第六届湖南省研究生数学建模竞赛

题 目 黄桃采摘贮藏模型统计分析

摘 要：

黄桃的机械指标、人工评分结果主要受采摘时机、存储方式、贮藏时间等因素的影响，构建机械指标到人工评分的关联模型、贮藏条件到黄桃品质的预测模型对黄桃产业具有重要意义。本文通过神经网络模型、回归分析分别辨识这两个模型，依此给出人工评分、贮藏条件和有损测试数量的建议。

问题 1 的数学实质是构建机械指标到人工评分的关联模型，本文采用了神经网络模型、对抗网络模型。首先，本文采用了 spearman 系数、kendall 系数和 CE 信息熵等指标分析了桃子的评价体系参量之间的相关性，结果表明：硬度与色泽、质地、味道存在弱相关性，这与人工打分的随机性差异性有关。其次，利用智能方法计算了机械指标、人工评分判别方式的结果，结果表明：在 0.1 的显著水平下，两种判别方式可以认为两种评判方式一致。这意味着果实的综合品质及性状与感官评价是一致的，未来可以用机械指标代替人工评分，依此节省人工成本。根据模型分析和数据比较可以发现，“专家经验”过于容易满足，本文建立的模型比“专家经验”更加客观实际，而“普遍认知”是符合的。

对于问题 2，用问题 1 建立的定量模型，可以通过 2020 年测量样本的机械指标，得到对应的人工指标预测数据，进而根据人工评分标准判断样本是否优质。结果发现，优质黄桃大多集中在贮藏前期，说明经过一段时间贮藏，整体优质率会下降，这是符合直观经验的。

对于问题 3，经过数据处理，分情况分别给出了贮藏时间与所有关心的量之间的关系，系，采用回归分析可以构建模型。由于贮存温度只有两种，因此采用分情况讨论的方法。经过回归模型结果表明，回归模型中的参数显著非零，故自变量（辨识贮藏条件）到因变量（黄桃品质）的显著性很强。同时，冷库中的黄桃保质期长，约为 17 天、风味评分下降较快，冷柜中的黄桃保质期虽然较短，只有约 4 天，但是风味评分下降很慢，这与“普遍认识”是一致的。预测黄桃贮藏期时拟使用中心试验设计方法，这样计算回归参数能够使回归模型显著性很强。经过有限次有损试验测得机械指标后，标定模型。改变贮藏时间参数，计算估计的机械指标值，根据问题 1 所建模型推算对应的人工评分，进一步判断是否优质。若存在一个贮藏时间，使得对应人工评分正好低于 80，则该贮藏时间为保质期。为了能更好地标定模型，采用中心试验设计方法，分析得总共需要 10 次有损测试，依次节省试验成本。

关键词：相关性分析 一致性分析 生成对抗网络 均匀试验设计方法

•

黄桃采摘贮藏模型统计分析

摘要

黄桃的机械指标、人工评分结果主要受采摘时机、存储方式、贮藏时间等因素的影响，构建机械指标到人工评分的关联模型、贮藏条件到黄桃品质的预测模型对黄桃产业具有重要意义。本文通过神经网络模型、回归分析分别辨识这两个模型，依此给出人工评分、贮藏条件和有损测试数量的建议。

问题 1 的数学实质是构建机械指标到人工评分的关联模型，本文采用了神经网络模型、对抗网络模型。首先，本文采用了 spearman 系数、kendall 系数和 CE 信息熵等指标分析了桃子的评价体系参量之间的相关性，结果表明：硬度与色泽、质地、味道存在弱相关性，这与人工打分的随机性差异性有关。其次，利用智能方法计算了机械指标、人工评分判别方式的结果，结果表明：在 0.1 的显著水平下，两种判别方式可以认为两种评判方式一致。这意味着果实的综合品质及性状与感官评价是一致的，未来可以用机械指标代替人工评分，依此节省人工成本。根据模型分析和数据比较可以发现，“专家经验”过于容易满足，本文建立的模型比“专家经验”更加客观实际，而“普遍认知”是符合的。

对于问题 2，用问题 1 建立的定量模型，可以通过 2020 年测量样本的机械指标，得到对应的人工指标预测数据，进而根据人工评分标准判断样本是否优质。结果发现，优质黄桃大多集中在贮藏前期，说明经过一段时间贮藏，整体优质率会下降，这是符合直观经验的。

对于问题 3，经过数据处理，分情况分别给出了贮藏时间与所有关心的量之间的关系，系，采用回归分析可以构建模型。由于贮存温度只有两种，因此采用分情况讨论的方法。经过回归模型结果表明，回归模型中的参数显著非零，故自变量（辨识贮藏条件）到因变量（黄桃品质）的显著性很强。同时，冷库中的黄桃保质期长，约为 17 天、风味评分下降较快，冷柜中的黄桃保质期虽然较短，只有约 4 天，但是风味评分下降很慢，这与“普遍认识”是一致的。预测黄桃贮藏期时拟使用中心试验设计方法，这样计算回归参数能够使回归模型显著性很强。经过有限次有损试验测得机械指标后，标定模型。改变贮藏时间参数，计算估计的机械指标值，根据问题 1 所建模型推算对应的人工评分，进一步判断是否优质。若存在一个贮藏时间，使得对应人工评分正好低于 80，则该贮藏时间为保质期。为了能更好地标定模型，采用中心试验设计方法，分析得总共需要 10 次有损测试，依次节省试验成本。

关键词: 相关性分析　回归分析　生成对抗网络　回归试验设计方法

一、问题重述

黄桃是一种商业价值较高的水果，然而容易腐败，必须要把握好采摘时机、运输时间、贮藏方式等才能够在风味最好的时候上架商铺。要把握好上述因素，首先要建立黄桃品质的评判因素。

专家经验显示，当果实硬度 > $1.5kg/cm^2$ 和 TSS 含量下降率 < 20% 时算优质黄桃。人工打分时，色泽、质地和味道三项指标 ≥ 80 时为优质黄桃，根据普遍认知，人工指标要求更苛刻一些。因此如何利用这些指标来对黄桃更科学的评分，是研究的重点问题。

问题 1 是要对机械指标、人工指标之间的关系进行分析，并给出两种方法的评判一致性。然后要构建黄桃品质分析的定量模型，并将结果与“专家经验”和“普遍认知”相比较，分析其差异。问题 2 是根据定量模型预测 2020 年的几个样本黄桃是否优质。问题 3 是根据所给数据确定贮藏环境和贮藏时间对黄桃品质的影响，推荐合适的贮藏环境和贮藏时间，并判断要经过多少次有损测试才能预测贮藏期。

