

数学建模竞赛学习笔记

从理论到实战

谷文军

数学建模。

前言

目 录

序	i
前言	iii
 第一部分 数学建模基础	 1
第一章 绪论	3
1.1 绪论	3
第二章 数学建模的思想与方法	5
2.1 建模步骤	5
2.1.1 问题提出	5
2.1.2 量的分析	5
2.1.3 模型假设	5
2.1.4 模型建立	5
2.1.5 模型求解	5
2.1.6 模型分析	6
2.1.7 模型检验	6
2.1.8 模型应用	6
2.2 建模方法	6
2.2.1 机理分析法	6
2.2.2 系统识别建模法	6
2.2.3 仿真建模法	6
2.2.4 相似类比建模法	6
2.3 分类与特点	7
2.3.1 数学建模的分类	7
2.3.2 数学建模的特点	7
2.4 数学建模能力的培养	8

第二部分 数学建模方法9

第三章 初等方法建模11

3.1	勾股定理与黄金分割率	11
3.1.1	黄金分割应用于高跟鞋问题	11
3.1.2	黄金分割在其他领域的应用	11
3.2	九宫图	11
3.2.1	九宫图问题的提出	11
3.2.2	九宫图问题的求解	12
3.3	椅子稳定问题	12
3.3.1	问题引入与建模准备	12
3.3.2	模型假设	13
3.3.3	模型建立	13
3.3.4	模型求解	14
3.4	商人过河问题	14
3.4.1	问题引入	14
3.4.2	模型分析	15
3.4.3	模型建立	15
3.4.4	模型求解	15
3.5	图论方法与网络模型	16
3.5.1	图论的起源	16
3.5.2	图的概念	16
3.5.3	哥尼斯堡七桥问题	16
3.6	层次分析方法	17
3.6.1	引子	17
3.6.2	层次分析法	18
3.6.3	层次分析法的基本步骤	18
3.6.4	假期旅游案例	21
3.7	双层玻璃问题	22
3.7.1	问题的提出	22
3.7.2	量的分析	22
3.7.3	模型假设	23
3.7.4	模型建立	23
3.7.5	模型分析与求解	24

第四章 微积分与微分方程方法建模27

4.1	Malthus 人口模型	27
4.1.1	Malthus 人口论	27

4.1.2	基本概念	27
4.1.3	模型假设	28
4.1.4	模型建立与求解	28
4.1.5	应用案例	28
4.1.6	Logistic 阻滞增长模型	29
4.2	细菌的繁殖数学模型	31
4.2.1	问题的提出	31
4.2.2	模型假设	31
4.2.3	模型建立	31
4.2.4	模型求解	32
4.3	传染病流行的控制模型	32
4.3.1	传统的传染病流行控制模型	32
4.3.2	改进的传染病流行控制模型 (SI 模型)	33
4.3.3	病人得到治愈的 SIS 模型	34
4.4	价格数学模型	35
4.4.1	价格数学模型的定义	35
4.4.2	价格数学模型分类	35
4.4.3	价格数学模型的分析	35
4.4.4	价格数学模型的建立	35
4.4.5	结论	36
4.5	湖泊污染减退模型	36
4.5.1	模型的背景介绍	36
4.5.2	问题的提出	37
4.5.3	模型假设	37
4.5.4	模型的建立与求解	37
4.5.5	结论	38
第五章	线性规划模型	39
5.1	线性规划模型实例	39
5.2	线性规划问题	39
5.3	求解线性规划问题的基本思想	39
5.4	线性规划问题的几何解释和图解法	39
5.5	整数线性规划问题	39
5.6	线性规划的对偶理论	39
5.7	非对称形式的对偶线性规划问题	39
5.8	两铁路平板车的装货问题	39

第六章 对策模型	41
6.1 对策模型的引入	41
6.2 对策模型的基本理论	41
6.3 矩阵对策模型	41
6.4 矩阵对策模型实例分析	41
6.5 鞍点存在定理	41
6.6 混合对策模型	41
第七章 决策模型	43
7.1 决策模型的概念及分类	43
7.2 风险型决策	43
7.3 不确定型决策	43
第三部分 数学建模思想	45
第八章 预测与预报	47
8.1 灰色预测模型	47
8.1.1 灰色预测模型的介绍	47
8.1.2 GM(1,1) 模型 (Gray Model)	47
8.2 微分方程预测	52
8.3 回归分析预测	52
8.4 马尔科夫预测	52
8.5 时间序列预测	52
8.6 小波分析预测	52
8.7 神经网络预测	52
8.7.1 BP 神经网络	52
8.8 混沌序列预测	54
第九章 评价与决策	55
9.1 模糊综合评价	55
9.2 主成分分析	55
9.3 层次分析法 (AHP)	55
9.4 因子分析	55
9.5 数据包络 (DEA) 分析法	55
9.6 秩和比综合评价法	55
9.7 优劣解距离法 (TOPSIS)	55
9.8 投影寻踪综合评价法	55
9.9 方差分析与协方差分析	55

第十章 聚类 and 判别	57
10.1 距离聚类 (常用)	57
10.2 关联性聚类 (常用)	57
10.3 层次聚类	57
10.4 密度聚类	57
10.5 其他聚类	57
10.6 贝叶斯判别 (统计判别方法)	57
10.7 费舍尔判别 (训练样本较多)	57
10.8 模糊识别 (分好类的数据点较少)	57
第十一章 关联与因果	59
11.1 灰色关联分析方法	59
11.2 Sperman 或 kendall 等级相关分析	59
11.3 Person 相关	59
11.4 Copula 相关	59
11.5 典型相关分析	59
11.6 标准化回归分析	59
11.7 生存分析 (事件史分析)	59
11.8 格兰杰因果检验	59
第十二章 优化与控制	61
12.1 线性规划、整数规划、0-1 规划	61
12.2 非线性规划与智能优化算法	61
12.3 多目标规划	61
12.4 动态规划	61
12.5 网络优化	61
12.6 排队论与计算机仿真	61
12.7 模糊规划	61
12.8 灰色规划	61
第四部分 数学建模算法	63
第十三章 蒙特卡洛算法	65
第十四章 数据处理算法	67
第十五章 规划类算法	69
第十六章 图论算法	71

第十七章 计算机算法	73
第十八章 智能优化算法	75
第十九章 网格算法与穷举算法	77
第二十章 一些离散化算法	79
第二十一章 数值分析算法	81
第二十二章 图像处理算法	83
 第五部分 数据分析	 85
第二十三章 数据预处理	87
23.1 数据清洗	87
23.1.1 缺失值分析	87
23.1.2 缺失值处理	87
23.1.3 异常值分析	88
23.1.4 异常值处理	89
23.2 数据集成	89
23.2.1 实体识别	89
23.2.2 冗余属性识别	89
23.3 数据变换	90
23.3.1 简单函数变换	90
23.3.2 数据规范化	90
23.3.3 连续属性离散化	91
23.3.4 属性构造	91
23.4 数据规约	92
23.4.1 属性规约	92
23.4.2 数值规约	93
 第二十四章 数据特征分析	 95
24.1 分布分析	95
24.1.1 定量数据的分布分析	95
24.1.2 定性数据的分布分析	95
 第六部分 数据可视化	 97
第二十五章 可视化基础	99

25.1	绪论	99
25.2	字体	99
25.3	颜色	99
25.3.1	可视化色彩的运用原理	99
第二十六章 统计图表		101
26.1	类别比较型图表	102
26.1.1	柱形图系列	102
26.1.2	条形图系列	102
26.1.3	克利夫兰点图	102
26.1.4	坡度图	102
26.1.5	南丁格尔玫瑰图	102
26.1.6	径向柱图	102
26.1.7	雷达图	102
26.1.8	词云图	102
26.2	数据关系型图表	102
26.2.1	散点图系列	102
26.2.2	曲面拟合	102
26.2.3	等高线图	102
26.2.4	散点曲线图系列	102
26.2.5	瀑布图	102
26.2.6	相关系数图	102
26.3	数据分布型图表	102
26.3.1	统计直方图	102
26.3.2	核密度估计图	103
26.3.3	数据分布图表系列	103
26.3.4	二维统计直方图	103
26.3.5	二维核密度估计图	103
26.4	时间序列型图表	103
26.4.1	折线图	103
26.4.2	面积图	103
26.4.3	日历图	103
26.4.4	量化波形图	103
26.5	局部整体型图表	103
26.5.1	饼图	103
26.5.2	圆环图	103
26.5.3	马赛克图	103

26.5.4	华夫饼图	103
26.5.5	块状/点状柱形图系列	103
26.6	高维数据型图表	103
26.6.1	高维数据的变换展示	103
26.6.2	分面图	103
26.6.3	矩阵散点图	103
26.6.4	热力图	103
26.6.5	平行坐标系图	103
26.6.6	RadViz 图	103
26.7	地理空间型图表	103
26.7.1	不同级别的地图	103
26.7.2	分级统计地图	103
26.7.3	点描法地图	103
26.7.4	带柱形的地图	103
26.7.5	等位地图	103
26.7.6	点状地图	103
26.7.7	简化示意图	103
26.7.8	邮标法	103

第二十七章 数值参数 105

第七部分 LATEX 论文写作 107

第二十八章 论文 109

28.1	题目类型	109
28.2	读题/选题	109
28.3	论文写作时间安排	109
28.4	论文的评选	109
28.5	题目	110
28.6	摘要	110
28.7	正文	110
28.7.1	问题重述	110
28.7.2	问题假设	111
28.7.3	符号说明	111
28.7.4	问题分析与模型准备	111
28.7.5	模型建立与求解	111
28.7.6	模型灵敏性分析	112
28.7.7	模型的讨论与评价	112

28.7.8 模型的改进与推广	112
28.8 参考文献	113
附录 A 代码	115

PART I

第一部分

数学建模基础

1.1 绪论

数学建模的思想与方法

2.1 建模步骤

8 步建模法

2.1.1 问题提出

- 了解实际问题的背景 (属于哪一个领域)
- 明确数学建模的目的 (解决什么问题)

2.1.2 量的分析

- 收集数学建模的必要信息 (相关数据和参考资料)
- 分析研究对象的主要特征 (内在机理或输入输出)

2.1.3 模型假设

根据所研究的对象特征及建模目的,抓住问题本质,忽略次要因素,对问题做出合理的简化假设 (基本符合实际情况,能够用数学语言描述问题)。

2.1.4 模型建立

根据假设,用数学语言、符号描述出研究对象的内在规律,并建立包含常量、变量等的数学模型,可以是函数表达公式、数学方程、数据表格、算法或图形甚至是一段文字描述。

2.1.5 模型求解

采用各种计算方法对所建立的数学模型进行求解,可能是求函数的极值、求方程的解、算法或图形的实现等。

2.1.6 模型分析

- 对求解结果进行数学上的分析
- 包括结果的误差分析（误差是否在允许范围内）
- 统计分析（结果是否符合特定的统计规律）
- 模型对数据的灵敏度分析（模型的结果是否会因数据的微小变化而发生大的变化）
- 对假设的鲁棒性分析（robustness）（模型的结果是否对某一假设非常依赖？）等

2.1.7 模型检验

- 将求解结果和分析结果翻译回到实际问题之中，与实际现象、实际数据进行比较，检验是否与实际吻合。
- 如果吻合较好，则模型及其结果可以应用于实际。如果吻合不好，则需要对模型进行修正。
- 此时问题常常出现在模型假设上（假设是否合理，是否忽略了重要因素而保留了次要因素），所以应对模型假设进行修正或补充，然后重新建模。

2.1.8 模型应用

当模型经过检验已成为一个具有合理性和实用性的模型后，即可以用来解决实际问题。

2.2 建模方法

2.2.1 机理分析法

在对研究对象内部机理分析的基础上，利用建模假设所给出的建模信息或前提条件及相关领域知识、相应的数学工具来构造模型。

2.2.2 系统识别建模法

对系统内部机理不清楚的情况下，利用建模假设或对系统进行实际测试所得到的数据信息，再运用数学方法确定模型形式，借助于概率论和数理统计来辨识参数，构造模型。

2.2.3 仿真建模法

利用各种仿真方法建模。

2.2.4 相似类比建模法

借助于相似原理和事物之间的类比关系进行建模的方法，即根据不同对象之间的某些相似性，借用移植领域的数学模型构造数学模型的方法。

2.3 分类与特点

2.3.1 数学建模的分类

- 按建模的数学方法划分

初等模型、数学规划模型、微分方程模型、差分方程模型、概率统计模型、图论模型、模糊模型和灰色模型等。

- 按建模中变量特点划分

确定性模型与随机性模型、静态模型与动态模型、线性模型与非线性模型、离散模型与连续模型等

- 按应用领域划分

人口模型、交通模型、环境模型、规划模型、生态模型、资源模型等

- 按建模目的划分 描述模型、预测模型、优化模型、决策模型、控制模型等

- 按照对问题的了解程度划分

白箱模型、黑箱模型、灰箱模型

2.3.2 数学建模的特点

- 逼真性和可行性

模型越逼真就越复杂,应用起来费用就越高,常与取得的效益不成正比,所以需要对逼真性与可行性进行折中

- 渐进性

数学模型通常不会是一次就成功的,往往需要反复修正,逐渐完善

- 健壮性

对于已经建好的数学模型,当观测数据有微小的改变或者模型结构及参数发生微小变化时,模型求解的结果也随之发生微小的变化

- 可移植性

数学模型是现实对对象抽象化的产物,它可能与其他领域或其他事物有共性,常常好多领域不同事物却有几乎相同的数学模型

- 非预制性

建模时遇到的问题往往事先没有答案,因此必须创新,产生新方法、新概念

- 条理性

从建模角度出发,人们对现实对象分析应该全面、深入,更具有条理性。即使建模失败,对解决研究实际问题也是有利的

- 技艺性

建模与其说是一门技术,不如说是一种技艺很强的技巧艺术,建模期间经验、想象力、洞察力、判断力以及直觉灵感起的作用往往比数学知识更大

- 局限性

由于建模时往往把显示对象简化、近似、假设,因此当模型应用到实际时就必须考虑被忽略的简化因

素,于是结论往往是相对的、近似的。另外,由于人类认识能力受科学技术以及数学本身发展水平的限制,至今还有不少实际问题没有建立出有价值实用的数学模型

2.4 数学建模能力的培养

- 数学知识的积累
- 多看、多学数学建模案例
- 培养观察能力和用数学解决问题的思想
- 需要丰富的想象力与敏锐、深刻的洞察力
- 兴趣是学习的动力,努力培养建模兴趣
- 与计算机紧密关联,学会使用相关软件
- 注意团队意识和团结协作
- 学会类比,做到“由此及彼和由彼及此”,培养发散思维能力
- 培养自学能力,能快速获取新知识,并能学以致用
- 从杂乱无章的各种信息中快速挑选收集有用信息,利用图书馆、网络查找相关资料

PART II

第二部分

数学建模方法

初等方法建模

3.1 勾股定理与黄金分割率

3.1.1 黄金分割应用于高跟鞋问题

原理:人的下肢长占身高的 0.618 时,满足最佳黄金比例。

计算黄金分割最简单的方法:

设某人身高为 h cm, 下肢长 h_1 cm, 选择高跟鞋的鞋跟高度为 x cm, 则由建模原理得:

$$\frac{x + h_1}{x + h} = 0.618 \quad (3.1)$$

解得:

$$x = \frac{0.618h - h_1}{0.382} \quad (3.2)$$

3.1.2 黄金分割在其他领域的应用

医学与 0.618 的联系

- 外界环境温度为人体温度的 0.618 倍时,人会感到最舒适
- 养生之道:动与静是一个 0.618 的比例关系,即四分动六分静是人们认为的最佳养生方式
- 吃饭六七成饱,几乎不生胃病

建筑学与 0.618 的联系

黄金矩形

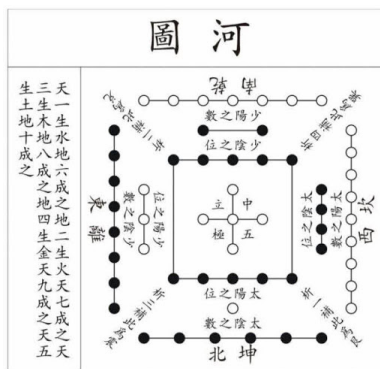
3.2 九宫图

3.2.1 九宫图问题的提出

九宫图又称为三阶幻方,出自西汉(公元前 206-公元 25)学者戴德编纂的《大戴礼》,源自河图洛书。



(a) 九宫图



(b) 河图



(c) 洛书

图 3.1: 九宫图及其起源

河图由十种花点组成,分别代表 1 – 10 这 10 个数,两种花点构成一组,分别布局在东、西、南、北、中五个位置上,每组花点所表示的两个数的差都是 5。

洛书由九种花点组成,分别代表 1 – 9 这 9 个数,各数的位置排列奇巧,奇偶交替变化,纵横六线及两条对角线上三数之和都为 15。

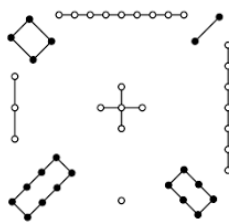
3.2.2 九宫图问题的求解

口诀一

九宫者,法以灵龟。二四为肩,六八为足。左三右七,戴九履一,五居中央。

口诀二

九子斜列,上下对易,左右相更。四维挺出。



(a) 九宫图解一

	1			9			9							
	4	2			4	2			4	2		4	9	2
7	5	3	→	7	5	3	→	3	5	7	→	3	5	7
	8	6			8	6			8	6		8	1	6
	9				1				1					

(b) 九宫图解二

图 3.2: 九宫图解

3.3 椅子稳定问题

3.3.1 问题引入与建模准备

一把四条腿的椅子放在不很平坦的地面上,是否是稳定的?

