实验3 推荐系统和社会网络

Web信息处理与应用 2021 / 011179.01

实验背景

豆瓣 (www.douban.com) 是一个中国知名的社区网站，以书影音起家，用户可以在豆瓣上查看感兴趣的电影、书籍、音乐等内容，还可以关注自己感兴趣的豆友。

研究者们通过爬虫获取了一些用户和电影、书籍、音乐的交互数据，以及用户之间相互关注的社交数据。本次实验中，我们要为豆瓣社区的用户**推荐音乐**。

实验任务

根据豆瓣数据集，结合课程中**推荐系统 (第14节) 、社会网络 (第15节)** 的内容，查阅相关资料，合理设计模型，为每个用户提供**音乐方面的个性化推荐(Personalized Recommendation)**，生成Top-N列表。

数据集

本次实验的数据集是豆瓣的一部分数据，共计5个文件，包含交互与评分、社交、跨域等信息。基本信息如下：

DoubanMusic.txt

每一行是一位用户交互过的音乐数据，用`\t`分隔。第一个元素是用户ID，之后是该用户交互过的音乐列表，列表的每个元素表示一首音乐，其中音乐ID、该用户给该音乐的打分用`,`区分开。已作了至少以下预处理：

1. 为提高数据质量，保证每个用户至少交互过5首音乐，每首音乐至少被10个用户交互过，因此不必考虑冷启动 (cold start) 问题；
2. 有的用户交互过该物品，但是没有打分(或打分后又取消了)，此时评分被标记为`-1`。书籍和电影数据同理。若不选做基于评分的推荐 (Rating-based Recommendation) ，你可以直接忽略这部分数据。

如下示例表示，用户0为音乐9503打了5分，为音乐5605打了5分..... 用户1为音乐17590打了5分，为音乐11377打了5分.....用户2虽和音乐14871、音乐7687、音乐17528都进行过交互，但评分数据都缺失了.....

```
0   9503,5  5605,5  19381,5 19273,4 17351,4
1   17590,5 11377,5 92,5    21535,5 6936,5
2   14871,-1    7687,-1 7083,-1 17528,-1
3   15194,-1    3112,-1 1518,-1 4122,-1 3345,-1 5869,-1 6006,-1
4   7671,4  6494,4  16411,4 10957,4 612,-1
......
```

DoubanBook.txt & DoubanMovie.txt

数据集还包含了用户的书籍和电影交互数据，它们的语义和音乐数据完全相同，你需要注意：

1. 都只涵盖了音乐数据中出现过的用户；
2. 不再作数据过滤，值得一提的是：某些用户只交互过音乐，从未交互过书籍和电影；
3. 若不选做跨域推荐 (Cross-Domain Recommendation) ，你可以直接忽略这部分数据。

DoubanSocial.txt & DoubanSocialFull.txt

每一行是一个关注和被关注的单向关系，用`\t`分隔。你需要注意：

1. 如果关注者**和**被关注者用户都在音乐数据中，则该社交关系将包含在`DoubanSocial.txt`中；
2. 如果关注者**或**被关注者用户出现在音乐数据中，则该社交关系将包含在`DoubanSocialFull.txt`中。可见，`DoubanSocial.txt`是`DoubanSocialFull.txt`的子集；
3. 未出现在音乐数据中的用户都用`-1`表示;
4. 若不选做社交推荐 (Social Recommendation)，你可以直接忽略这部分数据。

如下示例可以看出：

1. `DoubanSocial.txt`的第1行表示用户12127关注了用户2069；
2. `DoubanSocialFull.txt`的第1行表示某个未在音乐数据中出现的用户关注了用户9560，第6行表示用户12127关注了某个未在音乐数据中出现的用户。

```
# DoubanSocial.txt
12127   2069
12127   11324
12127   4363
12127   1269
12127   19022
7670    1269
7670    15414
......
# DoubanSocialFull.txt
-1  9560
-1  20030
-1  4340
-1  3884
-1  11185
12127   -1
12127   -1
12127   -1
12127   -1
12127   2069
......
```

