数学建模内容参考模板

摘要

摘要是全文最重要的部分，也是阅卷人首先看到的部分，阅卷人会只根据摘要将文章分成三六九等，所以如果不认真写摘要的话你就会有大麻烦，请务必留至少两小时用于摘要的打磨，将其控制在 1 面!!!

针对问题一：在写摘要的时候请搞清楚，首先摘要的第一段是题设的背景，你需要随便胡扯几句，但别扯太多，毕竟要把摘要控制在一面很难！然后写完第一段后从第二段开始就要开始介绍你对每个题目的理解、过程以及求解与评价，语言尽量精炼，不要啰嗦，再强调一次：那么多内容的摘要压缩在一面非常难！下面我将给大家演示如何对于自己已经完成的“模型建立与求解”部分进行摘要描述。

针对问题二，本文基于可视化和假设检验对所给数据集进行数据分析。首先对产品的销售价格和销售量的关系进行探究，本文分别通过**斯皮尔曼相关系数**对相关性进行定量描述，求得 ρ=-0.2946，反映出**销售价格和销售量的相关性较弱**。接着对区域与销量的关系进行探究，通过**方差检验**得知**不同地区对订单的需求量有显著差异**，并通过直方图探究出不同区域产品的需求量的不同特性。然后，本文对产品的销售方式与需求量的关系进行探究，通过 **Mann-Whitney U** 检验得出**线上销售与线下销售的销售量存在显著差异**。最后，通过**单样本 Wilcoxon 符号秩检验**将时间序列整体与促销活动单日的销量相比较，得出**促销活动单日的销量与平时的销量有显著差异**。为了将各特征的分布以及定量分析所得出的结论直观化，本文利用小提琴图、箱线图、直方图等对相关特征进行可视化，结果与定量分析一致。

针对问题三，我认为摘要关于每个题的内容应该有以下特点：整段的写作应该是循序渐进的，比如使用“首先本文 XXX”、“接着本文 XXX”、“然后本文 XXX”以及“最后本文 XXX”这样的形式；**关键的方法和结论**应该是加粗显示，LaTeX 在文中的加粗的代码想必大家已经看到了；段落中除了写建模过程中我们用了什么方式做了什么，还应该描述结论——定性的结论要直接给出，定量的结论如果较多可以概括性描述。

针对问题四，最后说一说关于这个模板，我认为数学建模竞赛的结果与论文的关联性是非常强的，所以不应该一味追求技术的实现，应该尤其重视文章的撰写与排版，LaTeX 作为一个非常经典的排版工具有一定的上手门槛，如果大家使用好了可以给文章加分不少，但是务必在比赛前多尝试，尤其是把自己之前的论文套入模板看看是否会出问题，否则比赛的时候如果花大量的时间摆弄 LaTeX，那就是得不偿失了。

关键词：随机森林；方差选择法；Voting Classifier；层次聚类分析；决策树

1 问题重述

1.1 问题背景

数学建模比赛论文是要我们解决一道给定的问题，所以正文部分一般应从问题重述开始，一般确定选题后，写手就可以开始写这一部分了——毕竟这个部分不需要等编程手和建模手掰扯。

这部分的内容是将原问题进行整理，将问题背景和题目分开陈述即可，所以基本没啥难度——问题背景就是把赛题给出的具体背景简要的阐释并加上具体的理解，问题提出就是把 3（4）个小题用自己的话复述一遍。

本部分的目的是要吸引读者读下去，所以文字不可冗长，内容选择不要过于分散、琐碎，措辞要精练。注意：在写这部分的内容时，绝对不可照抄原题！（论文会查重）应为：在仔细理解了问题的基础上，用自己的语言重新将问题描述一遍。语言需要简明扼要，没有必要像原题一样面面俱到。

1.2 问题提出

下面我将给出 2022 年数学建模国赛 C 题的“问题提出”示例供大家参考。现有一组由专家给出的关于玻璃的数据，要求通过分析与建模解决下面若干问题：

问题 1：分析这些玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系以及文物样品表面有无风化化学成分含量的统计规律，并对风化前的化学含量进行分析。

问题 2：依据附件数据分析两种玻璃的分类规律；对于每个类别选择合适的化学成分进行亚类划分，给出划分方法及结果，并分析分类结果的合理性和敏感性。

问题 3：分析未知类别玻璃文物的化学成分，鉴别其所属类型并对分类结果的敏感性进行分析。

问题 4：针对不同类别的玻璃文物样品，分析其化学成分之间的关联关系并对其差异性进行比较。

2 问题分析

2.1 问题一的分析

从实际问题到模型建立是一种从具体到抽象的思维过程，问题分析这一部分就是沟通这一过程的桥梁，因为它反映了建模者对于问题的认识程度如何，也体现了解决问题的雏形，起着承上启下的作用，也很能反应出建模者的综合水平。

这部分的内容应包括：题目中包含的信息和条件，利用信息和条件对题目做整体分析，确定用什么方法（并不是具体的算法和模型）建立模型，假如我后面要对数据进行清洗、特征缩放、特征编码然后建模，我不能在这里说我要怎么处理缺失值、异常值，也不能说我要用独热编码、要用数据归一化，更不能说我要用 SVM、要用树模型啦，而是

应该说“本文首先对数据集进行合理的数据预处理与特征工程，并建立分类模型实现预测……”。一般是每个问题单独分析一小节，分析过程要简明扼要，不需要放结论。

建议在文字说明的同时用图形或图表（例如流程图）列出思维过程，这会使你的思维显得很清晰，让人觉得一目了然！接下来我给出 2023 年数学建模国赛 C 题的“问题分析”示例供大家参考。