二、问题分析

首先要对附件给的数据进行预处理。题目给了 2018 年、2019 年和 2020 年的数据，不同时间采摘，用两种方式贮存，再隔不同时间进行机械指标测量和人工评分。第一，机械指标的测量是有损的，因此显然机械指标测量完毕后不能继续对同一个桃进行人工评分。因此数据之间是无法一一对应上的，无法对同一个桃同时进行机械指标和人工评分的分析，因此只能分析对应批次黄桃的整体性质。

第二，尽管人工评分具有一定主观性和随机性，但是作为商品，其优质与否必须以人的感受为准，因此对于黄桃实际是否优质，应当主要以人工打分为准。即当色泽、质地和味道三项指标 ≥ 80 时为优质黄桃。这一点应当作为一个标定量来衡量黄桃评价模型是否准确合理。

第三，附件中出现了很多额外数据，比如大小、重量等，而且这些额外数据在某些附件中有缺失，本文的处理方式是，对于每一批次都有的完整数据，都进行分析，对于缺失数据，将其忽略。总的来说，要考虑的机械指标有：带皮硬度、去皮硬度、TSS 含量、TSS 下降率、色差。

第四，TSS 含量方面，下降率 < 20% 为优质黄桃，然而经过计算，下降率全部都满足条件，而且变化很没有规律，有时 TSS 含量随贮藏时间增加反而增加，因此可能会对结果产生一定影响。

2.1 问题 1 的分析

首先将一个批次的所有数据（带皮硬度、去皮硬度、TSS 含量、TSS 下降率、色差 L、a、b、C、H、色泽、质地、芳香和味道）分别求平均作为一组数据。将 2018 年和 2019 年的所有数据整合到一起，将得到一个 86×13 的矩阵。每个变量的数据为一个 86 维的向量，对 13 个变量分别做相关性分析，求各类相关性系数进行综合分析。由于机械指标存在冗余，对其进行主成分分析，由此可分析各个指标之间的关系。通过 kendall 和谐系数方法比较两种评判方法的一致性。

为衡量机械指标与人工评分之间的关系，可以构建神经网络模型，并以此判断黄桃是否优质。根据这个结论可以进一步分析模型与“专家经验”和“普遍认知”之间的差异。

2.2 问题 2 的分析

问题 2 希望根据冷柜贮藏下的样本数据判断黄桃是否优质。由问题 1 可知，机械指标非常容易满足，不应该直接以该指标来判断黄桃是否优质。

与此同时，我们发现在问题一中的模型在初始测试集上的准确率高达 81.25% 因此在几个基本的假设之下本文由机械指标推算出人工指标，再用人工指标来推断是否优质，这样考虑相对来说比较合理。

2.3 问题 3 的分析

首先应当建立各项因素随贮藏时间变化的模型，包括人工评分（判断是否优质）、风味评分（判断味道好坏）、硬度、TSS 含量。然后设计试验根据测得机械指标标定上面建立的模型，改变贮藏时间参数，计算估计的机械指标值，根据问题 1 所建模型推算对应的人工评分，进一步判断是否优质。若存在一个贮藏时间，使得对应人工评分正好低于 80，则该贮藏时间为保质期。

而该方法的关键在于标定模型，拟采用中心组合试验设计方法，能够使模型更显著。

下图为论文整体框架：

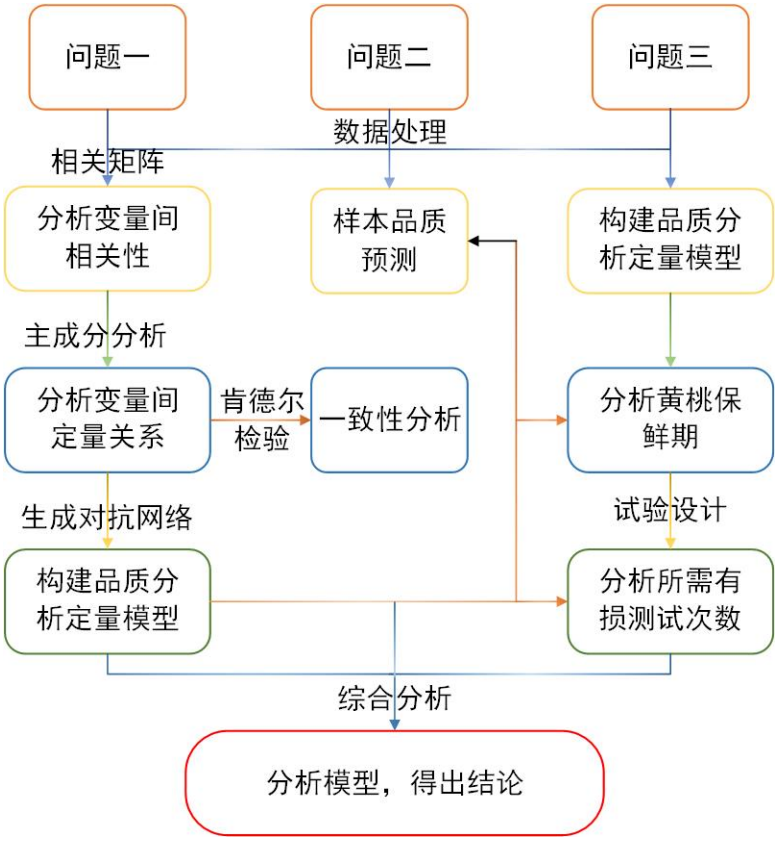

图 1: 论文整体框架图

三、模型假设

3.1 基本假设

1. 数据文件中只出现几次的数据（即不完整数据），直接将其舍弃，不作为影响因子出现。

2. 由于人工指标和机械指标的数据无法一一对应（不是一个黄桃），因此本文对一个批次中对黄桃的抽样测量结果的平均值，假设这个平均值能够作为该批次整体质量的一个评估。

3. 假设每一年情况相似，即 2018 年数据可以与 2019 年数据合并使用，且可以作为 2020 年数据预测的依据。

4. 假设人工评分测试员评分标准基本一致。

3.2 数据预处理

首先进行数据处理，根据假设 2 首先进行平均处理。考虑到 2019 年的数据比 2018 年的数据更加具有可靠性，在问题 3 中需要将 2018 年数据和 2019 年数据合并的时候为其赋权重分别为 0.2 与 0.8。