不很平坦的定义: 肉眼可见的地面是平坦的

椅子不稳定的原因:

- 椅子的四条腿的长度可能不一样
- 地面不平坦
- 椅子的四个底角与地面之间有距离 $h, h > 0$

三个点确定一个平面,对于相对比较平坦的地面来讲,总可以保证三条腿同时着地,因而只需再使其余的一条腿也能完全着地即可。

量的分析

- 椅子的脚和地面的距离作为变量。变量为 0,意味着椅子的脚与地面没有距离,椅子稳定。
- 设计旋转角度的连续函数

3.3.2 模型假设

假设 1: 椅子四条腿一样长,椅脚与地面接触处视为一个点,四脚的连线呈正方形 $ABCD$,四个椅脚的坐标系中对称;

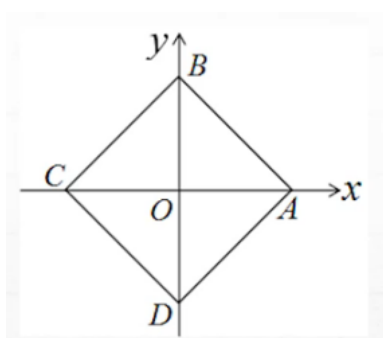


图 3.3: 椅脚坐标系

假设 2: 地面高度是连续变化的,沿任何方向都不会出现间断,如台阶,即地面可视为数学上的连续曲面;

假设 3: 对于椅脚的间距和椅腿长度而言,地面是相对平坦的,使椅子在任何位置至少有三只脚同时着地。

3.3.3 模型建立

构造表示距离的函数:

- 设 $f(\theta)$ 是 A, C 两椅脚与地面的距离之和
- $g(\theta)$ 是 B, D 两椅脚与地面的距离之和
- θ 是椅子绕中心点 O 旋转角度

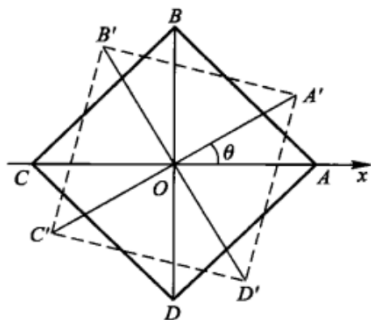


图 3.4: 椅脚坐标系

- 由假设 2, $f(\theta)$ 和 $g(\theta)$ 都是连续函数;
- 由假设 3, 椅子在任何位置至少有三只脚同时着地, 即对任意的 θ , $f(\theta)$ 和 $g(\theta)$ 中至少有一个为零函数。
- 在椅子处于初始不稳定的位置时, 即 $t = 0$ 时, 不妨设 $g(0) = 0, f(0) > 0$ 。

椅子在不很平坦地面上稳定问题的数学命题:

已知 $f(\theta)$ 和 $g(\theta)$ 是 t 的连续函数, 对任意 θ , 满足条件 $f(\theta) \cdot g(\theta) = 0$, 且 $g(0) = 0, f(0) > 0$, 则至少存在一个 θ_0 , 使 $f(\theta_0) = g(\theta_0) = 0$, 即在旋转椅子的时候, 可能在多个位置上都是稳定的。

3.3.4 模型求解

证明:

将椅子逆时针旋转 $\frac{\pi}{2}$, 则对角线 AC 和 BD 的位置相互交换, 且 $f(\theta)$ 和 $g(\theta)$ 是闭区间 $[0, \frac{\pi}{2}]$ 上的连续函数。

由上面数学命题, 一定有下列式成立:

$$\begin{cases} g(0) = 0, f(0) > 0 \\ g(\frac{\pi}{2}) > 0, f(\frac{\pi}{2}) = 0 \end{cases} \quad (3.3)$$

令 $h(\theta) = f(\theta) - g(\theta)$, 显然 $h(\theta)$ 也是闭区间 $[0, \frac{\pi}{2}]$ 上的连续函数,

容易检验: $h(\theta)$ 满足条件 $h(0) > 0, h(\frac{\pi}{2}) < 0$ 。

由闭区间上连续函数的零点存在定理, 至少存在 θ_0 , 使得 $h(\theta_0) = 0$, 即 $f(\theta_0) = g(\theta_0)$ 。

再由命题中的条件“对任意 $\theta, f(\theta) \cdot g(\theta) = 0$ ”, 得知 $f(\theta_0) = g(\theta_0) = 0$ 。

3.4 商人过河问题

3.4.1 问题引入

三个商人各带一名随从渡河, 随从们密约: 在河的任一岸, 一旦随从人数比商人多, 就杀人越货。问题: 商人该如何设计渡河方案以安全渡河?

3.4.2 模型分析

允许状态集合

X_k : 第 k 次渡河前此岸的商人数, $X_k, Y_k = 0, 1, 2, 3$;

Y_k : 第 k 次渡河前此岸的随从数, $k = 1, 2, \dots, n$;

渡河过程的状态: $S_k = (X_k, Y_k)$

允许状态集合: $S = \{(x, y) | x = 0, y = 0, 1, 2, 3; x = 3, y = 0, 1, 2, 3; x = y = 1, 2\}$

渡河状态集合

U_k, V_k : 分别任第 k 次渡河船上的商人数与随从数; $(U_k, V_k = 0, 1, 2; k = 1, 2, \dots, n)$

决策: $d_k = (U_k, V_k)$

允许决策集合: $D = (u, v) | u + v = 1, 2$

状态转移律

$$S_{k+1} = S_k + (-1)^k d_k$$

3.4.3 模型建立

制定安全渡河方案归结为如下的多步决策模型:

求 $d_k \in D (k = 1, 2, \dots, n)$, 使 $s_k \in S$ 按状态转移律由 $S_1 = (3, 3)$ 经有限步 n 到达 $S_{n+1} = (0, 0)$

3.4.4 模型求解

使用图解法, 一共 16 个状态格点, 其中允许状态 S 有 10 个点, 允许决策 D 表示沿方格移动 1 或 2 格。

- (1). k 为奇数, 由此岸去彼岸, 这时此岸人数减少。移动方向为左/下
- (2). k 为偶数, 由彼岸到此岸, 这时此岸人数增多。移动方向为右/上
- (3). d_k 移动一格表示船上有 1 个人, 移动两格表示船上有 2 个人
- (4). 用实线表示从此岸去彼岸, 虚线表示从彼岸回到此岸

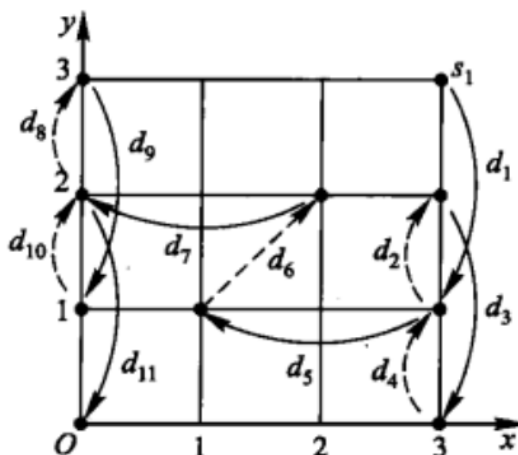


图 3.5: 安全渡河问题图解法

3.5 图论方法与网络模型

3.5.1 图论的起源

图论是组合数学的一个分支,起源于 1736 年欧拉的第一篇关于图论的论文,这篇论文解决了著名的“哥尼斯堡七桥问题”,从而使欧拉成为图论的创始人。

3.5.2 图的概念

图的定义

图是一个有序二元组 $G = (V(G), E(G))$, 其中 $V(G) = \{v_1, v_2, \dots, v_n\}$ 为顶点集 $V(G)$ 中的元素 v_i 称为顶点, $E(G) = \{e_1, e_2, \dots, e_m\}$ 称为边集, $E(G)$ 中的元素 e_k 叫做边。

- 顶点总数记位 $|V(G)|$, 边的总数记位 $|E(G)|$
- 若 $|V(G)| = n$, 则称 G 为 n 阶图
- 若顶点总数 $|V(G)|$ 与边的总数 $|E(G)|$ 均为有限数, 则称 G 为有限图

有向图的定义

若顶点集合 $V(G) \neq \emptyset$, 边集 $E(G) \cap V(G) = \emptyset$, 则称 $G = (V(G), E(G), \emptyset)$ 为一个有向图。

关联函数

若 $\Phi_G(e) = uv$, $\Phi_G(e)$ 称为关联函数, 表示边 e 与顶点 u 与 v 相关联, 又称顶点 u 与 v 相邻, u 是 e 的尾, v 是 e 的头, 即由 u 指向 v 。

边 e 也成为弧, 是由两个顶点组成的有序对, 通常用尖括号表示。例如: $\langle v_i, v_j \rangle$, v_i 称为弧尾, v_j 称为弧头。 $\langle v_i, v_j \rangle$ 和 $\langle v_j, v_i \rangle$ 是两条不同的有向边。

无向图的定义

若 G 的每条边头尾部分, 即 $\Phi_G(e) = uv = vu$, 则称 G 为无向图, 无向图中每条边均是顶点的无序对, 通常用圆括号表示, 例如: (v_i, v_j) 表示一条无向边, 且有 $(v_i, v_j) = (v_j, v_i)$ 。

3.5.3 哥尼斯堡七桥问题

柯尼斯堡七桥问题 (Seven Bridges of Königsberg) 是图论中的著名问题。这个问题是基于一个现实生活中的事例: 当时东普鲁士柯尼斯堡 (今日俄罗斯加里宁格勒) 市区跨普列戈利亚河两岸, 河中心有两个小岛。小岛与河的两岸有七条桥连接。在所有桥都只能走一遍的前提下, 如何才能把这个地方所有的桥都走遍?

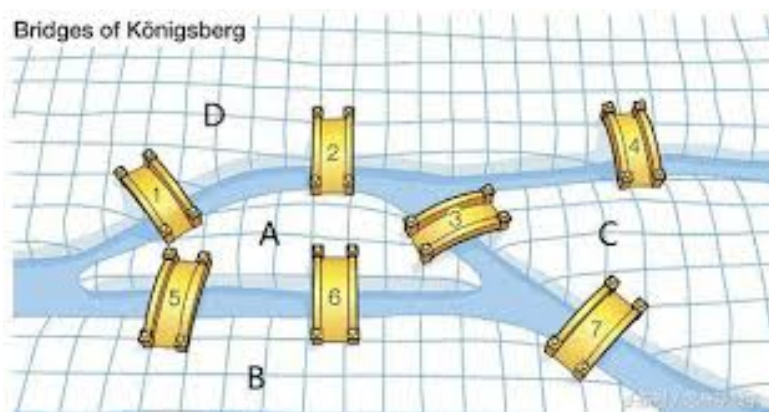


图 3.6: 哥尼斯堡七桥问题

我们将七桥抽象为无向图中的边, 四片陆地抽象为无向图中的点, 该无向图可以表示为 $G = \{V(G), E(G), \Phi_G\}$, 其中 $V(G) = \{A, B, C, D\}$, $E(G) = \{a, b, c, d, e, f, g\}$, $\Phi_G(a) = AB$, $\Phi_G(b) = BC$, $\Phi_G(b) = BC$, $\Phi_G(c) = \Phi_G(d) = AC$, $\Phi_G(e) = \Phi_G(f) = AD$, $\Phi_G(g) = BD$ 。

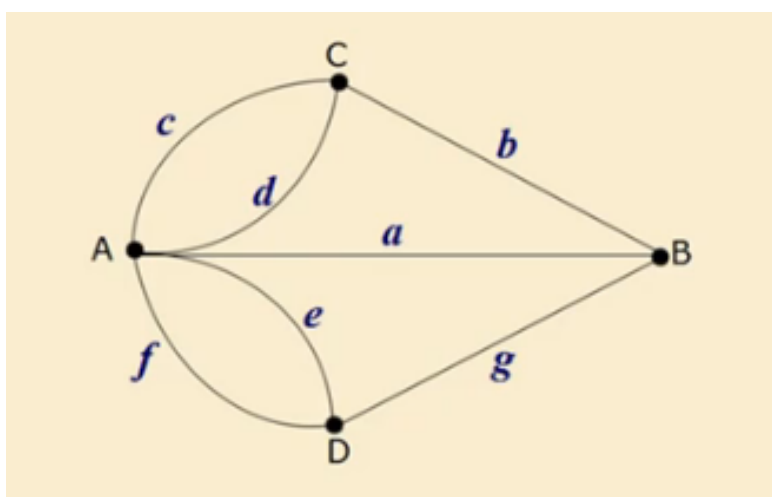


图 3.7: 哥尼斯堡七桥问题图结构

七桥问题的图论阐述

从七桥问题无向图中某个顶点出发, 遍历每条边恰好一次, 最后能否还回到原来的顶点处 (出发点)?