2.2 问题二的分析

针对问题二，第一小问要求分析各蔬菜品类的销售总量与成本加成定价的关系。首先为了便捷分析，将四个附件的数据信息合并；接着按日聚合数据并将销售量并求和，并取销售单价的均值；然后采用孤立森林模型清洗异常值。因数据未提供成本加成定价，所以首先计算出此数值。接着采用多项式回归拟合模型，输出销售总量与成本加成定价间的函数解析式，客观精确地分析出两者关系。

第二小问要求给出使商超受益最大的各品类蔬菜未来一周的的日补货总量和定价策略。首先本文采用三重指数平滑模型预测出计算收益所需要的每日批发价格。接着的通过置信区间得出日补货量与成本加成定价的取值范围。得到取值范围后，将每日收益公式的最大值点求出，从而得到使商超受益最大的的最优补货定价策略

2.3 问题三的分析

针对问题三，本问要求在给定对蔬菜数量、市场陈列量以及现实问题等约束的情况下制定 7 月 1 日的单品补货量以及定价的策略，使商超尽可能收益最大。首先需已知当日的单品批发价，由于仅需一日的数据，故直接取 6 月 30 日的单品批发价；接着需确定补货量与定价的约束条件，基于对前一周销售量与成本加成定价的描述性统计分析，并加以扰动生成各单品的约束条件；最后确定目标函数并建立优化模型并对其进行求解。

2.4 问题四的分析

针对问题四，本问要求给出需要采集的数据并推出这些数据与上题中求出的结果之间的关系。因此首先想出与补货和定价决策可能有关的数据，再将这些数据进行分析，找出最为合适的数学模型来判断所提供数据与原数据之间是否存在相互关联。本文采用了回归分析和时间序列模型，分别构建出商超销量与竞争对手销量关系的公式和市场价与销售价之间的关系对销售量的影响公式来优化补货和定价决策的制定。

刚刚说了大家应该在问题分析部分给一个流程图，当然流程图很大，放在这里太占空间，也不美观，因此我采取的方式是把流程图放在附录，并在问题分析部分作出索引。这里本来应该把关于流程图的文字介绍放在问题分析最后面，并用“如图 X”进行索引，索引的数字是蓝色的，评委很容易看出来！如图 ?? 。

这里提醒一下啊，LaTeX 中的所有编号都是自动排序的，不需要管，后面我会告诉大家如何插入图片表格，上面那个“杠 ref{label}”的意思就是索引某个图表的标签，“fig3”

是我给插入在附录的流程图设置的标签，如同姓名一样用作区分而没有实际意义。

3 模型假设

1. 假设所给数据真实可靠。
2. 本文认为在对样本进行采样过程中没有破坏样本的完整性。
3. 玻璃统计规律可以代表玻璃的一般规律，不随其他无关因素而改变。
4. 从玻璃内部的所取化学物质与表中所给物质吻合，不发生化学反应。
5. 针对不同玻璃，其颜色特征不会因光照等因素而发生改变。
6. 针对不同的化学成分，在风化前后的化学物质的比例稳定。
7. 化学成分含量发生变化也适用于变化前的模型规律。

4 符号说明

符号	含义
$i=1, i=2$	分别表示高钾、铅钡玻璃
j	表示表中从二氧化硅 (SiO_2) 到二氧化硫 (SO_2) 中第 j 类化学物质
$z=1, z=2$	分别表示风化前和风化后
x_1, x_2, x_3, x_4	分别表示纹饰、类型、颜色、风化表面
y_j	表示第 j 类化学物质的含量
$\overline{y_j}$	表示第 j 类化学物质的平均含量

5 问题一的建模与求解（公式与建模部分行文简介）

5.1 数学公式

模型建立部分是需要有大量的数学公式的，而很多小伙伴都没有 LaTeX 基础，这里就不得不提到 Axmath 或 Mathtype 了，队友可以先用 Axmath 或 Mathtype 在 Word 或者 WPS 中敲好公式，然后再用 Axmath 或 Mathtype 直接将公式转换为 Tex 代码。

数学公式的排版大致有两种类型，前者是行内公式，如 $\int_1^{+\infty} f(x)\mathbf{d}x$，这种情况一般适用于公式量不大的时候，一般就把公式当作文字放在文中了；另一种是行间公式，一般长公式、复杂公式就会单独占一行，比如这样：

$$\int_a^b f(x)\mathbf{d}x = a. \tag{1}$$

要注意的是所有行间公式结束的时候都必须要有标点符号（英文版本的），一般就是英文逗号或者英文句号，我是这样判断的：当行间公式后面跟着“其中 XXX 是 XX”这样的话时用逗号，其余情况全是句号。另外如果选用了英文句号，那么行间公式后面的文字段落与“杠 end{equation}”这一行中间就要空一行表示分段，也就是这是新的一段，但如果使用的是逗号则不能空行，表示这一段还没结束！

有的时候后文需要用到前文的数学公式可以通过交叉引用来避免重复内容，首先在行间公式中加一行“杠 label{name}”，这个 name 你可以任意取定但必须全文唯一，比如我下面的数学公式取定为“equation1”，然后我就可以说：如公式 2 所示：

$$\int_a^b f(x)\mathbf{d}x = a. \tag{2}$$

LaTeX 会自动给你编号并把数字的颜色设置为蓝色，评委很容易看出来，点击就可以跳转到这个公式！

有的时候可能会遇到特别长的公式，可以仿照公式 3 进行换行：

$$\begin{aligned} y = &(-2.014e-21)x^{12} + (2.417e-18)x^{11} - (1.275e-15)x^{10} + (3.892e-13)x^9 \\ &- (7.605e-11)x^8 + (9.957e-09)x^7 - (8.872e-07)x^6 + (5.356e-05)x^5 \\ &- 0.002138x^4 + 0.05374x^3 - 0.7827x^2 + 5.619x - 0.389. \end{aligned} \tag{3}$$

