCA和线性回归对附件的数据进行建模。附件的数据为美国1992年总统竞选各个 county 的投票情况，数据来源 http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/ 请将从pop.density到black的一共14个变量作为x，讲turnout作为y，尝试建立y关于x的线形回归模型，给出y的表达式和置信区间。（1）使用PCA+线性回归建模；（2）直接使用病态回归模型建模，比较两种方法的结果

要求：

1. 实现PCA 算法，具体要求如下

(1) 实现函数（以 MATLAB 函数为例）

```
function [pcs, cprs_data, cprs_c] = pca_compress(data, rerr)
```

其中输入输出变量含义如下

变量名	含义
data	输入的原始数据矩阵，每一行对应一个数据点
rerr	相对误差界限，即相对误差应当小于这个值，用于确定主成分个数
pcs	各个主成分，每一列为一个主成分
cprs_data	压缩后的数据，每一行对应一个数据点
cprs_c	压缩时的一些常数，包括数据每一维的均值和方差等。利用以上三个变量应当可以恢复出原始的数据

(2) 实现函数（以 MATLAB 函数为例）

```
function recon_data = pca_reconstruct(pcs, cprs_data, cprs_c)
```

其中输入输出变量含义如下

变量名	含义
pcs	各个主成分，每一列为一个主成分
cprs_data	压缩后的数据，每一行对应一个数据点
cprs_c	压缩时的一些常数，包括数据每一维的均值和方差等。利用以上三个变量应当可以恢复出原始的数据
recon_data	恢复出来的数据，每一行对应一个数据点

2. 线性回归相关函数请使用前两次作业自己编写的函数；如果对自己编写的

函数不置信，可以使用工具包中现成的线性回归函数进行辅助调试，但是最终请使用自己变写的函数进行线性回归

3. 关于PCA 部分资料or 代码在计算协反差矩阵时会除以n-1 而不是课程中介绍的n，在本次作业请以课件中为准进行实现。如果按照n-1 进行计算，需要在报告中说明+给出这两者的区别。

【解答】

一、 PCA+线性回归建模思路：

（1） 首先要判断自变量之间是否线性相关（判断是否是病态问题），这是通过判断XX^T是否有约为 0 的特征值来进行的，如果有这样的特征值则说明自变量之间存在线性相关关系；

（2） 如果（1）中确定该问题是病态问题，则利用主成分分析法对自变量进行压缩，通过函数 pca_compress( )的处理得到主成分 pcs；

（3） 做多元函数非病态拟合；

（4） 通过函数 pca_reconstruct( )的处理由主成分进行数据还原

（5） 进一步得到回归方程。

二、 结果对比及分析

使用病态回归模型建模的结果：

```
病态线性回归（显著性水平α=0.05）：
此问题是病态线性回归问题，需要从14维降至10维
F=208.45291，Fa=1.69496，F>Fa，即存在线性关系
回归方程为y =
19.5890644346
-0.0003788167 * x1
-0.0000021578 * x2
-0.0014961083 * x3
+0.6077045854 * x4
 +0.6804618566 * x5
 -0.0004206165 * x6
+0.3297450554 * x7
 +0.0002528184 * x8
 +0.1611176211 * x9
 -0.0001676262 * x10
-0.0961458256 * x11
+0.1556694170 * x12
 +0.0551412104 * x13
 -0.0304527729 * x14
置信区间为（y-10.72359，y+10.72359）
```

使用 PCA+线性回归建模的结果：

```
主成分分析（显著性水平α=0.05）：
利用主成分进行非病态的多元线性回归的中间结果：
此问题不是病态线性回归问题
F=292.21075，Fa=1.83375，F>Fa，即存在线性关系
回归方程为y =
0.0000000000
-0.2001058082 * x1
-0.2037152129 * x2
-0.2453011415 * x3
+0.0443960685 * x4
 +0.0577250009 * x5
 -0.1802209779 * x6
-0.2795066469 * x7
+0.1213597700 * x8
 -0.1915926382 * x9
-0.0042084432 * x10
置信区间为（y-1.40882，y+1.40882）
```

```
最终结果：
回归方程为y =
19.5890644346
-0.0003788167 * x1
-0.0000021578 * x2
-0.0014961083 * x3
+0.6077045854 * x4
 +0.6804618566 * x5
 -0.0004206165 * x6
+0.3297450554 * x7
 +0.0002528184 * x8
 +0.1611176211 * x9
 -0.0001676262 * x10
-0.0961458256 * x11
+0.1556694170 * x12
 +0.0551412104 * x13
 -0.0304527729 * x14
置信区间为（y-10.72359，y+10.72359）
```

可以看到两种方法得到的最终结果完全一致。两种建模方法都需要首先对变量进行归一化，得到单位正交矩阵，目的是获得一组单位正交基作为实际建模的变量，保证处理后没有互相线性相关的自变量。两种方法都是通过取XX^T的前m个特征值对应的特征向量来确定这组基，因此去病态处理就相当于主成分分析，后续都进行多元函数拟合，两者方法的一致性保证了结果的一致性。