在平均处理完成后，根据野点剔除准则对数据进行野值剔除：

野点剔除准则：

设数据集 $x_1, x_2, x_3, \cdots, x_m$ 为独立同正态分布的测量数据序列，且 $x_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, 其中最多有一个野点。设 α 为给定的显著性水平，$t_{1-\alpha/2}$ 为自由度 m-2 的 t 分布对应于 $1-\alpha/2$ 的右分位数，而且 $|x_j - \bar{x}| = max|x_{max} - \bar{x}|, |x_{min} - \bar{x}|$，若：

$$|x_j - \bar{x}| > \sqrt{\frac{m}{m-1}} t_{1-\alpha/2} s_j \tag{1}$$

其中 s_j^2 为异己方差，表达式为：

$$s_j^2 = \frac{1}{m-2} \sum_{i=1}^{m-1} (x_i - \bar{x})^2, i = i, i \neq j \tag{2}$$

则认为 x_j 为野点，应剔除，否则应保留。

四、模型的建立与求解

4.1 模型建立

4.1.1 问题 1

首先将 2018 年数据和 2019 年数据进行整合，生成一个 86×13 的矩阵 A，这样指标 s_i 可以用 $[A_{1i}, A_{2i}, \cdots, A_{86i}]$ 来表示，为一个向量。

衡量变量之间相关性的系数有 pearson 系数、spearman 系数、kendall 系数、CE 信息熵等，由于未知分布，此处相关性判定不能够用 pearson 系数。本文用到了 spearman 系数、kenadll 系数和 CE 熵系数来综合判断。

spearman 系数适用于不满足正态分布、总体分布未知的连续性变量。其系数定义为：

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3}$$

其中 d_i 为数据之间的秩差（即对初始数据按照大小排序后的原序列组合，再进行平方差求得），n 为数据量。

kendall 系数是对两个有序变量或两个秩变量之间相关程度的度量，定义为：

$$\tau_a = \frac{C - D}{\frac{1}{2}N(N-1)} \tag{4}$$

其中 C 表示拥有一致性的元素对数（两个元素为一对），D 表示 s_i, s_j 中拥有不一致性的元素对数。

2008 年，马健与孙增圻提出了 Copula 熵（Copula Entropy: CE）的概念 [1]，CE 是一种更高级的相关性度量，优点有：无模型假设、可处理非线性关系、统计独立性度量、单调变换不变性、在高斯情况下与相关系数等价。关于 CE 的计算用到了马健教授发到 PyPi 上的包 copent。

一致性检验采用肯德尔和谐系数的方法。肯德尔和谐系数是计算多个等级变量相关程度的一种相关量，此处用来计算两种评分标准之间的一致性程度。人工评分的排名很好确定，按照三项指标平均值大小来排。但是机械指标优劣关系不好判定，由于 TSS 下降率越小越好，对应的应该是 TSS 含量越高越好，因此这里机械指标的排名根据的是硬度的平均值加上 TSS 含量。根据两个排名数据可计算肯德尔和谐系数。

至于黄桃品质判定的定量模型，如前文所述，黄桃品质真正的好坏要以人工评价为准，因此对黄桃品质的定量模型应该是根据机械指标直接对黄桃进行打分并判断黄桃是否优质的模型。

模型 1　神经网络模型，用于拟合机械指标到人工指标之间的关系。

首先尝试了简单的二分类法及欧式几何的曲线回归，用台湾大学林智仁开发的 libsvm 库并未取得较好的准确率，这说明需要从非常规的非线性关系上考虑机械指标和人工评分之间的关系。因此本文拟建立神经网络模型来对这些指标进行拟合。本研究采用生成对抗网络模型（GAN）和两层线性神经网络模型分别训练，以 70 组数据作为训练数据，后 16 组数据作为检验数据，评估模型的准确性。

生成式对抗网络（GAN, Generative Adversarial Networks）是一种深度学习模型，是目前复杂分布无监督学习上最具有前景的方法之一 [2]。主要包括两个模型：生成模块和判别模块。在本研究中的 GAN 结构如下图所示：

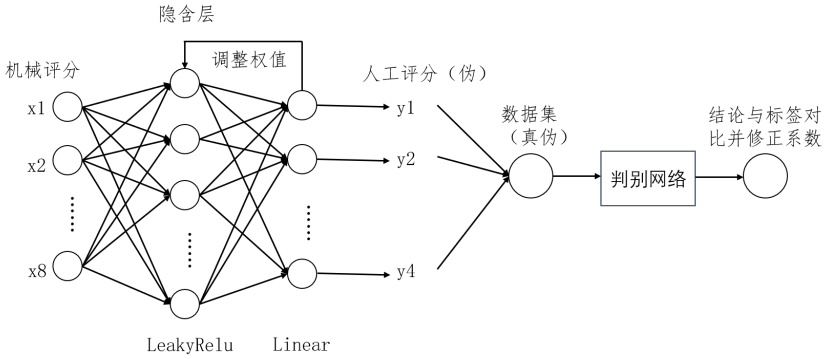

图 2: GAN 网络工作模式示意图

GAN 模型属于无监督学习，首先生成模型造数据，贴上标签 0，判别网络用原本的数据，贴上标签 1，然后对判别网络进行训练，让判别网络更加准确的同时也让生成网络能够更好地从 8 维机械指标数据生成 4 维人工指标数据。这样做的好处是能够忽略人工评分标签的影响，更加接近其根本的规律。

两层 BP 神经网络是直接以 8 个机械指标作为输入，4 个人工指标作为输出，根据人工指标的计算值与实际值的误差反向更新权值，经过反复调整来使数据拟合更准确，结构如下：

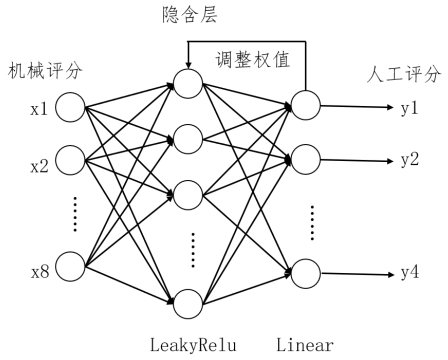

图 3: BP 网络工作模式示意图

模型参数设置上有值得注意的地方 [3]：1. 由于输入数据中的色差参数中有一个负值但并不是很大，我们希望减弱它的影响，因此采用 LeakyRelu 的激活函数，这样既可以达到上面提到非线性目的，也可以减弱负值参数的影响。2. 由于是向量作为输入输出时常采用 LP 范数，而对数据仅仅处理了平均值，为了保证权值向量的稀疏性，这里损失函数采用了 L1 范数。

