技术向：推荐学习推荐系统（深度思考，不是广告）

原创 机智的叉烧 CS的陋室 2019-02-05

Welcome to Planet Urf

英雄联盟 - Welcome to Planet Urf

往期回顾：

- 【NLP.TM】GloVe模型及其Python实现
- 【陋室推荐】| 2018-5-8
- Tensorflow入坑指南
- MLincubator：一种面向实验的机器学习框架
- 做算法，你要掌握这些

机器学习我大概是16年末17年初接触的，深入学习后转到了自然语言处理，借助自然语言处理自己基本对数据科学、深度学习一块的内容以及整个版图的分布情况有了比较全面的了解，本着毕业之前在科研任务完成的基础上先看深度，就业之后看广度的想法，遇到一定的进步瓶颈期，我发现了推荐系统，并且会做了一些分析和思考，决定后续在这块进行一定的探索和思考，希望和大家一起讨论讨论。

我先把结论抛出来，推荐系统是了解这个领域的一个非常优秀的窗口，也推荐在机器学习学完之后遇到瓶颈期的同学去了解并且学习。

模型角度的思考

从模型角度，推荐系统几乎涵盖了所有机器学习能够解决的问题的结构，并且尝试了各种特征工程方法，构建了很多重要的模型结构，而且这些模型十分完整，具有很强的借鉴意义。

从上游，推荐系统常涵盖大量特征，根据今日头条的特征工程来看，主要分为相关性特征（内容、用户及其匹配信息）、环境特征（时间和地点）、热度特征（热点信息等）和协调特征，从抽象层面，

有一些现实的、具体的特征，也有经过特殊化处理和转化的embedding特征，如果经过系统学习，就会对特征工程有更加深入、完整的理解。

在下游，到了模型层面，无论是低维线性的LR、非线性的RF、GBDT，还是更高级的DNN，甚至到目前被广泛使用的wide&deep，知识结构完整和详细，同时这些模型也经受住了实践的检验，例如而很多类似朴素贝叶斯、SVM等可能在学术界很火的模型，其实在现实中可能因为很多原因而难以被铺开使用，其实经历了一些知识筛选，这里捞一手，只是没有广泛使用，但是不代表不行哈，有的时候试一试说不定有意外效果。

应用角度的思考

从商业角度，推荐系统在算法层面具有很强的功能性。

在目前互联网作为主流的商业环境下，推荐系统或者其抽象模式其实起到了很大作用。互联网主要起到的是平台作用，以互联网为平台，构建起多方沟通的桥梁，例如淘宝对应卖家和卖家，美团对应商家和用户，去哪儿是酒店和旅客，头条是信息产出方和读者，而作为平台，除了要满足用户本身的需求（可以认为是食物链中的消费者），还要考虑到商家、酒店（可以认为是食物链的生产者）的利益，这样平台脆嫩巩固有流量，才能有进一步的转化，达到盈利，虽然人口盈利目前已经被广泛使用，尽管甚至有已经接近消耗殆尽的言论，但是在算法层面，人力层面，非常需要拥有推荐系统相关知识的人才，这是算法工程师、数据科学家等职业非常好的去处。

浅谈embedding

本想当做特殊的一篇来写，但是感觉内容不够，再在这里谈一下。我想说的是，embedding这块在未来会具有很强的应用价值和研究价值。

所谓embedding，中文翻译为嵌入，可能会比较抽象，最早我是在NLP里面听说的，w2v，GloVe等，但是到了推荐系统领域就有了很广的应用。这是一个什么东西呢，简单的可以理解为把一些不好作为计算的东西根据某些规则转化为可供计算的向量或者矩阵，在NLP领域，最多的就是词嵌入，而在推荐系统，当然就可以对用户或者是物品做类似的操作，另外还有地点（hex或者geohash）也可以，甚至是一些组合特征，其实都可以这么去做，这种方式的好处就是在特征工程层面，可以挖掘更多的信息，这些信息对模型的性能绝对是有好处的，而在目前，个人感觉还有很大的空间，因此大家可以密切关注，这是在除了强化、迁移甚至弱监督、半监督问题等大问题还有待深入研究的基础上的一个重要小问题，尤其是在应用层面，这个问题可能在很多研究问题中都不会作为重点，但是在应用，尤其是工业界的应用中，是一个不可避免的问题。

