南京理工大学

NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

2023 年数据科学导论期末课程项目

负责人：_____ 学　号：_____

电　话：_____

选　题：题目一：矩阵补全模型及其应用

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. 选定一个项目后，请删除其余不相干内容. *请务必按照格式规范撰写大作业.*
- 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.
- 数值试验题应同时提交纸质和电子版报告(包含程序并打包)，其中书面报告有详细的推导和数值结果及分析. 说明：尽管完成的方式、使用的工具、编程的技巧都很重要，但本项目强调理论分析、看重方法的使用、重视结果的分析…
- 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(包括纸质版、电子版源文件及程序). (4)2024 年 1 月 11 日 18:00 点前发(递交)至课代表处.
- 期末课程项目可分组完成(自由组合，不超过 3 人一组，可以 1 个人)，可以同学间相互讨论或者找老师答疑. 如果是多人组队，请填写附表 1，明确说明每人负责的部分和内容.
- 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. 格式请参考本科生毕业论文写作模板 (http://bysj.njust.edu.cn/Index.aspx).
- 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(包含但不限于：泰迪杯、全国大学生数学建模大赛、美国大学生数学建模竞赛、Kaggle 竞赛…)，具体咨询组内成员或相关老师.
- 请使用公式编辑器对项目内包含的公式进行编辑……
- 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 矩阵补全模型及其应用

【1】项目背景

基于凸优化的压缩感知技术及其相关的矩阵补全与矩阵恢复技术是近年来统计与机器学习领域内的研究热点, 广泛应用于图像处理、推荐系统和计算机视觉等众多领域. Netflix 电影评分挑战赛便是矩阵补全最经典的例子之一(Bennett and Lanning, 2007). Netflix 是一家著名的在线影片租赁公司, 2006 年该公司举办了首届电影评分挑战赛, 目的是提高系统向顾客推荐影片的能力. Netflix 提供的数据集有$n=17770$ 部影片(每部影片为 1 列), 顾客数为$m=480189$ (每个用户为 1 行). 顾客对影片的评分范围是$1\sim5$, 其中 1 是最差, 5 是最好. 在训练集中, 数据矩阵非常稀疏, 仅有 100 万(1%)个被评过分的元素. 这个比赛的目的就是要得出没被评价过的影片的得分, 以便更好地向顾客推荐影片.

矩阵填充的越准确, 为用户推荐的电影也就越符合用户的喜好. 由于影响用户对电影喜好的因素数目有限, 如电影的题材、演员、年代、导演等, 这个矩阵本质上是一个低秩矩阵. 在 2006 年, Netflix 采用的是 Cinematch 算法, 该算法在一个大的测试集上得到的均方根误差(Root Mean Square Error, RSME)为 0.9525. 这个比赛从 2006 年开始, 在那一年, 获得第一名的算法将 RSME 降低了至少 10%. 在 2009 年, 这个比赛的最终获胜者是一群研究人员, 他们称为 Bellkor's Pragmatic Chaos, 这是三个一起赢得比赛的小组的名称. 获胜算法采用了大量统计技术, 但同其他参与比赛的算法一样, SVD 起到了至关重要的作用.

与矩阵补全密切相关的另外一个研究方向是矩阵恢复, 该问题考虑如何从较大的但稀疏的误差中恢复出本质上低秩的数据矩阵. 在不同场合, 低秩矩阵恢复也被称为矩阵低秩稀疏分解(sparse and low-rank matrix decomposition) (即将一个矩阵分解为一个低秩矩阵和一个稀疏矩阵之和)、鲁棒主成分分析(robust principle componentanalysis, RPCA) 、低秩稀疏非相干分解 (rank-sparsity incoherence)等. 值得注意的是: 在低秩矩阵填充中, 矩阵中位置元素的位置是已知的; 而在低秩矩阵恢复中, 矩阵是完整的, 但是其中哪些元素受到了误差的破坏并不知道. 也就是说, 低秩矩阵恢复比低秩矩阵填充更具有挑战性, 或者说低

秩矩阵恢复要同时检测被破坏的元素的位置并恢复它们. 低秩矩阵填充和低秩矩阵恢复可以合称为低秩矩阵重建(low-rank matrix reconstruction). 本项目重点考察与矩阵补全相关的内容.