根据模型 1，用 2020 年测得样本的机械指标参数，可以推算出样本大致的人工评分，就可以以此进行是否优质的评判（≥ 80）。

4.1.2 问题 2

对于 2020 年数据，只有机械指标，采用“专家经验”不够客观准确，而问题 1 构建了机械指标到人工评分之间的关系，因此可以先根据模型 1，用机械指标估计其对应的人工评分，再进一步进行判断。

4.1.3 问题 3

问题 3 要得出各类因素随贮藏时间变化的关系，包括人工评分、风味评分、硬度、TSS 含量。拟采用回归模型分析其关系。

模型 2 回归模型。贮藏时间与人工平均评分的回归关系，以及贮藏时间与人工味道评分的回归关系。

由于总共只有两种贮藏方式，因此不妨分情况讨论，对 1 度冷柜和 8 度冷库两种情况分别讨

论，在对数据进行分析的基础上，建立二次回归模型：

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon \\ &= X\beta \end{aligned} \tag{5}$$

保质期不妨定义为采摘后根据人工指标判断黄桃品质非优质时对应的贮藏日期，模型 2 的作用是：1. 求出保质期，为商家推荐贮藏方式和时间。2. 研究冷柜和冷库保存风味差别，与“普遍认为”做比较。

模型 3 回归模型。分别求贮藏时间与硬度、TSS 含量的回归关系。

问题 3 要求新的一年里通过有限次有损测试，预测贮藏期。因此需要根据机械指标判断贮藏期的预测模型。首先建立模型 3，标定根据机械指标得到贮藏时间。

模型与模型 2 类似，先建立硬度随贮藏时间的模型，再建立 TSS 含量随贮藏时间变化的模型。根据模型 2 和模型 3 可以找到有限次数有损测试算出贮藏期的预测模型。

模型 4 贮藏期预测模型。

经过计算，模型 2 和模型 3 回归显著性都较强，因此可以认为贮藏时间作为自变量，根据模型 2 和模型 3 计算人工指标和机械指标是有参考价值的。根据有限次有损测试预测贮藏期的方案如下：

Step1：经过多次有损测试，标定模型 3 参数。测试次数需要在少的同时让模型 3 的显著性足够强，因此拟采用回归试验设计的二次中心组合试验设计方法。

关于试验设计，这里很自然地用到了回归试验设计。试验设计中的回归设计是普遍使用的试验设计方法，其他较为常用的试验设计方法还有优选法、正交设计法和均匀设计法等 [4]。

优选法可以节省许多试验次数，但他的缺点是适合于单因素分析，且优选法选择出来的最优局限性大，很可能把因素之间的好的搭配方案漏掉，当因素或水平数更多时，常常会得出错误的结论。回归设计是处理变量与变量之间关系的统计方法，主要找出变量之间的数学表达式，根据一个变量的取值预测或控制另一个变量的取值。正交设计是根据数理统计学原理，从大量的试验挑选适量的具有代表性的试验点，从而反映全面情况的试验结果，特点是均衡分散，整齐可比，缺点是因素或水平较多时试验次数会很大。均匀设计法可以最小化试验次数，获取数据间的本质关系，并不适合用到此处。