关于数学公式的问题就介绍这么多。

5.2 建模部分行文简介

下面我来介绍一下一个标准的“模型建立与求解（一级标题）”内容应该怎么写。

首先是前面的模型建立部分，我建议大家二级标题这么写：“基于 XXX 的 XXX”，当然有的时候前面还会有数据预处理之类的准备工作。模型建立部分应该主要是数学公式进行解释，毕竟这是数学建模比赛，要求用数学方法、计算机手段解决问题。首先你应该介绍一下你的数据集并给它设一个数学符号，然后将其带入到数学公式进行运算，直到能算出目标，当然这里并不是具体的目标，只是数学意义上的目标。

然后把前面所有的“基于 XXX 的 XXX”写完以后就该谢模型求解了，大家注意：每一个“基于 XXX 的 XXX”都是二级标题，模型求解也是二级标题（反正我是这么做的）。模型求解部分做两件事：把求解出的结果或者结论（比如机器学习模型预测结果、显著性检验的结果）以三线表或可视化的形式给出；把对模型的测试结果给出（比如混淆矩阵、ROC 图、MAE、MSE 等）。有的同学可能有疑问：那后面的“模型评价（一级标题）”是干嘛的，我觉得那里是整体的评价和推广，不是单个模型哈，当然每个人理解不一样也很正常，如果觉得对就可以相信自己。

6 问题二的建模与求解（交叉引用与图表排版简介）

6.1 交叉引用

大家在前面的第 5.1 节已经初步见识到了交叉引用的作用，它以蓝色的编号作为标志,让阅读者点击编号可以直接跳转到指定位置,而这样的位置可以是“章”、“节”、“图”、“表”等。交叉引用主要有两种情况，下面我一一说明：

6.1.1 图、表、章节的交叉引用

这一类的交叉引用仅需在代码后面加上“杠 label{name}”，这一点第 5.1 节中对数学公式的交叉引用我们已经介绍过，对于章节的交叉引用第 5.1 节也给了我们一个很好的示例。需要注意的是这个 name 必须是全文唯一！

进行 label 标注以后，我们可以在全文的任意地方进行引用，加上如上的代码就可以引用。

6.1.2 参考文献的交叉引用

这一类的交叉引用则是首先要在 chapter 文件夹中的第九个文件“9-参考文献.tex”中将参考文献按规范排版好，然后通过 [1] 进行索引。

6.2 插入图片

我在模板文件中已经设置好图片文件的储存路径，也就是与“数学建模竞赛论文模板参考.tex”同处一个文件夹的 Figures 文件夹，要想在 LaTeX 插入图片，首先要把图片放在这个文件夹中！接下来要做的就是复制我下面写好的代码，并修改三个地方即可。

这里要注意，插入图的时候尽量使用矢量图，因为这样的图任意放大、缩小都不会模糊。特别是基于 Python 实现数据可视化时，尽量导出 eps 格式的矢量图，如图 1 所示：

