ESMM

本文介绍 阿里妈妈团队 发表在 SIGIR'2018 的论文《Entire Space Multi-Task Model: An Effective Approach for Estimating Post-Click Conversion Rate》。文章基于 Multi-Task Learning 的思路，提出一种新的【CVR预估】模型——ESMM，有效解决了真实场景中CVR预估面临的**数据稀疏**以及**样本选择偏差(sample selection bias)**这两个关键问题。 实践出真知，论文一点也不花里胡哨，只有4页，据传被 SIGIR'2018 高分录用。

一、Motivation

CVR预估是十分重要的。对于广告商来说，现在更加重视的是OCPC(optimized cost-per-click), 更加重视广告的转化而非只有点击。对于平台来说，我们不仅希望用户点击商品(engagement)，还希望用户购买它(satisfation)，所以CTR、CVR都要重视，因此在最后排序item的时候，是按照CTR、CVR和其他各种指标加权求和来排序的。

不同于CTR预估问题，CVR预估面临两个关键问题：

1. **Sample Selection Bias (SSB)** 。转化是在点击之后才"有可能"发生的动作，传统CVR模型通常**以click数据为训练集，其中点击未转化为负例，点击并转化为正例**。但是训练好的模型实际使用时，我们训练的CVR模型是要用在**整个空间的样本(impression)**上的，而非只对点击样本进行预估。即是说，训练数据与实际要预测的数据**来自不同分布**，这个偏差对模型的泛化能力构成了很大挑战。
2. **Data Sparsity** 作为CVR训练数据的**click**样本远小于CTR预估训练使用的**impression**样本。

二、Model

介绍ESMM之前，我们还是先来思考一个问题——"**CVR预估到底要预估什么**"。想象一个场景，一个item，由于某些原因，例如在feeds中的展示头图很丑，它被某个user点击的概率很低，但这个item内容本身完美符合这个user的偏好，若user点击进去，那么此item被user转化的概率极高。CVR预估模型，预估的正是这个转化概率，**它与CTR没有绝对的关系，很多人有一个先入为主的认知，即若user对某item的点击概率很低，则user对这个item的转化概率也肯定低，这是不成立的。**更准确的说，**CVR预估模型的本质，不是预测"item被点击，然后被转化"的概率**（CTCVR），**而是"假设item被点击，那么它被转化"的概率**（pCVR）。

$$pCVR = p(conversion|click, impression)$$

这就是不能直接使用全部样本训练CVR模型的原因，因为咱们压根不知道这个信息：那些unclicked的item，"假设"他们被user点击了，它们是否会被转化。如果直接使用0作为它们的label，会很大程度上误导CVR模型的学习。所以这样就导致了Sample Selection Bias问题，即CVR模型的训练数据只能是clicked。

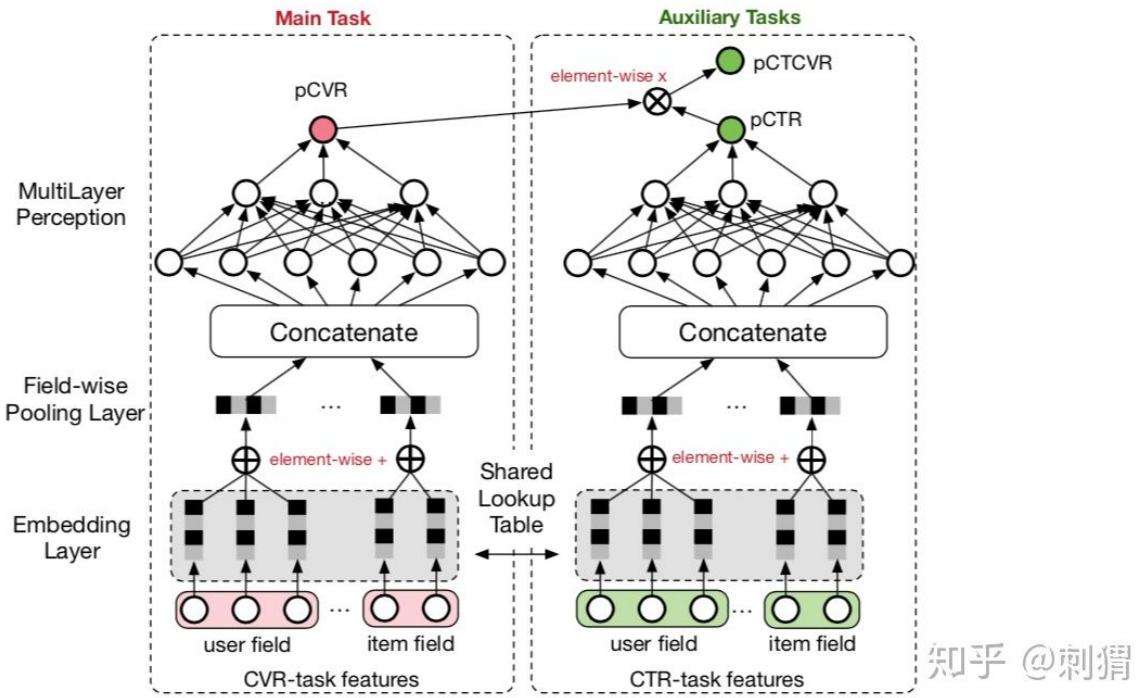
一些策略试图缓解CVR训练时样本空间只是clicked的这个问题，例如**从impression中对unclicked样本抽样做负例缓解SSB**，但是我们上面已经说过了，把unclicked item当成CVR预估的负样本，这样做是不对的。

