

**2020年兰州大学数学建模竞赛**

**中文标题**  混合物中特定成分的检测

队伍编号 1 1 1

学 院 信息科学与工程学院

专 业 数据科学

学生姓名 蔡琪琦

丁俊伟

曹延飞

**摘要**

这里写上摘要正文。

**关键字：**关键字1 关键字2

目录

[1. 问题重述 4](#_Toc45988533)

[2. 数据准备 4](#_Toc45988534)

[3. 模型假设 4](#_Toc45988535)

[4. 通用符号说明 4](#_Toc45988536)

[5. 问题一：给出判定特定成分存在的主要指标 4](#_Toc45988537)

[5.1 问题分析 4](#_Toc45988538)

[6. 问题二：成分存在与否的判定是否存在模糊区域 4](#_Toc45988539)

[6.1 问题分析 4](#_Toc45988540)

[7. 问题三：建立数学模型，判断测试数据 5](#_Toc45988541)

[7.1 问题分析 5](#_Toc45988542)

[8. 模型分析与评价 5](#_Toc45988543)

[8.1 模型优点 5](#_Toc45988544)

[8.2 模型改进 5](#_Toc45988545)

# 问题重述

为了检测某种混合物中是否含有特定成分，对该混合物的25000个样本进行了7项指标(记为V1,V2,…,V7)的检测。检测结果见附件“Data.xlsx”。数据文件包含两个表单。第一个表单是“训练数据”，包含20000个混合物样本，已知混合物中是否含有特定成分，其中1表示含有特定成分，0表示不含特定成分。第二个表单是“测试数据”，包含5000个混合物样本，未知混合物中是否含有特定成分。请完成以下三个任务：

1. 给出判定特定成分存在的主要指标；
2. 成分存在与否的判定是否存在模糊区域（对于7项指标），即在这些区域中很难判断成分是否存在。如果存在，试估计这一区域；
3. 建立数学模型，判断测试数据中每一个混合物是否含有特定成分。论文中给出前10个混合物的判定即可。

