

**机器学习大作业**

**【题目】**

基于机器学习的个性化推荐系统

课程：机器学习

姓名：xxx

学号：2000xxxxxx

班级：计科x班

日期：2025.5.30

1. 研究背景

是一个基于协同过滤的电影推荐系统，主要涉及推荐系统、协同过滤、数据挖掘、机器学习和用户行为分析等领域。

采用了基于用户的协同过滤方法，通过余弦相似度计算用户间的相似性，并结合数据可视化分析、用户偏好建模和时间序列分析等。

在推荐系统这个领域，主要存在冷启动、数据稀疏、可扩展性和推荐多样性等问题。针对这些问题，有了许多解决方案，比如混合推荐方法、矩阵分解技术、深度学习模型和实时推荐系统等。

我这个项目的创新点主要体现是实现了多维度推荐策略、可视化分析和个性化推荐三个方面。多维度推荐策略结合用户评分历史、电影类型偏好和时间因素；可视化分析包括用户评分分布分析、电影类型分布展示、推荐结果直观展示和用户相似度热力图；个性化推荐支持用户自定义偏好类型、时间范围筛选和动态调整推荐权重。

1. 模型方法

**详细**说明本工程所使用的模型方法和理论。

我的这个项目的核心模型基于协同过滤推荐算法，主要使用了基于用户的协同过滤方法。该方法的基础是"相似用户对相似物品有相似的偏好"，所以我们就额可以通过分析用户的历史行为数据，发现用户之间的相似性，从而为用户推荐其相似用户喜欢的物品。

我采用了余弦相似度作为度量标准在用户相似度计算方面，余弦相似度通过计算两个用户评分向量的夹角余弦值来衡量用户间的相似程度，我使用这种方法的优势在于能够有效处理评分尺度不一致的问题，并且对稀疏数据具有良好的鲁棒性，也就是说不会导致数据过滤稀疏而出现幻觉。

在推荐生成步骤中，我使用了加权评分预测方法。对于目标用户未评分的电影，通过计算其相似用户对该电影的评分加权平均来预测评分值。权重系数由用户相似度决定，相似度越高的用户，其评分对预测结果的影响越大。同时，我还引入了电影类型偏好权重和时间衰减因子，用来提升推荐的个性化和时效性。

到了数据处理阶段，就需要着重处理了，因为数据处理的好坏直接影响最终的结果。先对原始评分数据进行标准化处理，消除评分尺度差异；再通过时间戳分析，引入时间衰减因子，使得近期评分具有更高的权重；最后对电影类型进行one-hot编码，便于后续的特征提取和相似度计算，数据肯定是机器学习的基石部分，所以我选择了不同的处理方法

然后就是数据的可视化了，使用比较流行的matplotlib以及seaborn完成可视化任务，然后有很多不同类型的数据，每个数据最好都使用合适其展示的方法，比如我使用直方图展示用户评分分布，使用条形图展示电影类型分布，使用热力图展示用户相似度矩阵，使用折线图展示推荐结果的评分对比，帮助评估推荐效果。根据不同的得到的处理的数据，选择不同的图标加以展示。

在模型优化方面，需要使用多层次的优化策略。首先，通过引入电影类型偏好权重，提高推荐结果的多样性；其次，结合时间衰减因子，确保推荐结果具有时效性；最后，通过动态调整相似用户数量，在推荐准确性和计算效率之间取得平衡。在这些优化策略协同作用下，使得推荐系统能够更好地满足用户的个性化需求。

1. 系统设计

在完成这个个性化推荐系统的过程中，我采用了模块化的设计思路，将整个系统分成了数据加载、数据预处理、模型训练、推荐生成和可视化分析这五个主要部分。选择用main.py作为程序的入口文件，这样可以让整个推荐流程更加清晰和有条理。

文本

AI 生成的内容可能不正确。

在数据加载这部分，我写了一个loadData()函数来处理数据的读取和初始化工作。这个函数主要用来读取四个重要的数据文件：ratings.csv（记录用户对电影的评分）、movies.csv（电影的基本信息）、tags.csv（用户给电影打的标签）和links.csv（电影在其他网站上的链接）。读取完数据后，我还写了一个printData()函数来显示数据集的基本情况，比如每个数据集有多少条记录，方便后期调试。



