

《系统工程导论》第五次作业

主成分分析

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

- 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
- 关于 PCA 部分资料 or 代码在计算协方差矩阵时会除以 $n-1$ 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。

其他要求：

1. 独立完成作业，不能抄袭
2. 提交报告 pdf，命名为学号+姓名+第几次作业，不接受 word 版
3. 提交代码文件 or 讲代码附在报告后面；
4. python 请提交.py 文件，不要交.ipynb 文件
5. ddl：5 月 11 日晚 12 点