已有了回归模型 3，要高显著性地标定模型参数，就应当采取回归试验设计中的二次中心组合设计方法。在 4.2.3 中给出了具体试验方案以及最少需要多少次试验的分析。

Step2：改变贮藏时间，算出因变量硬度和 TSS 含量。然后根据模型 1，将机械指标转换为人工指标。

Step3：判断人工指标，当人工指标 < 80 时黄桃判断为不优质，因此可以认为保质期就是次人工指标对应的时间。

至此可以求出保质期。该模型的重点是在测试少的同时让模型 3 的显著性够强。

4.2 模型的求解

4.2.1 问题 1

经过分析后结果如下：

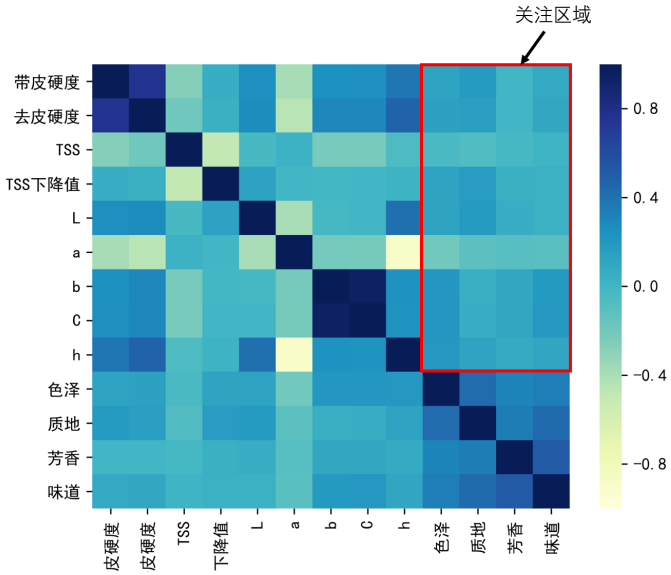

图 4: spearman 系数分析

kendall 系数分析：

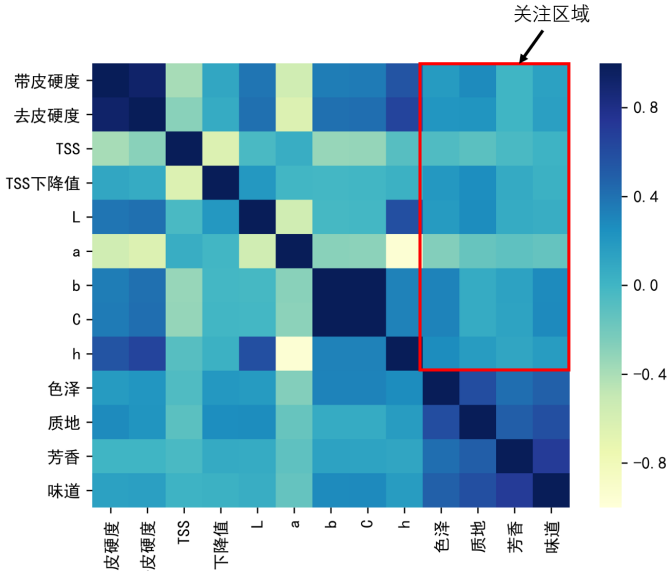

图 5: kendall 系数分析

CE 熵分析

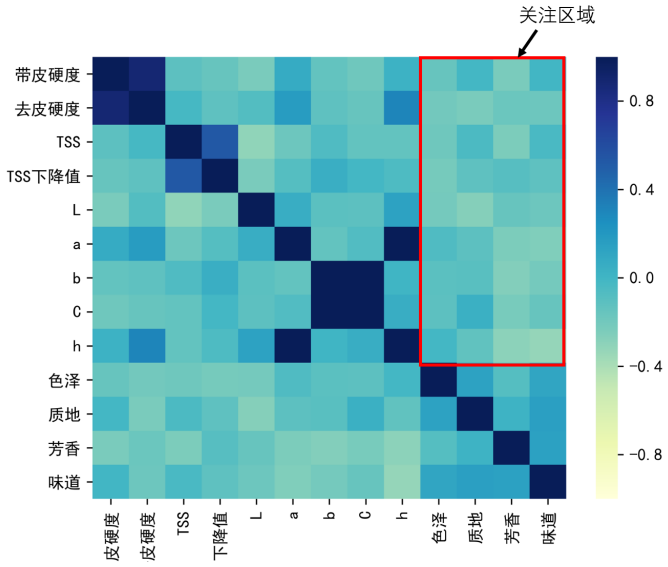

图 6: CE 熵分析

可见 spearman 系数分析和 kendall 系数分析结果比较一致，可以得到以下结论：1. 在关心的区域里果皮硬度、TSS 下降率和色泽的一些参数对色泽、质地、味道三个参数有一定相关性，而对第三个参数“芳香”相关性不强，这也可以解释人工评分为什么只以其余三者的平均值作为判据。2. CE 熵的分析显示果皮硬度、TSS 以及下降值对质地和味道有一定影响，而对色泽和芳香这几个参数相关性较差。3. CE 熵显示色差参数 L, a, b, C, h 与色泽和质地相关性比较明显，这也是符合常识的。

对机械 9 个机械指标做主成分分析，分别为：带皮硬度、去皮硬度、TSS、TSS 下降率、L、a、b、C、h，结果如下：

第一主成分：

$$Z_1 = 0.449x_1 + 0.471x_2 - 0.238x_3 + 0.148x_4 + 0.233x_5 + 0.063x_6 + 0.38x_7 + 0.38x_8 + 0.386X_8 + 0.383X_9 \quad (6)$$

第二主成分：

$$Z_2 = 0.173x_1 + 0.189x_2 + 0.58x_3 - 0.556x_4 + 0.36x_5 - 0.217x_6 - 0.149x_7 - 0.136x_8 + 0.247x_9 \quad (7)$$

第三主成分：

$$Z_3 = -0.123X_1 - 0.095X_2 + 0.266X_3 - 0.421X_4 - 0.419X_5 + 0.06X_6 + 0.511X_7 + .504X_8 - 0.182X_9 \quad (8)$$

第一主成分综合反映桃子的硬度指标；第二主成分主要反映桃子的 TSS 维持能力；第三主成分主要反映桃子色彩鲜艳程度。三个主成分的总方差解释达到 76.602%。

然而从图中可以看出，机械指标与人工评分之间的相关性并不是很强，这与人工打分的随机性主观性有一定关系。人工打分的随机性体现在不同人打分由于口味不同，权重标准不同，因此就算同一个黄桃也可能会有截然不同的结果，但是平均值仍具有一定的参考价值。

kendall 和谐系数验证的表格如下：

表 1: kendall 和谐系数检验结果

个案数	2
kendall W^a	0.611
卡方	103.872
自由度	85
渐进显著性	0.08

上表说明在 0.1 的显著性水平下可以认为两种评判方式一致。

根据简单数据处理发现，按照机械指标硬度 > $1.5kg/cm^2$，TSS 下降率 < 20% 可以发现，几乎所有批次的黄桃都能够满足条件，达到 98.84%。即按照机械指标基本所有黄桃都是优质黄桃。而根据人工评分的结果（色泽、质地、味道的平均值 ≥ 80）可以得到 58.14% 的优质率。这是可以理解的：人在进行评分的时候总是趋向于进行一定比例的优质和不优质的区分。根据神经网络的结果，结合测试数据发现，生成对抗网络最后有大约 75% 的判别准确率，BP 网络能达到 81% 左右。对于相关性很弱的数据达到如此准确率，应该说还是有一定效果的。

由于 BP 网络准确率更高，也更加稳定，因此最终决定采用 BP 网络的结果，优质黄桃定量模型为：根据机械指标的参数，代入已经训练好的 BP 网络中，再根据算出的人工评分求出色泽、质地、味道的平均分，大于 80 则为优质黄桃。

该模型与“专家经验”并不相同，根据机械指标优质率 98.84% 和人工指标优质率 58.14% 的结果就可以判断“普遍认知”是正确的。

4.2.2 问题 2

根据之前问题 1 所得 BP 神经网络模型，将 2020 年几组样本的数据代入可得人工评级结果：

表 2: 问题二模型预测结果

样本	品质	样本	品质	样本	品质
样本 A	优质	样品 D	优质	样品 G	优质
样本 B	非优质	样品 E	非优质	样品 H	非优质
样本 C	优质	样品 F	非优质	样品 I	非优质

在数据表中可以发现，样本 ADG 都是刚采摘下来的样本，理论上说优质的可能性最大，这一点符合的很好。样本 C 虽然采摘完之后放了很久，但是观察机械指标的值后发现，各项指标还是相对比较优秀的，因此作为个例，优质也是合理的。

4.2.3 问题 3

值得注意的是，模型 2 衡量贮藏时间与人工评分的回归关系时，由于同时有 2018 年和 2019 年的数据、不同采摘时间的数据，因此一个贮藏时间可以对应多个人工评分。此处对贮藏时间相同的数据取了平均。后面模型 3 也采用了类似的操作，不再赘述。

经过数据处理，给出人工评分与贮存方式（即贮存温度，分冷柜和冷库）和贮存时间之间的关系，以及人工风味评分与贮藏方式之间的关系。发现呈现明显的二次特征，如下图所示（已拟合）：