你可以在这里下载到本次实验的数据集：

> 链接：https://rec.ustc.edu.cn/share/ad301ae0-5a9c-11ec-9124-ab8f60474934
>
> 密码：exp3
>
> 有效期至：2022-02-28

模型评测方式和指标

[评测方式] 本次实验采用**留一法 (leave one out)** 进行评测。音乐数据已经被划分成了训练集和测试集。具体划分方式如下：设用户u交互过M个物品，则最后一个物品作为测试集 (用于在线评测系统)，其余$M-1$个物品作为训练集 (即`DoubanMusic.txt`)。

[评测指标]

在Top-N场景下，设整个数据集中共有N个物品$(N>M)$，对于用户u，你需要对训练集外的$N-M+1$个物品进行排序，生成推荐列表，使得测试集中的目标物品应当尽量靠前。

本次实验使用$HR@K$和$NDCG@K$作为指标，其中$K = 20, 100$。**因此，你只需要给出推荐列表的前100个物品(音乐)。**

Hit Ratio (HR) 是一个基于召回的指标，$HR@K$用来评估目标物品是否在包括在Top-K推荐列表中。

Normalized Discounted Cumulative Gain (NDCG) 是一个位置敏感的指标，目标物品在候选列表的前面，会被赋予一个更高的分数。

设你为用户u生成的推荐列表是$R_u = \{r_u^1, r_u^2, \dots, r_u^K\}$，其中$r_u^k$是用户$u$推荐列表第$k$个位置的物品，$T_u$是用户$u$在测试集中的目标物品，那么：

$$HR@K = \frac{1}{|U|} \sum_u I(|R_u \cap T_u|)$$

$$NDCG@K = \frac{1}{|U|} \sum_u \sum_{k=1}^{K} \frac{2^{I(|\{r_u^k\} \cap T_u|)} - 1}{log_2(k+1)}$$

其中U是用户集合，$I(x)$是一个指示函数，当$x > 0$为1，反之为0。

上述公式可能有些晦涩难懂，结合下面这个例子更容易理解。假设数据集中有3个用户、若干个物品，你为他们生成了Top-5推荐列表，于是：

用户	推荐列表	目标物品	$HR@K$(该用户)	$NDCG@K$(该用户)
0	$\{0, 1, 2, 3, 4\}$	1	1	$\frac{1}{log_2(2+1)} = 0.6309$
1	$\{0, 1, 2, 3, 5\}$	4	0	0
2	$\{5, 4, 3, 2, 1\}$	3	1	$\frac{1}{log_2(3+1)} = 0.5$

最终$HR@5 = \frac{1+0+1}{3} = 0.6667, NDCG@5 = \frac{0.6309+0+0.5}{3} = 0.3770$。

Baseline

TAs实现了一个较先进的、复杂度也较高的Baseline，对性能和时间进行综合评估，以保证实验可行。原则上，TAs不会提供代码。

Model	$HR@20$	$HR@100$	$NDCG@20$	$NDCG@100$
Baseline	0.2249	0.4183	0.1092	0.1442

在线评测系统

TAs为本学期课程实验搭建了一个简洁的在线评测系统，即https://mine.ustc.edu.cn/webinfolab3/homepage 。 你需要注意：

1. 该系统与实验2数据互不相通，使用学号登录，初始密码和学号一致，请尽早修改密码。
2. 在截止时间以前，你可以在系统上进行**最多10次提交**。只有提交符合要求的文件并返回评测结果(如上图)，才被视作一次有效提交，错误的文件格式等不会消耗提交次数。
3. 你不能看到他人的评测结果。

在线评测系统需要你提交一个文本文件 (必须是`.txt`为后缀) ，满足：

1. 每一行是一个用户的推荐列表，每一行是**推荐列表的用户ID**和**前100部音乐排序的ID列表**，用户ID和列表之间用`\t`隔开，列表各元素用英文逗号`,`隔开。
2. 文件行数和`DoubanMusic.txt`行数(即音乐数据中的用户数)相同，末尾不应该包含额外空行。

