107 | 广告回馈预估综述

2018-06-08 洪亮劼

AI技术内参

进入课程 >

讲述：初明明
时长 06:24　大小 2.94M

在上一篇的分享里，我们详细地讨论了广告系统的架构，熟悉了各个组件都是怎么运作的，特别是我们重点剖析了对于每一个广告请求，供应侧平台（SSP）、广告交易平台（ADX）、需求侧平台（DSP）以及数据处理平台（DMP）都扮演了什么样的角色。同时，我们介绍了对于用户信息的追踪和整合，业界的基本技术就是存储用户的 Cookie，以及慢慢催生的 Cookie 的整合技术。

今天，我们就来看一看整个计算广告领域最核心的一个问题：**广告回馈预估**。

什么是广告回馈预估

什么是广告回馈预估？广告回馈预估要解决什么问题？我们先来弄明白这个问题。

我们说过计算广告有两大应用领域：搜索广告和展示广告，以及围绕这些广告的生态系统。这些系统或者领域都希望达到一个最终的目的，那就是用户和广告进行**交互**，并且能够对广告所代表的服务或者产品产生印象，从而达成某种程度的**交易**。

这里的“交互”包括对传统广告的点击，也包括对视频广告的观看。而在和广告交互之后，用户对于广告所代表的服务或商品达成的“交易”，包括购买、订阅甚至是改变印象等等。那么，这一切和广告本身的交互以及和广告所代表的服务或者商品达成的交易，我们都通通称为“**回馈**”（Feedback）。

而我们所说的“**回馈预估**”，**就是要预测用户这种行为的可能性，或者说是概率**。也就是说，我们希望了解用户是不是有可能点击这个广告；有多大概率观看完这段视频广告；有多大可能去购买这个广告所代表的商品。

对广告的回馈概率进行有效估计是很多广告生态系统组件的一个核心需求。对于发布商来说，显示广告，不管是通过搜索结果还是通常的页面，都希望能够有用户交互从而带来收入。很多广告系统的收入模式就是依靠用户的点击从而让广告商为此支付一定费用给发布商。

因此，对于发布商来说，越是能准确估计这部分的点击率，就越能保证自己的收入。相应地，对于广告商来说，很有必要知道某一种类型的广告在哪个发布商能够带来更多的点击，从而能够有针对性地对于某个发布商进行投放。由此看来，广告回馈预估是一个非常重要的有待解决的技术问题。

广告回馈预估的普遍挑战和技术难点

既然广告回馈预估很重要，我们是不是可以直接用现成的机器学习工具就可以解决这个问题呢？这个问题有什么自身的特点，又有哪些挑战和技术难点呢？

在比较简单的设定下，广告回馈预估可以看做是某种**监督学习的任务**。在这类监督学习任务里，标签是用户的动作，例如点击或者观看，或者购买等。我们需要建立的是一个用户在某种上下文中对广告标签的一个模型。这里的上下文包括查询关键词、用户信息、广告信息以及一切其他有用的信息。

那么，在这样的设定下，广告回馈预估的核心挑战是什么呢？

核心挑战其实来自于稀疏的数据。

不管是在搜索广告中也好，还是在展示广告中也好，从平均的角度来说，相比于用户和正常的搜索结果或者展示结果（比如新闻内容等）的互动，用户与广告的互动要成倍地减少。有一项研究表明，在同样一个位置，广告的点击率可以是正常内容的十分之一、百分之一甚至是千分之一。也就是说，从概率的角度来看，用户普遍是不点击广告的。这个观察基本上是符合我们对用户的普遍理解的。但是，较少的点击数据造成的结果就是，从监督学习的角度来说，大量的数据点都是未交互的数据，只能当做**负例**来处理。

实际上，**在广告点击率预估的问题中，正例的数目常常是负例的百分之一或者千分之一**。这样造成的就是非常“**不均衡**”的数据集。而要想在不均衡的数据里中进行概率估计，往往都是一件困难的事情。

而购买事件相对于广告点击来说就更加稀少，这一点其实也很正常。在点击了广告之后，又有多少人真正会去购买这些产品呢？因此，提高广告的转化率，也就是交易发生的概率，往往就是更富有挑战的任务。

值得一提的是，在监督学习的框架中，除了数据问题以外，广告回馈预估还有**目标函数的挑战**。具体是什么情况呢？

在真实的系统中，我们需要在很多候选可能的广告中，选出最优的一个或者几个显示在页面上。从某种程度上来说，这更像是一个**排序问题**。同时，对于不少 DSP（需求侧平台）来说，广告排序的最终目的是进行“竞拍”（Auction）。因此，最后估算广告的点击率以后，还需要看广告的竞价，由此来对广告是否会赢得竞拍从而被显示在页面上进行一个更加全面的估计。很显然，和传统的推荐或者搜索比较，这些问题都要复杂许多。

广告回馈预估的算法和模型

广告回馈预估的难点和挑战来自两方面，一方面是稀疏的数据，会造成不均衡的数据集；一方面是目标函数的挑战。那么，广告回馈预估有哪些比较常见的算法或者模型呢？

我们接下来会对这一系列有关的算法和模型进行详细讨论。今天，我会带你从宏观上进行一下总结。

从最直接的监督学习的角度来看，广告回馈预估的一个常见算法就是把这个问题当做**二元分类问题**，并且直接利用“**对数几率回归**”（Logistic Regression）来对这个问题建模。实际上，直到今天，对数几率回归依然是广告回馈预估领域的重要方法。

第二类经常使用的就是**树模型**，特别是 **GBDT** 这个通常在搜索中使用的模型。我们前面已经提到了这类模型对于排序学习的作用。

第三类目前比较火热的领域就是**如何利用深度学习来对回馈预估进行建模。**这一类模型在最近几年有了比较大的进展。

总结

今天我为你介绍了广告系统中最核心的一个问题：广告回馈预估。一起来回顾下要点：第一，广告回馈预估就是预测“用户与广告的交互以及达成交易这种行为”的概率；第二，广告回馈预估有两方面的难点和挑战，分别来自数据和目标函数；第三，在这个领域有一些流行的模型，比如对数几率回归和数模型等。

最后，给你留一个思考题，当我们有不均衡数据集的时候，我们一般都有哪些解决方案？针对广告预估，是否需要对这些方案进行额外的处理呢？

欢迎你给我留言，和我一起讨论。

上一篇 106 | 广告系统架构
下一篇 108 | Facebook的广告点击率预估模型

精选留言 (1)

写留言

旭
2018-06-08

训练集数据不平衡一般有大样本欠采样、小样本过采样、重采样数据生成、修改评价指标权重等多种方法。