

- 基础知识

- 1. 数据预处理
 - 缺失值处理
 - 异常值处理
- 2. 怎么建模
- 3. 排版
- 4. 标题
- 5. 摘要

基础知识

1. 数据预处理

缺失值处理

go偶读
23年6月23日
计算方法里面学了
数学建模 | 数据预处理

三次样条插值函数
这个好顶

π 数学建模BOOM

□ 缺失值

- 比赛提供的数据，发现有些单元格是null或空的
- 缺失太多：例如调查人口信息，发现“年龄”这一项缺失了40%，就直接把该项指标删除
- 最简单处理：均值、众数插补
 - 定量数据，例如关于一群人的身高、年龄等数据，用整体的均值来补缺失
 - 定性数据，例如关于一群人的性别、文化程度；某些事件调查的满意度，用出现次数最多的值补缺失
 - 适用赛题：人口的数量、年龄、经济产业情况等统计数据，对个体精度要求不大的数据
- Newton插值法
 - 根据固定公式，构造近似函数，补上缺失值，普遍适用性强
 - 缺点：区间边缘处的不稳定震荡，即龙格现象。不适合对导数有要求的题目
 - 适用赛题：热力学温度、地形测量、定位等只追求函数值精准而不关心变化的数据
- 样条插值法
 - 用分段光滑的曲线去插值，光滑意味着曲线不仅连续，还要有连续的曲率
 - 适用赛题：零件加工，水库水流量，图像“基线漂移”，机器人轨迹等精度要求高、没有突变的数据

缺失太多数据，直接删掉

简单处理：

1. 对于定量，采用均值
2. 对于定性，采用众数

复杂处理:

1. newton插值法

缺点:会有边缘跳变

2. 样条插值法

一般两种方式一起用

异常值处理

数学建模 | 数据预处理

数学建模BOOM

异常值

- 样本中明显和其他数值差异很大的数据,例如一群人的身高数据中有个3米2的
- 正态分布 3σ 原则
- 数值分布在 $(\mu - 3\sigma, \mu + 3\sigma)$ 中的概率为99.73%, 其中 μ 为平均值, σ 为标准差
- 求解步骤: 1. 计算均值 μ 和标准差 σ ; 2. 判断每个数据值是否在 $(\mu - 3\sigma, \mu + 3\sigma)$ 内, 不在则为异常值
- 适用题目: 总体符合正态分布, 例如人口数据、测量误差、生产加工质量、考试成绩等
- 不适用题目: 总体符合其他分布, 例如公交站人数排队论符合泊松分布

画箱型图

- 箱型图中, 把数据从小到大排序。下四分位数 Q_1 是排第25%的数值, 上四分位数 Q_3 是排第75%的数值
- 四分位距 $IQR = Q_3 - Q_1$, 也就是排名第75%的减去第25%的数值
- 与正态分布类似, 设置个合理区间, 在区间外的就是异常值
- 一般设 $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ 内为正常值
- 适用题目: 普遍适用

3.2 大
2 上极限
1.6 上四分位数 Q_3
1.5 中位数 50%
1.2 下四分位数 Q_1
1.0 晶须
0.8 下极限

1. 正态分布 3σ 原则

2. 画箱型图:

1. $Q_1=25\%$ 处的值, $Q_3=75\%$ 处的值
2. $IQR=Q_3-Q_1$,
3. $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$

2. 怎么建模

摘要(最重要) 问题重述, 模型假设和符号说明, 模型建立与求解(最长), 模型的优缺点改进与改进方法, 参考文献和附录

□建模过程

• 一篇完整的数模论文

- 包括摘要（最重要）、问题重述、模型假设和符号说明、模型建立与求解（最长）、模型的优缺点与改进方法、参考文献和附录。

• 摘要：最重要

- 读者看完摘要，就知道论文研究的问题、用了什么方法、求得了什么结果，以及每一部分的[大致步骤](#)。

• 问题重述

- 将题目简述一遍即可，并不重要。注意不要复制粘贴，避免查重

• 模型假设和符号说明

- 好的假设能让你事半功倍
- 例如某一年太阳投影问题，影子长度与地球公转也有关系，但地球公转对影长的影响远远小于自转，可在模型假设里说明“忽略公转对影长的影响”
- 符号说明将论文中定义的重要符号列出表格说明即可