3.6 层次分析方法

3.6.1 引子

生活中一些问题很难用完全定量的数学模型来解决, 对这种复杂决策问题, 运用层次分析方法能够找到最佳的解决办法, 给出最优的决策。层次分析法 (Analytic Hierarchy Process 简称 AHP) 是将与决策有关的元素分解成目标、准则、方案等层次, 在此基础上进行定性和定量分析的决策方法。其基本思路与人对一个复杂的决策问题的思维、判断过程大体上是一样的。

3.6.2 层次分析法

将决策问题按总目标、各层子目标、评价准则直至具体的备选方案的顺序分解为不同的层次结构,然后使用求解判断矩阵特征向量的方法求得每一层次的各元素对上一层次某元素的优先权重,最后再运用加权的方法求出各备选方案对总目标的最终权重向量。其中权重最大者即为最优方案。

层次分析法中每一层的权重设置到最后都会直接或间接影响到结果,而且在每个层次中的每个因素对结果的影响程度都是量化的,非常清晰明确。这种方法擅长于对无结构特性的系统评价以及多目标、多准则、多时期等的系统评价。

层次分析法也具有**局限性**:

- 不能为决策提供新方案,层次分析法只能从原有的备选方案中选取最佳方案,因此可能会产生一个由于主观性或者自身创造能力不够而不是最优的方案。
- 定量数据少,主观因素多,不易令人信服。
- 指标过多时数据统计量大,且权重难以确定。当我们希望解决一个具有普适性的问题时,指标的选取数量很可能会随之增加。指标的增加意味着需要构造的层次更深、数量更多、规模更庞大的判断矩阵。
- 特征值和特征向量的精确求法比较复杂,随着指标的增加,阶数也随之增加,在计算上也变得越来越困难。

3.6.3 层次分析法的基本步骤

1. 建立层次结构模型

在深入分析实际问题的基础上,将有关的各因素按照不同属性自上而下地分解成若干层次,同一层的各个因素因为从属于上一层因素,因此对上层因素会产生较大影响,同时又会支配下一层因素或受到下层因素的影响。

在层次结构模型中,最上层为目标层,通常只有一个因素,最下层通常为方案层,中间可以有一个或几个层次,通常为准则或指标层。当准则过多时(例: > 9),应进一步分解出子准则层。准则层不宜过少,通常5-7个左右。

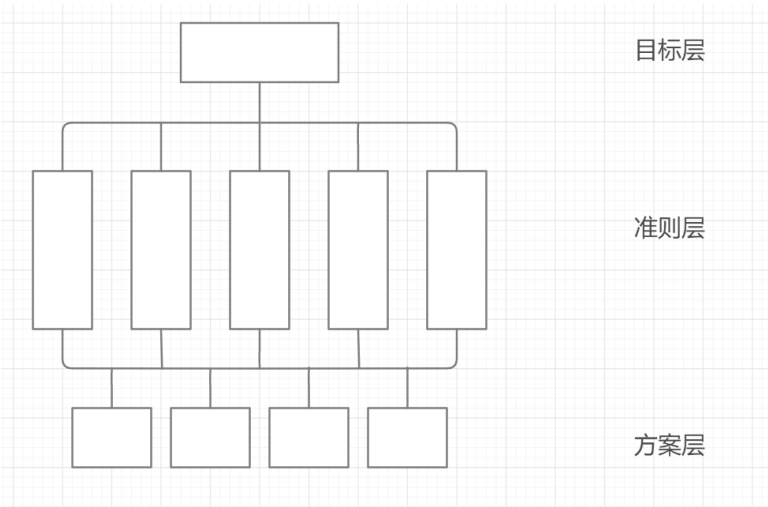


图 3.8: 层次结构模型

2. 构造成对比较矩阵

这一步是要比较层次结构模型的第二层各个因素对上一层因素的影响,从而确定它们对上层因素的影响作用中所占的权重。

设有 n 个因素 x_1, x_2, \dots, x_n 对上一层目标有影响直接确定它们对目标的影响程度不是很容易,所以每次取两个因素 x_i 与 x_j 比较。用 a_{ij} 表示 x_j 和 x_i 对上层目标的影响比,则 $A = (a_{ij})_{n \times n}$ 称为成对比较矩阵,又称为正互反矩阵。

$$A = \begin{bmatrix} \frac{x_1}{x_1} & \frac{x_1}{x_2} & \dots & \frac{x_1}{x_n} \\ \frac{x_2}{x_1} & \frac{x_2}{x_2} & \dots & \frac{x_2}{x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_n}{x_1} & \frac{x_n}{x_2} & \dots & \frac{x_n}{x_n} \end{bmatrix}_{n \times n}, a_{ij} > 0, a_{ii} = 1, a_{ij} = \frac{1}{a_{ji}} (i = 1, 2, \dots, n) \quad (3.4)$$

成对比较矩阵中,每一个 a_{ij} 的取值都是有一定的尺度和规范的,按照 Satty 的提议, a_{ij} 在 1-9 及其倒数中间取值。例如:

- $a_{ij} = 1$, 元素 i 与元素 j 对上一层次因素的重要性相同
- $a_{ij} = 3$, 元素 i 比元素 j 略重要
- $a_{ij} = 5$, 元素 i 比元素 j 重要
- $a_{ij} = 7$, 元素 i 比元素 j 重要的多
- $a_{ij} = 9$, 元素 i 比元素 j 及其重要
- $a_{ij} = 2, 4, 6, 8$, 元素 i 与 j 的重要性介于 1, 3, 5, 7, 9 之间

3. 计算权向量及一致性检验

对于每一个成对比较阵计算其最大特征根 $\lambda_{\max}(A)$ 及对应特征向量,利用一致性指标、平均随机一致性指标和一致性比率做一致性检验。

若检验通过,那么标准化特征向量即为权向量,若不通过,需要重新构造成对比较阵 A 。

由于成对比较矩阵是我们对复杂事物采取两两比较得到的矩阵,构造过程具有明显的主观性,不可能做到判断具有完全的一致性,难免有误差。所以需要成对比较矩阵进行一致性检验。

定义 1

如果一个正互反矩阵 A 满足 $a_{ij}a_{jk} = a_{ik} (i, j, k = 1, 2, \dots, n)$ 则称 A 为一致矩阵,简称一致阵

性质 1

如果矩阵 A 是一致阵,那么它的秩为 1,唯一非零特征根为 n 。

性质 2

一致阵的任一列(行)向量都是对应于特征根 n 的特征向量。

判别法

判别一个 n 阶矩阵 A 是否为一致阵,只要计算 A 的最大特征根即可。如果 A 不是一致阵,则可以证明 $\lambda_{\max}(A) > n$ 而且 $\lambda_{\max}(A) - n$ 越大,说明不一致程度越严重。

定义 2

设 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)^T$ 为正向量,称 α' 为 α 的标准化向量,其中

$$\alpha' = \left(\frac{\alpha_1}{\sum_{i=1}^n \alpha_i}, \frac{\alpha_2}{\sum_{i=1}^n \alpha_i}, \dots, \frac{\alpha_i}{\sum_{i=1}^n \alpha_i} \right) \quad (3.5)$$

一致性检验

设 A 为 n 阶成对比较矩阵

- **一致性指标**: $CI = \frac{\lambda - n}{n-1}$ 来表征一致性程度, 当 $CI = 0$ 时为一致阵, CI 越大, A 的不一致程度越严重。
- **平均随机一致性指标**: RI 来确定 A 的不一致程度的容许范围。对于固定 n , 随机构造成对比较矩阵 $A' = (a'_{ij})$ 其中 a'_{ij} 是从 $1, 2, \dots, 9$ 和 $1, \frac{1}{2}, \dots, \frac{1}{9}$ 中随机抽取的。这样构造的 A' 是不一致的, 它的 CI 相当大。如此构造相当多的 CI , 然后算出这些 A' 的平均值作为平均随机一致性指标的 RI

表 3.1: 随机一致性指标 RI 数值表

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

- **一致性比率**: $CR = \frac{CI}{RI}$, 当 $CR < 0.1$ 时 A 的不一致程度在容许范围内, 可用其标准的特征向量 α' 作为权向量, 否则需要重新调整判断矩阵 A

权向量的计算方法

- (1). 如果成对比较矩阵是一致矩阵, 则把它的列向量 (最大特征根对应的特征向量) 标准化得到的向量即为各个因素对上一层目标影响大小的权向量。
- (2). 如果成对比较矩阵不是一致矩阵, 而它的不一致程度又在容许的范围内, 则计算成对比较矩阵的最大特征根及其相对应的特征向量, 然后将其标准化, 令它作为各个因素对上一层目标影响大小的权向量。
- (3). 当成对比较矩阵的不一致程度很严重时, 需要重新构造或修正成对比较矩阵。

运用方根法近似计算最大特征值与相应的特征向量

- (1). 计算判断矩阵每一行元素的乘积 $W_i = \prod_{j=1}^n a_{ij} (i, j = 1, 2, \dots, n)$
- (2). 计算 W_i 的 n 次方根 $\bar{W}_i = \sqrt[n]{W_i}$
- (3). 对向量 $\bar{W}_i = (\bar{W}_1, \bar{W}_2, \dots, \bar{W}_n)$ 做标准化处理 $a_i = \bar{W}_i \div \sum_{j=1}^n \bar{W}_j (i = 1, 2, \dots, n)$, 得到 $B = (a_1, a_2, \dots, a_n)^T$ 即为所求的特征向量
- (4). 计算最大特征值 $\lambda_{\max}(A) = \frac{1}{n} \sum_{i=1}^n \frac{(AB)_i}{a_i}$, 其中

$$AB = \begin{bmatrix} (AB)_1 \\ (AB)_2 \\ \vdots \\ (AB)_N \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

上式中, $(AB)_i$ 表示向量 AB 的第 i 个元素

4. 计算组合权向量并做组合一致性检验

这是层次分析法的最后一个步骤, 需要计算最下层对最上层目标的组合权向量, 并做组合一致性检验。若检验通过, 则可按照组合权向量表示的结果进行决策, 否则需要重新考虑模型或重新构造那些一致性比率较大的成对比较矩阵。

3.6.4 假期旅游案例

问题描述

假如你在准备一趟假期旅游,有三个景点 A、B、C 供你选择,你将根据景点费用、居住条件、饮食和旅途等准则比较这三个候选景点,并最终决策选择其中一个景点。

层次结构模型

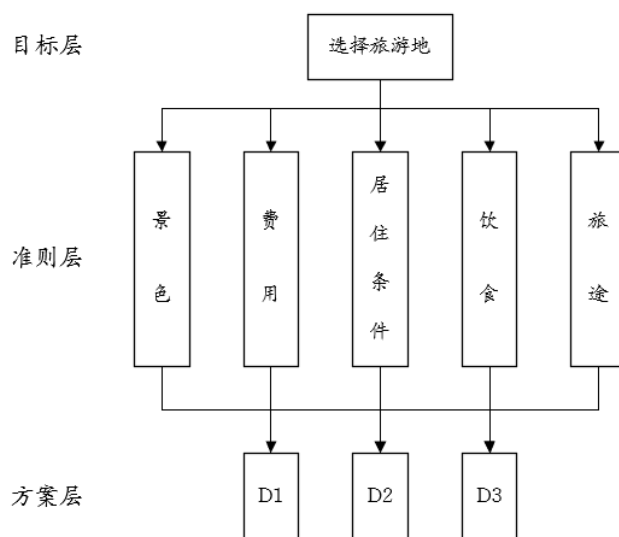


图 3.9: 旅游地选择层次分析模型

构造成对比较矩阵

设景色、费用、居住条件、饮食和交通便利五个因素分别为 X_1, X_2, X_3, X_4, X_5 , 假设成对比较矩阵如下:

$$A = \begin{bmatrix} 1 & \frac{1}{2} & 4 & 3 & 3 \\ 2 & 1 & 7 & 5 & 5 \\ \frac{1}{4} & \frac{1}{7} & 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{5} & 2 & 1 & 1 \\ \frac{1}{3} & \frac{1}{5} & 3 & 1 & 1 \end{bmatrix}$$

其中 $a_{21} = 2$ 表示费用 X_2 与景色 X_1 对选择旅游地这个目标的比是 2 : 1, 说明对费用看的更重要一些, 其他同理。

计算权向量与一致性检验

运用上一节的方法计算可得最大特征值 $\lambda_{max} = 5.073$, 一致性指标 $CI = \frac{\lambda_{max} - n}{n - 1} = \frac{5.073 - 5}{5 - 1} = 0.018$, 查表得 $RI = 1.12$, 最终可得一致性比率 $CR = \frac{CI}{RI} = \frac{0.018}{1.12} = 0.016 < 0.1$, 因此通过了一致性检验。

计算特征根 $\lambda_{max} = 5.073$ 对应的特征向量并对其标准化, 得 $\alpha = \alpha^{(2)} = (0.263, 0.475, 0.055, 0.099, 0.110)^T$, 为了跟后面得步骤进行区分, 这里记位 $\alpha^{(2)}$ 。

从向量中可以看出费用最为重要, 其次是景色和旅途, 再次是饮食和居住条件。

计算组合权向量并做一致性检验

三个方案分别针对五个准则构造成对比较矩阵,如下:

$$B_1 = \begin{bmatrix} 1 & 2 & 5 \\ \frac{1}{2} & 1 & 2 \\ \frac{1}{5} & \frac{1}{2} & 1 \end{bmatrix} \quad B_2 = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{8} \\ 3 & 1 & \frac{1}{3} \\ 8 & 3 & 1 \end{bmatrix} \quad B_3 = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 1 & 3 \\ \frac{1}{3} & \frac{1}{3} & 1 \end{bmatrix} \quad B_4 = \begin{bmatrix} 1 & 1 & 4 \\ \frac{1}{3} & 1 & 1 \\ \frac{1}{4} & 1 & 1 \end{bmatrix} \quad B_5 = \begin{bmatrix} 1 & 1 & \frac{1}{4} \\ 1 & 1 & \frac{1}{4} \\ 4 & 4 & 1 \end{bmatrix}$$

B_1, B_2, B_3, B_4, B_5 表示三种旅游方案 D_1, D_2, D_3 分别对准则层中的五个因素 (景色、费用、居住条件、饮食和旅途) 的影响程度的比较矩阵。例如 B_3 中的 $b_{23} = 3$ 表示方案 D_2 与 D_3 相比居住条件好坏的程度之比。

接着分别计算各矩阵的最大特征根以及相应的权向量,再经过标准化得到标准化向量,然后我们还需要对他们分别进行一致性检验,结果如下表:

表 3.2: 数值表

k	1	2	3	4	5
$\alpha_k^{(3)}$	0.595	0.082	0.429	0.633	0.166
	0.277	0.236	0.429	0.193	0.166
	0.129	0.682	0.142	0.175	0.668
λ_k	3.005	3.002	3	3.009	3
CI_k	0.003	0.001	0	0.005	0
CR_k	0.052	0.002	0	0.009	0

显然,这 5 个矩阵都通过了一致性检验,其中 $\alpha_k^{(3)}$ 即为组合权向量,表示第二层对第一层以及第三层对第二层各个因素的综合影响。

最终我们需要的组合权向量 $\alpha^{(3)} = (\alpha_1^{(3)}, \alpha_2^{(3)}, \alpha_3^{(3)})^T = (0.300, 0.246, 0.456)^T$, 即表示三个景点 D_1, D_2, D_3 中分别占的比重,所以可知方案 D_3 占有更高的比重,因此应该选择方案 D_3 。

3.7 双层玻璃问题

3.7.1 问题的提出

来自北方的同学可能会知道, 身边很多的玻璃窗都是双层的, 两层玻璃之间是真空的空隙。那么这样的工艺有什么好处呢? 它可以有效的阻止室内温度想室外的扩散。那么当建筑物室内外的热传递过程处于热力学平衡状态时, 这种构造形式的双层玻璃窗究竟能比单层玻璃窗阻止多少热量损失? 两层玻璃窗之间的距离控制在多少为好?

