**《大数据分析与处理》实验报告**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **年级、专业、班级** | |  | | | **姓名** |  |
| **实验题目** | **实验二 基于奇异值分解的推荐算法** | | | | | |
| **实验时间** |  | | **实验地点** |  | | |
| **实验成绩** |  | | **实验性质** | **□验证性 ☑设计性 □综合性** | | |
| 教师评价：  □算法/实验过程正确； □源程序/实验内容提交 □程序结构/实验步骤合理；  □实验结果正确； □语法、语义正确； □报告规范；  其他：  评价教师签名： | | | | | | |
| 一、实验目的   * 在真实数据上通过奇异值分解实现推荐算法，了解奇异值分解在推荐算法上的实际应用。 * 学习如何拆分数据集，了解一般算法的优化过程与调参过程。 | | | | | | |
| 三、实验过程或算法（源程序）  本次实验内容为实现基于奇异值分解的电影推荐算法，实验数据集选自MovieLens中的ml-1m.zip。数据集中包含四个文件,分别为：历史评分文件“ratings.dat”、电影信息文件“movies.dat”、用户信息文件“users.dat”以及“README”文件。文件中具体字段含义请查看“README”文件。  实验原理如下：  “ratings.dat”中数据格式为： UserID::MovieID::Rating::Timestamp。基于以上数据生成一个*M* × *N*的评分矩阵A，每行代表一个user （即UserID），每列代表一个item（即MovieID），其中的元素表示user对item的打分Rating，Rating数值为从1到5，若user未对item打分则元素值为0，即需要预测的值。  基于评分矩阵A，SVD的计算公式如公式(1)所示，  (1)  其中，表示用户和语义间的关系，表示电影和语义间的关系，则。*M*表示user数量，*N*表示item数量，*K*表示隐式语义维度，是本次实验中需要调优的超参数。  从向量的角度看公式(1)，可以得到预测公式如公式(2)所示，  (2)  其中表示矩阵*P*的*u*个行向量，*qi*表示矩阵*Q*的第*i*个列向量，公式(2)给出了用户*u*对电影*i*预测评分的计算公式，而预测值与真实值之间的误差公式如公式(3)所示，  (3)  其中*rui*表示真实评分，即评分矩阵*A*中的非零元素。由单个到总体，算法的总误差平方和如公式(4)所示，  (4)  算法的优化目标即尽可能让*SSE*值变小，从而让预测值尽可能接近真实值，需要注意的是，公式(4)关于*u*和*i*求和时只考虑的部分，即用户实际评过分的部分。优化算法需要确定优化的方向，这里优化算法采用梯度下降法，因此需要对公式(4)进行求导。公式(4)中存在两个变量：*puk*和*qki*。*SSE*在变量*puk*的梯度如公式(5)所示（详细求导过程略），  (5)  按照梯度下降法，更新方向沿负梯度方向。令更新步长（学习速率）为，两变量的更新公式如公式(6)和公式(7)所示，  (6)  (7)  至此，得到了SVD算法的计算过程及参数更新过程。*puk*和*qki*的初始值可以随机初始化为1~5之间的数。本实验算法采用的评价指标可选择推荐算法常用的RMSE（均方根误差）。  具体实验步骤如下：   1. 数据预处理。   从“ratings.dat”中读取数据，并预处理为适用的格式。   1. 将“ratings.dat”分为训练集、验证集与测试集。   训练集是算法在进行训练时用到的数据，train\_size = 80%；验证集是作为每次训练后进行验证时用到的数据，我们可以通过验证来优化SVD算法中的K值（隐式语义维度）以及学习率参数*η*，val\_size = 10%；测试集用于测试最终算法的效果，test\_size = 10%。  同学们需自己对“ratings.dat”文件进行拆分，拆分方法如下：从每一个用户的所有评分中随机无放回地选择test\_size比例的数据作为测试集，再选择val\_size比例的数据作为验证集，其余作为训练集，保证各集间不存在交叉数据。   1. 基于SVD的推荐算法实现。   根据实验原理实现基于SVD的推荐算法。   1. 进行算法调参，得到最终的预测结果。   利用测试集对K、*η*进行算法调参，绘制不同K、*η*值对应的验证集RMSE变化曲线，确定最终的参数，并给出测试集上的最终结果。   1. （选作）将SVD推荐方法与一些基本的推荐算法进行对比，如以某个用户或电影的平均评分值、众数、中位数作为预测值，以TopK推荐评价指标作为评分指标（如Precision@K、Recall@K、F1@K等），给出对比结果。 | | | | | | |
| 四、实验结果及分析和（或）源程序调试过程  本实验数据处理，主要基于numpy,scikit-surprise, scikit-learn等python第三方库函数展开实验。具体步骤如下： 数据预处理及数据集划分 本节主要是对数据进行读取和格式化处理。同时也将数据进行打乱并按照训练集：验证集：测试集=8:1:1的比例对数据集进行划分   基于SVD的推荐算法实现 按照本实验的相关指导，实现了梯度下降的SVD。其基本思想是随机初始化P,Q两个矩阵，然后通过train\_data对P,Q两个矩阵进行梯度下降优化。从而完成了SVD算法的计算过程及参数更新。    在运行了累计20个epoch后，程序运行结果如图所示。其中可以发现loss的变化量已经很小了。已经趋近于收敛。   对SVD进行算法调参 所选取的不同K、*η*在进行10个epoch后，对应的RMSE值。    绘制不同的K，η最后100个loss记录的图像。    对同一个K值，取更大的lr，下降的会更加迅速。但是loss曲线可能因为学习率过大存在抖动。  对于同一个lr值，在K值很小为5的时候，SVD由于拟合能力不强从而导致下降较为缓慢且抖动很大。而当K变得原来越大到30的时候，容易造成模型的过拟合从而导致loss曲线不断抖动。  综上我们最终考虑取K=25，同时lr动态变换。  **4其他方法**  本节主要使用了scikit-surprise第三方库进行代码的编写。 基于基线评级的基本协同过滤算法 协同过滤算法(collaborative filtering，CF)：CF算法的汇总的是所有的<user,item>行为对，类似于朋友推荐，比如用户A和用户B都喜欢差不多的东西(item相似)，用户B喜欢某样东西，但是用户A还没有喜欢，那么此时就将用户B喜欢的item推荐给用户A。(User-Based CF)。还有一种协同推荐，即对比数据(item)，发现itemA和itemB类似(即被差不多的users喜欢)，就把某user的所有喜欢的item的类似item过滤出来作为候选推荐给该user。  基线评级：用户对商品的打分不仅取决于用户和商品间的某种关系，还取决于用户和商品独有的性质，将这些性质通过基线评分(baseline estimates)或者说是偏置(bias)来表示。  代码实现：    结果展示：   基于非负矩阵分解的协同过滤算法 非负矩阵分解：使分解后的所有分量均为非负值(要求纯加性的描述)，并且同时实现非线性的维数约减。  代码实现：    结果展示：   对比 根据上文结果，三种方法的RMSE如下表所示：   |  |  |  |  | | --- | --- | --- | --- | | Method | SVD | KNNBaseline协同过滤 | NMF协同过滤 | | RMSE | 0.8770 | 0.8587 | 0.9181 |   通过对RMSE等指标的对比，发现KNNBaselin(0.8587)效果优于SVD(0.8770)优于NMF(0.9181)算法。  注意：因为本实验主要针对电影评分值的RMSE进行性能评价，没有具体的离散类别评判，在询问老师后，确定本实验无需对以TopK推荐评价指标（如Precision@K、Recall@K、F1@K等）进行性能评价。 思考 在具体实验中，我们并没有严格采取一个固定的学习率。而是动态地选择了学习率。在最开始的1个epoch中，我们采取了学习率为0.001作为预热。而在后面的epoch中我们将学习率调整到0.01，并在最后五个epoch将学习率再次降到0.001从而使SVD趋向收敛。这样动态既使收敛更快，同时也不会由于学习率过大出现inf或者nan的情况。 | | | | | | |