• 摘要：总的概述，结构清晰的摘要很重要

• 问题重述：把问题简单重复一遍即可

• 模型假设和符号说明：省略次要因素，关注主要因素，便于计算

• 模型建立：先建立一个简单的模型，再根据条件一步一步去约束模型

• 模型的建立

- 一组公式，和对公式中每个变量的解释，就是一个模型

• 先查阅资料，[看着资料，用自己的话复述一个简单的模型](#)，再根据题目中的约束条件去一步步修改模型，[把题目中的变量带入模型中去](#)

3.2 两个指定顶点之间最短路问题的数学表达式

假设有关图有 n 个顶点，现需要求从顶点 1 到顶点 n 的最短路。设 $W = (w_{ij})_{n \times n}$ 为赋权邻接矩阵，其分量为

$$w_{ij} = \begin{cases} w(v_i v_j), & v_i v_j \in E \\ \infty, & \text{其它} \end{cases}$$

决策变量为 x_{ij} ，当 $x_{ij} = 1$ ，说明弧 $v_i v_j$ 位于顶点 1 至顶点 n 的路上；否则 $x_{ij} = 0$ 。其数学规划表达式为

$$\min \sum_{v_i v_j \in E} w_{ij} x_{ij}$$

$$\text{s.t. } \sum_{\substack{j=1 \\ v_i v_j \in E}}^n x_{ij} - \sum_{\substack{j=1 \\ v_j v_i \in E}}^n x_{ji} = \begin{cases} 1, & i = 1 \\ -1, & i = n \\ 0, & i \neq 1, n \end{cases}$$

$$x_{ij} = 0 \text{ 或 } 1$$

- 模型求解: 采用编程求解

已关注
数学建模 | 究竟怎么建模

模型的求解

- 例如上文我们所建立的最短路径模型, 查阅资料可知单源最短路径的常用算法是Dijkstra算法, 那么模型的求解过程可以**把资料上的用自己的话复述一遍**:

采用Dijkstra算法求解: (以下内容可以百度或从常见资料里找到)

- 从起始点开始, 将起点放进集合N中, 查找所有与其相连的节点及到达下一节点的花费, 并且记录下来;
- 接下来选择花费最短的一条路径, 到这条最短路径指向的节点去, 把这个点也放进集合N中, 然后查找所有与这个节点相连的其他不在集合N中的点, 并且也计算到达下一点所需要的花费并记录下来。保存花费最小的一条记录;
- 继续选择花费最短的路径重复执行第2步, 一直到所有的点都已有了最短路径, 完毕。

- 需要注意的是, 必须**根据赛题解释清楚**“起始点”在本问题中究竟是什么、算法里的“节点”在本题中的实际意义是什么、最短路径意味着什么

Bilibili

- 模型优缺点与改进方法: 非必要, 写上最好, 实在写不上可以不屑
- 参考文献: 知网中引用, 选择第一个

- 参考文献

- 格式一定要规范
- 知网检索结果右侧有引用按钮, 打开后复制即可

引用

GB/T 7714-2015 格式引文 [1]常世杰,袁铭润.基于Matlab的热网换热站最短分布问题研究[J].山西建筑,2021,47(16):104-105+139.

MLA格式引文 [1]常世杰, and 袁铭润.“基于Matlab的热网换热站最短分布问题研究.”*山西建筑* 47.16 (2021):104-105+139. doi:10.13719/j.cnki.1009-6825.2021.16.040.

APA格式引文 [1]常世杰 & 袁铭润.(2021).基于Matlab的热网换热站最短分布问题研究.*山西建筑*(16), 104-105+139. doi:10.13719/j.cnki.1009-6825.2021.16.040.

知网研学 (原E-Study) | EndNote | NoteExpress | Refworks | NoteFirst | 更多引用格式 >>

- 附录: 别直接抄代码, 变量名字换一换

- 附录

- 附录里要写出正文中求解时用到的代码
 - 一定不要把网上搜到的代码直接复制粘贴!!!
 - 把查到的代码里**变量名换一换**就不会被查重
 - 曾出现过参加国赛, 在省内被推到国奖, 但查重发现代码是复制的, 结果被取消获奖并官网通报的先例

APA格式引文 [1]常世杰, & 袁铭润.(2021).基于Matlab的热网换热站最短分布问题研究.*山西建筑*(16), 104-105+139. doi:10.13719/j.cnki.1009-6825.2021.16.040.