3.7.2 量的分析

符号说明

表 3.3: 符号表

符号	说明
T_1	室内温度
T_2	室外温度
d	单层玻璃厚度
l	两层玻璃之间的空气厚度
T_a	内层玻璃的外侧温度
T_b	外层玻璃的内侧温度
k	热传导系数
Q	热量损失

3.7.3 模型假设

1. 热量的传播过程只有传导没有对流,即窗户的密封性能良好,两层玻璃之间的空气是不流动的
2. 室内温度和室外温度保持不变,热传导过程已处于稳定状态即沿热传导方向,单位时间内通过单位面积的热量是常熟
3. 玻璃材料均匀,热传导系数是常数

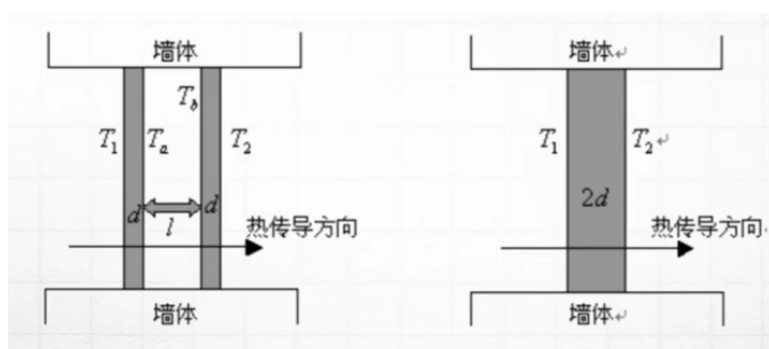


图 3.10: 玻璃热传导示意图

3.7.4 模型建立

热力学传导定律

厚度为 d 的均匀介质,两侧温度差为 ΔT ,则单位时间由温度高的一侧向温度低的一侧通过单位面积的热量 Q 与温度差 ΔT 成正比,与介质的厚度 d 成反比,即

$$Q = k \frac{\Delta T}{d} \quad (3.6)$$

其中, k 为热传导系数。

双层玻璃的热量流失

由热传导方程遵从的物理定律得知

$$Q = k_1 \frac{T_1 - T_a}{d} = k_2 \frac{T_a - T_b}{l} = k_1 \frac{T_b - T_2}{d} \quad (3.7)$$

由 (3.7) 式, 可知

$$T_1 - T_a = T_b - T_2 = \frac{dQ}{k_1}, T_a - T_b = \frac{lQ}{k_2} \quad (3.8)$$

从而得到

$$T_1 - T_2 = (T_1 - T_a) + (T_a - T_b) + (T_b - T_2) = \left(\frac{2d}{k_1} + \frac{l}{k_2}\right)Q \quad (3.9)$$

由此可知

$$Q = \frac{k_1(T_1 - T_2)}{d(s+2)}, s = h\frac{k_1}{k_2}, h = \frac{l}{d} \quad (3.10)$$

单层玻璃的热量流失

对于厚度为 $2d$ 的单层玻璃窗户, 其热量流失为

$$Q' = k_1 \frac{T_1 - T_2}{2d} \quad (3.11)$$

单层玻璃窗和双层玻璃窗热量流失比较

由 (3.7) 式和 (3.11) 式, 可知:

$$\frac{Q}{Q'} = \frac{2}{s+2} < 1 \quad (3.12)$$

由此可见, 双层玻璃窗热量流失一定是小于单层玻璃窗的。

已知不流通、干燥空气的热传导系数是 $k_2 = 2.5 \times 10^{-4}(\text{J}/\text{cm} \cdot \text{s} \cdot \text{度})$, 常用玻璃的热传导系数是 $k_1 = 4 \times 10^{-3} - 8 \times 10^{-3}(\text{J}/\text{cm} \cdot \text{s} \cdot \text{度})$, 于是 $\frac{k_1}{k_2} \in [16, 32]$ 。

在分析双层玻璃窗比单层玻璃窗可减少多少热量损失时, 我们做最保守的估计, 即取 $\frac{k_1}{k_2}$ 的最小值 16, 由 (3.7) 式和 (3.11) 式可得:

$$\frac{Q}{Q'} = \frac{1}{8h+1}, h = \frac{l}{d} \quad (3.13)$$

3.7.5 模型分析与求解

比值 $\frac{Q}{Q'} = (8h+1)^{-1}$ 反映了双层玻璃窗在减少热量损失的功效, 它只与 $h = \frac{l}{d}$ 有关。

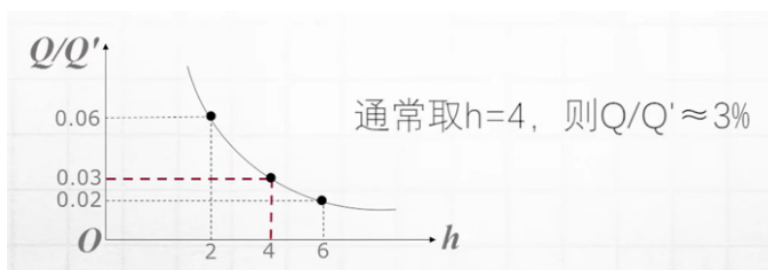


图 3.11: 热量损失比重示意图

通过描述曲线, 我们可以观察到 $h = 4$ 时, $\frac{Q}{Q'} \approx 3\%$, 即 $Q \approx 3\%Q'$, 也就是说双层玻璃窗热量损耗是单层玻璃窗的 3%, 换句话说双层玻璃窗能够阻止 97% 的热量, 要比单层玻璃窗的功效更好。从图中可以看到, 从

$h = 4$ 往后, 曲线的走势趋于平缓, 减少热量损失的能效不明显, 因此实际操作时可以建议操作者, 双层玻璃窗之间的厚度和玻璃之间的比值控制在 4 倍即可。

微积分与微分方程方法建模

4.1 Malthus 人口模型

4.1.1 Malthus 人口论

Malthus 指数增长模型

反映人口增加与食物增加速度之间的关系,是刻画单种群模型的典型代表,被世界公认为“马尔萨斯人口论”

马尔萨斯人口理论主要论点

生活资料按 $1, 2, 3, 4, 5, \dots$ 的算术级数增加,而人口是按 $1, 2, 4, 8, 16, \dots$ 几何级数增长,因此生活资料的增加赶不上人口的增长,这是自然的永恒的规律。

重要假设

大规模种群的个体数量是时间的连续可微函数。

设 $x(t)$ 是 t 时刻的人口总数量,种群个体总数的变化主要受出生、死亡、迁入和迁出等因素的影响。假设种群是孤立的,即不考虑迁入迁出因素,那么种群的繁衍发展就不受其他生物种群的影响。

4.1.2 基本概念

人口自然增长率

一定时期内人口自然增加数与同一时期平均人口之比,即:

$$\text{人口自然增长率} = \frac{\text{一年内人口自然增长率}}{\text{年平均人口数}} \quad (4.1)$$

人口平均增长率

一定时期内人口增加量与期初数和所用时间乘积之比,即:

$$\text{人口平均增长率} = \frac{\text{期末人口总数} - \text{期初人口总数}}{\text{期初人口数} \times \text{所用时间}} \quad (4.2)$$

人口增长率

人口平均增长率在所用时间趋于零时的极限,即:

$$r(t) = \lim_{\Delta t \rightarrow 0} \frac{x(t + \Delta t) - x(t)}{\Delta t \times x(t)} \quad (4.3)$$

其中 $x(t)$ 为 t 时刻的人口总数量。

4.1.3 模型假设

- 假设 1: 人口发展过程比较平稳 (这是建模工作的基本要求);
- 假设 2: 人口数量为时间的连续可微函数, 即人口的取值在整数集合上的离散变量, 不是连续的量。但是由于通常人口数量很庞大, 为了运用微积分工具, 将离散问题做连续化处理;
- 假设 3: 人口增长率是与时间 t 无关的常数 r , 该假设是对欧洲百余年人口数据做统计研究作出的, 是一种近似。

4.1.4 模型建立与求解

根据假设 (3), 人口增长率 r 为与时间无关的常数, 即任意时刻 t , 均有:

$$\lim_{\Delta t \rightarrow 0} \frac{x(t + \Delta t) - x(t)}{\Delta t \times x(t)} = r (r \text{ 为常数}) \quad (4.4)$$

再根据假设 (2), $x(t)$ 为时间的连续可微函数, 可得 $x(t)$ 满足以下微分方程:

$$\begin{cases} \frac{dx}{dt} = rx \\ x(0) = x_0 \end{cases} \quad (4.5)$$

运用分离变量法, 容易求得上述常微分方程的解为

$$x(t) = x_0 e^{rt} \quad (4.6)$$

容易看出, 当 $t \rightarrow \infty$ 时, 人口总数将趋于无穷大, 即

$$\lim_{t \rightarrow \infty} x(t) = \lim_{t \rightarrow \infty} x_0 e^{rt} = \infty \quad (4.7)$$

4.1.5 应用案例

以美国 1790 – 1930 年的人口统计资料为数据依据:

年份	实际人口数	Malthus指数增长模型计算结果	计算误差(%)
1790	3.9	—	—
1800	5.3	—	—
1810	7.2	7.2	0.0
1820	9.6	9.8	2.0
1830	12.9	13.3	3.1
1840	17.1	18.1	5.7
1850	23.2	24.6	5.9
1860	31.4	33.4	6.3
1870	38.6	45.4	17.5
1880	50.2	61.6	22.8
1890	62.9	83.8	33.2
1900	76.0	113.8	49.8
1910	92.0	154.7	68.1
1920	106.5	210.2	97.4
1930	123.2	285.7	131.9

图 4.1: 美国 1790-1930 年人口统计数据 单位:百万

通过这张表,我们可以看出,在短期内的表现 Malthus 人口模型的拟合还是比较吻合的,但是随着时间的推移,长期来看,误差越来越大。为什么会出现这样的误差呢? 我们慢慢来分析。

要想构造 Malthus 人口模型,我们首先就要通过给出的资料来计算人口增长 5 率。假设增长率 r 是常数,我们将 $x(0) = 3.9, x(1) = 5.3$,代入人口指数增长模型中,计算人口增长率 r :

$$r = \frac{\ln x(1) - \ln x(0)}{t} = \frac{\ln 5.3 - \ln 3.9}{10} = \frac{1.66771 - 1.36098}{10} \approx 0.03067 \quad (4.8)$$

结果分析:

种群规模和时间跨度小,符合实际;种群规模和时间跨度大,误差较大。若 $r > 0$,当 $t \rightarrow +\infty$ 时,有 $x(t) \rightarrow +\infty$,因此,该模型不能用于对人口的长期预测。

原因就出在假设 3, r 是常数没有考虑有限的资源对种群的增长会产生阻滞作用。因此如果要更好的预测长期的种群增长情况,我们一定要考虑地球有限的资源会给种群增长带来什么样的影响。因此我们需要对 Malthus 人口模型进行修正。

人口的增加必定会消耗有限的资源,因此有限的资源必定会限制人口的增长,这种现象必将促使人口增长率的下降,也就是说,人口增长率是人口数量的递减函数。我们将据此对 Malthus 模型进行修正,这种修正的结果是我们将得到一个另外的反映人口变动的模型,叫做 Logistic 模型。

4.1.6 Logistic 阻滞增长模型

P.F.Verhulst 引入环境的最大容量常数 X_{max} ,简记为 X_m ,表示自然资源和环境条件下所能容许的最大人口数量,而且为了进一步的考虑资源环境对种群的规模增长所产生的这种阻滞的影响作用,Verhulst 又增加了以下几条假设:

- 假设 4:确定的环境内的资源供给为常数,且对每个个体的分配是均等的,即表明当种群规模(密度)增大时,每个个体食物的平均分配量必然减少,从而导致种群增长率降低。
- 假设 5:种群个体的增长率是 $x(t)$ 的线性减函数,即为:

$$r - r \frac{x(t)}{x_m} = r_1 - \frac{x(t)}{x_m} (x_m > x(t) > 0) \quad (4.9)$$

当 $x(t) = X_m$ 时,种群规模不再增大, X_m 代表环境所能容许的种群最大数量,得到种群增长的 Verhulst 阻滞增长模型:

$$\begin{cases} \frac{dx}{dt} = r \left[1 - \frac{x(t)}{x_m} \right] x(t) \\ x(0) = x_0 \end{cases} \tag{4.10}$$

其中 $1 - \frac{x(t)}{x_m}$ 叫做 Verhulst 因子, $-\frac{r}{x_m}x^2(t)$ 反映种群密度对种群规模增长的抑制作用,称为密度制约。不考虑密度制约时就是原来的 Malthus 模型

Logistic 模型的解

模型 (4.10) 即为 Logistic 模型,其解为:

$$x(t) = \frac{x_m}{1 + (\frac{x_m}{x_0} - 1)e^{-rt}} \tag{4.11}$$

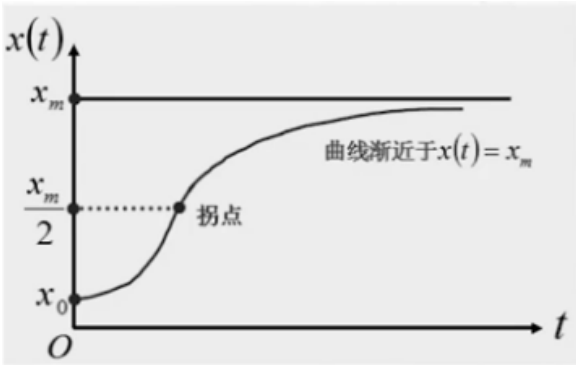


图 4.2: Verhulst 人口模型

人口总量与人口对时间变化率的关系:

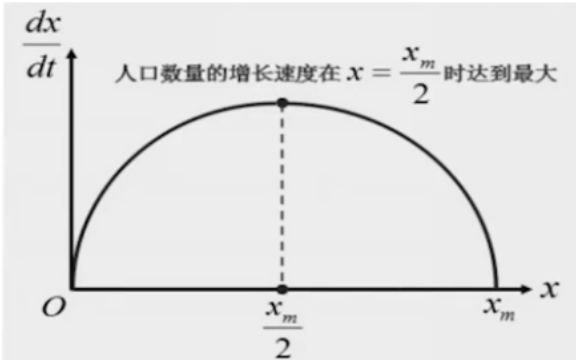


图 4.3: 人口总量与人口对时间变化率的关系

由图可知, $0 - \frac{X_m}{2}$ 时,增长率是单调递增的,当 $x = \frac{X_m}{2}$ 时,增长率达到最大,当 $x = X_m$ 时,增长率为 0。

4.2 细菌的繁殖数学模型

4.2.1 问题的提出

已知某时刻某种细菌的数量,预测在任意时刻 t 这种细菌的数量变化趋势,建立细菌繁殖的数学模型。

4.2.2 模型假设

由实验可知,细菌繁殖的速度应该和培养基是否充足。当培养基充足时,细菌繁殖的速度 V 与当时已有的数量 A_0 成正比,即 $y = kA_0$ ($k > 0$ 为比例常数)

- 假设某种细菌的个数按照指数方式增长,给出如下表:

表 4.1: 细菌繁殖数据表

天数	细菌个数
5	936
10	2190

- 假设在每一个小时间段 $[\frac{(i-1)t}{n}, \frac{it}{n}]$ ($i = 1, 2, \dots, n$) 内细菌的繁殖速度不变,且在各小段时间内只繁殖一次。

按照假设,我们要求的是:

- (1). 开始时细菌的个数可能是多少?
- (2). 若继续以现在速度增长下去,假定细菌无死亡,那么 60 天后细菌的个数应该是多少?

4.2.3 模型建立

首先我们要量化这个数学模型,设细菌总数为 y ,为了计算出 t 时刻细菌的个数,我们需要将时间间隔 $[0, t]$ 分成 N 等分,通过对细菌数量的变化来着手研究细菌的繁衍速度问题。

细菌的繁殖过程可视为连续变化的,在很短的一段时间内数量的变化很小,繁殖的速度可以近似地看作不变。由假设 (2) 在第一段时间 $[0, \frac{t}{n}]$ 内,细菌繁殖的数量为 $\frac{kA_0t}{n}$

第一段时间末细菌的数量为 $A_0(1 + \frac{kt}{n})$;

第二段时间末细菌的数量为 $A_0(1 + \frac{kt}{n})^2$;

依此类推,到最后一段时间末细菌的数量为

$$y = A_0(1 + k\frac{t}{n})^n \quad (4.12)$$

结论 时间间隔无限细分 (即 $n \rightarrow \infty$) 时,可求得精确值,经过时间 t 后细菌的总数是:

$$\begin{aligned} A_0 \lim_{n \rightarrow \infty} \left(1 + k \cdot \frac{t}{n}\right)^n &= A_0 \lim_{n \rightarrow \infty} \left(1 + \frac{kt}{n}\right)^n \\ &= A_0 \lim_{n \rightarrow \infty} \left[\left(1 + \frac{1}{x}\right)^x\right]^{kt} \\ &= A_0 \left[\lim_{x \rightarrow \infty} \left(1 + \frac{1}{x}\right)^x\right]^{kt} \\ &= A_0 e^{kt} \end{aligned} \quad (4.13)$$

至此我们得到细菌繁殖服从生长函数模型:

$$y = A_0 e^{kt} \quad (4.14)$$

其中, k 为繁衍比例系数, t 为时间, A_0 为细菌初始数量。

4.2.4 模型求解

由公式 (4.14) 及题目所给数据表 4.1 得:

$$\begin{cases} 936 = A_0 e^{5k} \\ 2190 = A_0 e^{10k} \end{cases} \quad (4.15)$$