沿用上面的例子，你提交的文件内容应该如下：

```
0   0,1,2,3,4
1   0,1,2,3,5
2   5,4,3,2,1
```

提示和建议

- 磨刀不误砍柴工。TAs强烈建议在实验初期先对数据集进行充分的统计分析 (可参考知识库的数据统计)，结合数据规模、算法复杂度以及潜在的内存占用量，合理设计模型。
- 为了有效利用宝贵的提交次数，TAs建议大家在模型训练时，将数据集再划分训练/验证/测试集，如把每个用户的倒数第二、第一个物品分别作为验证集和测试集，然后在线下比较你所设计的模型。
- 一些模型可能训练时间较长，你可能需要预留足够的时间来训练模型。GPU可能对你加速这一过程有帮助，详见知识库中的GPU计算资源部分。

提交和评分

提交内容

你需要提交一个压缩包(`.zip`或`.rar`为后缀)。请以如下文件目录结构组织相关文件结构：

```
exp3/
|----src/
    |----model_A
    |----model_B
    |----...
    |----utils
|----submit/
    |----best_result_model_A.txt
    |----best_result_model_A.png
    |----best_result_model_B.txt
    |----best_result_model_B.png
    |----...
|----report.pdf
|----requirement.txt
|----readme.md
```

各目录/文件具体要求如下：

- `src`目录下放置你的源代码文件。其中至少有一个文件夹，每个文件夹包括你设计的一个模型的源代码，如`model_A`、`model_B`等。文件夹也可以自由命名，文件内容可以自行组织。若多个模型之间存在公共依赖，可以存放在`utils`文件夹中，但请在`readme.md`或`report.pdf`中说明。
- `submit`目录放置你的每个模型的最优结果所对应提交的`.txt`文件和在线评测系统的截图，如`best_result_model_A.txt`和`best_result_model_A.png`。以在线评测系统返回的指标作为比较依据，而不是线下结果。
- `report.pdf`文件是你提交的实验报告，具体见下面关于实验报告的详细说明。
- `requirements.txt`文件仅当你使用Python时提供，用于记录所有依赖包及其精确的版本号，方便TAs在复现结果的时候重新部署环境。你可以在Terminal通过`pip freeze >requirements.txt`命令生成。若你使用了Anaconda，可以通过`conda list --export > requirements.txt`命令生成。
- `readme.md`文件应该包含你的源代码的运行环境、编译运行方式，以及对关键函数的说明。和实验有关的其它内容，也可以在这里说明。你可以参考https://www.freecodecamp.org/news/how-

to-write-a-good-readme-file/ 。

特别注意：

- 你可以不提供你的模型文件，但应保证TAs可以完全复现你的实验结果。若你认为有必要，也可以将模型文件上传到睿客网、Google Drive等网盘中，然后在`readme.md`或`report.pdf`中说明，但请**务必不要**放到压缩包中。
- 若你要使用GitHub等代码寄存服务平台，在实验截止时间以前请将你的仓库设置为**private repository**。当然，TAs非常鼓励你在之后发挥开源精神，将它转为public repository。

实验报告

实验报告至少有以下要求：

- 你需要说明你对数据集的统计分析、预处理方式。
- 你需要尽量详细地阐述你的模型细节，绘制模型结构图、合理贴代码等可能会达到事半功倍的效果。
- 模型的设计往往是迭代的，你可以由模型简单到复杂逐步阐述你对模型迭代的动机，最好能附上你的消融实验的结果，以证明你的模型迭代是有效的。(由于在线评测次数有限，这里显然可以是线下结果)
- 你在实验报告中需要附上所设计模型在在线评测系统上的最终结果，以截图方式呈现。
- 如果你参考了他人代码 (特别是基于知识库开源库或其它Github仓库再进行二次开发) 、技术文档、论文等，请务必在实验报告中**规范引用**。
- 你最好在实验报告中也注明组员的学号和姓名。

