**统计建模过程**

下面我们分两步进行统计建模和分析

第一步：品种分类。

针对品质指标对品种进行分类:首先对这些指标进行标准化，即每个数据 值减掉其对应指标的均值、然后除以标准差，对标准化后的数据进行分类(采用K-Median分类)。分三类：1表示不适宜、2表示基本适宜、3表示适宜。

第二步：统计建模。

采用kruskal相关性检验、字典法构造基以及lasso变量选择结合的方法，最终建立出既能控制变量个数又能较好拟合数据的模型。具体来说：

1. 设定阈值0.1，用kruskal相关性检验选择出满足条件的几个自变量，不妨暂时记为。
2. 构造字典：由变量集合经过添加A的每一元素的平方项、任意两项的交叉项构成集合记为B（例如，若p=2，则），

记，B与C合并即得到由字典法构造的基。

1. 用lasso在字典中选择重要的变量

通过调整阈值，重复上述三个步骤，建立符合要求的模型。

**花生油**

组分：1含水量、2含油量、3酸价、4棕榈酸、6油酸、7亚油酸、9花生酸、13O/L、14不饱和脂肪酸。



拟合准确率：0.90

验证准确率：0.74

**花生酱**

组分： 脂肪含量、蛋白质含量、天门冬氨酸、 蛋氨酸、 精氨酸、 油酸、 O/L



拟合准确率1。

验证准确率0.73。

**凝胶性**

组分：粗纤维、天门冬氨酸、异亮氨酸、亮氨酸、谷氨酸、苏氨酸、精氨酸



拟合准确率：0.863。

验证准确率：0.857。

**溶解性**

组分：粗纤维、天门冬氨酸、亮氨酸、谷氨酸、苏氨酸、精氨酸



拟合准确率：0.94。

验证准确率：0.89。