数据可视化效果预览

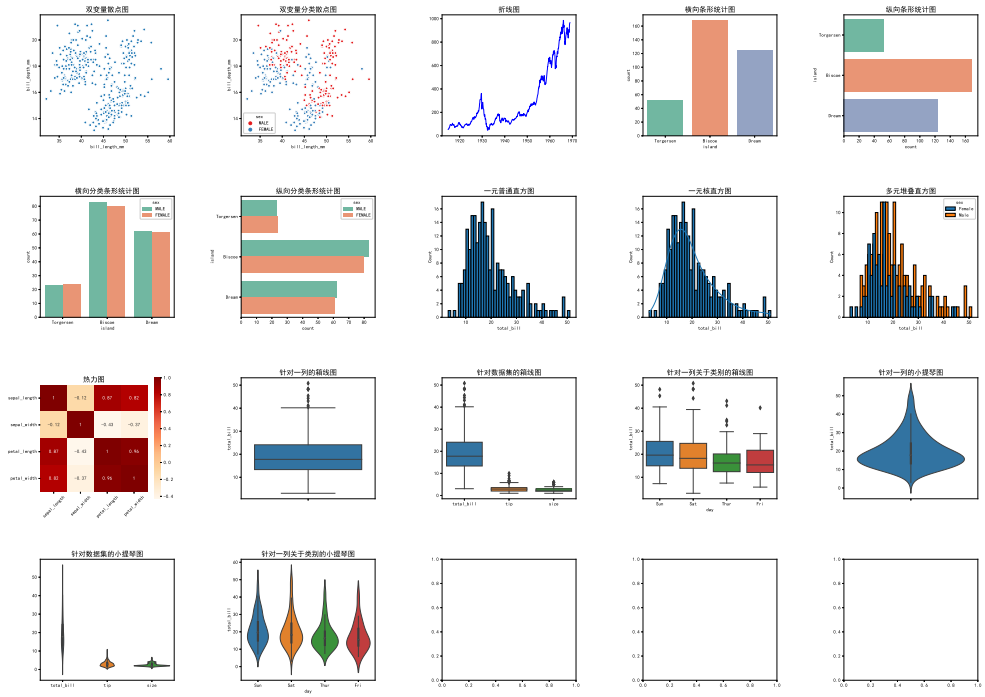

图 1. 矢量图示例

矢量图也可以是 pdf 格式，比如本文的流程图——图 6 就是 pdf 格式的。如果你希望在一行放多张图，可以参照图 2 , 3 ：

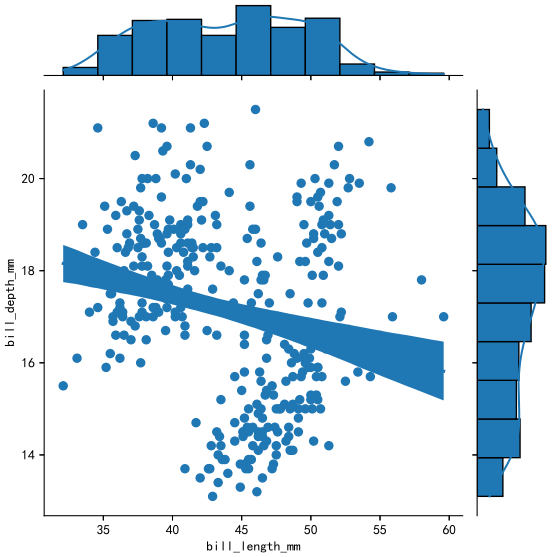

图 2. 左图

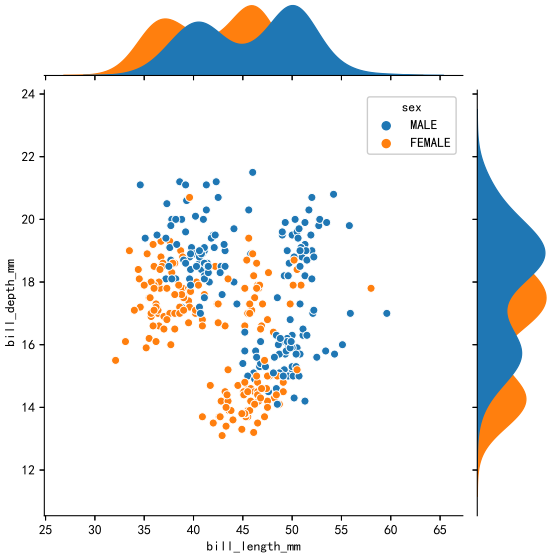

图 3. 右图

各位可以尝试一下，就会发现确有其事。

6.3 插入表格

论文中使用的表格一般是以三线表的形式呈现，三线表用处极多，必须掌握，而且在 LaTeX 中三线表是最麻烦的！我当初国赛通宵排版差点被三线表熬死，幸好孙竹清出手帮我搞定了，如表 1 :

表 1. 这是表格标题。

符号	含义
$i=1, i=2$	分别表示高钾、铅钡玻璃
j	表示表中从二氧化硅 (SiO_2) 到二氧化硫 (SO_2) 中第 j 类化学物质
$z=1, z=2$	分别表示风化前和风化后
x_1, x_2, x_3, x_4	分别表示纹饰、类型、颜色、风化表面
y_j	表示第 j 类化学物质的含量
$\overline{y_j}$	表示第 j 类化学物质的平均含量

当然有时候也会遇到行合并单元格的情况，列合并的就比较少了，下面给出具体例子如表 ?? :

表 2. 对照组与实验组前、后测时间 Mann-Whitney U 检验

	前测	平均值	标准差	p 值	后测	平均值	标准差	p 值
加法	对照组	某数据	某数据	某数据	对照组	某数据	某数据	某数据
	实验组	某数据	某数据		实验组	某数据	某数据	
减法	对照组	某数据	某数据	某数据	对照组	某数据	某数据	某数据
	实验组	某数据	某数据		实验组	某数据	某数据	

我们在写论文时很可能会遇到极大量的计算结果需要放在三线表里，这个时候就很费心了，怎么办呢？在 LaTeX 中，你可以使用 sidewaystable 环境来创建一个横向表格，并且当表格横向放置不够时，自动将其旋转为竖向放置。你需要导入 rotating 宏包来使用这个环境。如表 3 所示：

表 3. 这是一个横向的大表格

列 1	列 2	列 3	列 1	列 2	列 3 列 1	列 2	列 3 列 1	列 2	列 3 列 1	列 2	列 3	列 1
内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1
内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1
内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1
内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1
内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1
内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1
内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1
内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1
内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1
内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1
内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1	内容 2	内容 3	内容 1

7 问题三的建模与求解（23 华中赛 B 题问题 3 示例）

7.1 数据预处理

7.1.1 数据清洗与特征编码

本问要求对根据用药信息和患者信息对给药后 3 分钟以内的 IPI 数据进行预测，通过观察附件 2 发现只需要提取出 IPI005、IPI1、IPI015、IPI2、IPI025、IPI3 的相关数据，并从原始数据集中筛选出了需要的特征作为特征空间——包括性别、年龄、身高、体重、有无手术史、是否吸烟、是否酗酒、镇静药名称、镇静药诱导剂量、有无追加镇静、镇静药总剂量、镇痛药总剂量。

与上文数据预处理类似地，基于 Pandas 查找缺失值并删除一些对模型预测没有意义的特征，包括手术说明、既往史说明、ASA 评分等，之后删除少量含有缺失值的行。对特征空间中的数值特征进行特征缩放中的数据归一化，以消除其量纲；对分类特征进行编码以转化为离散特征——这些特征包括性别、有无手术史等。

7.1.2 数据降维

由于数据集中有大量的分类特征导致数据集整体较为离散，模型难以充分提取特征，且容易过拟合。使用主成分分析法对处理后数据进行降维，当前数据共有 14 列，通过将方差解释比例可视化可得下图：