知网研学 (原E-Study) | EndNote | NoteExpress | Refworks | NoteFirst | 更多引用格式 >>

3. 排版

□ 颜值即正义

- 各级标题与正文层次分明
 - 一般标题级别不超过三级,
 - 正文中文字体设置宋体、英文Times New Roman
- 正文排版紧凑, 看起来充实, 没有大片空白
 - 避免图片过大导致出现大片空白, 且不要留有空行
- 表格与图片
 - 表格用标准的三线表
 - 表的标题放在表的上方, 图名放在图的下方
- 公式编辑
 - 推荐mathpix, 或用word的公式编辑器
 - 公式需要解释清楚每个变量的意义; 重要的公式后面带有编号

(关注公众号:数学建模BOOM, 回复 模板)

$$u(x, y) = \frac{1}{2\pi^2} \int_0^\pi \int_{-\infty}^{+\infty} \frac{a p(r, \theta)}{a r d \cos(\varphi - \theta)} dr d\theta \quad (9)$$

式中, θ 表示X射线的法线和x轴正向间的夹角, 满足 $0 \leq \theta \leq \pi$; r 表示X射线与坐标原点之间的距离。

由radon变换可实现在已知CT系统接收信息的情况下可求得被测物体的吸收率、几何形状等; 同理, 在已知物体吸收率时也可由radon逆变换可求得接收信息。而radon变换是建立在准确安装CT系统的前提下的, CT系统是否准确安装将影响每一次测量与求解。因此, 在图像重建之前, 对CT系统的进行参数标定至关重要。

5.2 对问题一的求解

5.2.1 笛卡尔坐标系的确定

在旋转中心的位置与旋转角度的求解中, 需要建立坐标体系对模板的几何信息进行量化处理。本文建立以椭圆中心为标准的笛卡尔坐标体系, 并通过数据处理对椭圆与圆的重要几何转换信息进行标记。其中, 椭圆的几何中心坐标为圆心O(0, 0), 椭圆上顶点坐标为A(0, 40), 椭圆右顶点坐标为B(15, 0), 小圆的圆心几何坐标为M(45, 0)。

5.2.2 探测器单元之间的距离 l_0 的确定

当X射线为180°方向时, 即平行于椭圆的长轴, $i = PH_{min}$, $a_i = 180^\circ$, 区域1为对应的椭圆短轴 d_1 的投影长度。通过观察附件2中的数据可知 PH_{max} 的非0区间为 (b_{160}, b_{277}) , 即从探测器第169单元至第277单元, 此时探测器间距 $l_i = (277 - 169)l_0 = 108l_0$, $i = PH_{max}$ 。由模板示意图可知椭圆短轴 $d_1 = 30mm$, 则 $d_1 = l_i = 108l_0$, 即 $l_0 = 0.2778mm$ 。

• 标题一般不超过三级

• 中文设置宋体

• 表格用三线表

• 表的标题放在表的上方, 图片名字放在图片的下方。

• 重要公式加编号, 每个公式加解释说明

4. 标题

• 格式: 基于xxx模型/算法的xxx问题的研究

- 基于xxx模型/算法的xxx问题研究
 - xxx模型就是正文里的核心模型
 - xxx问题就是赛题的中心词
 - 不要使用过多的修饰词

样例：

- ◆ 基于一维热传导方程的炉温曲线机理问题研究
- ◆ 基于单目标优化模型的CT系统成像问题研究
- ◆ 基于多目标优化模型的系泊系统问题研究
- ◆ 基于动态规划补贴的出租车资源配置问题研究
- ◆ 基于差分方程和元胞自动机的交通阻塞问题研究
- ◆ 基于回归分析的长江水质预测问题研究

- 要求：读完标题，就能知道建立的是说明模型

- 简明扼要，符合规范，便于检索
- 字数限制在一行内
- 不要写公式，非公知的缩写

5. 摘要

- 基本要求
 - 是对论文的总结概括，
 - 让评阅老师读完摘要，就知道文本解决的问题，建立的模型和求解结果
- 注意事项
 - 不要加表格
 - 3/4页到1页，就差不多刚刚好一页，不要留太多空白