【2】方法陈述

问题一: 结合搜集到的资料, 论述1~2种矩阵补全模型及其算法.

问题二: 尝试与课程中习的方法进行比较, 并论述其异同.

【3】案例实战

1. 教材案例回顾: 在电商领域, 好的推荐算法能带来更高的销售业绩. 例如, 淘宝或京东的网店使用产品推荐技术向潜在用户推荐产品. 在各种在线视频网站中, 推荐算法也具有很大的应用空间, 可以根据用户对电影的评分数据来对其进行电影推荐.

本案例提供一份来自Grouplens的电影评分数据集, 记录了671名用户对9125部电影的评分信息. 数据集包括movies和ratings两部分: movies部分记录电影的基本信息; ratings部分记录用户对电影的评论情况, 共有100004个样本, 每个样本包含4个特征. 特征的具体信息如表1所示.

表 1　ratings 数据集特征

特征名称	特征说明
userId	用户编号
movieId	电影编号
rating	电影评分
timestamp	时间戳

请使用教材中提及的关联规则挖掘算法, 根据电影评分数据集, 找出频繁项集以及关联规则, 对用户进行电影推荐. *(数据集可在教材网站获取)*

2. 矩阵补全模型下的推荐结果

现在我们将注意力转向矩阵补全模型, 并使用1~2种矩阵补全方法将上述MovieLens 数据集进行补全, 并对用户进行电影推荐.

3. 不同方法间的比较

(1) 请选择合适的评价指标，对矩阵补全及关联规则方法给出的结果进行分析;

(2) 为验证你的结论，请在不同数据集下探究二者之间的差异．部分数据集如下所示:

Jester: https://eigentaste.berkeley.edu/dataset/

MovieLens: https://grouplens.org/datasets/movielens/

Netflix 数据集见附件一.

【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

案例三代码

课程项目二 分类、集成模型及其应用

【1】项目背景

在回归模型中假设响应变量是定量的, 但很多情况下响应变量却是定性的. 例如, 眼睛的颜色是定性变量, 取值蓝色、棕色或绿色. 定性变量也称为分类变量, 两者的统计含义是一样的. 预测一个观测的定性响应值也指对观测分类, 因为它涉及将观测分配到一个类别中. 一些分类方法先从预测定性变量不同类别的概率开始, 将分类问题作为概率估计的一个结果. 从这个角度上看, 分类与回归方法有许多类似之处.

《数据科学导论》课程中已习得几种分类方法, 如逻辑回归、K 近邻、决策树、支持向量机、随机森林和 AdaBoost 等. 现有分类器中, 支持向量机(support vector machine, SVM)是一种起源于机器学习社区的流行方法, 它是使用单一分类函数对二分类问题进行最大间隔分类的典型例子, 在实际应用中性能良好. 在很多问题中, 数据通常包含两个以上的类别. 尽管 SVM 在二分类上取得了成功, 但如何将其应用于多类别分类问题仍然是一个挑战. 文献中使用 SVM 进行多分类的方法可分为两大类: 第一类方法先训练一系列的二值支持向量机, 再将结果组合起来进行多类别分类. 例子包括一对一(one-versus-one)和一对多(one-versus-rest)方法. 尽管这种方法在概念和实现上都很简单, 但它也有自身的缺点. 第二类方法是在一个优化问题中同时考虑所有 K 个类. 常用的方法是用 K 个分类函数来表示对应优化中的 K 个类, 并根据哪个分类函数最大给出最终的预测结果. 目前, 在第二类框架下提出了许多同步多分类支持向量机分类器. 本项目重点关注支持向量机与其他方法的比较, 以及如何使用 SVM 处理多分类问题.

【2】方法陈述

问题一: 请对现有的 3~5 种分类方法进行总结, 试搜寻每类方法的优缺点(SVM 需要着重介绍).

问题二: 尝试借助课程及网络, 检索 SVM 及其变体模型, 并对其相关机理进行论述.

【3】案例实战

1. 教材案例回顾：乳腺癌诊断

早期的乳腺癌检测主要检查乳腺组织的异常肿块. 如果能利用机器学习算法, 通过乳腺肿块的检测数据自动进行诊断, 将会给医疗系统带来很大的益处: 不仅能够大大提高检测效率, 还可以降低误判的风险.

