《系统工程导论》

第五章作业

主成分分析 PCA

姓名：卢志

班级：自 43 班

学号：2014011497

2016 年 4 月 11 日

1. 题目一

(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

答：按照作业要求，题目一部分详见作业代码，不作具体说明。

2. 题目二

2.1 题目描述

利用上面编写的函数，以及前一章作业中编写的函数，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，数据说明如下。

Name	Labels	Storage
county		character
state		character
pop.density	1992 pop per 1990 miles^2	double
pop	1990 population	double
pop.change	% population change 1980-1992	double
age6574	% age 65-74 1990	double
age75	% age >= 75 1990	double
crime	serious crimes per 100000 1991	double
college	% with bachelor's degree or higher of those age>=25	double
income	median family income 1989 dollars	double
farm	farm population % of total 1990	double
democrat	% votes cast for democratic president	double
republican	% votes cast for republican president	double
Perot	% votes cast for Ross Perot	double
white	% white 1990	double
black	% black 1990	double
turnout	1992 votes for president / 1990 pop x 100	double

数据来源： http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/

请将从 pop.density 到 black 一共 14 个变量作为 x，将 turnout 作为 y，试建立 y 关于 x 的线性回归模型,给出 y 的表达式和置信区间(为书写方便，可以在有明确说明的情况下只给出 y 表达式中的系数和置信区间的半长度)。

提示：可以利用所学的知识检查自变量之间是否有线性相关关系，利用 pca 对自变量进行压缩后即可认为消除了自变量之间的线性相关。

2.2 解题过程和思路

根据题意，可以归纳出解题的过程如下：

1）首先需要判断 n(n=14)个自变量之间是否线性相关，从而判断是否为病态问题。而这一步骤中，根据第四章黑箱建模的结论，相当于判断矩阵 XX^T 是否接近奇异。而判断的方式是求出其特征根，根据阈值判断是否有约为 0 的特征根。如有，则说明 n 个自变量之间存在线性相关。

2）如果是病态问题，则需要对自变量进行压缩，即使用 PCA 主成分分析法对自变量 X 进行压缩，根据函数 pca_compress 的定义，可以生成维数较小的矩阵

pcs，即为主成分 Y。由 PCA 方法的原理，此时的 Y 中元素之间是线性无关的。可以进行后续处理。

3）利用非病态的多元线性回归方法，以 Y 为自变量，y 为因变量，做多元回归函数非病态拟合。

4）最后，将 Y 反变换为 X，得到 y 关于 X 的回归方程，解答完毕。

注意：在求解过程中需要进行一定的规范化和反规范化处理，不能忽略。

2.3 计算结果

本次作业重点理解病态线性回归拟合问题和主成分分析法之间的关系，在这里，使用第四章作业的代码对问题进行分析，计算结果如下图所示：

```
方法一：使用带病态处理方法的回归模型建模：
此问题是病态线性回归问题！由14维降至10维
√显著性水平 α = 0.050 条件满足
回归方程为 y =
19.589064434611402
- 0.000378816652531 * x1
- 0.000002157812840 * x2
- 0.001496108306704 * x3
+ 0.607704585360747 * x4
+ 0.680461856553134 * x5
- 0.000420616476178 * x6
+ 0.329745055391723 * x7
+ 0.000252818439003 * x8
+ 0.161117621057526 * x9
- 0.000167626173550 * x10
- 0.096145825596999 * x11
+ 0.155669416960756 * x12
+ 0.055141210429796 * x13
- 0.030452772857135 * x14
置信区间为 (y-10.724 , y+10.724)
```

而利用 2.2 节思路可对问题进行主成分分析，再进行非病态的多元线性回归处理。计算结果如下：

```
方法二：使用主成分分析方法去线性相关建模：
√显著性水平 α = 0.050 条件满足
回归方程为 y =
19.589064434611377
- 0.000378816652531 * x1
- 0.000002157812840 * x2
- 0.001496108306704 * x3
+ 0.607704585360747 * x4
+ 0.680461856553135 * x5
- 0.000420616476178 * x6
+ 0.329745055391722 * x7
+ 0.000252818439003 * x8
+ 0.161117621057526 * x9
- 0.000167626173550 * x10
- 0.096145825596999 * x11
+ 0.155669416960756 * x12
+ 0.055141210429796 * x13
- 0.030452772857136 * x14
置信区间为 (y-10.724 , y+10.724)
```

对比分析：由以上结果可知，两种方法得到的结果完全一致。究其原因，这是必然的。二者都是在过程中进行了变量的归一化，期间得到了单位正交矩阵，目的是获得一组单位正交基作为实际建模的变量，由此保证建模过程中没有自变量互相线性相关。而且取的都是 XX^T 的前 m 个特征值所对应的特征向量，所以说去病态处理就相当于主成分分析，后续均进行多元函数拟合处理。所以说二者的方法是基本一致的。

2.4 遇到的问题和解决方案

1. **问题**：对题意理解不明，PCA 和病态回归到底是不是一样的原理？

   解决：通过阅读讲义上的推导，并且加以理解，其实二者原理一样，相当于把病态回归中的 Z 作为主成分 Y 而已。但是两者的目的不同，一般而言，PCA 用来数据压缩，更关注压缩后的主成分点集。而病态回归重在回归函数的获得。

2. **问题**：在主成分分析方法中，进行线性回归之后，如何将其映射回 x 域。一

开始的时候，我认为这是很难做到的，因为变换矩阵Q为(n×m)，是将n维变量变为m维变量。直接的思路是将其直接乘以Q的逆，进行反变换。但是在编程中求解一个非方阵的逆是没有途径的。

解决：在仔细研究原理后，我惊奇地发现原来$QQ^T = \mathrm{I}$，这是根据PCA变换矩阵每一列向量都是单位正交列向量的性质，其左乘或右乘其自身的转置，得到的都是单位矩阵，所以其转置即为其逆矩阵。基于此，可以得到n维向量。

3. **问题**：选取阈值时不知如何选取？

   解决：因为矩阵维数的不同会使阈值发生变化，不同的阈值对于舍弃的特征根的数量造成影响，从而导致降维不一致。所以需要结合实际矩阵的大小，调整阈值。

2.5 作业总结

本次作业的实质是数据压缩，将更高维的成分因素通过PCA的方法压缩为较低维的主成分因素。而在编程实现时，还要求我们实现压缩和反压缩两个过程。

在实现中也会遇到一定的困难，因为这两个过程之间并不是严格互逆的。因为压缩时，往往会丢掉一些接近于0的特征值，但这些特征值不是真的为0，所以出现了一定的损失。但这些损失是可以容忍的，这是为了更进一步降低维度，拥有更好的变量独立性。

通过这一次作业中的对比试验，我还对比了PCA压缩回归和基于病态的拟合回归之间的差别。正如胡老师上课所说的，这两者的原理是基本相同的，所以最后的结果也是相同。其实，换句话说，主成分分析就是处理病态回归问题的方法！在本次作业中，得到了很好的验证。

2.6 原始代码

见工程文件，主函数为HW5.m，点击运行即可。其余函数功能可见代码注释。