两个辅助函数：

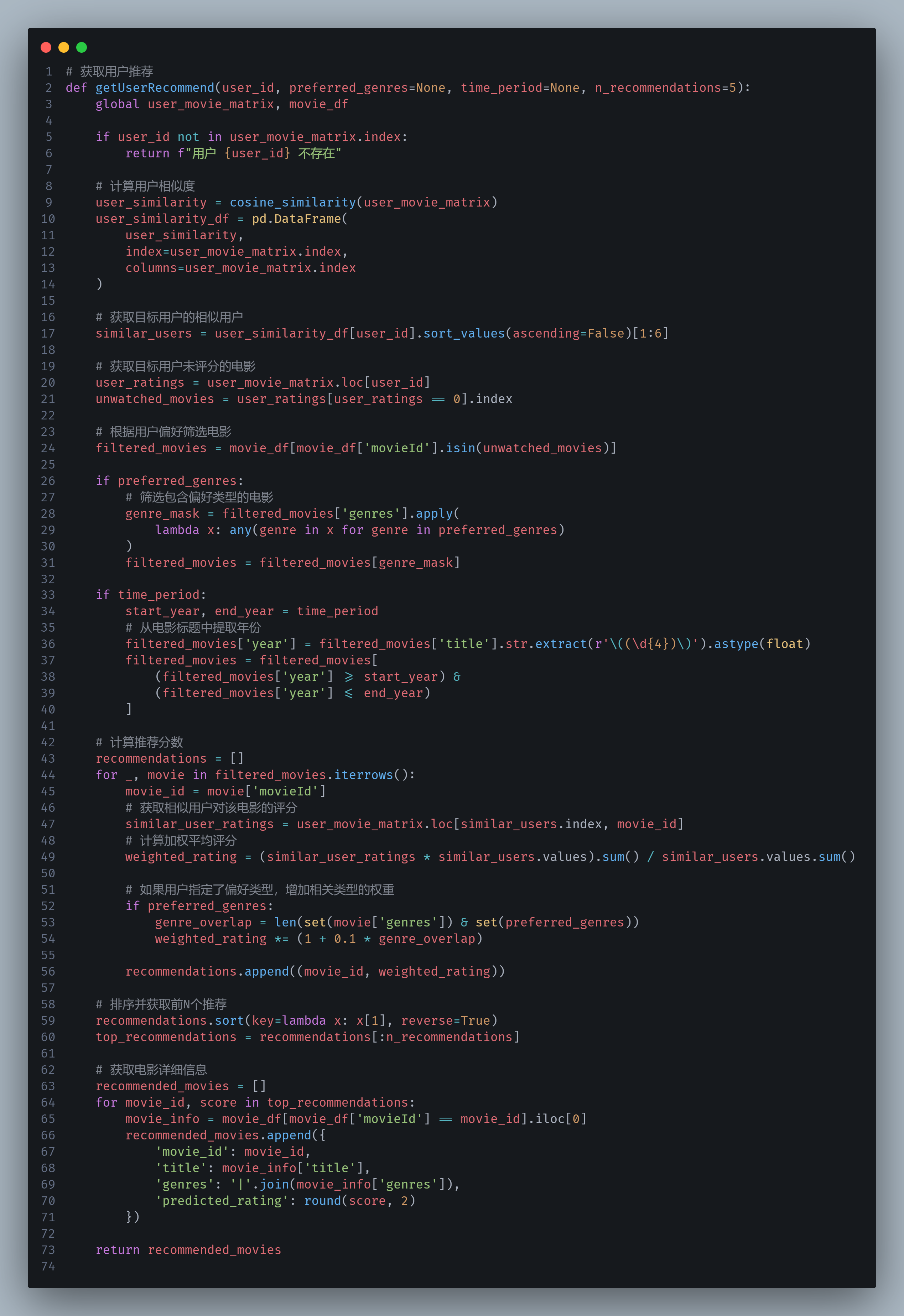


数据预处理这部分我放在了initData()函数里，主要在这个函数里面做了三件事：第一，把电影类型从字符串变成列表，这样后面筛选电影类型的时候会更方便；第二，用pivot函数把用户评分数据重新整理成一个矩阵，行是用户，列是电影，没有评分的地方用0填充；第三，把时间戳转换成成能看懂的日期格式，然后这样的话在后面分析时间的时候会更直观。

文本

AI 生成的内容可能不正确。

模型训练的核心代码都在getUserRecommend()函数里，就是这个这个函数实现了基于用户的协同过滤算法。我先用cosine\_similarity计算了用户之间的相似度，然后找出和目标用户最相似的5个用户，再找出这些用户看过但目标用户没看过的电影，最后根据用户选择的类型和时间范围来筛选电影，计算每部电影的预测评分。



在生成推荐结果的时候，我用了多层的筛选和排序方法。先根据用户选择的类型和时间范围筛选一遍电影，然后计算每部电影的预测评分，这个评分会考虑用户相似度和类型匹配度，最后把评分最高的几部电影作为推荐结果。这些都在recommendations列表里实现。