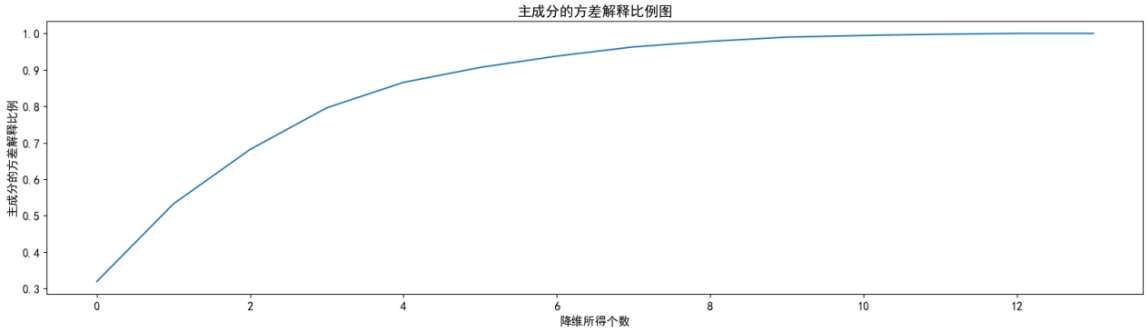

图 4. 累积方差解释比例图

通过上图发现前四个主成分即可表示样本 80% 的信息，因此保留了 4 个主成分，对 4 个主成分特征重新导入到数据集，便于后面利用监督学习中的回顾方法。

7.2 岭回归与支持向量机回归的加权平均预测

7.2.1 模型建立

（一）岭回归模型

1. 岭回归

设有一个线性回归模型：

$$y = \overrightarrow{X}\overrightarrow{\beta} + \overrightarrow{\varepsilon}. \tag{4}$$

其中，$\overrightarrow{X}$ 是 $n \times p$ 的自变量矩阵，$\overrightarrow{\beta}$ 是 p 维系数向量，y 是 n 维因变量向量，$\overrightarrow{\varepsilon}$ 是 n 维误差向量。

引入 L2 正则化项后，岭回归的目标函数为：

$$\min_{\overrightarrow{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \overrightarrow{x}_i^T \overrightarrow{\beta} \right)^2 + \lambda \sum_{j=1}^{p} \overrightarrow{\beta}_j^2 \right\}. \tag{5}$$

其中，λ 是正则化强度超参数，控制正则化项的权重大小。岭回归的解可以用闭式解表达式表示：

$$\widehat{\overrightarrow{\beta}} = \left(\overrightarrow{X}^T \overrightarrow{X} + \lambda \overrightarrow{I} \right)^{-1} \overrightarrow{X}^T y. \tag{6}$$

其中，$\overrightarrow{I}$ 是 $p \times p$ 的单位矩阵。当 $\lambda = 0$ 时，岭回归就退化成了普通的线性回归；当 λ 取值较大时，正则化项的影响就越大，模型的系数就越接近于 0。

2. **支持向量机模型**

对于数据集 $D = \{(\overrightarrow{x}_1, y_1), (\overrightarrow{x}_2, y_2), \cdots, (\overrightarrow{x}_m, y_m)\}$，，得到一个回归模型 $f(\overrightarrow{x})$ 与 y 尽可能接近.SVR 问题可形式化为：

$$f(\overrightarrow{x}) = \overrightarrow{\omega}^T \Phi(\overrightarrow{x}) + b. \tag{7}$$

$$\overrightarrow{\omega} = \sum_{i=1}^{m} (\widehat{\alpha}_i - \alpha_i) \Phi(\overrightarrow{x}_i). \tag{8}$$

$$f(\overrightarrow{x}) = \sum_{i=1}^{m} (\widehat{\alpha}_i - \alpha_i) \kappa(\overrightarrow{x}, \overrightarrow{x}_i) + b. \tag{9}$$

其中 $\kappa(\overrightarrow{x}_i, \overrightarrow{x}_j) = \Phi(\overrightarrow{x}_i)^T \Phi(\overrightarrow{x}_j)$ 为核函数.$\overrightarrow{\omega}, b$ 为模型参数，$\widehat{\alpha}_i \geq 0, \alpha_i \geq 0$ 是拉格朗日乘子. 利用高斯核函数求解，公式为：

$$\kappa(\overrightarrow{x}_i, \overrightarrow{x}_j) = \exp\left(-\frac{\|\overrightarrow{x}_i - \overrightarrow{x}_j\|}{2\sigma^2} \right). \tag{10}$$

7.2.2 模型求解

首先基于 Scikit-Learn 建立三分钟内各时间点的 IPI 数值的两种预测模型——岭回归模型和 SVR 模型，设定岭回归模型在 i 样本的预测值为 $h_1(\overrightarrow{x}_i)$，SVR 模型的预测值为 $h_2(\overrightarrow{x}_i), i = 1, 2, \cdots, m$，构建集成学习器 H 包含两个基学习器h_1, h_2。

对于题目中由已知的数据集进行建模,分别需要对六个标签——IPI005、IPI1、IPI015、IPI2、IPI025 以及 IPI3 进行回归分析。加权平均法是一种常见的回归任务的模型融合方法，利用加权平均法进行结合，其中原理如下：

$$\omega_i = \frac{MSE_i}{MSE_1 + MSE_2}, i = 1, 2. \tag{11}$$

$$H(\overrightarrow{x}) = \frac{1}{2}\sum_{i=1}^{2} \omega_i h_i(\overrightarrow{x}). \tag{12}$$

其中,ω_1, ω_2 分别表示两个模型的权重,通过计算两种模型集成后组成的最优模型 $H(\overrightarrow{x})$。

对于两个基学习器和基于加权平均法的融合模型，本文分别在测试集上进行测试，基于 MAE 和 MSE 对模型泛化能力进行评价，所得如下：

表 4. 三种模型在测试集上的测试评价

标签名称	模型	MAE	MSE
IPI005	岭回归	0.1447	0.0423
	SVR	0.1418	0.0417
	模型融合	0.1432	0.0419
IPI1	岭回归	0.2376	0.1020
	SVR	0.2272	0.1153
	模型融合	0.2322	0.1066
IPI015	岭回归	0.3669	0.1630
	SVR	0.3567	0.1808
	模型融合	0.3570	0.1651
IPI2	岭回归	0.2666	0.1103
	SVR	0.2363	0.1174
	模型融合	0.2509	0.1101
IPI025	岭回归	0.1680	0.0679
	SVR	0.1663	0.0682
	模型融合	0.1671	0.0678
IPI3	岭回归	0.1810	0.0798
	SVR	0.1788	0.0810
	模型融合	0.1797	0.0800

