课程大作业：推荐系统

大作业概述

本大作业要求大家复现基于深度学习的推荐系统。希望通过本次大作业，锻炼大家阅读论文、对论文进行文献综述和算法总结、复现相关代码的能力。也希望通过本次大作业，让大家熟悉推荐系统的基于深度学习方法的实现，了解深度学习环境的搭建、模型的构建、训练、测试等过程，从而为今后的学习科研等奠定一定的基础。

任务介绍

（1） 阅读所附的 3 篇推荐系统相关的论文，进行文献综述和算法总结。

（2） 根据“xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems”论文（下文称之为：要复现的论文），进行实验复现或部分复现；

（3） 结合 PyTorch 中有关 xDeepFM 的教程，复现相关工作及结果。参考链接：https://github.com/shenweichen/DeepCTR-Torch

具体要求

请仔细阅读下述大作业的具体要求，并遵照要求完成大作业。

（1）阅读所提供相关论文，进行文献综述

【基本要求】15 分

重点阅读“要复现的论文”，关注其中引言和相关背景的介绍部分，并结合大作业所提供的其他论文对相关领域（Recommendation）的主流方法进行文献综述。

在进行文献综述时，**可以以时间为序**，根据相关领域的发展脉络，介绍各个时期的典型文献以及他们所提的主要方法的概述（基本原理、优缺点等）；**也可以以类别为纲**，结合相关领域的不同典型方法，分别介绍相关方法的基本原理、代表性文献、方法的优缺点等。

【加分项】≤5 分

结合相关领域的发展趋势，对“要复现的论文”之后发表的论文（比如 2017 年以后）的典型方法进行文献综述。

（2）根据所阅读的相关论文进行算法总结

【基本要求】15 分

（1）对相关任务（Recommendation）所要解决的问题给出形式化的定义。比如通过形式化的数学方法给出相关任务的输入、输出的定义，相关任务的概率函数表达；假设使用神经网络解决该问题，神经网络建模函数的意义、神经网络的训练目标（损失函数）的定义等。

（2）对“要复现的论文”中所给出的算法或模型进行归纳总结，分别介绍相应的算法或模型的主要模块，特别是要给出相应模块的功能、原理、方法等。

【加分项】≤5 分

（1）结合相关领域的发展趋势，对“要复现的论文”之后发表的论文的典型算法或模型进行总结，并介绍相应的算法或模型的主要模块。

（2）算法总结能体现相关领域的发展脉络和趋势：指出相关算法提出时拟解决的问题（也即提出的动机 Motivation），能指出相关算法的局限或者不足（也即能指出问题 Problem），能指出后续算法对前续算法的改进等。

（3）熟悉有关编程平台与深度学习框架，根据指定论文复现对应方法

【基本要求】30 分

（1）根据所选择的深度学习框架和平台，对“要复现的论文”中的方法进行复现。

（2）可以使用网上已有的开源代码，也可以自己对相关代码进行整合改进。

（3）无论是开源代码、还是整合改进，均必须对所写（所使用）的代码的关键功能模块使用中文进行“详细”的注释，解释该模块的功能、基本方法等。请关注数据预处理的流程、创建特征列、模型源码。

（4）在实验报告的“复现流程”一节，详细介绍实验过程，包括所用的数据、数据的预处理、模型的实现、模型的参数说明、模型的训练过程（包括训练所用的参数）、模型的测试过程（包括测试所用的参数）等。同时要给出各流程对应的代码文件名及函数名。

