南京理工大学

NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

数据科学导论期末课程项目

姓　名：　　学　号：

选　题：　题目一：几种重要的非线性回归模型

专　业：　数据科学与大数据技术

学　院：　网络空间安全学院

南京理工大学基础前沿交叉中心

年　月　日

期末课程项目注意事项及要求

- 本课程期末考核包含三个项目，需完成其中任意一个，请谨慎选择.
- 考核内容：侧重数据科学导论课程中算法及模型的理解与应用.
- 数值试验题应该同时提交书面报告和程序(打包)，其中书面报告有详细的推导和数值结果及分析.
- 成绩评定：平时作业约占考核总体的 60%，项目大作业占比 40%. 迟交一天(24 小时)打折 10%，不接受晚交 4 天的作业和项目，任何时候理由都不接受.
- 评分细则：(1)完成要求的任务. (2)鼓励提出原创性的方法并予以实现. (3)提交一份完整的项目报告(包括源文件及程序). (4)12 月 16 日 18:00 点前发给课代表.
- 期末课程项目应独立完成，可以同学间相互讨论或者找老师答疑. 鼓励开源交流，禁止直接抄袭. 有讨论或从其它任何途径取得帮助，请列出来源.
- 可选加分项目：(1)代码部分所使用的函数自己编写，非调包. (2)将项目报告以 slides 形式展示给他人，包括老师与同学. (3)在 GitHub 中将代码分享给需要使用的人，非上传至盈利组织(如某文库、某丁网等).
- 挑战项目：DataHub 成员类项目(持续更新…)，具体咨询组内成员或相关老师.
- 本项目及其相关内容在未经授课教师准许，请勿随意上传至网络，仅限于校内交流合作使用.

目　　录

课程项目一 几种重要的非线性回归模型

【1】项目背景

回归分析是一种通过建立模型来研究变量之间相互关系的密切程度、结构状态及进行模型预测的有效工具，在工商管理、经济、社会、医学和生物学等领域应用十分广泛. 课程中我们讲述了线性回归模型，包含 Lasso、岭回归及最小二乘. 在实际的回归分析中，很多数据中输入特征和目标特征并非呈现线性关系. 此时，简单的线性回归并不能很好地拟合数据. 本项目重点关注几种常用的非线性回归模型.

【2】方法陈述

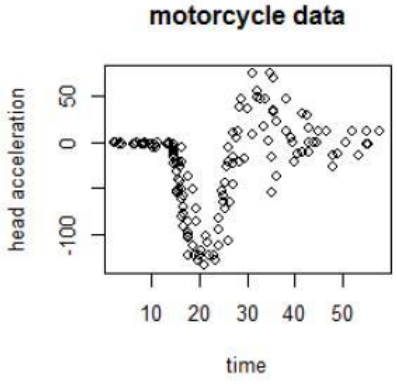
局部多项式回归、样条函数回归、小波，任选感兴趣的两种即可.

问题一：尝试介绍局部多项式回归模型，分析其理论机理，重点介绍局部线性回归模型及其中需要注意的问题(如窗宽与核函数的选取).

问题二：基于网络对样条回归进行介绍，了解其工作的机理，并对样条与最小二乘及局部线性回归之间的差异进行简要分析.

【3】案例实战

1. 下图展示的是一个关于摩托车碰撞试验的数据集的散点图，为 R 程序包 MASS 中 mcycle 数据集. 该数据集由一系列摩托车事故实验中头部加速度的数据测量值构成，用于测试碰撞头盔的性能. 从下图中可以看到，加速度与时间呈现明显的非线性关系. 试根据[2]中的方法建立适当的回归模型，并对所建模型进行解释.



2. 医疗费用预测(教材课后习题)

保险公司通常需要募集比花费在受益者的医疗服务上更多的年度保费. 因此, 精确预测医疗费用对保险公司具有重要价值. 本案例提供的数据集是从美国人口普查局的人口统计资料整理得出. 数据集共有 1338 个样本, 包含 7 个特征. 特征的具体信息如下表所示.

表 1 医疗费用数据集特征

特征名称	特征说明
age	受益者年龄
sex	保单持有人性别
bmi	身体健康指数
children	保险计划中所包含的孩子/受抚养者的数量
smoke	被保险人是否吸烟
region	受益人的居住地
charges	已结算的医疗费用

请将“charges”作为目标特征, 构建回归模型并分析拟合的结果, 预测受益者的平均医疗费用. 数据集可在网络上自行下载.