分析上表可知，基于三种模型在测试集上的评价指标，对于这六种标签预测评价指标都接近 0，说明这三种模型的预测效果都十分优良。从标签种类分析，可以发现对于标签 IPI005 的预测效果最佳，对于标签 IPI015 的预测效果最差。不同标签预测效果整体上的优良性从最优到优的排序为 IPI005、IPI025、IPI3、IPI1、IPI2、IPI015。

分析三种模型发现无论是岭回归模型还是 SVR 模型都不能保证是性能最优，而将模型融合可以弥补单个模型的不足，提高准确率。同时提升模型的鲁棒性，降低由于数据的随机性导致的模型波动，得到的预测结果更准确，泛化能力更强，模型稳定性更强。模型融合是提高预测准确性、提高模型鲁棒性、降低模型波动性的一种有效方法。

7.2.3 灵敏度分析

为了评估两个不同的模型（Ridge 回归和支持向量机）对输入数据的敏感性，这里通过施加一些高斯噪声（从 0 到 0.5 的比例），原理为

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{13}$$

其中，x 表示随机变量的取值，μ 表示期望值，σ 表示标准差。改变测试数据集的特征，来模拟模型的偏差，以确定模型的鲁棒性。然后根据分段函数：

$$error_{ij}=\begin{cases}0 & ,(0,c)\\ \overline{y}_{ij}+f(x_j) & ,(c,0.5)\end{cases} \tag{14}$$

其中 $\overline{y}_{ij}$ 为第 i 个特征的第 j 个的样本值，$f(x_j)$ 表示加入的高斯噪声，c 表示噪音阈值。重新预测，并计算预测结果与真实结果之间的均方误差。下文以 IPI3 为例对模型灵敏度进行分析（其余见附件），得到数据可视化图如下所示：

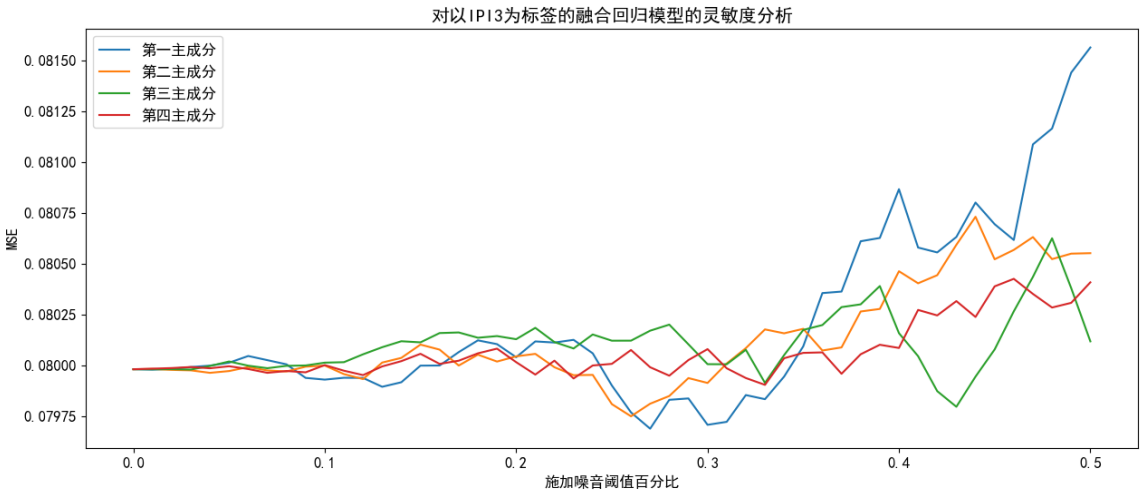

图 5. 对以 IPI3 为标签的融合回归模型的灵敏度分析

从上图可以看出，对第一主成分、对第二主成分、对第三主成分、对第四主成分分别不断施加噪音，可以直观地看出 MSE（均方差）值在这个区间波动，四个特征在施加

噪音百分比在 30% 以内时 MSE 均无明显波动，说明该模型具有一定的可靠性和稳定性性能优良。另外五个标签的灵敏度分析图详见附录 A。

8 模型评价和改进

8.1 模型优点

1. 在问题一中对文物表面是否风化与类型、颜色、纹饰关系分析过程中，不仅对于单变量之间进行了分析，还进一步用树模型进行了多变量与单变量的分析，同时利用互信息进一步对结果进行检验，提高模型的合理性。

2. 本文做了大量图表来统计分析数据特点，直观的对比出两类玻璃风化前后的化学成分的变化量以及各类玻璃化学成分之间的关系。

3. 在文本中多次对模型进行调参，利用混淆矩阵和多个指标检验，提高了模型的准确性。

4. 使用强分类器，构建 Voting 集成算法，得到一个完美模型，得到结果可信度高。

8.2 模型缺点

1. 做编码时由于颜色样本有 7 个非叙述类别，而数据集中存在大量的分类特征，没有找到合理高效的特征编码方式。

9 参考文献

[1] 伏修锋, 干福熹. 基于多元统计分析方法对一批中国南方和西南地区的古玻璃成分的研究 [J]. 文物保护与考古学 2006（04）.

