CTR预估模型中的正负样本问题

原创 Thinkgamer 搜索与推荐Wiki 2020-06-17

收录于话题
#推荐相关笔记

32个

> 目前推荐系统中给用户进行推荐大部分都是基于CTR预估来做的，CTR预估中很重要的一环便是正负样本的选择，那么不同业务场景下我们如何定义正负样本、如何控制正负样本的比例、正负样本选择有哪些技巧？虽然这些只是模型训练中的一环，但却也扮演着重要的角色。

这篇文章简单聊一下上边提到的问题，如何你对这有什么想法和意见，欢迎在评论区留言，一起沟通。

分析业务场景

不同业务场景下对应的kpi也是不同的，那么模型训练的目标也是不一致的，比如kpi是点击率，那么模型训练的目的就是增加推荐的准确性，提升用户的准确率；如果kpi是交易额，那么模型训练的目的就要考虑用户的下单率和物品单价，不能仅仅考虑点击。

但是有些kpi指标并不太容易直接量化成模型训练的目标，比如内容平台的用户停留时长、用户阅读率、用户活跃度度等，但可以通过对业务指标进行分析，简单量化成模型训练的目标。之前做过一段时间用户消息push，个性化消息的推送目标是吸引用户进入app，从而增加日活、月活，这个时候就是考虑的点击，因为用户一旦点击了，便可以进入平台，转变成活跃用户。

所以在做推荐的时候，要搞清楚出我们的业务目标是什么，从而制定合适的模型目标，而不是盲目追寻大众，看大家都在做ctr，都是用点击率进行模型训练的，就盲目的使用点击率。

正负样本的定义

一般使用skip-above思想，即用户点过的Item之上，没有点过的Item作为负例 （假设用户是从上往下浏览Item，且会把正负样本选择限定在同一个业务场景中）。

对于视频或者音频节目而言，分成几个种类：

- 喜欢的节目：用户当天播放过的节目
- 历史的节目：用户在过去的一段时间内播放过所有节目
- 曝光的节目：一段时间内对用户曝光的节目。

由此，正样本可以定义为用户当天播放过的节目，也就是“喜欢”。负样本则有两种选择方案：

- （1）负样本指的是对用户曝光过的节目，但是用户至始至终都没有播放过，也就是说该节目并不在“历史”和“喜欢”两个分类里面
- （2）负样本指的是在整个抽样的池子里面，但是用户至始至终都没有播放过，也就是说该节目并不在“历史”和“喜欢”这两个分类里面

但是一般情况下，我们会选择给用户曝光但是用户没有进行播放的节目作为负样本。

对于电商平台而言，如果kpi的指标是GMV，那么我们的正负样本的选择则会很大的不同。点击率预估时正样本可以是用户点击的商品，负样本可以是给用户曝光但是用户没有点击的商品。目标是GMV时，我们要在保证用户点击的情况下促进用户进行下单，继而增加GMV，这时候正样本如果还是点击的话就没有什么意义了，简单讲，正样本可以是用户点击且下单的商品，负样本可以是用户点击但没有下单的商品。但这时候又会引出另外一个问题：样本稀缺。

阿里之前有一篇论文讲述就是上边提到的问题的解决算法，感兴趣的可以阅读：**【论文】Entire Space Multi-Task Model**

正负样本的选择技巧

- 对于正样本的选择不能简单认为用户产生行为就是正样本，比如用户误点击，当然对于视频类平台，如果用户被视频标题和图片吸引，产生观看行为但是观看之后发现并不是想看的，这时候就会退出，如果单纯的认为观看就是正样本，也会引起模型训练的误差，此时有效的做法可以是设定一定的阀值或者一定的观看比例才能够反映用户是否喜欢该节目。比如YouTube的视频节目，不止有“订阅”，“添加到”，“分享”，还有能够反映用户喜好的“like”（顶一下），“dislike”（踩一下）。有的时候顶一下可能不足以反映用户是否喜欢，但是踩一下基本上可以确定该用户不喜欢这个视频节目。除了“like”和“dislike”，对于其余的一些APP或者视频网站，还会有其余的操作，比方说评论，分享，收藏，下载等操作。这些操作从某些层面上也会看出用户是否喜欢该节目。

- 无论是什么平台，用户的活跃度分布都是一个长尾分布，越活跃的用户对应的人数越少，但是其所占的行为越多。这种情况下，如果不考虑用户活跃度去筛选正负样本，难免活跃用户所占的权重就会增大，此时有效的解决办法是针对每个用户提取相同的正负样本。

- 《美团机器学习实战》一书提到，它们在 feed 场景中采用了Skip Above的方式来提高效果。具体来讲就是根据用户最后一次点击行为的位置，过滤掉最后一次点击之后的展示，可以认为用户没有看到，也可以保留最后一次点击之后的少数几个。笔者认为也可以进行尝试。

- 针对同一个内容在不同时间对同一个用户曝光多次的情况，这时候训练集中可能会出现同一用户对同一内容点击与不点击并存的情况，如果多次曝光的间隔非常短，考虑只使用其中的一次曝光数据。

- 平台内会存在一些恶意行为的用户，此时可以进行相应的识别，模型训练时去掉这些恶意用户的行为数据。

- 补充一点划分训练集和测试集的两种方法：

  - 随机拆分（比如整个样本集拆分为训练集和测试集）
  - 按时间维度拆分（比如样本集 按天分区，前5天的数据作为训练集，接下来2天的数据作为测试集）

正负样本的比例选择

正负样本不均衡问题一直伴随着算法模型存在，样本不均衡会导致：对比例大的样本造成过拟合，也就是说预测偏向样本数较多的分类。这样就会大大降低模型的泛化能力。往往accuracy（准确率）很高，但auc很低。

正负样本不均衡问题的解决办法有三类：

- 采样处理——过采样，欠采样
- 类别权重——通过正负样本的惩罚权重解决样本不均衡的问题。在算法实现过程中，对于分类中不同样本数量的类别分别赋予不同的权重
- 集成方法——使用所有分类中的小样本量，同时从分类中的大样本量中随机抽取数据来与小样本量合并构成训练集，这样反复多次会得到很多训练集，从而训练出多个模型。例如，在数据集中的正、负样本分别为100和10000条，比例为1：100，此时可以将负样本随机切分为100份，每份100条数据，然后每次形成训练集时使用所有的正样本（100条）和随机抽取的负样本（100条）形成新的训练数据集。如此反复可以得到100个模型。然后继续集成表决

当然常用的是上述集成方法，但是不是直接进行使用，试想训练100个模型进行表决，离线会很麻烦，而且线上使用也不现实。

所以通常会进行修改使用，一般情况下在选择正负样本时会进行相关比例的控制，假设正样本的条数是N，则负样本的条数会控制在2N或者3N，即遵循1:2或者1:3的关系，当然具体的业务场景下要进行不同的尝试和离线评估指标的对比。

当然上述的前提是正负样本的选择。好了至此，完事，如果你看到了这里，麻烦点个【在看】，可以的话进行分享，让更多的人看到，毕竟知识是没有界限的。