解此方程组,得 $A_0 = 400, k = 0.17$. 即开始时细菌个数为 400 个 按此速度增长下去,则 60 天后细菌个数为:

$$y(60) = 400e^{60 \times 0.17} \approx 10761200(\text{个})$$

4.3 传染病流行的控制模型

通常某种传染病的大面积流行,例如伤风、流感等,最常见的传播方式是带菌患者通过空气、食物和饮食等渠道,把病菌传播给健康人,形成传播锁链。

4.3.1 传统的传染病流行控制模型

假设 t 时刻的病人数目 $x(t)$ 是时间 t 的连续可微函数,每天每个病人有效接触 (足以使被接触者致病) 的人数为常数 k ,考察 t 到 $t + \Delta t$ 发病人数的增加,再假设 $t = t_0$ 时有 x_0 个病人,即得微分方程

$$x(t + \Delta t) - x(t) = k \cdot x(t) \cdot \Delta t \quad (4.16)$$

$$\frac{dx}{dt} = kx(t), x(t_0) = x_0 \quad (4.17)$$

将这个模型对比前面学习过得 Malthus 指数增长模型,可以看出二者非常相似,这个模型得解与 Malthus 指数增长模型的解是完全一样的

模型的结果分析

随着时间 t 的增加,患病人数 $x(t)$ 将无限增加,不符合实际,模型建立失败

原因分析

病人有效接触的人群中既有健康人也有已经发病的人,只有健康人才可以被传染而成为患病的人,建模需要区分人群。

4.3.2 改进的传染病流行控制模型 (SI 模型)

模型假设

- 在疾病传播期内,所考察地区的总人数 N 不变,既不考虑生死,也不考虑迁移。
- 把人群 N 分为易感染者和已感染者两类,简称健康者和病人。在 t 时刻这两类人在总人数中所占的比例分别记作 $x(t)$ 和 $y(t)$ 。
- 每个病人每天有效接触的平均人数是常数 k , k 称为日接触率。假设当病人与健康者进行有效接触时,可使健康者受到感染成为病人。

模型建立

每个病人每天可使 $kx(t)$ 人感染成为病人,病人数为 $N \cdot y(t)$,每天有 $kx(t)Ny(t)$ 个健康者被感染。

- 病人数 $N \cdot y(t)$ 对时间的增长率为

$$\frac{d(Ny)}{dt} = k \cdot x(t) \cdot N \cdot y(t) \quad (4.18)$$

- 初始时刻 $t = 0$,病人的比例记位 $y(0) = y_0$

得到改进的 SI 模型:

$$\begin{cases} \frac{dy}{dt} = k \cdot y(t) \cdot [N - y(t)] \\ y(0) = y_0 \\ N = x(t) + y(t) = 1 \end{cases} \quad (4.19)$$

这依然是一个微分方程的初值问题,其实它就是本章第一节所讲的改进的 Logistic 模型,即阻滞增长模型

模型求解

$$y(t) = \frac{N}{1 + \left(\frac{N}{y_0} - 1\right) \cdot e^{-kt}} \quad (4.20)$$

结果分析

负指数函数 e^{-kt} ($k > 0$) 起到了衰减作用,因此 t 越来越大(发病持续的时间逐渐增加)时, e^{-kt} 将越来越小,因此 $1 + \left(\frac{N}{y_0} - 1\right) \cdot e^{-kt}$ 也将越来越小,从而 $y(t)$ 随 t 的增大单调增加。因此当 $t \rightarrow \infty$ 时, $\lim_{t \rightarrow \infty} y(t) = N$ 。

- 当已感染者达到总人数的一半, 即 $y = 0.5$ 时, $\frac{dy}{dt}$ 达到最大值, 很容易求出这个时刻为 $t_m = k^{-1} \ln\left(\frac{1}{y_0-1}\right)$, 表明此时刻病人增加的速度最快。可以认为这是医院门诊量最大的时刻, 预示着传染病爆发高潮的到来, 需要医疗卫生部门高度关注。

- 由 $t_m = k^{-1} \ln\left(\frac{1}{y_0-1}\right)$ 得知, t_m 与 k 成反比例关系, 原因是模型中的日接触率表示该地区的医疗卫生水平的高低, k 越小, 表明卫生水平越高。因此, 改善保健设施, 提高卫生与防御水平可以推迟传染病爆发高潮的到来时间。

- 由 SI 模型解的表达式 (4.20) 可以看出, $y(t)$ 随 t 单调增加, $t \rightarrow \infty$ 时有 $\lim_{t \rightarrow \infty} y(t) = N$, 这说明所有的人最终都将被传染, 全部成为病人, 不符合实际情况。

4.3.3 病人得到治愈的 SIS 模型

在上一节的结果分析指出, SI 模型解的表达式不符合实际情况。为什么这么说呢? 因为 SI 模型没有考虑到病人还可以被治愈成为健康者, 存在思考上的缺陷。有些传染病被治愈后会使患者的免疫力非常低, 因此可以假设没有免疫性, 于是病人被治愈后成为健康者, 但是健康者还有可能再次被感染, 再次成为病人。这个模型称为 SIS 模型。

模型假设

- 在疾病传播期内, 所考察地区的总人数 N 不变, 既不考虑生死, 也不考虑迁移。
- 把人群 N 分为易感染者和已感染者两类, 简称健康者和病人。在 t 时刻这两类人在总人数种所占的比例分别记作 $x(t)$ 和 $y(t)$ 。
- 每天被治愈的病人数占病人总数的比例为常数 λ , 称为日治愈率。病人治愈后成为仍可被感染的健康者。 $\frac{1}{\lambda}$ 是这种传染病的平均传染期。

模型建立

$$\begin{cases} N \frac{dy}{dt} = k \cdot N \cdot x(t) \cdot y(t) - \lambda \cdot N \cdot y(t) \\ x(t) + y(t) = 1 \end{cases} \quad (4.21)$$

经过整理, 得到:

$$\begin{cases} \frac{dy}{dt} = k \cdot y(t) \cdot [N - y(t)] \\ y(0) = y_0 \end{cases} \quad (4.22)$$

定义 $\sigma = \frac{k}{\lambda}$, 其中 k 是每个病人有效接触 (足以使被接触者致病) 的人数, $\frac{1}{\lambda}$ 是这种传染病的平均传染期。 σ 的含义是整个传染期内的每个病人有效接触的平均人数, 简称接触数。

模型求解分析

利用定义 $\sigma = \frac{k}{\lambda}$, SIS 模型可以改写成:

$$\frac{dy}{dt} = -ky(t) \left[y(t) - \left(1 - \frac{1}{\sigma} \right) \right] \quad (4.23)$$

SIS 模型的结论

接触数 $\sigma = 1$ 是一个阈值 (临界值), 讨论如下:

- $\sigma > 1$ 时, $y(t)$ 的增减性取决于 y_0 的大小, 其极限值 $\lim_{t \rightarrow \infty} y(t) = 1 - \frac{1}{\sigma}$ 随 σ 的增加而增加。
- 当 $\sigma \leq 1$ 时, 病人比例 $y(t)$ 越来越小, 最终趋于零。这是由于传染期内经有效接触从而使健康者变为病人数不超过原来病人数的缘故。

4.4 价格数学模型

4.4.1 价格数学模型的定义

价格数学模型是运用数学的方法对价格形成和价格变动规律所作的描述。最常见的是投入-产出价格数学模型

4.4.2 价格数学模型的分类

- 从价格形成的内在机制上来描述经济变量间的数量关系
- 从经济事物表面现象上探索经济变量间的内在联系
- 组合模型

4.4.3 价格数学模型的分析

一些概念

- 均衡价格是指商品的供给与需求相等时的市场价格
- 实际的市场价格通常不会等于均衡价格, 并且随时间不断变化

假设在某一时刻 t , 某商品的价格与该商品的均衡价格有差别, 此时会存在供需差, 供需差必将促使价格变动

4.4.4 价格数学模型的建立

设某商品在时刻 t 的销售价格为 $p(t)$, 市场对该商品的需求量和它对市场的供给量分别为 $D(p)$ 、 $S(p)$, 它们均为价格 p 的函数。

在 t 时刻价格 $p(t)$ 对时间 t 的变化率 $\frac{dp}{dt}$ 与该商品在同一时刻市场超额需求量 $D(p) - S(p)$ 成正比, 即

$$\frac{dp}{dt} = k[D(p) - S(p)] \quad (k > 0, \text{为比例系数}) \quad (4.24)$$

由经济学原理可知, 商品供给量 $S(p)$ 是价格 p 的单调递增函数, 而市场需求量 $D(p)$ 是价格 p 的单调递减函数。假设

$$S(p) = a + bp, D(p) = c - dp \quad (4.25)$$

其中 a, b, c, d 均为正的常数。当供给量 $S(p)$ 与需求量 $D(p)$ 相等时,对应的价格便是均衡价格 p^* 。
那么当供给量与需求量相等时

$$a + bp = c - dp$$

这表示两条曲线 $S(p) = a + bp$ 与 $D(p) = c - dp$ 的交点恰好为供求平衡时的均衡价格 p^* ,即

$$p^* = (c - a) \div (d + b) \quad (4.26)$$

当商品供不应求时,即 $S(p) < D(p)$, 商品价格要上涨

当商品供过于求时,即 $S(p) > D(p)$, 商品价格要下降

将 (4.25) 式代入 (4.24) 式中,得

$$\frac{dp}{dt} = \alpha(p^* - p) \quad (4.27)$$

其中, $\alpha = k(b + d) > 0$, 另方程 (4.27) 得通解是:

$$p(t) = p^* + Ce^{-\alpha t} \quad (4.28)$$

假设初始价格为 $p(0) = p_0$, 代入上式,得 $C = p_0 - p^*$, 从而得到价格表达式

$$p(t) = p^* + (p_0 - p^*)e^{-\alpha t} \quad (4.29)$$

由 $\alpha > 0$ 时,当 $t \rightarrow \infty$ 时, $e^{-\alpha t} \rightarrow 0$, 因而价格将趋近于均衡价格,即 $p(t) \rightarrow p^*$

4.4.5 结论

随着时间不断延续,市场价格将逐渐趋近于均衡价格 p^*

4.5 湖泊污染减退模型

4.5.1 模型的背景介绍

随着世界经济的迅猛发展,水污染问题日益加剧,水体的污染最突出的问题就是富营养化。根据联合国环境规划署 (UNEP) 的一项调查表明全球范围内的湖泊和水库 30%-40% 都有不同程度的富营养化现象。

湖泊富营养化的特点

- 发展速度快
- 危害程度大
- 治理过程难
- 修复历时长

数学模型可以综合反映系统特征,因此建立该问题的数学模型是针对湖泊污染管理工作种特别有用的手段。

湖泊富营养化研究模型的种类

1. **经验回归模型**: 经验回归模型是在多年实测水质浓度资料及相关环境资料, 如生物数据的基础上, 进行多元回归分析建立起的经验模型。大多用来描述叶绿素和磷与透明度之间的关系。也可用来预测浮游植物生物量, 藻类平均与最大生物量之间的关系。由于自身的缺点, 通常只在数据不太理想或建立复杂模型前用作初步的半定量估计。
2. **营养盐模型**: 主要用来研究引起富营养化的营养物质, 主要是碳、氮、磷, 一般淡水环境中三者的比率为 106 : 16 : 1
3. **浮游植物动力学模型**: 浮游藻类的生长是富营养化的关键过程, 研究氮、磷负荷与浮游藻类生产力的相互作用和关系是揭示湖泊富营养化形成机理的主要途径
4. **生态系统动力学模型**: 研究湖泊的生态结构、功能、时空演变规律及其物理、化学、生物过程对水生生态系统的影响及其反馈机制

4.5.2 问题的提出

如果在某个时刻污染物质停止进入湖里, 那么需要多长时间, 能使湖水的污染浓度下降到污染刚刚停止时污染浓度的 5%?

4.5.3 模型假设

1. 把湖泊视为一个单一的流入、流出系统。
2. 湖水中的污染物质均匀分布在水域里, 不存在局部湖水中污染物浓度高于其他区域的现象。
3. 湖水体积视为一个常数。
4. 模型中各个变量是连续变化并且充分光滑的。
5. 不考虑生物学自身反应或变化导致的污染物在水域中的自行消亡现象, 即污染物质除了流出湖水外, 不会因自身发生某些生物化学反应、沉积、死亡或其他各种方式而自尽自灭。

4.5.4 模型的建立与求解

假设某湖水的体积 V 不变, 记 $x(t)$ 为任一时刻 t 每立方米湖水所含污染物质的数量 (即污染程度的度量), r 为每天流出湖里的水量 (立方米), 则由假设 3 (湖水的体积是常数) 得知 r 也是每天流入湖里的水量

解决该方法的问题依赖以下事实

污染物质对时间的变化率 = 单位时间内流入湖中的污染物质 - 单位时间内流出湖中的污染物质, 即

$$\frac{d}{dt}[x(t) \cdot V] = 0 - r \cdot x(t) \quad (r, V \text{ 为常数}) \quad (4.30)$$

得到污染减退的数学模型, 如下

$$V \frac{dx}{dt} = -rx \quad \text{或} \quad \frac{1}{x} \cdot \frac{dx}{dt} = -\frac{r}{V} \quad (4.31)$$

利用分离变量法,再积分,求得该问题的解为

$$x(t) = x(0)e^{-\frac{r}{V}t} \quad (4.32)$$

当 $x(t) = 5\%x(0)$ 时,有 $5\%x(0) = x(0)e^{-\frac{r}{V}t}$,两边取对数,整理得 $t_{0.05} \approx \frac{3V}{r}$

4.5.5 结论

本模型的建立是在非常简单的假设条件下的一个很粗糙的模型。事实上江河湖海的污染问题是各种各样、复杂多变的。污染物质的减退不仅仅靠湖水多年来自然流动,蒸发和渗漏也是不可忽视的因素,另外不同的污染物质对湖水污染的影响大小也是存在差异的。

线性规划模型

5.1 线性规划模型实例

5.2 线性规划问题

5.3 求解线性规划问题的基本思想

5.4 线性规划问题的几何解释和图解法

5.5 整数线性规划问题

5.6 线性规划的对偶理论

5.7 非对称形式的对偶线性规划问题

5.8 两铁路平板车的装货问题

对策模型

6.1 对策模型的引入

6.2 对策模型的基本理论

6.3 矩阵对策模型

6.4 矩阵对策模型实例分析

6.5 鞍点存在定理

6.6 混合对策模型

决策模型

7.1 决策模型的概念及分类

7.2 风险型决策

7.3 不确定型决策

PART III

第三部分

数学建模思想

预测与预报

8.1 灰色预测模型

8.1.1 灰色预测模型的介绍

灰色预测模型实际上是建立在微分方程的基础上的。灰色的概念是相对白色和黑色而言的,即部分信息已知,部分信息未知。灰色模型就是对既含有已知信息又含有不确定信息的系统进行预测,就是对在一定范围内变化的、与时间有关的灰色过程进行预测。

灰色预测对原始数据进行生成处理来寻找系统变动的规律,并生成有较强规律性的数据序列,然后建立相应的微分方程模型,从而预测事物未来发展趋势的情况。

灰色建模的宗旨是将数据列简称微分方程模型,但是由于微分方程只适合连续可微函数,而时间序列数据非连续,更谈不上可微性,因此灰色预测建模得到的是近似微分方程,称之为灰色微分方程

8.1.2 GM(1,1) 模型 (Gray Model)

模型介绍

GM(1,1) 是使用原始的离散非负数据列,第一个‘1’表示微分方程是一阶的,后面的‘1’表示只有一个变量。GM(1,1) 模型通过一次累加生成削弱随机性的较有规律的新的离散数据列,然后通过建立微分方程模型,得到在离散点处的解经过累减生成的原始数据的近似估计值,从而预测原始数据的后续发展。

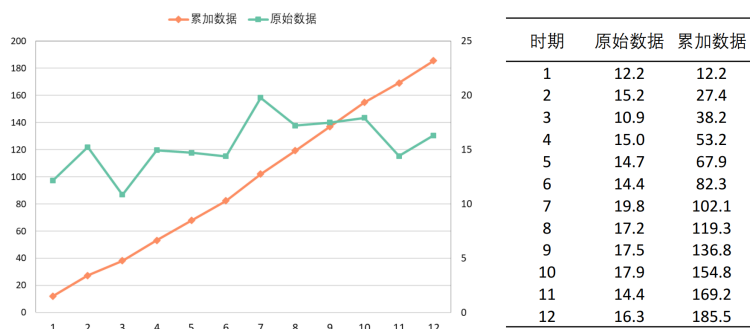


图 8.1: GM(1,1) 模型原理示意图

模型原理

设 $x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n))$ 是最初的非负数据列, 我们对其进行一次累加得到新的生成数据列 $x^{(1)}$, 记为 $x^{(0)}$ 的 1-AGO(Accumulating Generation Operator) 序列:

$$x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(n))$$

其中 $x^{(1)}(m) = \sum_{i=1}^m x^{(0)}(i), m = 1, 2, \dots, n$.

