006 | Google的点击率系统模型

2017-10-16 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 10:06 大小 4.63M

广告是很多互联网公司的重要收入来源，比如 Google、Facebook、微软、阿里巴巴、百度、腾讯等。以 Facebook 为例，它的 2017 年第一季度财报显示，公司总营收为 78.4 亿美元，这其中 98% 的收入来自广告。同样，在这些公司内部，都有着完善的广告系统来支撑其广告业务。

当然，大型广告系统的成功需要依靠很多相互协调的子系统和组件。今天我要和你聊的是广告系统里最基础的一个子系统，也是整个广告系统的核心功能之一——**点击率预估系统。**点击率预估，顾名思义就是根据环境和广告的类型，来估计用户有多大的可能性点击当前的广告。这个预估值会用于广告系统的其他组件，比如对广告主（投放广告的客户）的计费模块。因此，点击率预估的准确性和实时性就变得十分重要。

今天和你分享一篇广告点击率预估文献史上非常重要的论文，它来自 Google 广告团队，标题是《工程实践视角下的广告点击率预估》（“Ad Click Prediction: a View from the Trenches”）。

论文背景

这篇论文发表于 KDD 2013 年的工业论文组，在短短几年时间里就获得了近 200 次的文章引用数，不少公司争相研究其中的内容，希望能够复制类似的算法和技术。

这篇文章的作者群多达 16 人，他们都是来自 Google 西雅图、匹兹堡、硅谷以及剑桥等地办公室的研究人员和工程师，文章的致谢部分也有 9 人。可见整个论文以及里面的技术的确是团队协作的结果。

这里面有两位作者值得介绍一下。第一位是论文的第一作者布兰登（H. Brendan McMahan）。布兰登早年在卡内基梅隆大学计算机系获得博士学位。他的博士生导师是戈登（Geoff Gordon）以及布卢姆（Avrim Blum），这两位都是卡内基梅隆大学机器学习界的权威教授。布兰登本人长期对优化算法有深入的研究，这篇论文的重要核心算法就来自于他的研究成果。

文章的另外一位作者斯卡利（D. Sculley）从塔夫茨大学（Tufts University）博士毕业之后，一直在 Google 的匹兹堡分部工作，并着手研究大规模机器学习系统，其中重要的代表性研究成果是如何把回归问题和排序问题结合起来（发表于 KDD 2010 年）。斯卡利曾经是一个著名的开源大规模机器学习软件包 sofia-ml 的作者，里面实现了一个大规模版本的 RankSVM，一度受到关注。

在线逻辑回归（Logistic Regression）

文章首先讲解的是点击率预估的核心算法。因为 Google 要处理的数据集非常庞大，不管是样本数量还是样本的特征数都是百亿级别的，所以选用什么样的算法至关重要。2013 年，也就是这篇论文发表的时候，当时大规模深度学习的环境还没有完全成熟起来，Google 的科学家和工程师选择了**逻辑回归**，这是一个非常传统但也非常强大的线性分类工具。

我们这里简单回顾一下逻辑回归模型。

逻辑回归是要对二元分类问题进行建模，模型的核心是通过一组（有可能是非常巨大规模的）特征以及所对应的参数来对目标的标签进行拟合。这个拟合的过程是通过一个叫逻辑转换或函数来完成的，使得线性的特征以及参数的拟合能够非线性转换为二元标签。

普通的逻辑回归并不适应大规模的广告点击率预估。有两个原因，第一，数据量太大。传统的逻辑回归参数训练过程都依靠牛顿法（Newton's Method）或者 L-BFGS 等算法。这些算法并不太容易在大规模数据上得以处理。第二，不太容易得到比较稀疏（Sparse）的答案（Solution）。也就是说，虽然数据中特征的总数很多，但是对于单个数据点来说，有效特征是有限而且稀疏的。

我们希望最终学习到的模型也是稀疏的，也就是对于单个数据点来说，仅有少量特征是被激活的。传统的解法，甚至包括一些传统的在线逻辑回归，都不能很好地解决答案的稀疏性问题。

这篇文章提出了用一种叫**FTRL**（Follow The Regularized Leader）的在线逻辑回归算法来解决上述问题。FTRL 是一种在线算法，因此算法的核心就是模型的参数会在每一个数据点进行更新。FTRL 把传统的逻辑回归的目标函数进行了改写。

新的目标函数分为三个部分：第一部分是一个用过去所有的梯度值（Gradients）来重权（Re-Weight）所有的参数值；第二部分是当前最新的参数值尽可能不偏差之前所有的参数值；第三个部分则是希望当前的参数值能够有稀疏的解（通过 L1 来直接约束）。从这三个部分的目标函数来看，这个算法既能让参数的变化符合数据规律（从梯度来控制），也能让参数不至于偏离过去已有的数值，从而整个参数不会随着一些异常的数据点而发生剧烈变化。

在算法上另外一个比较新颖的地方，就是对每一个特征维度的学习速率都有一个动态的自动调整。传统的随机梯度下降（Stochastic Gradient Descent）算法或是简单的在线逻辑回归都没有这样的能力，造成了传统的算法需要花很长时间来手工调学习速率等参数。