【4】案例代码

（请将案例代码粘贴在此处, 并给出必要注释）

案例一代码

案例二代码

课程项目二 网络模型中的社区发现

【1】项目背景

在过去的几十年里，网络数据的数量和对相关统计推断工具的需求在快速增长. 网络数据分析的主要研究课题之一是从单一的观察网络中找出隐藏的社区. 简单地讲，网络社区指的是相互靠近的个体更容易连接的现象，因此边缘密度在同一个社区内和社区之间都是不同的. 当前人们的生活中充满了各种各样的网络，包括社会网络、生物网络、通信网络等. 研究者希望通过网络社区划分加深对复杂网络的了解. 例如，将社区检测用于了解社会行为、蛋白质与蛋白质的相互作用、基因表达、 个性化产品推荐、网页排序等. 社区检测或图聚类的最基本目标是将图的节点划分为几个内部连接更紧密的簇. 自 20 世纪 80 年代以来，对社区检测方法的研究已有极大的发展，在机器学习、网络科学、社会科学和统计物理学等不同领域，学者提出了各种不同的网络社区检测的方法. 本项目给出几个基于网络社区检测方法的应用.

【2】方法陈述

问题一：谱聚类是一种重要的聚类算法，它可以利用网络节点间的相似性，通过对图的分割来识别节点的社区属性. 现尝试介绍谱聚类的机理及其使用方法，并借助网络寻找其在社区发现的应用.

问题二：随机分块模型是一种具有区块结构的随机网络生成模型，同时也是一种简单、常用的网络分析模型. 它假设网络中节点间的连接是服从铂努利分布的，且网络节点间的连接是相互独立的，其中区块内各点的链接概率相同. 请借助网络对该模型进行学习，并介绍其机理.

【3】案例实战

1. 政治博客数据集(见附件 Pol_Blogs_CSV 文件)

数据集网址一: http://konect.cc/networks/dimacs10-polblogs/

数据集网址二:

http://www.casos.cs.cmu.edu/computational_tools/datasets/external/polblogs/index11.php

政治博客数据是由 Adamic 和 Glance 于 2005 年收集并分析的. 该数据集包含了 2004 年美国总统选举前不久的 1000 多个网络日志的快照, 其中节点是网络日志, 边是超链接. 节点被标记为自由派或保守派, 这可以被视为两个定义明确的社区. 请尝试使用随机分块模型对这一数据集进行建模, 使用谱聚类算法对该数据集进行社区聚类, 并对所建模型及聚类结果给出合理的解释.

2. Twitter 好友网络分析及社区发现(教材课后习题)

目前, Twitter、Facebook、Google+和微博都提供了获取用户社交网络数据的 API 或者 APP, 例如 Facebook 的 Netvizz 和 myFnetwork, 它们提供了可用于社交网络分析的原始数据. 本案例提供了通过 Twitter 社交网络 API 提取的 Twitter@wiredUK 的社交网络数据. 该数据集共有 254 个节点和 3834 条边. 请使用 Gephi 工具或 ggplot 将该社交网络进行可视化, 并进行社区发现分析. 请注意给出最终结果的合理性分析.

【4】案例代码

（请将案例代码粘贴在此处, 并给出必要注释）

案例一代码

案例二代码

课程项目三 2022 年高教社杯全国大学生数学建模竞赛 C 题

【1】项目背景

全国大学生数学建模竞赛为 DataHub 小组重点关注的比赛. 此竞赛创办于 1992 年, 每年一届, 1994 年被教育部列为全国大学生四大赛事之一, 目前已列入“高校竞赛评估与管理体系”目录（位列第五）, 为传统的五大赛事之一（其余分别为中国创新创业大赛、“挑战杯”全国大学生课外学术科技作品竞赛、中国大学生计算机设计大赛、全国大学生英语竞赛）. 目前该项竞赛已成为全国规模最大、在国内外具有重要影响的基础性学科竞赛之一. 该竞赛是面向全国大学生的群众性科技活动, 旨在激励学生学习数学的积极性, 提高学生建立数学模型和运用计算机技术解决实际问题的综合能力, 培养创造精神及合作意识. 本次大赛由中国工业与应用数学学会主办, 共有来自中国、英国、马来西亚等国家与地区的 1606 所院校/校区、54257 队（本科 49424 队、专科 4833 队）、16 万多人报名参加. 此项目重点关注与数据分析及挖掘相关的 C 题.