令 $z^{(1)}(m) = \delta x^{(1)}(m) + (1 - \delta)x^{(1)}(m - 1), m = 2, 3, \dots, n$ 且 $\delta = 0.5$

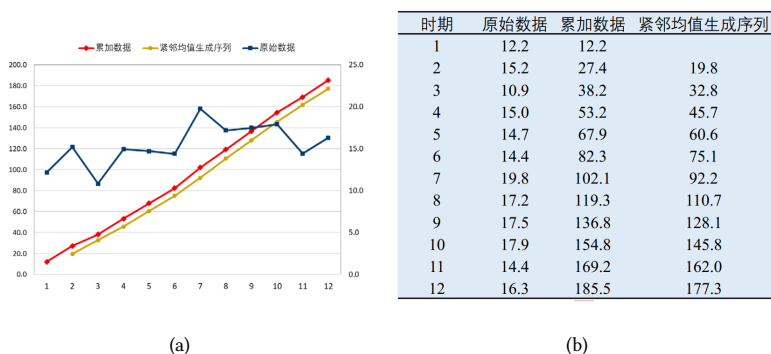


图 8.2: GM(1,1) 模型原理示意图

我们称方程 $x^{(0)}(k) + az^{(1)}(k) = b$ 为 GM(1,1) 模型的基本形式 ($k = 2, 3, \dots, n$)。其中, b 表示灰作用量, $-a$ 表示发展系数。

我们引入矩阵形式:

$$u = (a, b)^T, Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix}, B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(n) & 1 \end{bmatrix}$$

于是, GM(1,1) 模型 $x^{(0)}(k) + az^{(1)}k = b$ 可表示为

$$Y = Bu$$

我们可以利用最小二乘法得到参数 a, b 的估计值为

$$\hat{u} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = (B^T B)^{-1} B^T Y \quad (8.1)$$

看起来很复杂,其实就是将 $x^{(0)}$ 序列视为因变量 y , $z^{(1)}$ 序列视为自变量 x , 进行回归。

$$x^{(0)}(k) = -az^{(1)}(k) + b \Rightarrow y = kx + b \quad (8.2)$$

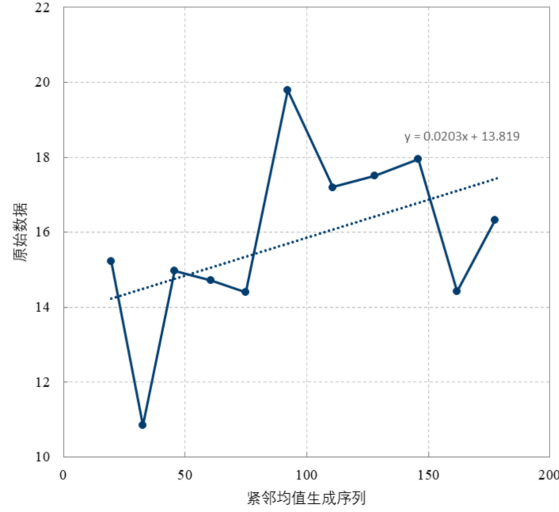


图 8.3: 紧邻均值生成序列拟合示意图

利用 OLS 估计我们能得到 \hat{a} 和 \hat{b} , 即 $x^{(0)}(k) = -\hat{a}z^{(1)}(k) + \hat{b}$ ($k = 2, 3, \dots, n$)

$$x^{(0)}(k) = -\hat{a}z^{(1)}(k) + \hat{b} \Rightarrow x^{(1)}(k) - x^{(1)}(k-1) = -\hat{a}z^{(1)}(k) + \hat{b}$$

$$x^{(1)}(k) - x^{(1)}(k-1) = \int_{k-1}^k \frac{dx^{(1)}(t)}{dt} dt \quad (\text{牛顿-莱布尼茨公式})$$

$$z^{(1)}(k) = \frac{x^{(1)}(k) + x^{(1)}(k-1)}{2} \approx \int_{k-1}^k x^{(1)}(t) dt \quad (\text{定积分的几何意义})$$

$$\int_{k-1}^k \frac{dx^{(1)}(t)}{dt} dt \approx -\hat{a} \int_{k-1}^k x^{(1)}(t) dt + \int_{k-1}^k \hat{b} dt = \int_{k-1}^k [-\hat{a}x^{(1)}(t) + \hat{b}] dt$$

其中微分方程 $\frac{dx^{(1)}(t)}{dt} = -\hat{a}x^{(1)}(t) + \hat{b}$ 被称为 GM(1,1) 模型的白化方程; GM(1,1) 模型的基本形式 $x^{(0)}(k) + az^{(1)}(k) = b$ 则被称为灰色微分方程。

对于白化方程, 如果我们取初始值 $\hat{x}^{(1)}(t)|_{t=1} = x^{(0)}(1)$, 我们可以求出其对应的解为:

$$\hat{x}^{(1)}(t) = \left[x^{(0)}(1) - \frac{\hat{b}}{\hat{a}} \right] e^{-\hat{a}(t-1)} + \frac{\hat{b}}{\hat{a}} \quad (8.3)$$

$$\text{所以 } \hat{x}^{(1)}(m+1) = \left[x^{(0)}(1) - \frac{\hat{b}}{\hat{a}} \right] e^{-\hat{a}m} + \frac{\hat{b}}{\hat{a}}, m = 1, 2, \dots, n-1$$

由于 $x^{(1)}(m) = \sum_{i=1}^m x^{(0)}(i)$, $m = 1, 2, \dots, n$, 所以我们可以得到:

$$\hat{x}^{(0)}(m+1) = \hat{x}^{(1)}(m+1) - \hat{x}^{(1)}(m) = (1 - e^{\hat{a}}) \left[x^{(0)}(1) - \frac{\hat{b}}{\hat{a}} \right] e^{-\hat{a}m}, m = 1, 2, \dots, n-1 \quad (8.4)$$

如果要对原始数据进行预测, 只需要在上式取 $m \geq n$ 即可。由上式也可以看出, 对于序列 $x^{(1)}(k)$, GM(1,1) 模型的本质是有条件的指数拟合: $f(x) = C_1 e^{C_2(x-1)}$

最小二乘估计原理 (OLS)

公式 (8.1) 的推导如下:

$$y_i = kx_i + b + u_i$$

$$\hat{k}, \hat{b} = \arg \min_{k,b} \hat{u}_i^2 = \arg \min_{k,b} \sum_{i=1}^n (y_i - kx_i - b)^2$$

$$\text{令 } L = \sum_{i=1}^n (y_i - kx_i - b)^2 = [y_1 - kx_1 - b, y_2 - kx_2 - b, \dots, y_n - kx_n - b] \begin{bmatrix} y_1 - kx_1 - b \\ y_2 - kx_2 - b \\ \vdots \\ y_n - kx_n - b \end{bmatrix}$$

$$\text{令矩阵 } Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \beta = \begin{bmatrix} b \\ k \end{bmatrix}$$

$$\text{则 } X\beta = \begin{bmatrix} b + kx_1 \\ b + kx_2 \\ \vdots \\ b + kx_n \end{bmatrix}, Y - X\beta = \begin{bmatrix} y_1 - kx_1 - b \\ y_2 - kx_2 - b \\ \vdots \\ y_n - kx_n - b \end{bmatrix}$$

$$\text{所以有 } L = (Y - X\beta)^T(Y - X\beta) = (Y^T - \beta^T X^T)(Y - X\beta) = Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta$$

$$\text{则 } \hat{\beta} = \begin{bmatrix} \hat{b} \\ \hat{k} \end{bmatrix} = \arg \min_{\beta} (Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta)$$

$$\frac{\partial L}{\partial \beta} = -X^T Y - X^T Y + 2X^T X\beta = 0 \Rightarrow X^T X\beta = X^T Y$$

$$\text{所以 } \hat{\beta} = (X^T X)^{-1} X^T Y$$

矩阵求导知识参考[常用的向量矩阵求导公式](#)

准指数规律的检验

要想较好的应用灰色预测模型,必须了解:

1. 数据具有准指数规律,这是使用灰色预测建模的理论基础。
2. 累加 r 次的序列为: $x^{(r)} = (x^{(r)}(1), x^{(r)}(2), \dots, x^{(r)}(n))$, 定义级比: $\sigma(k) = \frac{x^{(r)}(k)}{x^{(r)}(k-1)}, k = 2, 3, \dots, n$, 光滑比: $\rho(k) = \frac{x^{(0)}(k)}{x^{(1)}(k-1)}$
3. 如果 $\forall k, \sigma(k) \in [a, b]$, 且区间长度 $\delta = b - a < 0.5$, 则称累加 r 次后的序列具有准指数规律
4. 具体到 GM(1,1) 模型中, 我们只需要判断累加一次后的序列 $x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(n))$ 是否具有准指数规律。

根据上述公式: 序列 $x^{(1)}$ 的级比 $\sigma(k) = \frac{x^{(1)}(k)}{x^{(1)}(k-1)} = \frac{x^{(0)}(k) + x^{(1)}(k-1)}{x^{(1)}(k-1)} = \frac{x^{(0)}(k)}{x^{(1)}(k-1)} + 1$; 原式序列 $x^{(0)}$ 光滑比 $\rho(k) = \frac{x^{(0)}(k)}{x^{(0)}(1) + x^{(0)}(2) + \dots + x^{(0)}(k-1)}$, 假设 $x^{(0)}$ 为非负序列 (生活中常见的时间序列几乎都满足非负性), 那么随着 k 的增加, 最终 $\rho \rightarrow 0$ 。

因此要使得 $x^{(1)}$ 具有准指数规律, 即 $\forall k$, 区间长度 $\delta < 0.5$, 只需要保证 $\rho(k) \in (0, 0.5)$ 即可, 此时序列

$x^{(1)}$ 的级比 $\sigma(k) \in (1, 1.5)$ 。实际建模中,我们要计算出 $\rho(k) \in (0, 0.5)$ 的占比,占比越高越好(一般前两期 $\rho(2)$ 和 $\rho(3)$ 可能不符合要求,我们重点关注后面的期数)

发展系数与预测情形的关系

GM(1,1) 适用情况和发展系数的大小有很大的关系,一般情况下,发展系数越小,预测的越精确。

GM(1,1) 模型的评价

使用 GM(1,1) 模型对未来的数据进行预测时,我们需要先检验 GM(1,1) 模型对原数据的拟合程度(对原始数据还原的效果),一般有两种检验方法:

1. 残差检验

- 绝对残差: $\varepsilon(k) = x^{(0)}(k) - \hat{x}^{(0)}(k), k = 2, 3, \dots, n$
- 相对残差: $\varepsilon_r(k) = \frac{|x^{(0)}(k) - \hat{x}^{(0)}(k)|}{x^{(0)}(k)}, k = 2, 3, \dots, n$
- 平均相对残差: $\bar{\varepsilon}_r = \frac{1}{n-1} \sum_{k=2}^n |\varepsilon_r(k)|$

如果 $\bar{\varepsilon}_r < 20\%$,则认为 GM(1,1) 对原数据的拟合达到一般要求。

如果 $\bar{\varepsilon}_r < 10\%$,则认为 GM(1,1) 对原数据的拟合效果非常不错。

2. 级比偏差检验

首先由 $x^{(0)}(k-1)$ 和 $x^{(0)}(k)$ 计算出原始数据的级比 $\sigma(k)$:

$$\sigma(k) = \frac{x^{(0)}(k)}{x^{(0)}(k-1)} (k = 2, 3, \dots, n)$$

再根据预测出来的发展系数 $(-\hat{a})$ 计算出相应的级比偏差和平均级比偏差:

$$\eta(k) = \left| 1 - \frac{1 - 0.5\hat{a}}{1 + 0.5\hat{a}} \frac{1}{\sigma(k)} \right|, \bar{\eta} = \frac{1}{n-1} \sum_{k=2}^n \eta(k)$$

如果 $\bar{\eta} < 0.2$,则认为 GM(1,1) 对原数据的拟合达到一般要求。

如果 $\bar{\eta} < 0.1$,则认为 GM(1,1) 对原数据的拟合效果非常不错。

灰色预测的使用场景

1. 数据是以年份度量的非负数据(如果是月份或者季度数据一定用时间序列模型);
2. 数据能经过准指数规律的检验(除了前两期外,后面至少 90% 的期数的光滑比要低于 0.5);
3. 数据的期数较短且和其他数据之间的关联性不强(小于等于 10,也不能太短),如果数据期数较长,一般使用传统的时间序列模型。

8.2 微分方程预测

8.3 回归分析预测

8.4 马尔科夫预测

8.5 时间序列预测

8.6 小波分析预测

8.7 神经网络预测

8.7.1 BP 神经网络

误差逆传播算法

给定训练集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}^l$, 即输入示例由 d 个属性描述, 输出 l 维实值向量。如下所示:

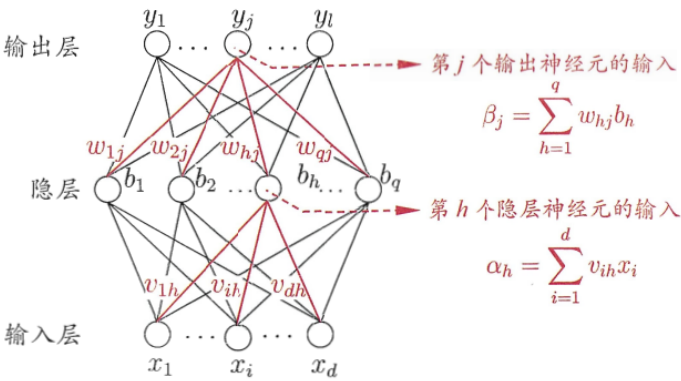


图 8.4: BP 神经网络及算法中的变量符号

图中给出了一个拥有 d 个输入神经元、 l 个输出神经元、 q 个隐层神经元的多层前馈网络结构, 其中输出层第 j 个神经元的阈值用 θ_j 表示, 隐层第 h 个神经元的阈值用 γ_h 表示. 输入层第 i 个神经元与隐层第 h 个神经元之间的连接权重为 v_{ih} , 隐层第 h 个神经元与输出层第 j 个神经元之间的连接权重为 w_{hj} . 记隐层第 h 个神经元接收到的输入为 $\alpha_h = \sum_{i=1}^d v_{ih}x_i$, 输出层第 j 个神经元接收到的输入为 $\beta_j = \sum_{h=1}^q w_{hj}b_h$, 其中 b_h 为隐层第 h 个神经元的输出. 假设隐层和输出层神经元都是用 sigmoid 函数 $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$ 作为激活函数.