在这篇文章中，利用的是**"impression->click->conversion"**这个行为链条，从而可以在**整个空间(Entire Space)**上进行训练及预测。也就是说，ESMM方法同时解决了样本选择偏差和数据稀疏的问题。ESSM涉及到同时预测CTR与CVR两个任务，因此属于**多任务学习**的范畴。其中，主任务当然就是预测CVR（click的条件下，conversion的概率）；**辅助任务**有两个：一是CTR（impression的条件下，click的概率），二是CTCVR（impression的条件下，click并且conversion的概率）。可以看出，CTR和CTCVR都是针对全空间impression的item来说的，并不存在选择偏差的问题。

pCVR(click后conversion的概率)和CTR、CTCVR有什么关系呢？下面这个公式表明，**pCVR * pCTR = pCTCVR**.

$$\underbrace{p(z\&y=1|\boldsymbol{x})}_{pCTCVR}=\underbrace{p(z=1|y=1,\boldsymbol{x})}_{pCVR}\underbrace{p(y=1|\boldsymbol{x})}_{pCTR}, \quad (1)$$

其中 z, y 分别表示conversion和click。既然CTR和CTCVR这两个任务是使用全部impression样本的，那为啥不绕个弯，通过这学习两个任务，再根据上式**隐式地学习CVR任务**呢？这样，训练出的CVR就也是ESMM正是这么做的，具体结构如下：



仔细观察上图，留意以下几点：

1）**共享Embedding。**CVR和辅助任务CTR使用相同的特征embedding，这是为了解决数据少的问题。实际上，多任务学习的一个目的就是为了解决某些任务**数据稀疏**的问题，这也是一种迁移学习的思想，即CTR预估和CVR预估在representation层是相似的，利用好CTR中更丰富的数据集，把CTR中学习到的东西迁移到CVR中去。

2）**隐式学习pCVR** 啥意思呢？这里pCVR（粉色节点）仅是网络中的一个**variable，没有显示的监督信号。**pCVR的更新完全是由pCTCVR这个辅助loss来做的。

具体地，反映在目标函数中：

$$L(\theta_{cvr},\theta_{ctr})=\sum_{i=1}^{N}l(y_i,f(\boldsymbol{x}_i;\theta_{ctr}))+\sum_{i=1}^{N}l(y_i\&z_i,f(\boldsymbol{x}_i;\theta_{ctr})*f(\boldsymbol{x}_i;\theta_{cvr})),$$

其中，第一项是CTR的cross-entropy loss,第二项是CTCVR的cross-entropy loss。利用**CTCVR和CTR**的监督信息来训练网络，【隐式】地学习CVR，这正是ESMM的精华所在，至于这么做的必要性以及合理性，本节开头已经充分论述了。

再思考下，ESMM的结构是基于"乘"的关系设计——pCTCVR=pCVR*pCTR，是不是也可以通过"除"的关系得到pCVR，即 pCVR = pCTCVR / pCTR ？例如分别训练一个CTCVR和CTR模型，然后相除得到pCVR，其实也是可以的，但这有个明显的缺点：真实场景预测出来的pCTR、pCTCVR值都比较小，"除"的方式容易造成**数值上的不稳定**（就是精度问题）。作者在实验中对比了这种方法。

三、Experiment

实验设置

1. 对比方法：
   - BASE——图1左部所示的CVR结构，训练集为click item；
   - AMAN——从unclicked样本中**随机抽样作为负例**加入点击集合(这种方法之前说过，其实是不对的，因为unclicked并不代表click之后转化率不高)；
   - OVERSAMPLING——对click中的正例（conversion样本）过采样；
   - DIVISION——分别训练CTR和CVCTR，相除得到pCVR；
   - ESMM-NS——ESMM结构中CVR与CTR部分不share embedding。
2. 上述方法/策略都使用NN结构，relu激活函数，嵌入维度为18，MLP结构为360*200*80*2，adam优化器 with
$$\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$$
。
3. 按时间分割，1/2数据训练，其余测试

