上海理工大学光电信息与计算机工程学院

**《推荐系统》实验一**

****

|  |  |
| --- | --- |
| **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**

[实验一 User-based Pearson-CF算法实现 5](#_Toc152052378)

[1. 实验要求与环境 5](#_Toc152052379)

[1.1 实验要求 5](#_Toc152052380)

[1.2 实验环境 5](#_Toc152052381)

[2. 实验过程 5](#_Toc152052382)

[2.1 数据输入 5](#_Toc152052383)

[2.2 用户间相似性计算与排序 5](#_Toc152052384)

[2.3 用户评分预测 6](#_Toc152052385)

[2.4 评估与可视化 6](#_Toc152052386)

[2. 实验结果 6](#_Toc152052387)

[4. 个人收获 8](#_Toc152052388)

实验一 User-based Pearson-CF算法实现

## 实验要求与环境

### 1.1 实验要求

1. 选择MovieLens small 数据集；
2. 实现基于用户Pearson相似性的协同过滤（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

使用pandas库，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10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 用户间相似性计算与排序

读取用户的电影评分数据和电影信息后，将评分数据转换成双重字典数据结构。能够通过用户的userId找到movieId，再通过movieId找到对应的rating。设置一个函数用来找到每个用户对哪些电影评过分，之后才能通过用户之间的共同评分的电影来计算用户之间的相似性。基于Pearson相关系数计算用户之间的相似度，为每一对用户生成一个相似度分数。

Pearson相似度计算公式：

接下来为每个用户确定与其相似度最高的用户，并将其他用户按相似度降序排列，获得指定数量的邻居。

### 2.3 用户评分预测

设计函数用于对一组用户-物品对进行评分预测。它接受数据集的折叠、用户-物品-评分数据字典、所有用户的平均评分数据双重字典、用户配对信息和指定邻居数量。

在函数中，首先计算目标用户的平均评分（如果存在），然后根据邻居的评分和相似度计算出预测评分。如果计算结果不是一个数字（计算时有None），则返回目标用户的平均评分。最后，根据相似用户的邻居信息进行评分预测，将不同邻居数量要求下的预测结果存储在一个双重字典中。

基于Pearson相似性的协同过滤评分预测公式：

### 2.4 评估与可视化

得到预测得分之后，与用户真实的评分对比，计算预测得分和真实值之间的误差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库的函数，绘制MAE随邻居数量提升而变化的折线图，并以PDF格式保存。

## 实验结果

图表

描述已自动生成

Fig.1 Line Chart of MAE vs. Neighbors

如Fig.1所示，在从5至100的40个不同邻居数量下，模型的MAE逐渐下降。在Neighbors从5增加到100的过程中，MAE的下降趋势逐渐减小。邻居数量达到70后，MAE基本不再减小，开始在0.661左右波动。

从上述实验中，我们可以得出以下结论：在一定范围内增加Neighbors数量会提高模型的性能，降低MAE。然而，增加Neighbors的数量对改进模型性能的效果在某个点开始减弱，而且随着Neighbors数量的增加，性能的提升幅度逐渐减小。当Neighbors为80时，MAE达到0.668134，而当Neighbors为85时，MAE略微升高为0.668127，这说明在这个数据集中，Neighbors数量的微小变化可能对性能没有显著影响，或者MAE已经收敛到一个较低的值。

总之，从这组数据中可以得出，增加Neighbors数量可以改善模型的性能，但在某个点后性能提升会减小，因此需要权衡性能和计算成本，选择适当数量的Neighbors来进行建模

## 4. 个人收获

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。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，最开始的时候没有理解原理便开始做实验，效果不好，后来我深入掌握了协同过滤的核心思想和算法流程，学习了如何使用皮尔逊相关系数来衡量用户之间的相似度，以及如何使用k折交叉验证法来评估模型的准确性和稳定性，再写代码就快很多了。通过这个部分，我感受到了计算机科学的严谨性和创新性，激发了我对科研的兴趣和热情。

在编程实现方面，最开始并没有很清晰的思路，都是想到哪写到哪，没有太注意规范。后来经过同学指教，我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。

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，下次更快更好的完成实验。