备注：

1、教师在布置需撰写实验报告的实验前，应先将报告书上的“实验题目”、“实验性质”、“实验目的”、“实验项目内容”等项目填写完成，然后再下发给学生。

2、教师在布置需撰写报告的实验项目时，应告知学生提交实验报告的最后期限。

3、学生应按照要求正确地撰写实验报告：

* 1. 在实验报告上正确地填写“实验时间”、“实验地点”等栏目。
  2. 将实验所涉及的源程序文件内容（实验操作步骤或者算法）填写在“实验过程或算法（源程序）”栏目中。
  3. 将实验所涉及源程序调试过程（输入数据和输出结果）或者实验的分析内容填写在“实验结果及分析和（或）源程序调试过程”栏目中。
  4. 在实验报告页脚的“报告创建时间：”处插入完成实验报告时的日期和时间。
  5. 学生将每个实验完成后，按实验要求的文件名通过网络提交（上载）到指定的服务器所规定的共享文件夹中。每个实验一个电子文档，如果实验中有多个电子文档（如源程序或图形等），则用WinRAR压缩成一个压缩包文档提交，压缩包文件名同实验报告文件名（见下条）。
  6. 提交的实验报告电子文档命名为：“年级（两位数字不要“级”字）专业（缩写：计算机科学与技术专业（计科）、网络工程专业（网络）、信息安全专业（信息）、物联网工程（物联网））班级（两位数字）学号（八位数字）姓名实验序号（一位数字）．doc。如学号为20115676、年级为2011级、专业为“计算机科学与技术”专业、班级为“02班”、姓名为“王宇”的学生，完成的第一次实验命名为： 11计科02班20115676王宇1．Doc，以后几次实验的报告名称以此类推。

4、教师（或助教）在评价学生实验时，应根据其提交的其他实验相关资料（例如源程序文件等）对实验报告进行仔细评价。评价后应完成的项目有:

1. 在“成绩”栏中填写实验成绩。每个项目的实验成绩按照五级制（优、良、中、及格、不及格）方式评分，实验总成绩则通过计算每个项目得分的平均值获得（平均值计算时需将五级制转换为百分制优=95、良=85、中=75、及格=65、不及格=55）。
2. 在“教师评价”栏中用符号标注评价项目结果（用√表示正确，用×表示错误，用≈表示 半对半错）。
3. 在“教师评价”栏中“评价教师签名”填写评价教师（或助教）姓名。将评价后的实验报告转换为PDF格式文件归档。
4. 课程实验环节结束后，任课教师将自己教学班的实验报告文件夹进行清理。在提交文件夹中，文件总数为实验次数×教学班学生人数（如，教学班人数为90人，实验项目为5，其文件数为：90×5=450）。任课教师一定要认真清理，总数相符，否则学生该实验项目不能得分。最后将学生提交的实验报告刻光盘连同实验成绩一起放入试卷袋存档。