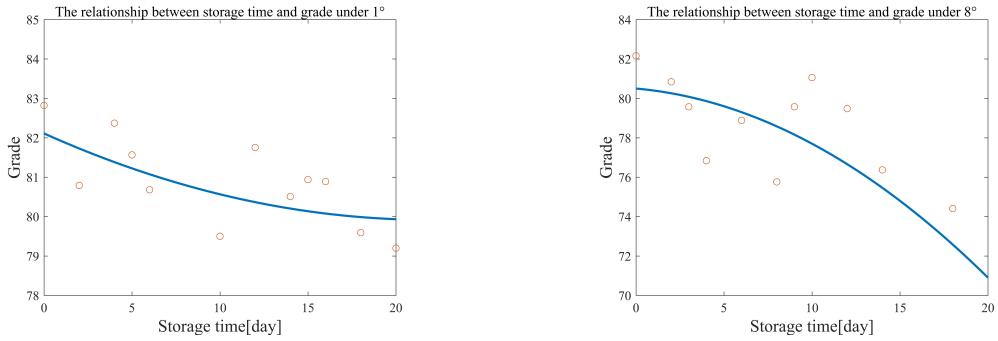

图 7: 不同贮藏方式条件下人工评分随贮存时间变化关系

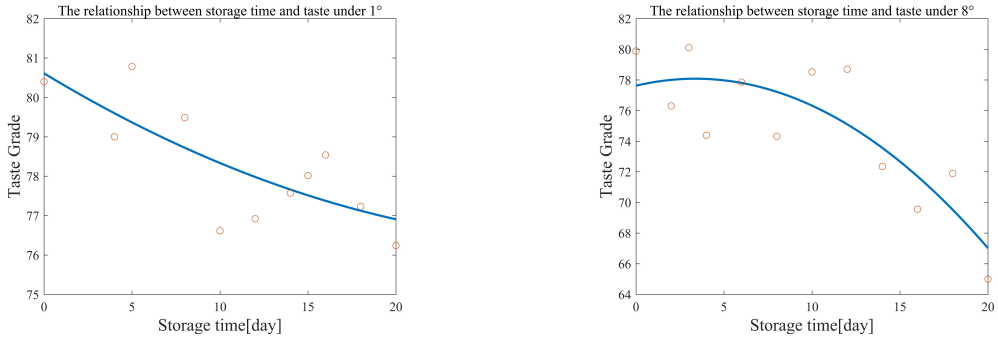

图 8: 不同贮藏方式条件下人工风味评分随贮存时间变化关系

根据线性回归模型：

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon \\ &= X\beta \end{aligned} \tag{9}$$

此处 y_i 为第 i 组的评分，x_i 表示贮藏时间，单位天。则参数 $\hat{\beta}$ 估计值为 [5][6]：

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{10}$$

根据人工评分随贮藏时间的变化模型，分别计算保质期。保质期的计算按照人工评分的拟合曲线低于 80 的值来算。值得注意的是，图 7 中拟合曲线在 0 时刻位置并不相同，那么为何还是可以通过拟合曲线来判断贮存期呢？观察可以发现，虽然两条拟合曲线在 0 时刻值并没有对上，但是散点的品质得分值基本上是差不多的，拟合曲线只是拟合一个整体趋势，即大部分黄桃在某个贮藏时间大概处于什么样的水平，而并不是精确的。因此模型预测贮藏期（保质期）也是大部分黄桃的贮藏期，是有参考意义的。

经过计算，1 度冷库中贮存贮藏期约为 17 天，8 度冷柜中贮存贮藏期约为 4 天。

结合人工风味随贮藏时间变化的模型发现，冷库保存贮藏期会变长，但是前期风味评分下降很快，而放入冷柜中贮藏，虽然贮存期变短，但前期风味评分下降很慢，拟合曲线在一开始甚至

有上升趋势。这与“普遍认为”完全一致。值得注意的是，虽然冷库保存一开始风味评分下降很快，但是后期风味评分下降会变缓，这是因为保质期长。

模型 3 处理结果如下：

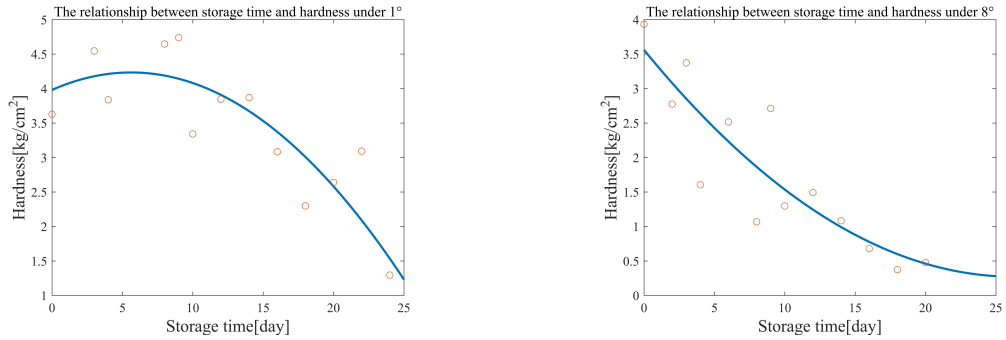

图 9: 不同贮藏方式条件下硬度随贮存时间变化关系

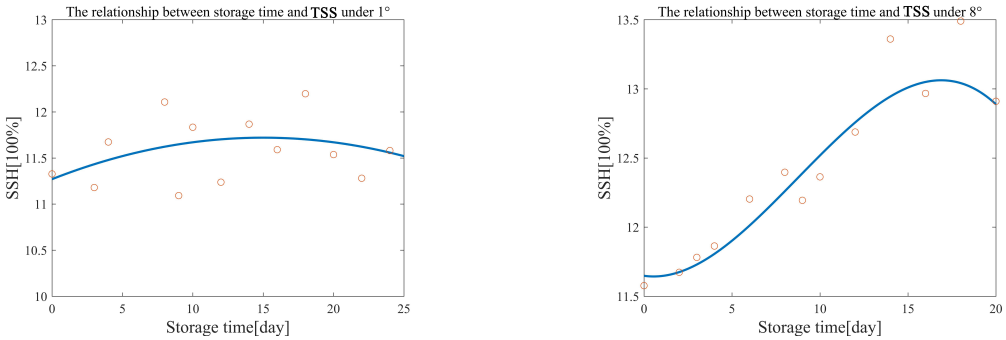

图 10: 不同贮藏方式条件下 TSS 含量随贮存时间变化关系

对上面的模型进行拟合优度检验，考察模型对数据的拟合程度，利用下式计算回归模型可决系数：

$$r^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \tag{11}$$

注意这里 ESS 表示回归平方和，TSS 表示总离差平方和而不是题目里的 TSS 含量，RSS 表示随机平方和。

计算模型的可决系数可得到下表：

注意到除了人工评分与贮藏时间的拟合不够好外，其余模型都拟合较好，而人工评分如前文所述有一定随机性，因此拟合不够好也可以接受。TSS 在冷柜中的模型采用了三次项，这是因为采用二次的时候可决系数很小，而采用三次就能有很大提升，所以这里用三次。而其余拟合不够好的几组公式是因为采用三次项也不能有很大提升，同时再考虑更高次项不仅会增加模型的复杂度，还会产生过拟合的问题，因此最终模型如上表所示。