再者，这个东西容易形成技术壁垒，因为业务类型和数据本身的原因，技术壁垒建立的快，容易形成新的竞争力，后来的只能去了解一些相似而不同的线路绕路突破，在这，流量是一块大小几乎是固定的蛋糕，谁的多了，别人就少了。

推荐系统的学习建议

自学有段时间了，入门是从项亮的《推荐系统实践》开始，然后是七月在线的《推荐系统实战》的课，后面开始就是推荐系统的论文和案例分析了，这个领域目前的状态有这么几个特点：

- 知识结构尚未完善，边界不清楚，想要学习也比较困难
- 有知识，也不好实践，开源代码主要都是小算法向，数据集有限，需要自己总结和比较大动手的量
- 正因为还算是刀耕火种的时代，基线都不好快速实现的时代，所以快速储备达到60分其实就很有竞争力，当然，由于是刀耕火种的时代，所以达到60分的难度也会高很多
- 刚开始有一些比较好的书籍，github上也有一定的资源，不能说丰富，但是感觉足够吧
- 企业需求不少，无论是创业公司还是一些已经站稳脚跟的公司

学习建议

因此，针对这些现状我对学习这个领域有如下建议吧：

- 基础知识要求过硬，机器学习领域的东西最好提前弄懂弄扎实再开始
- 最好有比较好的自学能力，自己能在网络上、github上查资料，不是歧视，如果是还停留在问别人一些百度就能找到问题的答案的同样学建议还是提前学习如何查资料、查文献会比较好，因为很多资料都需要自己去搜集
- 多动手，多看论文，多看case
- 这个领域我觉得可以分为这两块，基础模型和case，在这个领域里面，case的学习非常重要，因为每一个case的背景都会有一些区别，积累case可以为你的思维提供很多经验。

推荐基本入门的书吧，还有推荐欢迎留言！

- 项亮《推荐系统实战》，入门基础，但是整本书基本都在将协同过滤及其变式，有点局限，协同过滤是基础但是现在不仅仅有协同过滤哈
- 牛温雅《用户网络行为画像》，估计是本冷门书吧，讲了很多和推荐系统有关但容易忽略的现实问题，在用户画像构建的思考上感觉写的很详细
- 黄昕《推荐系统与深度学习》，腾讯大佬出品，很实用的一本书，算是新书吧，最近在看感觉写得很不错，还挺全面，还有代码。
- 闫泽华《内容算法》，不算新书，内容很通俗，基本是概述，比较适合产品去读吧，但是算法开发的也最好看看，推荐系统的实际应用场景很复杂，很多问题要考虑，并经这套东西很可能就是一个公司的生命了，最简单的例子就是头条了，头条的推荐绝对很厉害。

- 剩下就是看case啦，推荐看看这几个公司的案例，另外论文也推荐的：阿里（尤其是淘宝）、头条、YouTube、Aribnb、Google、美团、滴滴，滴滴之所以放在里面，主要原因就是他的调度系统其实也是推荐系统，嗯，大家仔细想想哈哈。

后续内容

后续我会出一些有关这方面的内容啦，整理整理自己学的东西，有一些好玩的案例和论文也和大家探讨探讨。

我是叉烧，欢迎关注我！

叉烧，北京科技大学数理学院统计学研二硕士（保研），本科北京科技大学信息与计算科学、金融工程双学位毕业，硕士期间发表论文4篇，学生一作2篇，1项国家自然科学基金面上项目学生第2参与人，参与国家级及以上学术会议4次，其中，1次优秀论文。曾任去哪儿网大住宿事业部产品数据，美团点评出行事业部算法工程师。

微信 zgr950123

邮箱 chahsaozgr@163.com

微信个人公众号
CS的陋室