**一、线性模型**

**1、共线性数据**

**指模型中的自变量之间存在较强的线性关系**

**容易出现多重共线性的场景：**

**1、数据样本量的不足会导致多重共线性。**

**2、多个变量之间都基于有同趋势的数据。**

**3、多个变量之间存在着近似线性的关系。**

**危害：**

**多重共线性的存在不仅会导致模型的过拟合，而且还会导致回归模型的稳定性和准确性大大的降低。**

**2、岭回归**

**岭回归(英文名：ridge regression, Tikhonov regularization)是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。**

**最小二乘法：**

IMG_256

邻回归：

IMG_256

注：指由样本值求得的估计值与待估参数的真值之间有系统误差，其期望值不是待估参数的真值。对于参数θ的所有值的偏差都等于零的估计量称为无偏估计。

**3.交叉验证**

**交叉验证(Cross-Validation)是用于防止模型过于复杂而引起的过拟合，有时亦称循环估计， 是一种统计学上将数据样本切割成较小子集的实用方法。于是可以先在一个子集上做分析， 而其它子集则用来做后续对此分析的确认及验证。 一开始的子集被称为训练集(training set)。而其它的子集则被称为验证集(validation set)，一般与测试集区(test set)分开。**

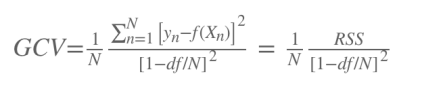
**K-fold Cross Validation(记为K-CV)**

**将原始数据分成K组(一般是均分)，将每个子集数据分别做一次验证集，其余的K-1组子集数据作为训练集，这样会得到K个模型，用这K个模型最终的验证集的分类准确率的平均数作为此K-CV下分类器的性能指标.K一般大于等于2，实际操作时一般从3开始取，只有在原始数据集合数据量小的时候才会尝试取2. 而K-CV 的实验共需要建立 k 个models，并计算 k 次 test sets 的平均辨识率。在实作上，k 要够大才能使各回合中的 训练样本数够多，一般而言 k=10 (作为一个经验参数)算是相当足够了。**

**留一交叉验证(LOO-CV)**

**留一交叉验证，名字和交叉验证类似，事实也是这样。留一交叉验证，其折数就是样本的个数，设样本个数为m，需要重复m次，最后取平均值。（计算成本较高。需要m次训练和预测，对于大数据集来说，通常不可行）。**

**广义交叉验证（Generalized Cross-Validation ,GCV）**

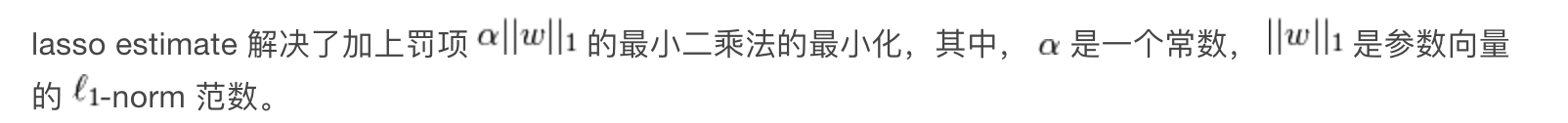


**（df是指用于估计模型的自由度数值）**

**4.Lasso**

**Lasso 是拟合稀疏系数的线性模型。 它在一些情况下是有用的，因为它倾向于使用具有较少参数值的情况，有效地减少给定解决方案所依赖变量的数量。**

IMG_256



**5.Akaike 信息判据**

衡量统计模型拟合优良性(Goodness of fit)的一种标准

在一般的情况下，AIC可以表示为：

IMG_256

k是所拟合模型中参数的数量，L是对数似然值,n是观测值数目

参数越少，AIC值越小，模型越好

样本数越多，AIC值越小，模型越好

这和调整的R方思路一致，即对变量多的模型加重惩罚力度

让n为观察数，SSR(SUM SQAURE OF RESIDUE)为残差平方和，那么AIC变为： AIC=2k+nln(SSR/n)

**BIC准则**

AIC为模型选择提供了有效的规则，但也有不足之处。当样本容量很大时，在AIC准则中拟合误差提供的信息就要受到样本容量的放大，而参数个数的惩罚因子却和样本容量没关系（一直是2），因此当样本容量很大时，使用AIC准则选择的模型不收敛与真实模型，它通常比真实模型所含的未知参数个数要多。BIC（Bayesian InformationCriterion）贝叶斯信息准则是Schwartz在1978年根据Bayes理论提出的判别准则，称为SBC准则(也称BIC)，弥补了AIC的不足。SBC的定义为：

IMG_256