多元统计分析—R 与 Python 的实现

回归

吴喜之

June 28, 2019

交叉验证

所谓交叉验证就是用一部分数据建立模型, 这部分数据集合称为训练集 (training set); 用另一部分数据检验该模型对这部分数据的预测精度, 这部分数据集合称为测试集 (testing set). 交叉验证中预测精度不好的模型肯定不是好模型.

训练集和测试集的确定方法有很多. 常用的一种就是m 折交叉验证 (m-fold cross validation), 它把数据随机分成 m 份 (折), 各折的观测值数目尽量相等. 然后做 m 次验证, 每次轮流用 1 份数据作为测试集, 其余 $m-1$ 份作为训练集, 用训练集建模, 再用根据训练集所建立的模型拟合测试集. 这样, 在 m 次验证之后得到 m 次预测误差, 最终得到交叉验证的平均误差.

交叉验证的误差度量

回归交叉验证的误差通常基于残差平方和 (SSE) 或均方误差 (MSE). 记训练集的自变量矩阵为 $\boldsymbol{X}_{train}$, 训练集的因变量为 $\boldsymbol{y}_{train}$, 记测试集的自变量矩阵为 $\boldsymbol{X}_{test}$, 测试集的因变量为 $\boldsymbol{y}_{test}$.

交叉验证的残差平方和为 (下面假定进入测试集的观测值总数为 n[2]):

$$SSE_{cv} = \sum_{i=1}^{n} \left(y_{test\,i} - \hat{y}_{test\,i}\right)^2,$$

而交叉验证的均方误差为:

$$MSE_{cv} = \frac{1}{n} SSE_{cv} = \frac{1}{n} \sum_{i}^{n} \left(y_{test\,i} - \hat{y}_{test\,i}\right)^2.$$

[2]如果原数据样本量为 n, 进行 K 折交叉验证, 那么每个观测值都会轮流作为测试集成员出现一次, 这时测试集的总样本量也是 n. 如果把数据仅仅分成一个训练集和一个测试集, 那么原数据样本量和测试集的样本量就不一样了.

NMSE 和 R^2 如果没有模型, 最朴素的预测就是均值, 那时, 可用均值 $\bar{\boldsymbol{y}}_{test}$ 来代替上面 SSE_{cv} 或 MSE_{cv} 的 $\hat{\boldsymbol{y}}_{test}$(分别记为 SST_{cv} 和 MST_{cv}). 我们可以用 SST_{cv} 或 MST_{cv} 作为除数来将 SSE_{cv} 或 MSE_{cv} 标准化, 得到标准化均方误差 (normalized mean square error, NMSE)

$$NMSE_{cv} = \frac{SSE_{cv}}{SST_{cv}} = \frac{MSE_{cv}}{\mathsf{MST}_{cv}} = \frac{\sum_{i=1}^{n} \left(y_{test\,i} - \hat{y}_{test\,i}\right)^2}{\sum_{i=1}^{n} \left(y_{test\,i} - \bar{y}_{test}\right)^2},$$

同样, 也可以得到另一个等价指标 R^2_{cv}:

$$R^2_{cv} = 1 - NMSE_{cv} = 1 - \frac{\sum_{i=1}^{n} \left(y_{test\,i} - \hat{y}_{test\,i}\right)^2}{\sum_{i=1}^{n} \left(y_{test\,i} - \bar{y}_{test}\right)^2},$$

该指标在经典回归中 (即没有测试集或训练集就是测试集时) 称为可决系数 (coefficient of determination)(按照机器学习术语也称为记分或得分 (score)).

显然，如果标准化均方误差 $NMSE_{cv} > 1$ 或者等价地 $R^2_{cv} < 0$，说明有模型还不如没有模型 (用朴素的均值作为预测值). 显然，交叉验证的这两个等价度量不仅仅适用于最小二乘回归，而且适用于任何模型，可以用于不同模型预测精度的比较.

对于例1的数据求标准化均方误差及记分, 我们做 $Z = 10$ 折交叉验证 (对于例1的数据, 每份 (折) 数据约 50 个观测值). 用函数cv 把数据行下标随机打乱, 输出 Z 个下标集. 然后用这个函数得到的 10 个下标集组成 10 个数据集, 轮流每次用 9 个数据集合并作为训练集, 再用一个作为测试集, 做十次拟合, 得到 $NMSE_{cv} = 0.2739$ 及 $R^2_{cv} = 0.7261$. 如果不用交叉验证, 即全部数据既是训练集又是测试集, 则得到 $NMSE = 0.256$ 及 $R^2 = 0.744$.