上海理工大学光电信息与计算机工程学院

**《推荐系统》实验三**

****

|  |  |
| --- | --- |
| **学　　院** | 光电信息与计算机工程学院 |
| **专　　业** | 数据科学与大数据技术 |
| **学生姓名** | 罗震宇 |
| **学　　号** | 2135060321 |
| **指导教师** | 艾均 |
| **成 绩** |  |

**目录**

[实验三 User-based Slope One算法实现 5](#_Toc152052520)

[1. 实验要求与环境 5](#_Toc152052521)

[1.1 实验要求 5](#_Toc152052522)

[1.2 实验环境 5](#_Toc152052523)

[2. 实验过程 5](#_Toc152052524)

[2.1 数据输入 5](#_Toc152052525)

[2.2 评分偏差计算 5](#_Toc152052526)

[2.3 用户评分预测 6](#_Toc152052527)

[2.4 评估与可视化 6](#_Toc152052528)

[3．实验结果 6](#_Toc152052529)

[4. 个人感想 7](#_Toc152052530)

实验三 User-based Slope One算法实现

## 实验要求与环境

### 实验要求

1. 选择MovieLens small 数据集；

2. 实现基于用户Slope One距离模型的协同过滤（CF）个性化推荐算法；

3. 读入Ratings数据并适当保存和索引；

4. 数据按折n验证分割成训练集和测试集；

5. 计算用户间相似性；

6. 用户邻居按照相似性进行排序；

7. 针对测试集中预测目标，预测用户在k个邻居时的可能评分；

8. 计算k个邻居时User-based Pearson-CF的平均绝对误差并画出误差曲线；

### 1.2 实验环境

处理器 AMD Ryzen 7 5800H with Radeon Graphics

内存 16 GB

系统 Windows 11

IDE PyCharm Professional 2023.2.4

## 2. 实验过程

### 2.1 数据输入

从指定文件地址读取电影评分数据和电影数据并随机打乱电影评分数据。将打乱后的电影评分数据采用K折交叉验证法进行划分，将数据分成5个折叠。返回这些折叠、原始的电影评分数据、电影数据以及用户ID。

### 2.2 评分偏差计算

读取用户的电影评分数据和电影信息后，将评分数据转换成User为列索引，Item为行索引的评分的Pandas DataFrame数据结构。基于评分偏差距离公式计算用户之间的评分偏差距离。

评分偏差公式：

### 2.3 用户评分预测

根据Slope One的协同过滤评分预测公式预测用户评分。

基于Slope One的协同过滤评分预测公式：

### 2.4 评估与可视化

本实验采用平均绝对误差(MAE)作为评估标准。

Mean Absolute Error：

通过调用Seaborn库函数，绘制MAE随邻居数量提升而变化的折线图，并以PDF格式保存。

## 3．实验结果

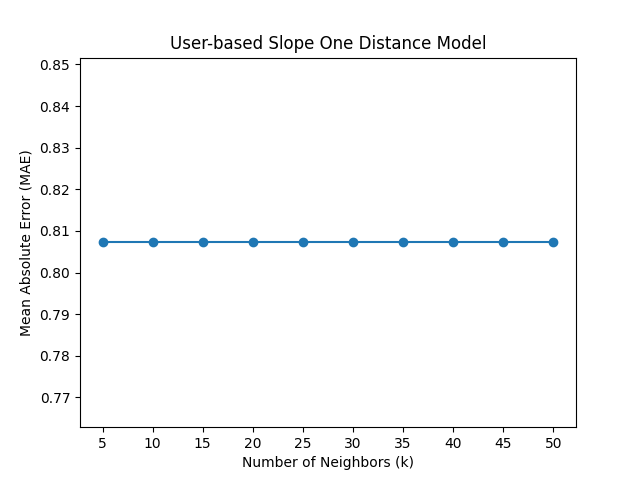


Fig.1 Line Chart of MAE vs. Neighbors

如Fig.1所示，与前两个实验的图不同，因为Slope One本身的特性与邻居数量无关，所以即便改变邻居数量，Slope One模型的MAE不会发生变化，结果为一条横线。

## 4. 个人感想

本次题目的目的是设计并实现一个基于协同过滤的推荐系统，利用用户的评分数据，为用户提供个性化的推荐服务，采用基于用户Slope One距离模型的协同过滤（CF）个性化推荐算法。我对本次题目的完成情况给予了96分。在完成本次题目的过程中，有以下几个方面的收获：

在理论学习方面，这个模型的原理较为清晰，学起来很快，但是实际上手发现还是和之前的不太一样，需要静下心来细想。如果直接按照前面的思路来进行操作得到的图是错误的。

在编程实现方面，采用了模块化的编程思想，将每一个步骤封装成一个函数，然后在主程序中调用，实现了代码的高效和可维护。我还注意了代码的规范和注释，提高了代码的可读性和可理解性。中途判断是否将同一用户取出来两次逻辑需要注意。其他还有一些算法流程上可以优化的点，例如将中间结果存进缓存而不是每次都重新计算。

在报告撰写方面，我遵循了上海理工大学论文的格式要求，详细地阐述了实验的背景、原理、过程和结果，较好地反映了本次实验的目的和意义。实验结果具有可重复性和科学性，可以作为推荐系统研究的一个参考案例。

整个过程中还是遇到了很多问题，例如每次运行代码时间都很长，导致debug速度较慢；对debug工具不够数量，前期浪费了很多时间等，希望自己能够汲取这次的经验教训，下次更快更好地完成实验。