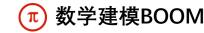
数学建模!快速入门

——带你临阵磨枪,突击国赛!

1-6 数据预处理

主讲人: 北海

数学建模 | 数据预处理 视频出自b站up主:数学建模B00 ® 数学建模B00M



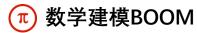
□缺失值

- 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

(该三种方法足够用,其他方法例如分段插值、Hermite插值就不再——介绍了)

关注公众号/B站:数学建模BOOM,带你玩转数学建模~ 交流群: 887602371

数学建模 | 数据预处理 视频出自b站up主:数学建模B00



□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 (μ-3σ,μ+3σ)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 (μ-3σ,μ+3σ)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 $[Q_1 1.5 * IQR, Q_3 + 1.5 * IQR]$ 内为正常值
- 适用题目: 普遍适用
- 异常数据处理方法与缺失值处理相同

