主成分分析在线性回归中的作用

何舜成

2015 年 5 月 11 日

1 PCA与线性回归去病态的关系

PCA是一种常见的数据降维、提取特征的方法，主要思想是消除不同变量之间的相关性，忽略影响较小的变量，从而消除一些冗余的，或者不重要的数据。当PCA应用在数据压缩上时，是一种有损压缩，不能精确还原原始数据。

PCA本质上是从数据的协方差矩阵入手，将协方差矩阵$\Sigma = X^T X/(n-1)$作正交分解，得

$$\Sigma = USU^T, where U^T U = I_n \tag{1}$$

剔除较小的特征值，将n维数据降维为m维，并得到变换矩阵U_m，使得

$$Z = U_m^T X \tag{2}$$

Z即为降维后的数据，保留了大部分信息，其误差由舍去的特征值与所有特征值之和的比值决定。压缩前应当设定该误差限。

多元线性回归中去除病态（亦即去除线性相关变量）是也是考虑$X^T X$的特征值，去除绝对值较小的特征值（例如小于0.1的特征值），以免在求逆时不稳定甚至无法求逆。

可以看到，PCA压缩数据和多元回归中病态的去除方法是相同的，都是去除协方差矩阵中的小特征值，只是目的分别是降低数据维度和去除线性相关。所以可以通过PCA求解病态线性回归问题。

2 解题步骤

依照以上分析，可以得到以下计算步骤：

（1）读入数据，确定X和Y，并将其分别作规范化处理，使其成为均值为0，标准差为1的数据；

（2）检验协方差矩阵$X^TX/(n-1)$的特征值，是否存在线性关系；

（3）对数据进行PCA压缩降维，去除线性相关；

（4）对降维后的数据进行线性回归；

（5）进行F检验，计算置信区间；

（6）得到最终表达式。

3 计算结果

以下检验均在显著性水平0.05下进行。

回归方程及置信区间：

$$\begin{aligned} y \quad = \quad & 19.5632 - 3.7900 \times 10^{-4}x_1 - 2.1670 \times 10^{-6}x_2 - 0.0010x_3 + 0.6070x_4 \\ & +0.6799x_5 - 4.1473 \times 10^{-4}x_6 + 0.3305x_7 + 2.5180 \times 10^{-4}x_8 + 0.1639x_9 \\ & +4.3848 \times 10^{-4}x_{10} - 0.0960x_{11} + 0.1540x_{12} + 0.0554x_{13} - 0.0309x_{14} \\ & \pm 10.7150 \end{aligned}$$

F检验结果表明可认为线性关系成立：

$$F = 324.7743 > F_\alpha = 1.8829 \tag{3}$$

PCA压缩选取的相对误差阈值为0.1，数据压缩率$\eta = 64.6\%$。

可以看到，最终的turnout与三个人口数据负相关，与老年人人口正相关，与犯罪率负相关，与受高等教育比例正相关，与收入正相关，与农业人口正相关，与民主党投票率正相关，与共和党投票率负相关，与Perot投票率正相关，与白人比例正相关，与黑人比例负相关。

这些因素里有许多是有关联的，如白人黑人的比例、犯罪率与受高等教育比例、人口数与人口密度，因此直接用所有因素作线性回归是不可取的，需要去除线性相关变量，这就是PCA在这其中的作用。