衡量指标：计算CVR和CTCVR的预估准确率。对于CVR,就是在click样本上，预测click的条件下conversion的概率；对于CTCVR，是在全空间上，计算click+conversion的概率，即pCTR*pCVR。

实验结果

如表1所示，ESMM显示了最优的效果。这里有趣的一点可以提下，ESMM是使用全部样本训练的，而CVR任务只在点击样本上测试性能，因此这个指标对ESMM来说是在biased samples上计算的，但ESMM性能还是很牛啊，说明其有很好的泛化能力。

Model	AUC(mean ± std) on CVR task	AUC(mean ± std) on CTCVR task
BASE	66.00 ± 0.37	62.07 ± 0.45
AMAN	65.21 ± 0.59	63.53 ± 0.57
OVERSAMPLING	67.18 ± 0.32	63.05 ± 0.48
UNBIAS	66.65 ± 0.28	63.56 ± 0.70
DIVISION	67.56 ± 0.48	63.62 ± 0.09
ESMM-NS	**68.25 ± 0.44**	**64.44 ± 0.62**
ESMM	**68.56 ± 0.37**	**65.32 ± 0.49**

表2. 在Public上的实验结果，AUC以%为单位

在Product数据集上，各模型在不同抽样率上的AUC曲线如图2所示，ESMM显示的稳定的优越性，曲线走势也说明了Data Sparsity的影响还是挺大的。

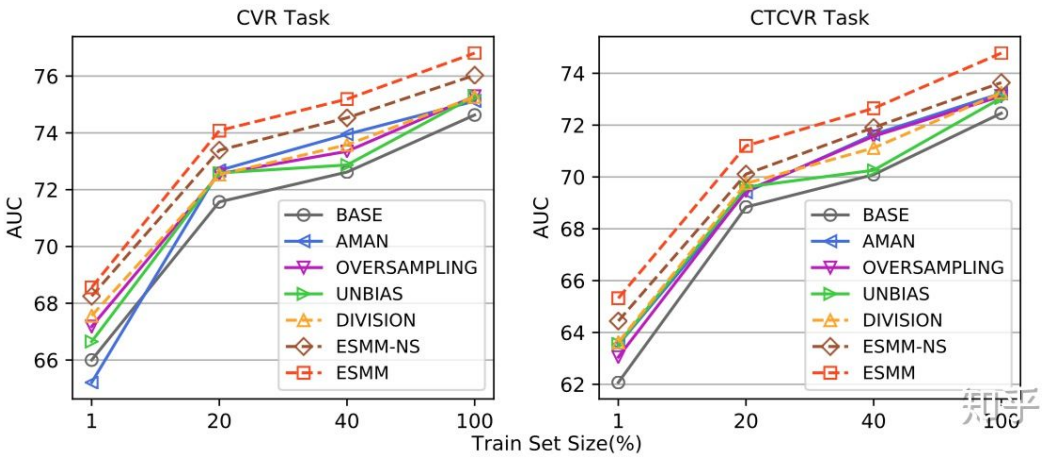


图2. 在Product上，各模型在不同抽样率上的AUC曲线

四、Discussion

1. ESMM 根据用户行为的"链条" -- "impression->click->conversion"，显示引入CTR和CTCVR作为辅助任务，“迂回” 学习CVR，从而在完整样本空间下进行模型的训练和预测，解决了CVR预估中的2个难题。
2. 可以把 ESMM 看成一个**新颖的 MTL 框架**，其中子任务的网络结构是可替换的，当中有很大的想象空间。至于这个框架的意义，这里引用论文作者之一@朱小强的描述：

> 据我所知这个工作在这个领域是最早的一批，但不唯一。今天很多团队都吸收了MTL的思路来进行建模优化，不过大部分都集中在传统的MTL体系，如研究怎么对参数进行共享、多个Loss之间怎么加权或者自动学习、哪些Task可以用来联合学习等等。ESMM模型的特别之处在于我们额外**关注了任务的Label域信息**，通过展现>点击>购买所构成的行为链，巧妙地构建了multi-target概率连乘通路。传统MTL中多个task大都是隐式地共享信息、任务本身独立建模，ESMM细腻地捕捉了契合领域问题的任务间显式关系，**从feature到label全面利用起来**。这个角度对互联网行为建模是一个较有效的模式，后续我们还会有进一步工作。