特别提醒: 若建模成员选择此题, 请勿照搬建模期间论文. 因此题已公布评价标准, 可供参考. 论文展示网址如下:

https://dxs.moe.gov.cn/zx/hd/sxjm/sxjmlw/2022qgdxssxjmjslwzs/

【2】问题重述: 古代玻璃制品的成分分析与鉴别

丝绸之路是古代中西方文化交流的通道, 其中玻璃是早期贸易往来的宝贵物证. 早期的玻璃在西亚和埃及地区常被制作成珠形饰品传入我国, 我国古代玻璃吸收其技术后在本土就地取材制作, 因此与外来的玻璃制品外观相似, 但化学成分却不相同.

玻璃的主要原料是石英砂, 主要化学成分是二氧化硅(SiO2). 由于纯石英砂的熔点较高, 为了降低熔化温度, 在炼制时需要添加助熔剂. 古代常用的助熔剂有草木灰、天然泡碱、硝石和铅矿石等, 并添加石灰石作为稳定剂, 石灰石煅烧以后转化为氧化钙(CaO). 添加的助熔剂不同, 其主要化学成分也不同. 例如, 铅钡玻璃在烧制过程中加入铅矿石作为助熔剂, 其氧化铅(PbO)、氧化钡(BaO)的含量较高, 通常被认为是我国自己发明的玻璃品种, 楚文化的玻璃就是以铅钡玻璃

为主. 钾玻璃是以含钾量高的物质如草木灰作为助熔剂烧制而成的, 主要流行于我国岭南以及东南亚和印度等区域.

古代玻璃极易受埋藏环境的影响而风化. 在风化过程中, 内部元素与环境元素进行大量交换, 导致其成分比例发生变化, 从而影响对其类别的正确判断. 如图 1 的文物标记为表面无风化, 表面能明显看出文物的颜色、纹饰, 但不排除局部有较浅的风化; 图 2 的文物标记为表面风化, 表面大面积灰黄色区域为风化层, 是明显风化区域, 紫色部分是一般风化表面. 在部分风化的文物中, 其表面也有未风化的区域.

图 1　未风化的蜻蜓眼玻璃珠样品

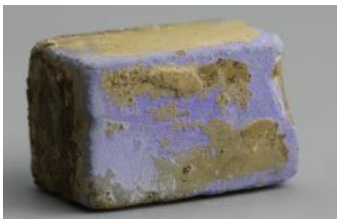


图 2　风化的玻璃棋子样品

现有一批我国古代玻璃制品的相关数据, 考古工作者依据这些文物样品的化学成分和其他检测手段已将其分为高钾玻璃和铅钡玻璃两种类型. 附件表单 1 给出了这些文物的分类信息, 附件表单 2 给出了相应的主要成分所占比例(空白处表示未检测到该成分). 这些数据的特点是成分性, 即各成分比例的累加和应为 100%, 但因检测手段等原因可能导致其成分比例的累加和非 100%的情况. 本题中将成分比例累加和介于 85%~105%之间的数据视为有效数据.

请依据附件中的相关数据进行分析建模, 解决以下问题:

问题 1 对这些玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系进行分析; 结合玻璃的类型, 分析文物样品表面有无风化化学成分含量的统计规律, 并根据风化点检测数据, 预测其风化前的化学成分含量.

问题 2 依据附件数据分析高钾玻璃、铅钡玻璃的分类规律; 对于每个类别选择合适的化学成分对其进行亚类划分, 给出具体的划分方法及划分结果, 并对分类结果的合理性和敏感性进行分析.

问题 3 对附件表单 3 中未知类别玻璃文物的化学成分进行分析, 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，并比较不同类别之间的化学成分关联关系的差异性.

附件

表单 1 玻璃文物的基本信息

表单 2 已分类玻璃文物的化学成分比例，其中

(1) 文物采样点为该编号文物表面某部位的随机采样，其风化属性与附件表单 1 中相应文物一致.

(2) 部位 1 和部位 2 是文物造型上不同的两个部位，其成分与含量可能存在差异.

(3) 未风化点是风化文物表面未风化区域内的点.

(4) 严重风化点取自风化层.

表单 3 未分类玻璃文物的化学成分比例

【3】模型代码

（请将案例代码粘贴在此处，并给出必要注释）

问题 1 代码

问题 2 代码

问题 3 代码

问题 4 代码