系统工程导论作业五

PCA

刘若涵 自 05 2020011126

1 方法原理

1.1 PCA 压缩 pca_compress(data, rerr)

1）样本数据规范化

$$\bar{x}_i(t)=\frac{x_i(t)-e(x_i)}{\sqrt{\delta^2(x_i)}} \quad \forall i,t$$

$$\bar{y}(t)=\frac{y(t)-e(y)}{\sqrt{\delta^2(y)}} \quad \forall t$$

2）计算出样本数据的协方差矩阵 XX^T 的特征值、特征向量并按照特征值从大到小排列λ_1到λ_n。

3）逐渐增加 m，计算 $\sum_{i=1}^{m}\lambda_i$，当 $\sum_{i=1}^{m}\lambda_i >= 1-rerr$ 时停止。

4）前 m 个特征值所对应的特征向量 q(1),q(2),...,q(m) 作为主成分方向，构成的矩阵为 pcs。

5）计算各样本数据在主成分方向上的投影

$$y(t)=[q(1),q(2),\cdots,q(m)]^T x(t)$$

压缩后的数据 cprs_data = $(pcs^T * \widetilde{X})^T$。

1.2 PCA 恢复 pca_reconstruct(pcs, cprs_data, cprs_c)

1）得到参数 EX,SX

2）由数据压缩公式得解压缩后 $\widetilde{X}_{recon}=\widetilde{X}*Q_m^T$。

2）去归一化得 recon_data = $\widetilde{X}_{recon}\cdot$ SX + EX = cprs_data * $pcs^T \cdot$ SX + EX。

2 结果分析

2.1 PCA 恢复

对比附件中原始数据 counties.xlsx 与恢复后数据表 recon_data.xlsx 可知，大部分数据成功恢复，量级正确且相对误差较小，个别极端数据恢复效果不太理想，与预期一致。

2.2 PCA + 线性回归模型与病态回归模型对比

1)PCA

alpha=0.500000,rerr=0.050000 情况下压缩数据为 10 维

Fα=0.934384,F=292.210755

F>Fα,x 与 y 存在线性关系

回归方程为：

y=19.589064+-0.000379*x1+-0.000002*x2+-0.001496*x3+0.607705*x4+0.680462*x5+-0.000421*x6+0.329745*x7+0.000253*x8+0.161118*x9+-0.000168*x10+-0.096146*x11+0.155669*x12+0.055141*x13+-0.030453*x14

置信区间为：[y - 3.687970,y + 3.687970]

2)病态线性回归

alpha=0.500000,rerr=0.050000 情况下存在病态问题, m=10

Fα=0.934384,F=292.210755

F>Fα,x 与 y 存在线性关系

回归方程为：

y=19.589064+-0.000379*x1+-0.000002*x2+-0.001496*x3+0.607705*x4+0.680462*x5+-0.000421*x6+0.329745*x7+0.000253*x8+0.161118*x9+-0.000168*x10+-0.096146*x11+0.155669*x12+0.055141*x13+-0.030453*x14

置信区间为：[y - 3.687970,y + 3.687970]

对比可知，两种方法得到的结果完全一致。

turnout = 19.589064 - 0.000379pop.density - 0.000002pop - 0.001496pop.change + 0.607705age6574 + 0.680462age75 - 0.000421crime + 0.329745college + 0.000253income + 0.161118farm - 0.000168democrat - 0.096146republican + 0.155669Perot + 0.055141white - 0.030453black

实际上，由第一部分“PCA 数据压缩”分析可知，PCA+线性回归与病态线性回归过程完全一致，只是某些步骤的描述不同，因此得到相同的输出结果是正确的。

2.3 协方差矩阵的计算

代码中，计算方差所用公式为 std(X, 0, 1)，其中 0 代表分母选择为 n-1。

根据统计学知识可知，当分母为 n-1 时为方差的无偏估计，分母为 n 时为方差的最大似然估计。本次编程中希望偏差最小，因此选择方差的无偏估计，而如果考量均方误差，则应选取最大似然估计。