[2] 司守奎，孙玺菁. 数学建模算法与应用（第 3 版）——北京：国防工业出版社，2021.4.

[3] 周志华著；李楠译. 集成学习：基础与算法——北京：电子工业出版社，2020.8.

[4] 司守奎，孙玺菁.Python 数学建模算法与应用——北京：国防工业出版社，2022.1.

[5] 王贺，刘鹏，钱乾著. 机器学习算法竞赛实战——北京：人民邮电出版社，2021.9.

[6] 何道江，黄旭东，张琼编. 数学建模优秀论文选编——北京：科学出版社，2020.11.

[7] 何伟，张良均主编. 机器学习原理与实践——北京：人民邮电出版社，2021.7.

[8] 孙玉林，余本国著.Python 机器学习算法与实践——北京：电子工业出版社，2021,9.

A 附录：关键代码实现

A.1 描述性统计分析

```
print('Hello World!')
```

A.2 数据可视化 1

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

plt.rcParams["font.sans-serif"] = ["SimHei"]
plt.rcParams['font.size'] = 12 # 字体大小
plt.rcParams['axes.unicode_minus'] = False # 正常显示负号

XXXXXX

plt.figure(figsize=(8, 7))
plt.scatter(df['item_price'], df['ord_qty'], s=10, c='red')
plt.xlabel('item_price')
plt.ylabel('ord_qty')
plt.title("产品价格-需求量散点图")
plt.show()
```

A.3 数据可视化 2

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

plt.rcParams["font.sans-serif"] = ["SimHei"]
plt.rcParams['font.size'] = 12 # 字体大小
plt.rcParams['axes.unicode_minus'] = False # 正常显示负号
```

```
XXXXXX

plt.figure(figsize=(8, 7))
plt.scatter(df['item_price'], df['ord_qty'], s=10, c='red')
plt.xlabel('item_price')
plt.ylabel('ord_qty')
plt.title("产品价格-需求量散点图")
plt.show()
```

A.4 数据可视化 3

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

plt.rcParams["font.sans-serif"] = ["SimHei"]
plt.rcParams['font.size'] = 12 # 字体大小
plt.rcParams['axes.unicode_minus'] = False # 正常显示负号