评分依据

在评分时，TAs会**依次**关注：

1. 个性化推荐模型的合理性、**新颖性**，实验设计、结果和分析的完整性和**说服力**，以及对应的**工作量**。
2. 实验报告的**完整性**和**可读性**、在线评测系统的指标。
3. 利用社交数据 (如有) 的合理性、有效性，对结果的提升。
4. 其它选做任务 (如有) 的完成情况。

特别注意：

- 同组的同学一般会得到相同分数。
- 没有规范引用他人代码、技术文档、论文会被扣分。
- 请务必独立完成，如果发现抄袭按零分处理。

提交程序

请于**2022年1月22日23:59:59 (UTC+8) 以前**提交到课程邮箱ustcweb2021@163.com。具体如下：

- 建议**两人一组**。单人也可，但没有优惠政策。
- 邮件标题以及压缩包命名为"学号1-姓名1-学号2-姓名2-实验3"格式，如"PB19111888-法外狂徒张三-PB19010999-懂法狂魔李四-实验3"。单人请按照"学号-姓名-实验3"格式。压缩包以附件形式上传。因未署名造成统计遗漏责任自行承担，你可以将邮件抄送你的队友。
- 若需要提交迭代版本(尽管并不鼓励)，请在上述格式基础上再增加"-第X次迭代"，如"PB19111888-张三-PB19010999-李四-实验3-第1次迭代"，并最好在邮件正文中说明情况。
- 迟交作业将不被接收。

FAQ

Q: 如果对实验有疑问、建议，怎样才能得到有效的反馈?
A: 本次实验主要由"本课程没有助教"和"上边那位也是助教"负责，你可以直接提问 (如通过QQ群) 。由于你们的问题往往具有共性，TAs鼓励大家在课程QQ群中直接提问，TAs也将及时回答(请合理@TAs)。如有未尽事宜，将通过**课程QQ群 (主要)** 、课程网站、在线评测系统等途径对本说明进行进一步更新。如果对课程有任何建议，可以通过QQ、QQ群、课程邮箱及时向TAs反馈，TAs会及时完善实验。

Q: 知识库一定要用吗?
A: 不一定。它们只是TAs出于善意的一些分享，是否使用和最终的评分无直接关系，在遇到困难时你可以将它当作**字典**使用。

Q: 算力不够怎么办?
A: TAs在知识库中提供了一些免费算力。需要说明的是，我们**不鼓励**大家使用收费GPU资源。TAs认为，即使你计划完成所有的可选任务，只要模型合理，免费算力资源并**不会**限制到你的发挥。

Q: 听说往届的学长学姐通过"奇技淫巧"找到了测试集，实在是太秀了，我是不是也可以尝试一下?
A: TAs也通过各种"奇技淫巧"对数据集进行了各种神秘莫测的处理，你大可以尝试一番。(手动微笑)

Q: 基于深度学习的模型效果上往往会超过传统模型，我一定要使用基于深度学习的模型吗?
A: **可以但不必要!** 模型的评测指标也只在最终评分中占一小部分比例。相较而言更希望大家能设计出更合理、有创新的模型，写出有说服力的实验报告，这与是否使用基于深度学习的模型没有必然关系。

Q: 社交、评分、跨域数据都要使用吗?
A: 不是。只有个性化推荐是必须完成的。社交推荐是可以尝试去探索的，其它可选任务是供学有余力且对推荐系统确实感兴趣的同学"消遣"用的。**拒绝内卷，从你我做起!**

Q: 用户交互序列是否具有顺序性？可否使用序列推荐 (Sequential Recommendation) 的模型?
A: 具有**日粒度**的时序性。但由于TAs对数据进行了过滤处理，时序特性可能会受到影响，**建议谨慎使用序列推荐模型。**

(Last updated on Jan 1, 2022 by TA Team)