回归结果发现显著性很强，因此可以认为回归函数可以很好地衡量两个机械指标与贮藏时间之间的关系。

表 3: 模型 2 和模型 3 的拟合结果

因变量	自变量	贮藏环境	公式	可决系数
人工评分	贮藏时间	冷库	$S_1 = 0.00456T^2 - 0.2T + 82.11$	0.459
人工评分	贮藏时间	冷柜	$S_2 = -0.02T^2 - 0.08T + 80.5$	0.516
品质评分	贮藏时间	冷库	$W_1 = -0.00424T^2 - 0.27T + 80.61$	0.626
品质评分	贮藏时间	冷柜	$W_2 = -0.04T^2 + 0.27T + 77.62$	0.711
TSS	贮藏时间	冷库	$TSS_1 = -0.002T^2 + 0.06T + 11.27$	0.526
TSS	贮藏时间	冷柜	$TSS_2 = -0.00065T^3 + 0.017T^2 - 0.018T + 11.648$	0.91
硬度	贮藏时间	冷库	$H_1 = -0.008T^2 + 0.09T + 3.98$	0.741
硬度	贮藏时间	冷柜	$H_2 = 0.00475x^2 - 0.25T + 3.56$	0.778

对于模型 4，问题 3 为预测贮藏期，需要进行多少次有损测试。根据之前模型 4 的描述，只需很好标定模型 3 就可以完成这个工作。

中心试验设计方法如下 [7]：

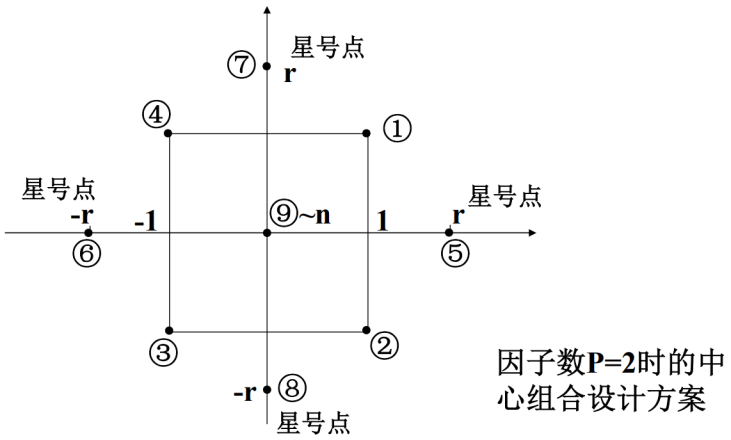

图 11: 中心试验设计方法原理图

中心组合设计方案由三部分组成：

1. 将编码值-1 与 1 看成每个因子的两个水平，同一次回归正交设计一样，采用二水平正交表安排试验，记其试验次数为 m_c。

2. 在每个因子的坐标轴上取两个试验点，该因子的编码值分别为 +r，-r，其他因子的编码值为 0，由于由 p 个因子，故这部分试验点共有 2p 个，常称这种试验为星号点。

3. 在试验区域中心进行 m_0 次重复试验，此时各因子的编码值均为 0，总试验次数等于：$m_c + 2p + m_0$。

此模型中 $m_c = 4, p = 2, m_0 = 2$，因此做十次试验，具体实验安排如下：

贮藏环境 /	贮藏时间/天	贮藏环境 /	贮藏时间/天
1	0	8	0
1	6	8	6
1	12	8	12
1	18	8	18
1	25	8	25

表 4: 中心组合试验设计

4.3 模型结果结论

问题 1 的结论是，机械指标、人工指标之间有一定相关性，但是相关性较弱。人工评分与机械评分虽然结果很大不同，但是在判断黄桃优质程度方面基本一致。机械评分非常容易满足，而人工评分更难一些。根据模型，“专家经验”不是很可靠，而“普遍认知”是没有问题的。

问题 2 利用模型 1 直接可以进行计算，结果表 2 给出。

问题 3 在 1 度冷库中黄桃保质期确实更长，模型中约为 17 天，但是同时风味也下降更快。在 8 度冷柜中黄桃保质期变短，约 4 天，但是风味却能保持较长时间。因此建议商家根据需求选择贮藏方式：若希望得到更久的贮藏时间，则采用冷库贮存的方式；若希望得到更好的风味，则采用冷柜贮存的方式。如果要按照本文模型预测下一年度的贮藏期，具体分析见 4.2.3 节，采用中心试验设计方法总共要做 10 次试验，试验的设计方法在 4.2.3 节中也已列出。

五、模型的分析与检验

模型 1 的 GAN 生成对抗网络模型和 BP 神经网络模型经过训练后对测试数据分别有约 75% 和 81% 的准确率，鉴于人工评分有很强的随机性（口味、心情等不同），这个准确率已经比较可观。根据观察发现，BP 网络稳定性和准确性都高于 GAN 对抗网络，这并不是说明 GAN 网络不行，而是因为 BP 网络此处用的是专门做拟合的网络，而且是有监督学习。而 GAN 主要应用在图像领域，而且是无监督学习，能够探寻变量之间的深层关系，在这里表现不尽如人意可能是因为训练样本不够多。

模型 2 给出了评分随贮藏时间变化的曲线，根据曲线可以看出，随贮藏时间的减小。整体品质是在下降的，这是符合常识的。最后贮藏期计算为 18 天和 4 天，也比较符合常识，说明应该尽量用冷柜进行保存。

模型 3 表 3 已经做了详细分析，拟合的都比较不错。值得注意的是模型 2，3 中出现三次项是因为三次项能比二次项误差减少非常多，而模型 1 中评分随贮藏时间的变化提高到三次，误差并不能减小很多；而如果再考虑更高次项，会导致模型过于复杂、产生过拟合等的问题，因此模型 2、3 采用表 3 所示的结果。

模型 4 的方法能够保证精确性的前提下最小化有损测试次数，方法比较可靠。

六、模型的评价、改进与推广

6.1 模型的优点

1. 采用两种神经网络模型对机械指标进行了分析，训练结果获得了较好的判断准确率，方法合适，结果相对可靠。

2. 对人工评分、硬度、TSS 含量随贮藏时间的变化做了很多处理，包括平均、处理野值、回归分析等，从结果上看，显著性较好。

3. 预测贮藏期的方法新颖。

6.2 模型的缺点

由于附件所给数据互相不对应，因此本文很多地方的数据直接采用了求平均的方法，这样的方法可能不够准确，也导致了训练样本数变少。

参考文献

[1] Jian, Ma, Zengqi, and Sun. Mutual information is copula entropy. *Tsinghua Science Technology*, 2011.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3:2672–2680, 2014.