1. 开头

- 模板

开头段模板：本文主要研究了XXX问题。根据XXX，利用XXX模型/方法，求解出XXX。

- 第一句：说清研究的问题
- 第二句：说清全文采用的模型/算法、采取的操作
- 开头段不要写详细的求解结果

- 注意事项

不要把求解过程完全写出来，留在后面写的

示例：

示例：17年国赛A题

针对问题二和问题三，根据数字图像处理技术和滤波器原理，利用不同空域特征的区域对应着频率平面中的不同区域的特点，在已知接收信息的情况下，利用傅立叶变换即可求得被测物体的吸收率、性状和位置等信息。根据CT系统正方形托盘的几何信息，可求得附件中10个点的位置对应的吸收率，依次为0.0757；-0.0325；-0.0380；-0.0015；1.9827；0.0023；0.0131；0.0125；-0.0204；0.0297。



2. 中间

- 模板

- 中间段模板：针对问题一，考虑/根据XXX，...，建立XXX模型/利用XXX方法，...求解出XXX。

- 注意事项+示例

一定要写清结果！！！

- 优化类、预测类和物理类的题目，要明确写清数值
- 要求提供建议或评价的题目，写要明确写清结论和数据依据，但不要有表格；数据过多可说明数据见附录

示例：17年国赛A题

针对问题二和问题三，根据数字图像处理技术和滤波器原理，利用不同空域特征的区域对应着频率平面中的不同区域的特点，在已知接收信息的情况下，利用傅立叶变换即可求得被测物体的吸收率、性状和位置等信息。根据CT系统正方形托盘的几何信息，可求得附件中10个点的位置对应的吸收率，依次为0.0757；-0.0325；-0.0380；-0.0015；1.9827；0.0023；0.0131；0.0125；-0.0204；0.0297。

3. 总结

- 注意事项

- 如果写完后摘要超过一页了，可以不写
- 不要累赘重复前面写过的内容
- 写一些本文的特色、自夸的语句

示例：17年国赛A题

定价模型考虑了会员密集程度、任务集中度、任务难易程度等因素。任务优化分配模型提高了任务的有效完成率，基于最大流的启发式算法计算精度高、运算时间短等优点，高效解决了拍照任务定价问题。

示例：16年国赛B题

本文的特色在于将机理分析与多目标规划相结合，运用熵权法将多目标问题转化为单目标问题，使得求解结果更加客观。此外，对于解空间较复杂的模型，设计了变步长搜索算法，在保证了求解的精度的同时，极大地提高了运算的时间复杂程度，为日后系泊系统的设计的发展提供了参考依据。

4. 关键字

- 关键词一般4~6个
- 使用的模型和算法、大家都知道的专业名词、问题的关键词
- 中间以空格分开

示例：16年国赛A题

系泊系统设计 多元非线性方程组 循环遍历法 层次分析法 优化模型
系泊系统设计 机理分析 最小二乘法 变步长搜索算法
系泊系统设计 刚体力学方程组 多重搜索算法 多目标优化
系泊系统设计 受力分析 悬链线 控制模型 多目标优化 遗传算法

首先是预测模型（1）神经网络预测模型（2）灰色预测模型（3）拟合插值预测（线性回归）（4）时间序列模型（5）马尔科夫模型（6）支持向量机模型（7）Logistic模型（8）组合预测模型（9）微分方程预测（10）组合预测模型 其次是评价模型（1）模糊综合评价法（2）层次分析法（3）聚类分析法（4）主成分分析评价法（5）灰色综合评价法（6）人工神经网络评价法（7）BP神经网络综合评价法（8）组合评价法 第三个是优化模型（1）规划模型，其中包括目标规划、线性规划、非线性规划、整数规划、动态规划（2）排队论模型（3）神经网络模型（4）现代优化算法，其中包括遗传算法、模拟退火算法、蚁群算法、禁忌搜索算法（5）图论模型（6）组合优化模型 第四个是分类模型（1）决策树（2）逻辑回归（3）随机森林（4）朴素贝叶斯 最后统计统计分析模型有：（1）均值T检验（2）方差分析（3）协方差分析（4）分布检

验 (5) 相关分析 (6) 卡方检验 (7) 秩和检验 (8) 回归分析 (9) Logistic回归
(10) 聚类分析 (11) 判别分析 (12) 关联分析 第二部分是十大算法 1、蒙特卡罗算
法 2、数据拟合、参数估计、插值等数据处理算法 3、线性规划、整数规划、多元规
划、二次规划等规划类问题 4、图论算法 5、动态规划、回溯搜索、分治算法、分支定
界等计算机算法 6、最优化理论的三大非经典算法：模拟退火法、神经网络、遗传算法
7、网格算法和穷举法 8、一些连续离散化方法 9、数值分析算法 10、图象处理算法