XXXXXX

plt.figure(figsize=(8, 7))
plt.scatter(df['item_price'], df['ord_qty'], s=10, c='red')
plt.xlabel('item_price')
plt.ylabel('ord_qty')
plt.title("产品价格-需求量散点图")
plt.show()
```

B　附录：图表（随便放的别当真）

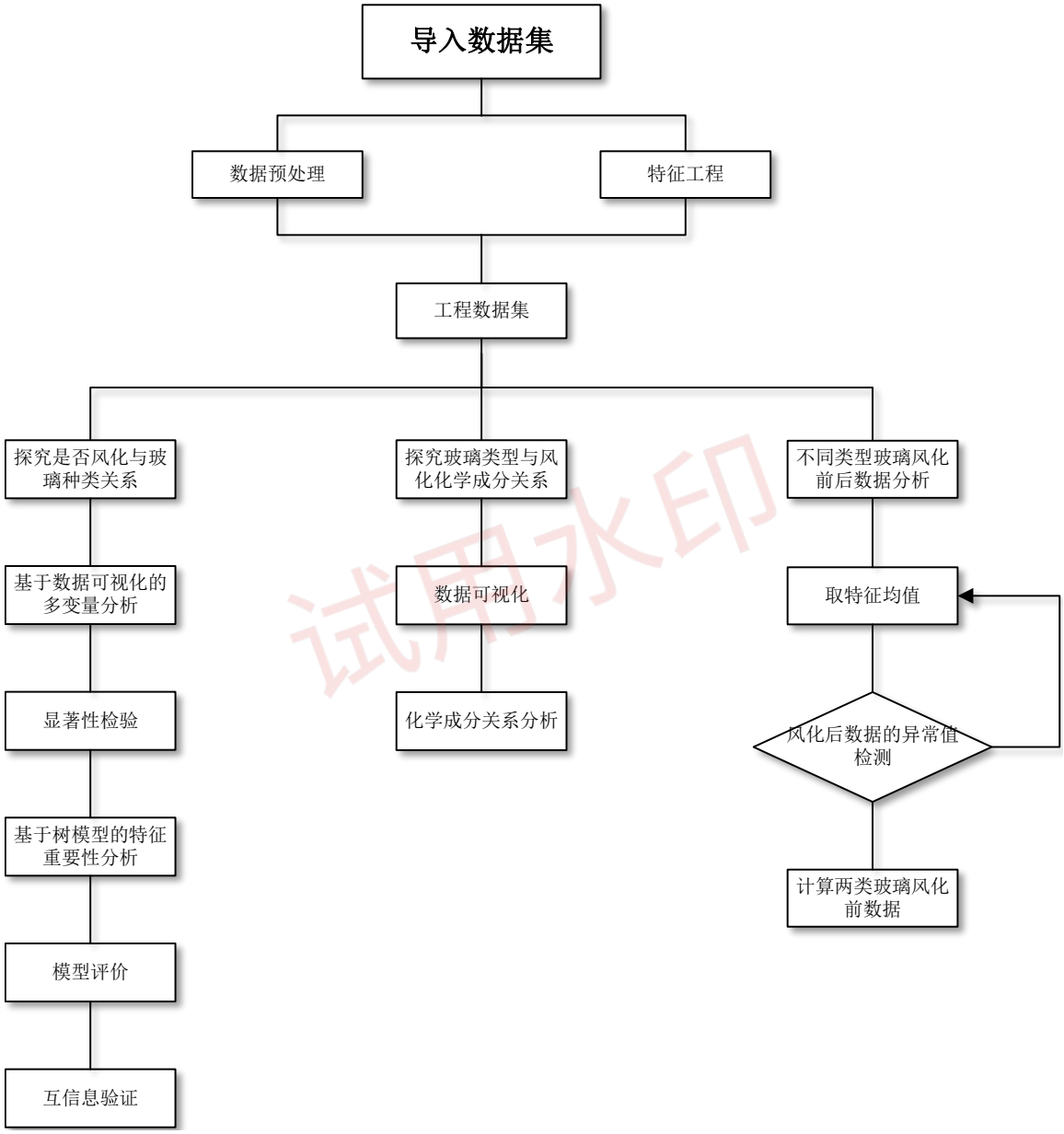

图 6. 附录图表 1

表 5. 附录图表 1

符号	含义
$i=1, i=2$	分别表示高钾、铅钡玻璃
j	表示表中从二氧化硅 (SiO_2) 到二氧化硫 (SO_2) 中第 j 类化学物质
$z=1, z=2$	分别表示风化前和风化后
x_1, x_2, x_3, x_4	分别表示纹饰、类型、颜色、风化表面
y_j	表示第 j 类化学物质的含量
$\overline{y_j}$	表示第 j 类化学物质的平均含量

C　附录：图表（随便放的别当真）

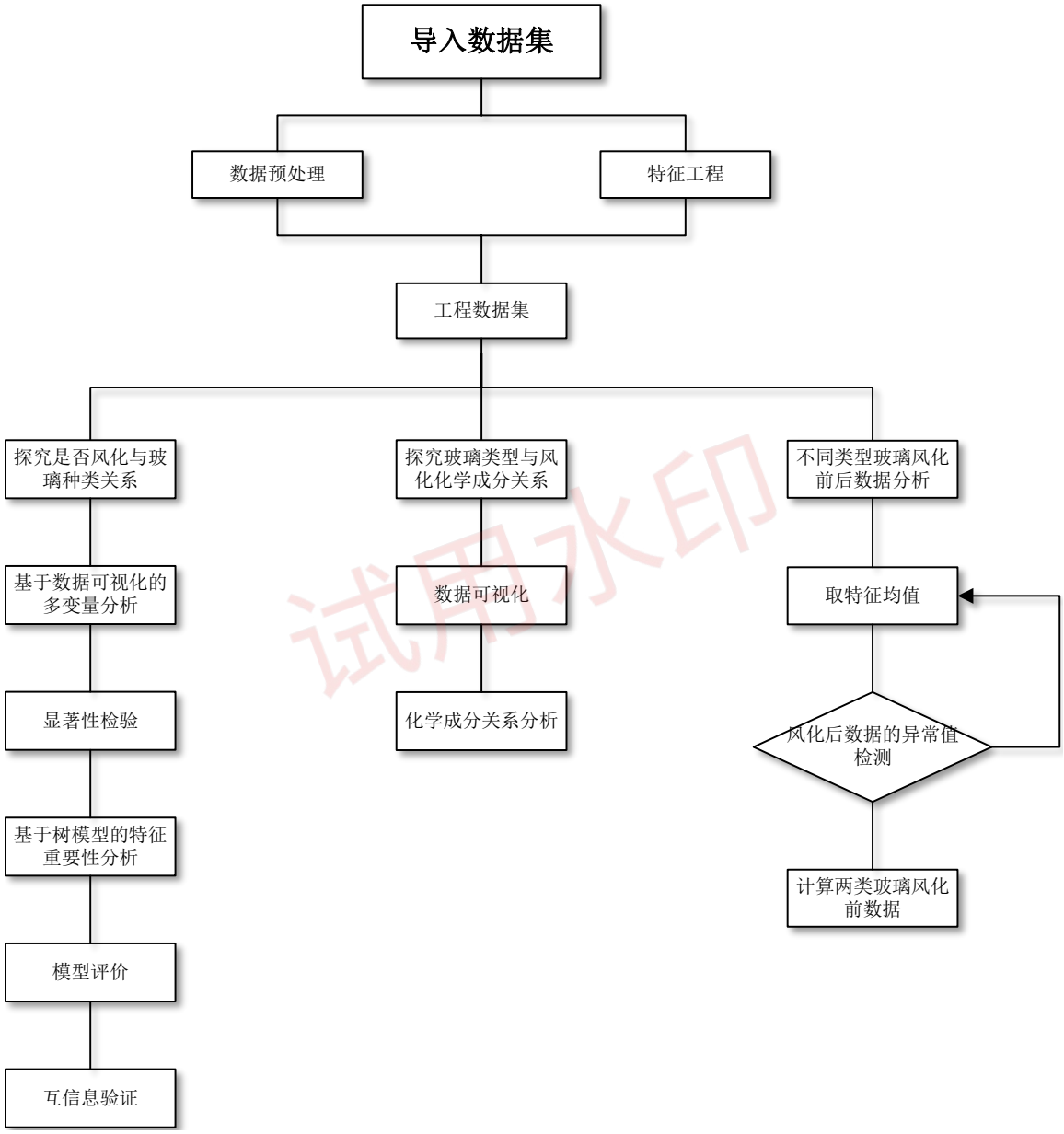

图 7. 附录图表 1

表 6. 附录图表 1

符号	含义
$i=1, i=2$	分别表示高钾、铅钡玻璃
j	表示表中从二氧化硅 (SiO_2) 到二氧化硫 (SO_2) 中第 j 类化学物质
$z=1, z=2$	分别表示风化前和风化后
x_1, x_2, x_3, x_4	分别表示纹饰、类型、颜色、风化表面
y_j	表示第 j 类化学物质的含量
$\overline{y_j}$	表示第 j 类化学物质的平均含量