[3] 李航. 统计学习方法. 清华大学出版社, 2012.

[4] 朱涛. 基于相似数据和均匀实验设计的磨损研究. PhD thesis, 东北大学, 2010.

[5] 吴翊, 汪文浩, and 杨文强. 概率论与数理统计. 概率论与数理统计, 2016.

[6] 吴孟达, 李兵, and 汪文浩. 高等工程数学. 高等工程数学, 2004.

[7] 林维宣. 试验设计方法. 试验设计方法, 1995.

附录

附录 1

介绍：支撑材料的文件列表

文件夹 GAN

BP。模型 1 的代码，运行环境 python3.6，运行软件 jupyter notebook

.ipynb_checkpoints——————环境文件

*_pycache_*——————————环境文件

data——————————————数据文件

gan.py————————————生成对抗式网络源码

twolayer.py——————————BP 网络源码

main.ipynb——————————主程序（直接运行）

*xunlian*_1.*pth*——————训练好的 BP 网络参数

*xunlian*_2.*pth*——————训练好的生成对抗式网络参数

文件夹 Correlation。模型 1 的代码，画相关性分析热力图

CE.py——————————————画变量间 CE 熵分析程序

Kendall&Spearman.py——画变量 *Kendall&Spearman* 系数分析程序

其余为数据文件。

文件夹 Regress。画回归分析图像

main.m————————————画回归分析散点图和拟合图程序

data.mat——————————处理后的数据文件

网络模型见附件，此处是画图模型

main.m：

```
load('data.mat')
% p = [82.11, -0.2, 4.56e-3];%D1R
% p = [80.5, -0.08, -0.02];%D8R
% p = [3.98, 0.09, -0.008];%D1JY
% p = [3.56, -0.25, 0.00475];%D8JY
% p = [11.27, 0.06, -0.002];%D1JS
% p = [11.648, -0.018, 0.017, -6.5e-4];%D8JS
% p = [77.62, 0.27, -0.04];%D8W
p = [80.61, -0.27, 0.00424];%D1W

% p = [81.305, -0.2, 4.56e-3];
% p = [81.305, -0.08, -0.02];%D8R

x = 0: 0.1: 24;
y = p(1)+p(2)*x+p(3)*x.^2;
% y = p(1)+p(2)*x+p(3)*x.^2+p(4)*x.^3;
plot(x, y, 'LineWidth', 2)
hold on
scatter(D1W(:, 2), D1W(:, 1))
% scatter(D1W(:, 1), D1W(:, 2))
set(gca,'fontsize',12,'Fontname','Times New Roman');
xlabel('\fontsize{16}Storage time[day]')
ylabel('\fontsize{16}TSS')
title('The relationship between storage time and taste under 1°')
xlim([0, max(D8J(:, 1))])
ylim([75, 82])
print -djpeg -r600 D1W.jpg
close
```

```
% a = 0;
% b = 20;
% e = 1;
% while e > 1e-5
%     temp = (a+b)/2;
%     value = p(1)+p(2)*temp+p(3)*temp.^2-80;
%     if value > 0
%         a = temp;
%     else
%         b = temp;
%     end
%     e = b-a;
% end
% a
```

CE.py：

```
from numpy.random import multivariate_normal as
mnorm
from openpyxl import load_workbook
import numpy as np
import copent
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
```

```
import matplotlib

matplotlib.rcParams['font.sans-serif'] =
['SimHei']
matplotlib.rcParams['font.family']='sans-serif'
#解决负号'-'显示为方块的问题
matplotlib.rcParams['axes.unicode_minus'] = False
#f1 = open('pearson.txt')
#f1 = open('spearman.txt')
x_tick=['带皮硬度','去皮硬度','TSS','TSS 下降值
','L','a','b','C','h','色泽','质地','芳香','味道']
y_tick=['带皮硬度','去皮硬度','TSS','TSS 下降值
','L','a','b','C','h','色泽','质地','芳香','味道']

def pearson(var):
    var1 = var[:, 0]
    var2 = var[:, 1]
    XMean = np.mean(var1)
    YMean = np.mean(var2)
    XSD = np.std(var1)
    YSD = np.std(var2)
```

```
    ZX = (var1-XMean)/XSD
    ZY = (var2-YMean)/YSD
    r = np.sum(ZX*ZY)/(len(var1))
    return r

wb = load_workbook('./前两年平均值.xlsx')
ws = wb['Sheet2']
result = np.zeros((13, 13))
for i in range(1, 14):
    for k in range(1, 14):
        #var1 = np.zeros((56, 2))
        var1 = np.zeros((86, 2))
        for ii in range(2, len(var1)+2):
            var1[ii-2, :] = [ws.cell(ii,
i).value,ws.cell(ii, k).value]
        result[i-1, k-1] = copent.copent(var1)
        #result[i-1, k-1] = pearson(var1)
pd_data=pd.DataFrame(result,index=y_tick,columns=
x_tick)
ax = sns.heatmap(pd_data, cmap="YlGnBu", vmin=-1,
vmax=1)
```

```
#plt.show()
plt.savefig('.\CE.png',dpi=600)
```

Kendall&Spearman.py：

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
import matplotlib

matplotlib.rcParams['font.sans-serif'] =
['SimHei']
matplotlib.rcParams['font.family']='sans-serif'
#解决负号'-'显示为方块的问题
matplotlib.rcParams['axes.unicode_minus'] = False
#f1 = open('pearson.txt')
#f1 = open('两年 spearman.txt')
```

```
x_tick=['带皮硬度','去皮硬度','TSS','TSS 下降值
','L','a','b','C','h','色泽','质地','芳香','味道']
y_tick=['带皮硬度','去皮硬度','TSS','TSS 下降值
','L','a','b','C','h','色泽','质地','芳香','味道']
f1 = open('两年 kendell.txt')
lines = f1.readlines()
result = np.zeros((13, 13))
A_row = 0
for line in lines:
    line = line.strip('\n')
    temp = line.split()
    for k in range(0, 13):
        temp[k] = temp[k].strip('*')
        result[A_row, k] = float(temp[k])
    A_row = A_row + 1
pd_data=pd.DataFrame(result,index=y_tick,columns=
x_tick)
f1.close()
ax = sns.heatmap(pd_data, cmap="YlGnBu", vmin=-1,
vmax=1)
#plt.show()
```

```
plt.savefig('kendell.png',dpi=600)
```