GS婵☆垪鈧磭鈧鈧冻缂氱弧<84>

Leiming Dong

2018-01-28

学习的参考文献是(Lorenz et al. 2011)。

最小二乘法不能估计全部的效应值，因为自由度不够（矩阵奇异）。即便自由度足够用，标记间共线性会过拟合模型。过拟合模型能夸大小幅波动，导致预测能力差。

于是提出各种统计模型，大致可分为4类： **收缩、变量选择、核和降维方法**。 模型的基本形式一样：

是基因型和表型的关联函数。模型都设法最小化一个特定的成本函数。最小二乘的成本函数是残差的平方和：。

## 岭回归（又叫随机回归BLUP）

遗传值是个标记效应值之和。

与最小二乘的区别是多了一个项，引入的目的是避免矩阵奇异，并且降低预测子间的共线性。 选择的标准是最小化模型误差；还有一个方法是假定标记效应从一个均值为0的正态分布中随机抽取的，然后解混合线性模型方程，是残差方差与标记效应方差的比值var(*e*)/var()。如果var()比var(*e*)小得多的话，会造成标记效应收缩到0，很大。 岭回归的成本函数是

第二项称为约束（constraint）、惩罚（penalty）或正则化（regularizer）。越大，大的收缩得越厉害。假设是标记效应同分布，指的是同样收缩到0，不是所有效应都相等。

## LASSO Least absolute shrinkage and selection operator

LASSO的成本函数：

与岭回归的共同点是都惩罚大的标记效应，二者的一般形式是：

*q*=1时是LASSO，*q*=2时是RR-BLUP，时，称为弹性网回归（elastic net）。因为正则化矩阵公式不同，LASSO的系数收缩比RR-BLUP的要强，而且一些系数收缩到0，产生一个稀疏模型，因此LASSO也能进行变量选择（所以有selection operator的意思）。LASSO的一个问题也是的取值问题。 Park and Casella (2008)提出贝叶斯LASSO。 在非贝叶斯LASSO中，解决方案允许有n - 1个非零回归系数。 这是一个问题，因为在密集的标记数据的情况下，没有理由为什么训练集中的个体数量应该限制具有非零效应的标记的数量。

## Reproducing kernel Hilbert spaces (RKHS) and support vector machine regression

RKHS：经典加性遗传模型结合核函数。核函数将预测变量转换为观测值之间的一组距离，以产生一个正定矩阵，用于线性模型。 \_\_输入空间和特征空间\_\_是这个文献中经常遇到的术语。 在GS的情况下， **输入数据是标记得分数据**，因此，输入空间是多维空间，其中每个个体的位置由其标记分数确定。 通过 **对输入数据应用内核函数将输入空间转换为特征空间**。 RKHS可以用矩阵符号表示：

是核心条目的矩阵，量化个体彼此之间的距离，就像量化系谱关系的加性关系矩阵，即个体之间的遗传距离。这里的区别在于，核函数包括一个或多个平滑参数（ **h**）以影响特征空间中的距离和输入空间中的距离之间的关系。向量可以被解释为可以使用标准混合模型方程估计的特征空间内的单独效应。CV或bootstrap方法可以用来确定h的最优值，例如在群体和性状之间的模型中允许很大的灵活性。RKHS方法可以在标准的定量遗传模型框架内解释。 由于内核方法包含了很大的灵活性，而且没有线性假设，所以它们可以 **同时捕获所有类型的非加性效应**。 **支持向量机也是核方法。其成本函数可称为 -intensive**

C是一个用来衡量误差对成本函数的贡献的常数；

Lorenz, Aaron J., Shiaoman Chao, Franco G. Asoro, Elliot L. Heffner, Takeshi Hayashi, Hiroyoshi Iwata, Kevin P. Smith, Mark E. Sorrells, and Jean-Luc Jannink. 2011. “Genomic Selection in Plant Breeding.” In *Advances in Agronomy*, 110:77–123. C. doi:[10.1016/B978-0-12-385531-2.00002-5](https://doi.org/10.1016/B978-0-12-385531-2.00002-5).