同时，因为每一个特征维度上特征数值的差异，造成了没法对所有特征选取统一的学习速率。而 FTRL 带来的则是对每一个维度特征的动态学习速率，一举解决了手动调整学习算法的学习速率问题。简单说来，学习速率就是根据每一个维度目前所有梯度的平方和的倒数进行调整，这个平方和越大，则学习速率越慢。

系统调优工程

很明显，光有一个比较优化的在线逻辑回归算法，依然很难得到最好的效果，还会有很多细小的系统调优过程。

比如文章介绍了利用**布隆过滤器**（Bloom Filter）的方法，来动态决定某一个特征是否需要加入到模型中。虽然这样的方法是概率性的，意思是说，某一个特征即便可能小于某一个值，也有可能被错误加入，但是发生这样事件的概率是比较小的。通过布隆过滤器调优之后，模型的 AUC 仅仅降低了 0.008%，但是内存的消耗却减少了 60% 之多，可见很多特征仅仅存在于少量的数据中。

文章还介绍了一系列的方法来减少内存的消耗。比如利用更加紧凑的存储格式，而不是简单的 32 位或者 64 位的浮点数存储。作者们利用了一种叫 q2.13 的格式，更加紧凑地存储节省了另外 75% 的内存空间。

此外，前面我们提到的计算每一步 FTRL 更新的时候，原则上都需要存储过去所有的梯度信息以及梯度的平方和的信息。文章介绍了一种非常粗略的估计形式，使得这些信息可以不必完全存储，让内存的消耗进一步降低。这部分内容可能并非对所有读者都有益处，然而我们可以看到的是，Google 的工程师为了把一种算法应用到实际中做出了非常多的努力。

另外，文章也特别提出，虽然大家都知道在点击率预估这样非常不对称的问题上（也就是正例会远远少于负例）需要对负样本进行采样，但是这里面需要注意的是直接采样会对参数的估计带来偏差。同时文章也提出了需要对模型的最后预测进行调整（Calibration），使得模型的输出可以和历史的真实点击率分布相近。这一点对于利用点击率来进行计费显得尤为重要，因为有可能因为系统性的偏差，预测的数值整体高出或者整体低于历史观测值，从而对广告主过多计费或者过少计费。

失败的实验

这篇文章难能宝贵之处是不仅介绍了成功的经验，还介绍了一些失败的或者是不怎么成功的实验结果，让后来的学者和工程师能够少走弯路。

比如著名的**Hashing Trick**，在这篇文章里，Google 的工程师们经过实验发现，特征经过哈希之后并没有显著降低内存而且模型的精准度有所下降，同时哈希也让模型变得不可解释，于是 Google 的工程师觉得没有必要对特征进行哈希。

另外一个热门的技术**Dropout**也被作者们尝试了，在 Google 的实验数据上并没有显著的效果。还有一个经常见到的技术，那就是对学到的参数进行归一化（Normalization），这

是让参数能够在一定的范围内不随便波动。遗憾的是，Google 的作者们也发现这个技术没有太大作用，模型的效果经常还会降低。

小结

今天我为你分享了这篇关于广告点击率预估的重要论文，你需要理解的核心要点有几个，一是 FTRL 模型的创新；二是这个模型如何应用到工业界的环境中特别是如何对内存的消耗进行调优；三是 Google 一系列失败尝试的总结。

总之，这篇论文是难得一见的工业界级别的科技论文分享。从 KDD 工业组的角度来说，很有借鉴意义；从业界贡献来说，除了广告之外，FTRL 也被广泛应用到推荐系统等领域。

最后，我们再来探讨个问题。假设你在负责公司的广告系统，那你应该如何判断自己的场景是不是应该使用 FTRL 呢？

欢迎你给我留言，和我一起讨论。

上一篇　005 | 数据科学家基础能力之系统

精选留言 (6)

写留言

Momo
2017-11-09

5

百亿特征是多大量文本特征onehot变换而来，从而导致大量特征值都是0，也就是稀疏问题

Classtag
2018-11-03

3

百亿特征是大部分都是onehot出来的吧，原始特征其实没那么多。

展开

帅帅
2018-10-20

2

模型真的是繁多；
不过坚信一个道理，模型容量越大，需要越多的数据、计算能力、架构能力；

从LR到GBDT+LR，到GBDT+FM、到WIDE&DEEP，现在又出现了FTRL，要学习的真的很多；...

展开

极客星星
2017-12-03

2

关于什么时候用ftrl 个人认为 如果数据量相对小 使用开源的LR库可以解决问题 就不需要用ftrl 否则 应该采用ftrl 因为它支持增量更新 有稀疏性 是在工业界得到充分验证的技术。此外 有个问题想咨询下洪老师 有没有什么论文讲工业界搜索广告特征工程方面的文章 或者洪老师能否介绍下经验 选取什么特征比较有效 谢谢

展开

作者回复: 目前并没有太系统的这类工作。主要是每一家的系统差别都很大，可能很多经验无法直接推广。

Keno Tu
2018-08-02

论文路径能提供下吗?

展开

Xuan
2017-10-24

百亿特征？这么多特征是怎么来的?
还有模型稀疏，这个怎么理解?