（4）给出自己复现方法所得的实验结果及实验分析

【基本要求】20 分

（1）明确相关任务（Recommendation）的评价指标（Evaluation Metrics），在实验报告中给出该评价指标，并对其进行解释说明。

（2）基于“要复现的论文”中的数据集，对上述复现的算法或模型进行实验，给出复现实验结果。

（3）将复现实验结果与“要复现的论文”中给出的结果进行比较，说明是否存在差异，并给出实验结果的分析；假设复现结果与论文中的结果存在差异，对可能的原因进行分析。

【加分项】≤10 分

（1）给出模型训练的可视化结果并对其进行分析。

（2）结合“要复现的论文”之后提出的新算法或模型，对其加以复现，并在实验报告“复现流程”一节中整合本部分复现过程，给出对应的实验结果和相应的分析。

（5）实验报告与海报展示

5.1 撰写实验报告

【实验报告】10 分

（评价实验报告撰写是否规范、内容是否全面丰富、逻辑是否清晰、重点是否突出）

（1） 实验报告可以以中文撰写、也可以以英文撰写。要求重点突出、逻辑清晰。

（2） 实验报告的格式参考正式的 paper，建议包括：

报告题目：推荐系统

个人信息：包括小组成员的姓名及学号；具体专业方向（不能只是电子信息）；电子邮箱

中文摘要及关键词

英文摘要及关键词

引言

1. 文献综述（这里 1 为建议编号，下同）（可细分为子章节，下同）

2. 算法总结

3. 评价指标

4. 数据集

5. 复现流程

6. 实验结果及分析

7. 结论

8. 所完成的加分项（以表格方式给出所完成的加分项，并给出实验报告中的对应子章节索引）

9. 成员分工及贡献比（以表格方式给出，可以按照具体要求中的项目划分，也可更加细分）

10. 心得体会

参考文献

附录：给出包括实验报告在内的大作业相关文件清单及相应说明（即上传到网络学堂的文件内容；代码可放于一个目录，并对该目录作说明）

（3） 实验报告的表格、图片等要给出相应的表题、图题，并顺序编号，并在正文中相应地方给出引用；参考文献应在正文中给出相应的引用。

5.2 准备海报展示

【海报展示】10 分

（老师/助教/同学互评的加权成绩：包括海报的美观度、工作亮点总结、汇报展示的效果等）

（1） 请每个小组准备一张海报，应当包括报告题目、小组成员姓名、学号、具体专业方向（不能只是电子信息）、电子邮箱等；

（2） 海报内容：除了基本算法/模型的介绍之外，应突出自己工作的亮点部分：可以是模型的亮点、实验结果的亮点、除了基本要求之外完成的加分项的亮点、甚至是实验报告撰写的亮点、实验结果呈现形式的亮点、心得体会的亮点等等，总之能够凸显自己工作特色的所有东西都可以作为亮点给出来。

（3） 完成大作业后，会花一次课程的时间，让大家在课堂上展示和介绍自己的海报（需打印海报）。

（4） 海报电子版需使用 pptx 格式准备，设置为 A0 大小。

（5） 海报电子版需在规定时间（具体时间请等待通知）之前上传到网络学堂。

（5）在截止日期前上传实验结果

将实验报告、海报电子版 pptx 文件、所复现代码以及相关说明文件（如代码运行环境需求说明、代码运行方法说明等），打包成一个 zip 文件上传到网络学堂。

请在截止日期前上传实验结果。否则将按以下公式扣分：

$$S' = S \times \min\left(0.85, 0.95^{D}\right)$$

其中，S'是迟交作业的评分，S是作业的原始得分，D是向上取整的迟交天数（超过 deadline 后即记为迟交一天）。例如：作业的 deadline 是 10 月 11 日，10 月 12 日补交的作业评分为原始作业得分的 85%，10 月 18 日补交的作业评分将被折合为原始作业得分的 69.8%。

有关资源

本次作业所提供的 3 篇论文和数据，提供清华云盘下载地址：（访问密码：bigdatathu）

https://cloud.tsinghua.edu.cn/d/72a3745d59bc4c2d84ae/

本次作业所用的数据集为 Criteo 广告数据集，下载后的完整数据集（4.5GB）在上述清华云盘链接的“Criteo”子目录下。

有关 Criteo 数据集

Criteo Dataset

潜心 edited this page on 1 Sep 2020 · 1 revision

Criteo Dataset

Criteo广告数据集是一个经典的用来预测广告点击率的数据集。2014年，由全球知名广告公司Criteo赞助举办Display Advertising Challenge比赛。但比赛过去太久，Kaggle已不提供数据集。现有三种方式获得数据集或其样本：