对训练样例 (x_k, y_k) , 假定神经网络的输出为 $\hat{y}_k = (\hat{y}_1^k, \hat{y}_2^k, \dots, \hat{y}_l^k)$, 即

$$\hat{y}_j^k = f(\beta_j - \theta_j) \tag{8.5}$$

则网络在 (x_k, y_k) 上的均方误差为

$$E_k = \frac{1}{2} \sum_{j=1}^l (\hat{y}_j^k - y_j^k)^2 \quad (8.6)$$

图 8.4 中的网络中有 $(d+l+1)q+l$ 个参数需确定: 输入层到隐层的 $d \times q$ 个权值、隐层到输出层的 $q \times l$ 个权值、 q 个隐层神经元的阈值、 l 个输出层神经元的阈值。BP 是一个迭代学习算法, 在迭代的每一轮中采用广义的感知机学习规则对参数进行更新估计, 任意参数 v 的更新估计式为

$$v \leftarrow v + \Delta v \quad (8.7)$$

下面我们以图 8.4 中隐层到输出层的连接权 w_{hj} 为例来进行推导。

BP 算法基于梯度下降 (gradient descent) 策略, 以目标的负梯度方向对参数进行调整。对式 (8.6) 的误差 E_k , 给定学习率 η , 有

$$\Delta w_{hj} = -\eta \frac{\partial E_k}{\partial w_{hj}} \quad (8.8)$$

注意到 w_{hj} 先影响到第 j 个输出层神经元的输入值 β_j , 再影响到其输出值 \hat{y}_j^k , 然后再影响到 E_k , 有

$$\frac{\partial E_k}{\partial w_{hj}} = \frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial w_{hj}} \quad (8.9)$$

根据 β_j 的定义, 显然有

$$\frac{\partial \beta_j}{\partial w_{hj}} = b_h \quad (8.10)$$

sigmoid 函数有一个很好的性质:

$$f'(x) = f(x)(1 - f(x)) \quad (8.11)$$

于是根据式 (8.6) 和 (8.5), 有

$$\begin{aligned} g_j &= -\frac{\partial E_k}{\partial \hat{y}_j^k} \cdot \frac{\partial \hat{y}_j^k}{\partial \beta_j} \\ &= -(\hat{y}_j^k - y_j^k) f'(\beta_j - \theta_j) \\ &= \hat{y}_j^k (1 - \hat{y}_j^k) (y_j^k - \hat{y}_j^k) \end{aligned} \quad (8.12)$$

将式 (8.12) 和 (8.10) 代入式 (8.9), 再代入式 (8.8), 得到 BP 算法中关于 w_{hj} 的更新公式

$$\Delta w_{hj} = \eta g_j b_h \quad (8.13)$$

类似可得

$$\Delta \theta_j = -\eta g_j \quad (8.14)$$

$$\Delta v_{ih} = \eta e_h x_i \quad (8.15)$$

$$\Delta v_h = -\eta e_h \quad (8.16)$$

式 (8.15) 和 (8.16) 中

$$\begin{aligned}
 e_h &= -\frac{\partial E_k}{\partial b_h} \cdot \frac{\partial b_h}{\partial \alpha_h} \\
 &= -\sum_{j=1}^l \frac{\partial E_k}{\partial \beta_j} \cdot \frac{\partial \beta_j}{\partial b_h} f'(\alpha_h - \gamma_h) \\
 &= \sum_{j=1}^l w_{hj} g_j f'(\alpha_h - \gamma_h) \\
 &= b_h (1 - b_h) \sum_{j=1}^l w_{hj} g_j
 \end{aligned} \tag{8.17}$$

学习率 $\eta \in (0, 1)$ 控制着算法每一轮迭代的更新步长, 若太大则容易震荡, 太小则收敛速度又会过慢. 有时, 为了做精细调节, 可令式 (8.13) 和 (8.14) 用 η_1 , 式 (8.15) 和 (8.16) 用 η_2 , η_1 和 η_2 不一定相等. 下图给出了 BP 算法的工作流程.

输入: 训练集 $D = \{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^m$;
学习率 η .

过程:

- 1: 在 $(0, 1)$ 范围内随机初始化网络中所有连接权和阈值
- 2: **repeat**
- 3: **for all** $(\mathbf{x}_k, \mathbf{y}_k) \in D$ **do**
- 4: 根据当前参数和式(5.3) 计算当前样本的输出 $\hat{\mathbf{y}}_k$;
- 5: 根据式(5.10) 计算输出层神经元的梯度项 g_j ;
- 6: 根据式(5.15) 计算隐层神经元的梯度项 e_h ;
- 7: 根据式(5.11)-(5.14) 更新连接权 w_{hj} , v_{ih} 与阈值 θ_j , γ_h
- 8: **end for**
- 9: **until** 达到停止条件

输出: 连接权与阈值确定的多层前馈神经网络

图 8.5: 误差逆传播算法

对每个训练样例, BP 算法执行以下操作:

1. 先将输入示例提供给输入层神经元, 然后逐层将信号前传, 直到产生输出层的结果;
2. 然后计算输出层的误差 (第 4-5 行), 再将误差逆向传播至隐层神经元 (第 6 行);
3. 最后根据隐层神经元的误差来对连接权和阈值进行调整 (第 7 行);
4. 迭代过程循环进行, 直到满足终止条件为止.

8.8 混沌序列预测

评价与决策

9.1 模糊综合评价

9.2 主成分分析

9.3 层次分析法 (AHP)

9.4 因子分析

9.5 数据包络 (DEA) 分析法

9.6 秩和比综合评价法

9.7 优劣解距离法 (TOPSIS)

9.8 投影寻踪综合评价法

9.9 方差分析与协方差分析

聚类和判别

10.1 距离聚类 (常用)

10.2 关联性聚类 (常用)

10.3 层次聚类

10.4 密度聚类

10.5 其他聚类

10.6 贝叶斯判别 (统计判别方法)

10.7 费舍尔判别 (训练样本较多)

10.8 模糊识别 (分好类的数据点较少)

关联与因果

11.1 灰色关联分析方法

11.2 Sperman 或 kendall 等级相关分析

11.3 Person 相关

11.4 Copula 相关

11.5 典型相关分析

11.6 标准化回归分析

11.7 生存分析（事件史分析）

11.8 格兰杰因果检验

优化与控制

12.1 线性规划、整数规划、0-1 规划

有约束,确定的目标

12.2 非线性规划与智能优化算法

12.3 多目标规划

柔性约束,目标含糊

12.4 动态规划

12.5 网络优化

多因素交错复杂

12.6 排队论与计算机仿真

12.7 模糊规划

范围约束

12.8 灰色规划

PART IV

第四部分

数学建模算法

蒙特卡洛算法

数据处理算法

规划类算法

图论算法

计算机算法

智能优化算法

网格算法与穷举算法

一些离散化算法

数值分析算法

图像处理算法

PART V

第五部分

数据分析

数据预处理

23.1 数据清洗

23.1.1 缺失值分析

数据的缺失主要包括记录的缺失和记录中某个字段信息的缺失,两者都会造成分析结果的不准确,以下从缺失值产生的原因及影响等方面展开分析。

缺失值产生的原因

- (1). 有些信息暂时无法获取,或者获取信息的代价太大。
- (2). 有些信息是被遗漏的。可能是输入时认为不重要、忘记填写或对数据理解错误等一些认为因素而遗漏,也可能是由于数据采集设备的故障、存储介质的故障、传输媒体的故障等人为原因而丢失。
- (3). 属性值不存在。在某些情况下,缺失值并不意味着数据有错误。对一些对象来说某些属性值是不存在的,如一个未婚者的配偶姓名、一个儿童的固定收入等。

缺失值的影响

- (1). 数据挖掘建模将丢失大量的有用信息。
- (2). 数据挖掘模型所表现出的不确定型更加显著,模型中蕴含的规律更难把握。
- (3). 包含空值的数据会建模过程陷入混乱,导致不可靠的输出。

23.1.2 缺失值处理

缺失值处理的方法可分为 3 类:删除记录、数据插补和不处理,常用的插补方法有:

插补方法	方法描述
均值/中位数/众数插补	根据属性值类型,用该属性取值的平均数/中位数/众数进行插补
使用固定值	将缺失的属性值用一个常量替换
最近邻插补	在记录中找到与缺失值样本最接近的样本的属性值插补
回归方法	根据已有数据和有关的其他变量建立拟合模型来预测缺失的属性值
插值法	利用已知点建立合适的插值函数 $f(x)$,未知值由对应点 x_i 求出的函数值 $f(x_i)$ 近似代替

表 23.1: 常用的插补方法

23.1.3 异常值分析

异常值分析是检验数据是否有录入错误以及含有不合常理的数据。忽视异常值的存在是十分危险的。

1. 简单统计量分析

可以先对变量做一个描述性统计,进而查看哪些数据是不合理的,最常用的统计量是最大值和最小值,用来判断这个变量的取值是否超出了合理的范围。如客户年龄的最大值为 199 岁,则该变量的取值存在异常。

2. 3σ 原则

如果数据服从正态分布,在 3σ 原则下,异常值被定义为一组测定值中与平均值的偏差超过 3 倍标准差的值。在正态分布的假设下,距离平均值 3σ 之外的值出现的概率为 $P(|x - \mu| > 3\sigma) \leq 0.003$,属于极个别小概率事件。

如果数据不服从正态分布,也可以用远离平均值的多少倍标准差来描述

3. 箱型图分析

箱型图提供了一个识别异常值的标准:异常值通常被定义为小于 $Q_L - 1.5IQR$ 或大于 $Q_U + 1.5IQR$ 的值。 Q_L 称为下四分位数,表示全部观察值中有四分之一的数据取值比它小; Q_U 称为上四分位数,表示全部观察值中有四分之一的数据取值比它大; IQR 称为四分位数间距,是上四分位数 Q_U 与下四分位数 Q_L 之差,其间包含了全部观察值的一半。

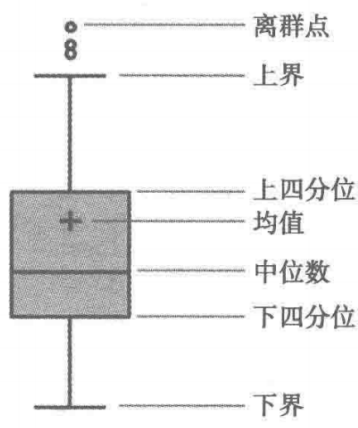


图 23.1: 箱型图检测异常值

一方面,箱型图依据实际数据绘制,没有对数据作任何限制性要求(如服从某种特定的分布形式),它只是真实直观地表现数据分布地本来面貌;

另一方面,箱型图判断异常值的标准以四分位数和四分位距为基础,四分位数具有一定的鲁棒性:多大 25% 的数据可以变得任意远而不会很大的扰动四分位数,所以异常值不能对这个标准施加影响;由此可见,箱型图识别异常值的结果比较客观,在识别异常值方面有一定的优越性。

23.1.4 异常值处理

异常值处理常用方法如下:

异常值处理方法	方法描述
删除含有异常值的记录	直接将含有异常值的记录删除
视为缺失值	将异常值视为缺失值,利用缺失值处理的方法进行处理
平均值修正	可用前后两个观测值的平均值修正该异常值
不处理	直接在具有异常值的数据集上进行挖掘建模

表 23.2: 常用的异常处理方法

23.2 数据集成

23.2.1 实体识别

实体识别是指从不同数据源识别出现实世界的实体,它的任务是统一不同源数据的矛盾之处,常见形式如下:

1. 同名异义

数据源 A 中的属性 *ID* 和数据源 B 中的属性 *ID* 分别描述的是菜品编号和订单编号,即描述的是不同的实体。

2. 异名同义

数据源 A 中的 *sales_dt* 和数据源 B 中的 *sales_date* 都是描述销售日期的,即 $A.sales_dt = B.sales_date$

3. 单位不统一

分别用不同的单位描述同一个实体

23.2.2 冗余属性识别

数据集成往往会导致数据冗余,例如

1. 同一属性多次出现;
2. 同一属性命名不一致导致重复

仔细整合不同源数据能减少甚至避免数据冗余不一致,从而提高数据挖掘的速度和质量。对于冗余属性要先分析,检测到后再将其删除。

有些冗余属性可以用相关分析检测。给定两个数值型的属性 A 和属性 B,根据其属性值,用相关系数度量一个属性在多大程度上蕴含另一个属性。

23.3 数据变换

23.3.1 简单函数变换

简单函数变换常用来将不具有正态分布的数据变换成具有正态分布的数据。在时间序列分析中,有时简单的对数变换或者差分运算就可以将非平稳序列转换成平稳序列。在数据挖掘中,简单的函数变换可能更有必要,比如个人年收入的取值范围为 10000 元到 10 亿元,这是一个很大的区间,使用对数变换将其进行压缩是常用的一种变换处理方法。常用的变换包括平方、开方、取对数、差分运算等,即:

$$x' = x^2$$

$$x' = \sqrt{x}$$

$$x' = \log(x)$$

$$\nabla f(x_k) = f(x_{k+1}) - f(x_k)$$

23.3.2 数据规范化

数据规范化也叫数据归一化,是数据挖掘的一项基础工作,不同评价指标往往具有不同的量纲,数值之间的差别可能很大,不进行处理可能会影响到数据分析的结果。为了消除指标之间的量纲和取值范围差异的影响,需要进行标准化处理,将数据按照比例进行缩放,使之落入一个待定的区域,便于进行综合分析。数据规范化对于基于距离的挖掘算法尤为重要。

1. 最小-最大规范化

最小最大规范化也成为离差标准化,是对原始数据的线性变换,将数值值映射到 $[0, 1]$ 之间。转换公式如下:

$$x^* = \frac{x - \min}{\max - \min} \quad (23.1)$$

离差标准化保留了原来数据中存在的关系。这种处理方法的缺点是若数值集中且某个数值很大,则规范化后各值会接近于 0,并且会相差不大。若将来遇到超过目前属性 $[\min, \max]$ 取值范围的时候,会引起系统出错,需要重新确定 \min 和 \max 。

2. 零-均值规范化

零-均值规范化也称标准差标准化,经过处理的数据的均值为 0,标准差为 1。转化公式为:

$$x^* = \frac{x - \bar{x}}{\sigma} \quad (23.2)$$

其中 \bar{x} 是原始数据的均值, σ 是原始数据的标准差。是当前用的最多的数据标准化方法。

3. 小数定标规范化

通过移动属性值的小数位数,将属性值映射到 $[-1, 1]$ 之间,移动的小数位数取决于属性值绝对值的最大值。转化公式为:

$$x^* = \frac{x}{10^k} \quad (23.3)$$

23.3.3 连续属性离散化

离散化的过程

连续属性得到离散化就是在数据的取值范围内设定若干个离散的划分点,将取值范围划分为一些离散化的区间,最后用不同的符号或整数值代表落在每个子区间中的数据值。所以离散化主要分为两个过程:确定分类数和如何将连续属性值映射到这些分类值。

常用的离散化方法

常用的离散化方法有等宽法、等频法和(一维)聚类:

1. 等宽法

将属性的值域分成具有相同宽度的区间,区间的个数由数据本身的特点决定,或者由用户指定,类似制作频率分布表。

2. 等频法

将相同数量的记录放进每个区间

3. 基于聚类分析的方法

一维聚类的方法包括两个步骤,首先将连续属性的值用作聚类算法(如 K-Means 算法)进行聚类,然后再将聚类得到的簇进行处理,合并到一个簇的连续属性值并做同一标记。聚类分析的离散化方法也需要用户指定簇的个数,从而决定产生的区间数。

23.3.4 属性构造

在数据挖掘的过程中,为了提取更有用的信息,挖掘更深层次的模式,提高挖掘效果的精度,我们需要利用已有的属性集构造出新的属性,并加入到现有的属性集合中。

比如,进行窃电诊断建模时,已有的属性包括供入电量、供出电量(线路上各大用户用电量之和)。理论上供入电量和供出电量应该是相等的,但是由于在传输过程中存在电能损耗,使得供入电量略大于供出电量,如果该条线路上的一个或多个大用户存在窃漏电行为,会使得供入电量明显大于供出电量。

因此,为了判断是否有大用户存在窃漏电行为,可以构造出一个新的指标——线损率,该过程就是构造属性。新构造的属性线损率按如下公式计算:

$$\text{线损率} = \frac{\text{供入电量} - \text{供出电量}}{\text{供入电量}} \times 100\% \quad (23.4)$$

线损率的正常范围一般在 3%~15%,如果远远超出该范围,就可以认为该条线路的大用户很可能存在窃漏电等用电异常行为。

23.4 数据规约

23.4.1 属性规约

属性规约通过属性合并来创建新属性维数,或者直接通过删除不相关的属性来减少数据维数,从而提高数据挖掘的效率、降低计算成本。属性规约的目标是寻找出最小的属性子集并确保新数据子集的概率分布尽可能地接近原来数据集的概率分布。属性规约常用方法如下表:

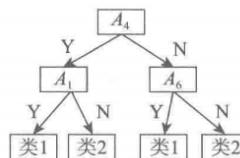
属性规约方法	方法描述	方法解析
合并属性	将一些旧属性合为新属性	初始属性集: $\{A_1, A_2, A_3, A_4, B_1, B_2, B_3, C\}$ $\{A_1, A_2, A_3, A_4\} \rightarrow A$ $\{B_1, B_2, B_3\} \rightarrow B$ \Rightarrow 规约后属性集: $\{A, B, C\}$
逐步向前选择	从一个空属性集开始,每次从原来属性集合中选择一个当前最优的属性添加到当前属性子集中。直到无法选择出最优属性或满足一定阈值约束为止	初始属性集: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\{\} \Rightarrow \{A_1\} \Rightarrow \{A_1, A_4\}$ \Rightarrow 规约后属性集: $\{A_1, A_4, A_6\}$
逐步向后删除	从一个全属性集开始,每次从当前属性子集中选择一个当前最差的属性并将其从当前属性子集中消去。直到无法选择出最差属性为止或满足一定阈值约束为止	初始属性集: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\} \Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow 规约后属性集: $\{A_1, A_4, A_6\}$
决策树归纳	利用决策树的归纳方法对初始数据进行分类归纳学习,获得一个初始决策树,所有没有出现在这个决策树上的属性均可认为是无关属性,因此将这些属性从初始集合中删除,就可以获得一个较优的属性子集	初始属性集: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  \Rightarrow 规约后属性集: $\{A_1, A_4, A_6\}$
主成分分析	用较少的变量去解释原始数据中的大部分变量,即将许多相关性很高的变量转化成彼此相互独立或不相关的变量	详见下面计算步骤

图 23.2: 属性规约常用方法

主成分分析是一种用于连续属性的数据降维方法,它构造了原始数据的一个正交变换,新空间的基底去除了原始空间基底下数据的相关性,只需要使用少数新变量就能够解释原始数据中的大部分变异。在应用中通常是选出比原始变量个数少,能解释大部分数据中的变量的几个新变量,即所谓主成分,来代替原始变量进行建模。

主成分分析的计算步骤如下:

1. 设原始变量 X_1, X_2, \dots, X_p 的 n 次观测数据矩阵为:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (X_1, X_2, \dots, X_p) \tag{23.5}$$

2. 将数据矩阵按列进行中心标准化。为了方便,将标准化后的数据矩阵仍然记位 X 。

3. 求相关系数矩阵 $R, R = (r_{ij})_{p \times p}, r_{ij}$ 的定义为:

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2 \sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}} \quad (23.6)$$

其中, $r_{ij} = r_{ji}, r_{ii} = 1$ 。

4. 求 R 的特征方程 $\det(R - \lambda E) = 0$ 的特征根 $\lambda_1 \geq \lambda_2 \geq \lambda_p > 0$ 。

5. 确定主成分个数 $m: \frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^p \lambda_i} \geq \alpha, \alpha$ 根据实际问题确定, 一般取 80%。

6. 计算 m 个相应的单位特征向量:

$$\beta_1 = \begin{bmatrix} \beta_{11} \\ \beta_{21} \\ \vdots \\ \beta_{p1} \end{bmatrix}, \beta_2 = \begin{bmatrix} \beta_{12} \\ \beta_{22} \\ \vdots \\ \beta_{p2} \end{bmatrix}, \dots, \begin{bmatrix} \beta_{1m} \\ \beta_{2m} \\ \vdots \\ \beta_{pm} \end{bmatrix} \quad (23.7)$$

7. 计算主成分:

$$Z_i = \beta_{1i}X_1 + \beta_{2i}X_2 + \dots + \beta_{pi}X_p, i = 1, 2, \dots, m \quad (23.8)$$

23.4.2 数值规约

数值规约指通过选择替代的、较小的数据来减少数据量, 包括有参数方法和无参数方法两类。

数据特征分析

24.1 分布分析

分布分析能揭示数据的分布特征和分布类型,对于定量数据,欲了解其分布形式是对称的还是非对称的,发现某些特大或特小的可疑值,可通过绘制频率分布表、绘制频率分布直方图、绘制茎叶图进行直观地分析;对于定性分类数据,可用饼图和条形图直观地现实分布情况。

24.1.1 定量数据的分布分析

对于定量变量而言,选择“组数”和“组宽”是做频率分布分析时最主要地问题,一般按照以下步骤进行。

1. 求极差;
2. 决定组距和组数;
3. 决定分点;
4. 列出频率分布表;
5. 绘制频率分布直方图;

遵循地主要原则如下:

1. 各组之间必须是相互排斥的。
2. 各组必须将所有数据包含在内。
3. 各组的组宽最好相等。

24.1.2 定性数据的分布分析

对于定性变量,常常根据变量的分类类型来分组,可以采用饼图和条形图来描述定性变量的分布。

饼图的每一个扇形部分代表每一类型的百分比或频数,根据定性变量的类型数目将饼图分成几个部分,每一部分的大小与每一类型的频数成正比;条形图的高度代表每一类型的百分比或频数,条形图的宽度没有意义。

PART VI

第六部分

数据可视化

可视化基础

25.1 绪论

25.2 字体

25.3 颜色

25.3.1 可视化色彩的运用原理

RGB 颜色模式

RGB 模式是颜色显示和图像处理中最常用的颜色空间。RGB 颜色模式使用了红 (red), 绿 (green), 蓝 (blue) 来定义所给颜色中红色、绿色和蓝色的光的量。在 24 位图像中, 每种颜色中的成分由 0 到 255 之间的数值表示。在位速率更高的图像中, 如 48 位图像, 值的范围更大。这些颜色成分的组合就定义了一种单一的颜色。

RGB 模式也被称为加色法混色模式。它是以 RGB 三色光互相叠加来实现混色的方法, 因而适合于显示器等发光体的显示。其混色规律是: 以等量的红、绿、蓝基色光混合。我们平时在绘图软件里调整颜色就是通过修改 RGB 颜色的三个数值来实现的。

HSL 颜色模式

大家平时在颜色选择中还会遇到一种颜色模式: HSL 模式, 其中 H 代表色相 (Hue)、S 代表饱和度 (Saturation)、L 代表亮度 (Lightness)。HSL 色彩模式是基于人眼的一种颜色模式, 是普及型设计软件的常用色彩模式, 具体如下

- 色相 H(hue): 代表的是人眼所能感知的颜色范围, 这些颜色分布在一个平面的色相环上, 取值范围是 $0^{\circ} \sim 360^{\circ}$ 的圆心角, 每个角度可以代表一种颜色。色相值的意义在于, 当不改变光感时, 可以通过旋转色相环来改变颜色。

- 饱和度 S(saturation): 是指色彩的饱和度, 它用 0 ~ 100% 的值描述了相同色相、明度下色彩纯度的变化。数值越大, 颜色中的灰度越少, 颜色越鲜艳, 呈现一种从理性(灰度)到感性(纯色)的变化。
- 亮度 L(Lightness): 是色彩的明度, 作用是控制色彩的明暗变化, 通常是从 0(黑)到 100%(白)的百分比来度量的, 数值越小, 色彩越暗, 越接近于黑色; 数值越大, 色彩越亮, 越接近于白色。

LUV 颜色模式

26.1 类别比较型图表

26.1.1 柱形图系列

26.1.2 条形图系列

26.1.3 克利夫兰点图

26.1.4 坡度图

26.1.5 南丁格尔玫瑰图

26.1.6 径向柱图

26.1.7 雷达图

26.1.8 词云图

26.2 数据关系型图表

26.2.1 散点图系列

26.2.2 曲面拟合

26.2.3 等高线图

26.2.4 散点曲线图系列

26.2.5 瀑布图

26.2.6 相关系数图

26.3 数据分布型图表

26.3.1 统计直方图

26.3.2 核密度估计图

26.3.3 数据分布图表系列

26.3.4 二维统计直方图

26.3.5 二维核密度估计图

26.4 时间序列型图表

26.4.1 折线图

26.4.2 面积图

26.4.3 日历图

26.4.4 量化波形图

26.5 局部整体型图表

26.5.1 饼图

26.5.2 圆环图

26.5.3 马赛克图

26.5.4 华夫饼图

26.5.5 块状/点状柱形图系列

26.6 高维数据型图表

26.6.1 高维数据的变换展示

主成分分析法

t-SNE 算法

26.6.2 分面图

26.6.3 矩阵散点图

26.6.4 热力图

26.6.5 平行坐标系图

26.6.6 RadViz 图

数值参数

PART VII

第七部分

LATEX 论文写作

28.1 题目类型

28.2 读题/选题

1. 题目每人各打印一份,分开读题,并做标记(在此期间,不查阅文献,不交流,要最原始的理解);
2. 40-60 分钟后,三人讨论(不反对,不辩论,各抒己见,提出思路和面对的困难);
3. 带着问题查文献,并确定各自的选题(各凭本领,获取信息,支持自己的选题);
4. 投票选题(激烈讨论,确定选题,查漏补缺,理清思路);
5. 安排分工,进入做题阶段。

28.3 论文写作时间安排

- 赛题发布的上午,分析题目,查找资料;
- 不论如何纠结,中午 12 点前确定所选题目;
- 下午开始动笔写论文,边写边分析;
- 第二天把模型构建好,并开始求解;
- 第三天中午 12 点前完成对所有问题的求解;
- 第三天下午 6 点前基本完成论文;
- 剩余的时间写摘要,改论文到最后一刻。

28.4 论文的评选

- 论文的评选好比“选美”,第一印象很重要:摘要、篇幅、排版、表述、怎么解决建模问题;
- 评定参赛队的成绩好坏、高低、获奖级别,竞赛论文是唯一依据;
- 答卷是竞赛活动成绩结晶的书面形式,一切全在纸上;

- 论文评选标准。
 1. 假设的合理性;
 2. 建模的创造性;
 3. 结果的正确性;
 4. 表述的条理性。

28.5 题目

- 题目是一篇论文给出的涉及论文范围与水平的第一个重要信息;
- 对题目的要求:简短精练、高度概括、准确得体、恰如其分;
- 论文题目一般不超过 20 字;
- 可以是关键字的经典组合,也可以是吸引眼球的热点问题,当然也可以直接用题目。

28.6 摘要

- 摘要是二等奖和三等奖的分界线
- 突出:算法、结论、创新点、特色、重点,废话全省略;
- 摘要是论文内容不加注释和评论的简短陈述、其作用是使读者不阅读论文即获得必要的信息;
- 论文摘要不要列举例证,不用图表,不给数学表达式,也不要自我评价;
- 保持在 500 字左右,行间距不超过一张纸。
- 摘要应包含以下内容:
 1. 数学模型的归类(在数学上属于什么模型),研究的目的和意义;
 2. 所用的数学知识、建模的思想、算法思想、模型及算法特点;
 3. 获得的基本结论和研究成果、突出论文的新见解及意义,主要的数值结果和结论要明确表现出来;
 4. 回答题目所问的全部问题,条理要清晰;
 5. 一般关键字 3-5 个,便于情报检索。

28.7 正文

28.7.1 问题重述

- 将原问题表达清楚,如果问题表述很长,数据很多,可以简洁的描述;
- 尽量用自己的语言整理归纳,篇幅不要过长,不要超过一页纸;
- 这里可以适当加入一些自己查到的背景资料或者前人的研究现状,但不要放自己主观的理解和见解,只是阐述问题及现状,把自己知道的和这个问题相关的资料陈述出来就行;
- 这里如果引用了一些参考文献,不要忘记备注在参考文献中。

28.7.2 问题假设

- 根据题目中的条件作出假设；
- 根据题目中要求作出假设；
- 关键性假设不能缺；
- 假设要切合题意、合理。

28.7.3 符号说明

- 要注意整篇文章符号一致；
- 可引用表格样式、隐藏边框（三线式）；
- 符号搭配协调，符合数学规则，大小写区分好，不用怪异符号；
- 如果符号太多，可放一些全文通用的符号，特殊符号可在文中解释。

28.7.4 问题分析与模型准备

- 问题分析与模型准备可以合成一个部分，主要作用是过渡，对问题要解决的重点及突破口进行分析，使接下来的建模不突兀；
- 同时可以将自己的理解、思考过程、思路进行巧妙的阐述，为下一部分模型的建立打下基础（表述多用“准备”怎么做，不要用“了”）；
- 这部分相当于一个引子，吸引阅卷老师继续读下去，所以文字不可太长，内容不要过于分散、琐碎、措辞要精练；
- 条理清晰，娓娓道来，吸引老师的眼球，让接下来建立的模型更顺理成章，方便理解。

28.7.5 模型建立与求解

模型的建立

- 这部分是正文的重头戏；
- 模型的建立可由简单到复杂建立多个模型（递进式），也可根据题目的要求，逐个问题建立模型（并列式）；一定要格外注意一点，各个模型之间的联系及变量的转换和公式的迭代（标号）；
- 模型要有特色与创新，注意改进和变通；
- 建模的同时也要考虑：会建也要会解，只建不解很难有说服力；
- 建立数学模型应当注意以下几点：
 1. 分清变量类型，恰当使用数学工具；
 2. 抓住问题本质，简化变量之间的关系；
 3. 建立数学模型时要有严密的数学推理；
 4. 用数学方法建模，模型要明确，要有数学表达式。

模型的求解

- 模型的求解,数学算法的选取会直接影响到结果;
- 算法的选取对接下来的误差分析及稳定性分析也有影响;
- 需要注意:
 1. 重要结论需要建立数学命题时,命题叙述要符合数学命题的表述规范,尽可能论证严密;
 2. 需要说明计算方法或算法的原理、思想、依据、步骤,若采用现有软件,说明采用此软件的理由,软件名称;
 3. 可以尝试多种算法、多软件处理,便于进行稳定性分析,同时验证结果的正确性。
 4. 计算过程,中间结果可要可不掉的,不要列出;
 5. 最终数值结果的正确性或合理性是第一位的,设法算出合理的数值结果;
 6. 题目中要求回答的问题,数值结果,结论,须一一列出;
 7. 结果表示:要集中,一目了然,直观,便于比较分析及评委查找;
 8. 数值结果表示:精心设计表格;可能的话,用图形表示更好;
 9. 如果在建模过程中,建模模型过于复杂,无法求出解析解,可以尝试求出数值解,但此操作之后必须进行灵敏性及稳定性分析
 10. 结论要及时清晰列出,评委就在这章里找结果。

28.7.6 模型灵敏性分析

- 对数值结果或模拟结果要进行必要的检验,若结果不正确,不合理或误差大时,要分析原因,对算法、计算方法、或模型进行修正改进;
- 必要时,要对模型进行稳定性分析、统计检验、误差分析,要对不同模型进行对比及实际可行性检验;
- SPSS 软件在统计及误差分析方面有一定的优势,可以考虑使用;
- 结果的连续性、唯一性、稳定性等都可以尝试入手。

28.7.7 模型的讨论与评价

- 评委特别喜欢这一部分,可以考察参赛者的研究深度;
- 模型的进一步讨论,理论归纳,科学性及现实意义;
- 模型的讨论即模型在稳定性分析的基础上,对模型的建立、求解及结果进行整理、归纳、讨论、拓展,可以查漏补缺,也可以将没实现或者没有考虑到的因素在此阐述,发散思考,拓展思路;
- 模型的评价要求我们突出优点,不回避缺点,客观公正;
- 可以认为是个小总结,重申你的结果。

28.7.8 模型的改进与推广

- 此部分并非一定要有的章节,只是在讨论的基础上进行少量计算或者延申,这部分不用写的很精彩,此部分只是根据题目的要求,使之更符合现实,更具有推广意义;

- 有的题目分几个问题,第一个问题让你建立个模型,然后让你计算出结果,之后第二个问题问你是否可以改进,得到更好的结果,如果遇到这种情况,第二问是正文的模型建模求解问题,而非此处的模型改进。

28.8 参考文献

- 参考文献要书写规范,可参考专业学术杂志;
- 在正文中提及或直接引用的材料、原始数据等来自于一些公开刊物的可在参考文献中列出;
- 参考文献需表明刊物作者的姓名、刊物名称、卷次、页码和出版日期;
- 参考文献反映出真实的科学依据,分清自己和别人的观点或成果、对前人科学成果的尊重,便于检索;
- 计算程序、详细的结果、详细的数据表格,可在附录中列出,但不要错,错的宁可不列;
- 主要结果数据,应在正文中列出,不怕重复。

代码

```
ng new appname --style scss --skip- install
```