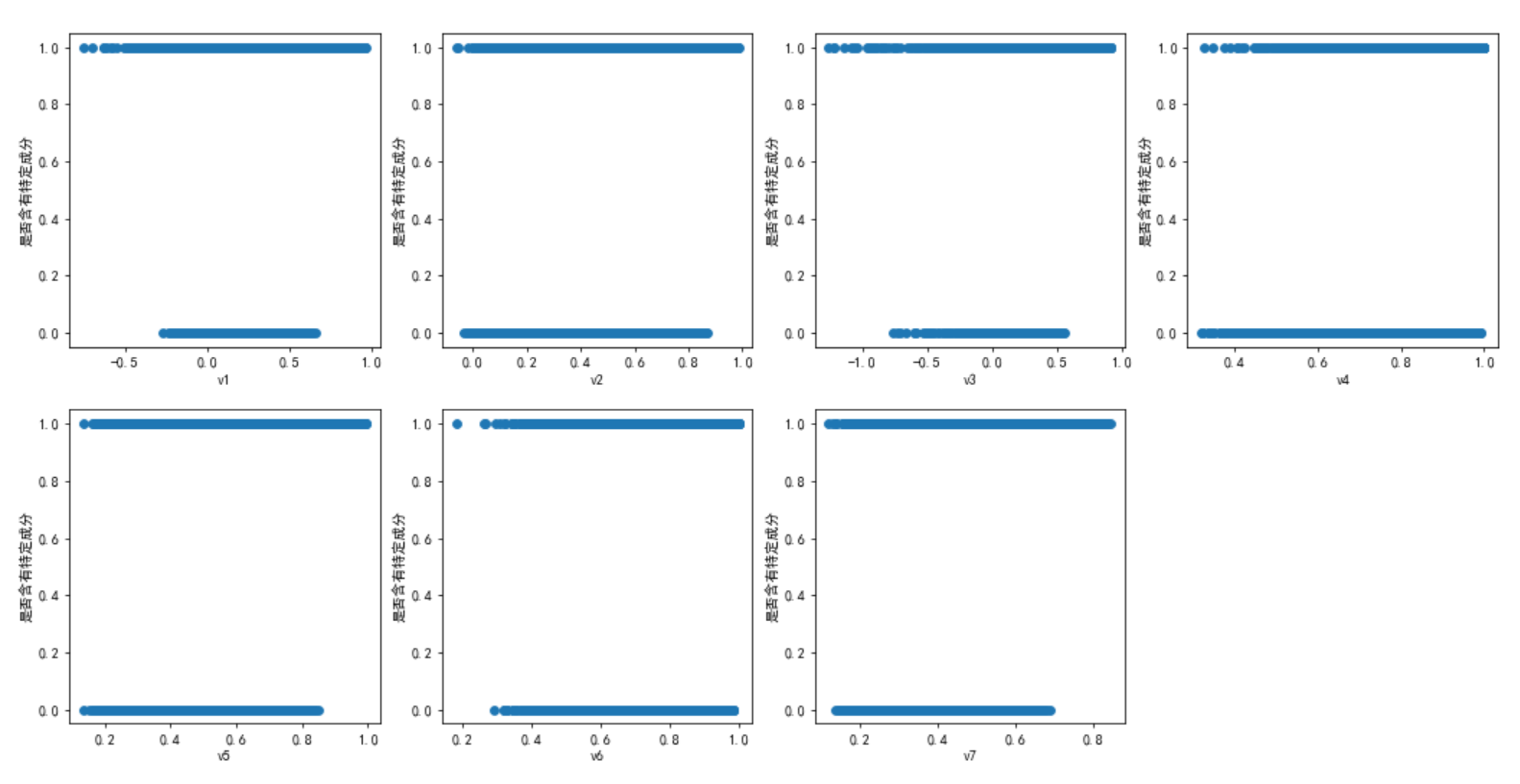
# 数据准备

### 数据来源

对某种混合物的25000个样本进行了7项指标(记为V1,V2,…,V7)的检测，判断是否含有特定成分。其中20000个样本作为训练数据，已知混合物中是否含有特定成分，1表示含有特定成分，0表示不含特定成分，5000个样本作为测试数据，未知混合物中是否含有特定成分。

### 数据清洗

在进行分析之前，需要先对数据进行清洗以及补值。因为数据集不存在缺失数据，而且通过分别绘制7项指标和是否含有特定成分的散点图可以看出，并不存在明显的异常离群值，故没有对原有数据进行任何操作。



# 模型假设

这里写模型假设。

# 通用符号说明

这里写符号说明。

# 问题一：给出判定特定成分存在的主要指标

## 问题分析

问题一要求给出判定特定成分存在的主要指标。首先选取合适的数据分析方法来对指标进行分析，本文采用的是数据挖掘方法；然后通过建立模型分析各项指标对判定结果的影响，本文采用逻辑回归模型；最后通过比较逻辑回归结果来得出判定特定成分存在的主要指标。

## 模型建立

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中提出隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。它是一类深层次的数据分析方法。

实施数据挖掘的步骤可以分为以下步骤：确定业务对象、数据准备、数据挖掘、结果分析、知识的同化。

### 确定业务对象

清楚地定义出业务问题，认清数据挖掘的目的是数据挖掘的重要一步。挖掘的最后结果是不可预测的，但要探索的问题应是有预见的。为了数据挖掘而挖掘往往带有盲目性，是不会成功的。

### 数据准备

①数据选择 搜索所有与业务对象有关的内部和外部数据信息，从中选择出适用于数据挖掘应用的数据。

②数据预处理 包括对冗余、与任务无关数据的删除，缺失数据、异常数据处理，数据标准化等等，数据预处理是数据挖掘成功的关键。

### 数据挖掘

对得到的经过转换的数据进行挖掘，也就是用分析模型进行分析计算。

### 结果分析

解释并评估结果，其使用的分析方法一般应随数据挖掘操作而定，通常还会用到可视化技术。

### 知识的同化

将分析所得的知识集成到业务信息系统的组织结构中去。

# 问题二：成分存在与否的判定是否存在模糊区域

## 问题分析

这里写问题2的分析。

# 问题三：建立数学模型，判断测试数据

## 问题分析

这里写问题3的分析。

## 模型建立

因变量Y是一个定性变量，只存在两种结果。基于因变量的特点，首先构建函数f(z)，使得经此函数变换后y的取值范围在[0,1]内。一个符合要求的函数为：

 (5.1)

其函数图像如图1所示。该函数自变量的取值范围是(-∞，+∞)，函数值的取值范围是[0,1]。



图1 f(z)函数图像

y的概率分布为：

 (5.2)

根据离散型随机变量期望值的定义，可得

 (5.3)

Y与x间的关系应满足：

E(*yi*) = ，i = 1,2,...,n (5.4)

其中， f(Xi,β) = (5.5)

在该式中，Xi’ = [1,Xi1,...,Xi7]，i = 1,2,...,n，β=[β0,β1,...,β7]’ 。β为七项指标组成的参数向量。

式11.3也可以表示为：

θi = P{yi = 1} = , i=1,2,...,n (5.6)

对模型5.6作变换可得：

y = ln =β0+β1X1+...+β7X7 (5.7)

所以最终的概率计算公式为：

P = (5.8)

# 模型分析与评价

## 模型优点

这里写模型优点。

## 模型改进

这里写模型改进。

**参考文献**

[1]迟畅. 基于主成分的回归和决策树对高血压数据的处理[D].吉林大学,2019.