数据集共有 569 个样本, 每个样本包含乳腺细胞的 30 个特征, 这 30 个特征都是由数字化细胞核的 10 个基础特征(包括半径、质地、周长、面积、对称性等)的均值、标准差及最大值构成. 请使用3~5分类方法(包含 SVM), 对乳腺癌数据进行分类, 并选择适当的指标对最终的分类结果进行评价.

注: 评价指标可参考教材附录 E.

2. 请将问题 1 中的分类器应用至下述数据集(二选一)

(1) 教材第四、五章中其余 6 个数据任选三个

(2) Gas、Optical、Vertebral 数据集(数据集见附件二)

强调分析不同分类器在不同数据集下的分类结果.

【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 2023 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. 2023 年, 来自全国及美国、澳大利亚、马来西亚的 1685 所院校/校区、59611 队(本科 54158 队、专科 5453 队)、近 18 万人报名参赛. 本科组一等奖获奖率约为 0.552%, 本科二等奖获奖比例约为 2.21%. 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/2023qgdxssxjmjslwzs/2023ctlw/

请勿沉溺于竞赛, 无法自拔

【2】问题重述: 蔬菜类商品的自动定价与补货决策

在生鲜商超中, 一般蔬菜类商品的保鲜期都比较短, 且品相随销售时间的增加而变差, 大部分品种如当日未售出, 隔日就无法再售. 因此, 商超通常会根据各商品的历史销售和需求情况每天进行补货.

由于商超销售的蔬菜品种众多、产地不尽相同, 而蔬菜的进货交易时间通常在凌晨3:00-4:00, 为此商家须在不确切知道具体单品和进货价格的情况下, 做出当日各蔬菜品类的补货决策. 蔬菜的定价一般采用“成本加成定价”方法, 商超对运损和品相变差的商品通常进行打折销售. 可靠的市场需求分析, 对补货决策和定价决策尤为重要. 从需求侧来看, 蔬菜类商品的销售量与时间往往存在一定的关联关系; 从供给侧来看, 蔬菜的供应品种在4月至10月较为丰富, 商超销售空

间的限制使得合理的销售组合变得极为重要.

附件1给出了某商超经销的6个蔬菜品类的商品信息；附件2和附件3分别给出了该商超2020年7月1日至2023年6月30日各商品的销售流水明细与批发价格的相关数据；附件4给出了各商品近期的损耗率数据. 请根据附件和实际情况建立数学模型解决以下问题:

问题1 蔬菜类商品不同品类或不同单品之间可能存在一定的关联关系，请分析蔬菜各品类及单品销售量的分布规律及相互关系.

问题2 考虑商超以品类为单位做补货计划，请分析各蔬菜品类的销售总量与成本加成定价的关系，并给出各蔬菜品类未来一周(2023年7月1-7日)的日补货总量和定价策略，使得商超收益最大.

问题3 因蔬菜类商品的销售空间有限，商超希望进一步制定单品的补货计划，要求可售单品总数控制在27-33个，且各单品订购量满足最小陈列量2.5千克的要求. 根据2023年6月24-30日的可售品种，给出7月1日的单品补货量和定价策略，在尽量满足市场对各品类蔬菜商品需求的前提下，使得商超收益最大.

问题4 为了更好地制定蔬菜商品的补货和定价决策，商超还需要采集哪些相关数据，这些数据对解决上述问题有何帮助，请给出你们的意见和理由.

附件1 6个蔬菜品类的商品信息

附件2 销售流水明细数据

附件3 蔬菜类商品的批发价格

附件4 蔬菜类商品的近期损耗率

注 (1) 附件1中，部分单品名称包含的数字编号表示不同的供应来源.

(2) 附件4中的损耗率反映了近期商品的损耗情况，通过近期盘点周期的数据计算得到.

【3】建模论文

【4】论文代码

（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

附表 1 参与人员

项目负责人及参与人员情况表

序号	姓名	性别	学号	专业	投入时间占比	任务分工	项目中职务
1							项目负责人
2							骨干
3							骨干

备注：本表需如实填写，可注明共同完成项目内容，将作为成绩评判的重要依据.