为了能直观地看到推荐效果，我写了四个可视化函数：visualize\_rating\_distribution()用来画评分分布图，visualize\_genre\_distribution()用来展示电影类型分布，visualize\_recommendations()用来对比推荐电影的评分，visualize\_user\_similarity()用来展示用户之间的相似关系。这些图都是用matplotlib和seaborn画的，因为与其看枯燥的数据，不如将数据可视化成一幅图片，这样就能方便快捷的知道每一项的差异和联系了

文本

AI 生成的内容可能不正确。

用户交互这部分我放在了getUserInp()函数里，这个函数会引导用户输入用户ID、喜欢的电影类型和时间范围。我还加了一些输入检查，比如用户ID必须在1-610之间，时间范围必须在1900-2024年之间，当然这些范围都是根据data文件夹下的数据推算得到的，这样可以避免用户输入错误的数据，我没有用trycatch处理下标溢出的情况，而是，只要用户给出的下标不对，我就反复提醒用户，知道数据正确为止。

在优化系统性能的时候，我想了几个办法：用pandas的向量化操作代替循环，用矩阵运算代替一个一个计算，用DataFrame存储中间结果避免重复计算。比如在计算用户相似度的时候用了cosine\_similarity函数，在筛选电影的时候用了布尔索引，这些都能让程序运行得更快，而上面这些方法都是官方所推荐的。

最后，我通过几个方面来评估这个推荐系统：看推荐结果的分布情况，分析用户之间的相似关系，检查推荐结果的类型和时间分布。这些评估指标都在可视化函数里实现了，比如在visualize\_recommendations()里可以看到推荐电影的评分分布，在visualize\_user\_similarity()里可以看到用户之间的相似关系。

1. 系统演示与分析

这里我们默认选择ID为1的用户：

日历

AI 生成的内容可能不正确。

可以看到之前说的测试函数能正常工作打印结果。

下来是选择电影类型：

文本

AI 生成的内容可能不正确。

我选择2，4（动作片和动画片）类型，当然也可以不选，那就失去了一个检索维度：

文本

AI 生成的内容可能不正确。

下来时选择指定的年份区间，这里我不选，那就是默认使用数据集中的全部的时间段：

文本

AI 生成的内容可能不正确。

可以看到，他成功的推荐了5部电影，并且进行了预测的评分。

然后四个可视化文件也已经生成了：

图表, 条形图

AI 生成的内容可能不正确。

图表, 条形图

AI 生成的内容可能不正确。

图表, 条形图

AI 生成的内容可能不正确。

图表, 树状图

AI 生成的内容可能不正确。

可以做进一步的分析。

五、对本门课的感想、意见和建议。

通过完成这个电影推荐系统的项目以及上课，我对机器学习有了更深入的理解。以前总觉得机器学习很神秘，现在通过实际动手做项目，发现其实很多概念都是可以理解和实现的。特别是在做数据预处理和可视化分析的时候，我学会了如何把复杂的数据变成直观的图表，这对我理解数据特征很有帮助。我觉得这门课最大的优点是理论和实践结合得很好。老师不仅讲解了推荐系统的基本原理，还让我们动手实现了一个完整的系统。在实现过程中，我遇到了很多问题，从无知的小白，一步步摸索，求助，最终完成了这个系统，收获还是很多的。

不过，我也觉得课程还可以有一些改进的地方。首先，希望老师能多讲一些实际应用案例，比如现在主流的推荐系统是如何工作的，这样我们能更好地理解理论知识。毕竟一直讲理论知识未免太枯燥了点。

**代码验收成绩**

|  |  |  |
| --- | --- | --- |
| 运用计算机科学与技术及其应用领域工程的思想与方法，从需求分析开始，独立完成机器学习系统的设计与开发，兼顾机器学习系统开发的社会评价。（10分） | 理解机器学习系统运行应遵守的社会规范，能够在机器学习系统开发实践中自觉遵守工程职业道德和规范，履行责任。（10分） | 理解机器学习系统开发相关的工程配置管理方法，掌握领域方向系统项目管理能力。（10分） |
|  |  |  |

**报告成绩**

|  |  |  |
| --- | --- | --- |
| CO1 运用计算机科学与技术及其应用领域工程的思想与方法，从需求分析开始，独立完成机器学习系统的设计与开发，兼顾机器学习系统开发的社会评价。（50分） | CO2 理解机器学习系统运行应遵守的社会规范，能够在领域方向系统开发实践中自觉遵守工程职业道德和规范，履行责任。（10分） | CO4 理解机器学习系统开发相关的工程配置管理方法，掌握领域方向系统项目管理能力。（10分） |
|  |  |  |

**总成绩：**

**指导教师评语（验收/报告）：**

**指导教师签名：**

**日期：**