2024春季大规模数据处理技术 第1次作业

2024.04.08

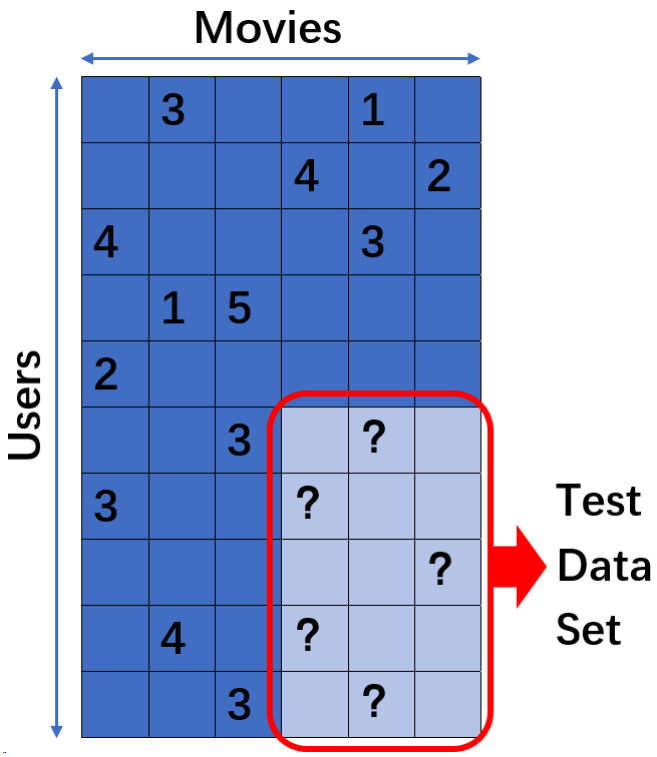
一、实验简介

本次实验旨在让同学们简单了解一下课堂上提到的传统的推荐系统算法并完成其具体实现。我们以MovieLens-1M作为原始数据集，在此基础上随机调整了用户和影片的ID以作区别，并移除了约1/10的数据作为我们的测试集。同学们需要使用我们调整后的数据集来实现用户对特定影片的打分预测，并提交原代码、预测结果以及实验报告。

二、数据集结构

数据集以矩阵表示，变量名为 `col_matrix`

- 行标：表示用户的ID，从0开始编号，共6040个用户（行）
- 列标：表示影片的ID，从0开始编号，共3952部影片（列）
- 矩阵的元素：表示用户对该影片的评分，以0-5分表示，0分为缺失，分数均为整数
- `col_matrix[4100:, 2700:]` 的部分为测试集（如下图淡蓝色的区域），已为同学们隐藏（设置为0），便于作业打分；其余部分数据公开，请同学们**自行划分训练集和验证集**



三、实验要求

1. 载入训练数据 `col_matrix.csv` ，在其上实现一种推荐算法（如UserCF，ItemCF等），预测淡蓝色部分的用户对影片的评分
2. 输出对测试区域 `col_matrix[4100:, 2700:]` 的预测并保存为 `test_prediction.csv` ，输出整数打分
3. 撰写实验报告，包括方法介绍、实现细节、机器配置、时间/内存开销、验证评估等内容

四、提交内容

1. 源代码
2. 报告，保存为pdf格式
3. 输出结果文件 `test_prediction.csv`

以上内容**打包为一个.zip压缩文件**，重命名为“<学号>_<姓名>.zip”（如“123033919999_张三.zip”）提交至canvas平台，**截止日期为4月22日23:59**，迟交将扣分

五、评分标准

本次实验的评分会根据报告内容、在测试集上的预测评估、算法开销以及代码内容等方面进行综合评分