算法在岗一年多的我，一些心得与收获

原创 对白 对白的算法屋 2022-05-03 18:36

收录于合集
#经验分享 50 #我的故事 15 #程序人生 9

大家好，我是对白。

上个月的工作一直都很忙，每天还有需求评审以及技术评审的会要开，我本人需要负责的业务也排期到了8月底，导致一直没有时间静下心来去梳理一些事情。

刚好最近放假了，所以想和大家聊一聊我的一些感触和想法。

在互联网公司里，**算法工程师主要分为两种：业务型算法和研究型算法**。研究型算法主要负责发论文、发专利和技术支持；业务型算法主要负责利用算法解决公司业务问题，提高业务收益。前几年像阿里达摩院、字节和腾讯AILab还有一些研究型算法的岗位，但由于给公司带来不了实际收益，因此很多这类最终也被合并到了业务部门中。

对于绝大多数算法工程师来说，可能都会担心自己的效果达不到业务指标。最近我也是，有好几个ddl需要上线，因此压力很大，经常写代码到晚上12点。不得不说，相比于开发岗，做算法的人真的很苦命。可能你花了两周的时间调整数据或优化模型，也不一定有效果。那这个问题如何才能规避呢？

在算法岗工作了一年多的我，踩过了无数这样的坑，也有了一些自己的心得，今天就毫无保留的分享给大家。

1、单机版验证效果，再转成分布式版。对于Pandas和Spark哪个更熟悉，我相信大多数人都会选择Pandas。一般离线或在线的模型，都是在Spark分布式集群上运行的，因此整个Pipeline都需要转成Spark。对于我来说，我一开始都会写单机版的模型，等单机版效果没有问题后，才会转成分布式版，这样总比一上来就写Spark代码，能节省不少时间。

2、所有中间输出过程都要落Hive表。以推荐架构举例，都要先经过数据预处理、再到召回、粗排、精排，最后是重排。这是一个很长的Pipeline，如果要在Spark集群上运行，很难排查出到底哪一个环节导致效果不好。因此我的习惯是，对于每一个中间过程，我都会将它创建成一个临时视图，然后将结果落到hive表里，这样如果是召回问题，我就优化召回，如果是排序问题，我就改排序。改完之后，注释掉之前已经输出中间结果的代码，直接读取hive表，就可以节省大量的代码运行时间。

3、模型效果不好，不一定是模型的问题。很多时候，当我们发现模型的效果不太好时，第一个想法就是更换模型，直接上最NB的。但事实是，你更换模型往往效果也不一定好，还费时间。为什么这样说，因为两个模型效果之间的差距可能只有2-3个百分点，以有监督和自监督来举例，当业务数据缺少时，我们往往会想通过无监督或自监督的方式去解决，但实际上，以我的经验判断，公司中的业务问题往往很复杂，学术界的那套自监督，很难取得什么效果。所以本质问题还是要去解决数据，数据稀疏那就想办法构造，数据无标注就自己标，这比自己调研了一圈学术界模型更加的靠谱和管用。

此外，还有一些做业务算法的小Tips，我准备后面写一篇完整的文章贡献给大家，以上三点大家可以先思考一下，如果觉得我说的对，记得给我点个赞哦~

您的“**点赞/在看/分享**”是我坚持的最大动力！

坚持不易，卖萌打滚求鼓励（ฅ>ω<*ฅ）

分享 收藏 点赞 在看

收录于合集 #经验分享 50

下一篇 · 在英特尔做了一年 AI 研发，真的很香！

喜欢此内容的人还喜欢

零基础教你学会视频剪辑，揭开抖音挣钱的秘密

AIRPHOTO