1. Criteo_sample.txt：包含在DeepCTR中，用于测试模型是否正确，不过数据量太少；
2. kaggle Criteo：训练集（10.38G）、测试集（1.35G）；(实验大部分都是使用该数据集)
3. Criteo 1TB：可以根据需要下载完整的日志数据集；

数据结构

数据集的特征如下所示：

- Label：标签，表示目标广告点击（1）或未点击（0）；
- I1-I13：13个数值特征，也称为计数特征；
- C1-C26：26个分类特征（稀疏特征），为了匿名化的目的，这些特性的值被散列到32位上；

	label	I1	I2	I3	...	C23	C24	C25	C26
0	0	1.0	1	5.0	...	3a171ecb	c5c50484	e8b83407	9727dd16
1	0	2.0	0	44.0	...	3a171ecb	43f13e8b	e8b83407	731c3655
2	0	2.0	0	1.0	...	3a171ecb	3b183c5c	NaN	NaN
3	0	NaN	893	NaN	...	3a171ecb	9117a34a	NaN	NaN
4	0	3.0	-1	NaN	...	32c7478e	b34f3128	NaN	NaN

该数据集已经为大家下载到上述清华云盘链接的“Criteo”子目录下，其原始下载链接为：

https://labs.criteo.com/2014/02/download-kaggle-display-advertising-challenge-dataset/

其他大家关心的问题

Q1：我的计算机计算能力有限，所提供的 Criteo 数据集太大，模型跑不动，怎么办？

A1：如果发现处理完整的 Criteo 数据集有困难，可以选取部分数据来进行实验，但不得少于全量数据的 5%，并在作业报告中说明数据选取的规则。另外，要注意，选取部分数据后，实验结果和原论文的结果性能会有差别。可在实验报告中进行分析（可参考 Q2）。

Q2：需要复现到和原论文一样的程度么？

A2：不需要。如果没有达到原论文的结果性能，也没有问题；但是最终成绩可能会参考这部分的情况，特别是大家一定要给出可能的原因分析。本大作业重点关注大家通过本次大作业能学到什么，而不是简单的一个实验结果。

Q3：可以使用 Tensorflow 的实现么？

A3：可以使用 Tensorflow 的实现。网上也有相应的资源，一个参考资源链接如下：

https://github.com/microsoft/recommenders/blob/master/examples/00_quick_start/xdeepfm_criteo.ipynb

相应的数据集在上述清华云盘下载连接的“Tensorflow”目录下。

Q4：代码会查重么？实验报告会查重么？

A4：代码不会查重，但是实验报告会查重。大家可以参考和下载已有的开源代码来完成大作业，但是一定要在实验报告中、说明文档中注明代码的来源。另外，必须要好好学习所下载的代码，要按照上述要求把整个实验过程和原理在实验报告中写清楚，不能只是跑了一个代码得到一个结果而已。

Q5：听说文献综述会影响实验报告的查重率？

A5：文献综述不是把别的论文中的内容简单拷贝粘贴过来，而是需要用自己的语言来对所阅读的论文进行总结，从论文所解决的问题、所提的方法、方法的优缺点等角度进行总结，具体要求见前述说明。通过自己的语言来进行文献综述，对实验报告查重率的影响应该可以控制得很小。

Q6：自己的电脑跑不起来论文中给定的数据集，课程会提供 GPU 资源么？或者换成小的数据集？

A6：由于学校没有为本课程配备服务器和 GPU 等计算资源，所以得同学们自己考虑去找计算资源。如果自己实验室没有 GPU 资源，大家可以考虑使用免费的 GPU 环境：google colaboratory（不过需要同学们自己去探索）；也有一些其他平台提供免费 GPU，大家可以去搜一下；另外，也可以问问身边的同学看能否帮忙。可以换小数据集（具体见 Q1）。

Q7：我遇到了问题怎么办？

A7：请放松随意地在课程群里提问，会有助教进行回答。

Q8：我能做完全部的加分项吗？

A8：非常鼓励有兴趣的同学自行尝试加分项的内容，但加分项最多只能累计 20 分。