**从最近的比赛学习CTR/CVR**

最近在玩kaggle的talking data的比赛，适逢IJCAI2018的比赛也是CTR问题，故将最近学习的东西整理下分享，欢迎大家拍砖。

CTR预估就是这样的任务：给定用户或某唯一标识，给定一个商品(Product)，给定了一定的环境，来看用户会不会买这个商品，买商品的概率有多高；或者说给用户推荐一个电影，用户会不会看这个电影，看的概率有多高。形式化的表示就是建模P(click|content)给定上下文，建模点击/购买/下载/使用的概率。

CTR是很多电商，社交，信息流涉及到盈利的非常关键的技术，各家都非常重视，CTR微小的提升，带来的就是可观的盈利。

CTR问题为什么是一个值得拿出来研究的问题，是因为问题的特殊性。

* 存在高维度的稀疏特征

比如从用户的维度看，大的电商其用户有上亿，商品种类也繁多，这样简单的独热编码就带来了上亿维的特征

* 很多离散性质的特征

比如其行为发生时间，行为发生的渠道/页面位置等

一般来讲，对于CTR问题解决方向有两个，一是通过精巧的特征工程结合简单的模型（LR），这需要非常之高的专家知识和特征工程能力。二是弱化特征工程，通过模型的复杂性来自动发现建模，比较有代表的是Facebook的GBDT+LR，FM，FFM，DNN等方法。深度学习风潮兴起之后，很多基于NN的方法，如谷歌的deep and wide, Deep FM, FNN等方法也取得了不错的效果。

参考 [https://github.com/wzhe06/Ad-papers](https://link.zhihu.com/?target=https%3A//github.com/wzhe06/Ad-papers)，wzhe06同学搜集的论文列表

除了非常厉害的特征工程助你取胜之外，下面有一些经典的论文和方法讲解。当然这只是一般套路，针对具体的数据需要你有自己的理解，活用方法和创新才能立于不败之地。

**0.FM/FFM**

FM和FFM是台大的童鞋们在Netflix的比赛中把名声打响，后来在CTR中一直有不错的效果，FM和FFM的设计，隐含了很多矩阵分解，以及深度学习中embedding的思想，可以说这种压缩降维表示的思想在高维稀疏数据中都有不错的效果，这篇博客对其原理有讲解。[https://blog.csdn.net/mmc2015/article/details/51760681](https://link.zhihu.com/?target=https%3A//blog.csdn.net/mmc2015/article/details/51760681" \t "_blank)不过我在实践中发现，libffm对于talking data这种数据量巨大，极不平衡的情况下没法收敛，具体的原因还在排查。

**1.FTRL**

[Ad Click Prediction a View from the Trenches](https://link.zhihu.com/?target=https%3A//github.com/wzhe06/Ad-papers/blob/master/CTR%2520Prediction/Ad%2520Click%2520Prediction%2520a%2520View%2520from%2520the%2520Trenches.pdf)

这篇论文是非常经典的FTRL方法，在taiking data的kernel上有Giba大佬的分享 [giba-darragh-ftrl-rerevisited](https://link.zhihu.com/?target=https%3A//www.kaggle.com/titericz/giba-darragh-ftrl-rerevisited)，  
FTRL还解决了在线学习的问题，在线学习是算法实用化非常关键的一点。[中文博客在这里有对原理的讲解](https://link.zhihu.com/?target=https%3A//blog.csdn.net/yz930618/article/details/75270869), 感兴趣的同学可以看一看。

**2.[Practical Lessons from Predicting Clicks on Ads at Facebook](https://link.zhihu.com/?target=https%3A//github.com/wzhe06/Ad-papers/blob/master/CTR%2520Prediction/Practical%2520Lessons%2520from%2520Predicting%2520Clicks%2520on%2520Ads%2520at%2520Facebook.pdf" \t "_blank)**

这篇文章也非常有名，是把GBDT的叶子结点送给LR，通过模型自动发现学习组合特征。在业界和竞赛上都有不错的效果，中文博客如下三篇对这个讲解不错，Xgboost使用简单的

new\_feature**=** xgb**.**predict(d\_test, **\***pred\_leaf**\*=**True)

就可以拿到GBDT的叶子结点特征，非常方便

[https://breezedeus.github.io/2014/11/19/breezedeus-feature-mining-gbdt.html#fn:fbgbdt](https://link.zhihu.com/?target=https%3A//breezedeus.github.io/2014/11/19/breezedeus-feature-mining-gbdt.html%23fn%3Afbgbdt)

[https://blog.csdn.net/dengxing1234/article/details/73739836](https://link.zhihu.com/?target=https%3A//blog.csdn.net/dengxing1234/article/details/73739836)

[https://blog.csdn.net/lilyth\_lilyth/article/details/48032119](https://link.zhihu.com/?target=https%3A//blog.csdn.net/lilyth_lilyth/article/details/48032119)

**3.一些NN方法**

NN的方法直觉思路是，通过embedding得到向量空间的表示，然后拿到表示后通过NN模型复杂的非线性完成分类。

在talking data一开始的时候，我在kernel写了个NN一把梭的baseline，后来有几个大佬在这上面做了一些优化，看起来比较简单。[https://www.kaggle.com/baomengjiao/embedding-with-neural-network](https://link.zhihu.com/?target=https%3A//www.kaggle.com/baomengjiao/embedding-with-neural-network" \t "_blank)

下面是一些比较经典的论文,这些论文很多在去年的腾讯计算广告计算大赛中第一名都有所使用，并且取得了单模型就能排到第三的很好的成绩，所以深度学习不容小觑。

* [Deep Neural Networks for YouTube Recommendations](https://link.zhihu.com/?target=https%3A//github.com/wzhe06/Ad-papers/blob/master/CTR%2520Prediction/Deep%2520Neural%2520Networks%2520for%2520YouTube%2520Recommendations.pdf)

youtebe使用DNN做推荐

* [Wide & Deep Learning for Recommender Systems](https://link.zhihu.com/?target=https%3A//github.com/wzhe06/Ad-papers/blob/master/CTR%2520Prediction/Wide%2520%2526%2520Deep%2520Learning%2520for%2520Recommender%2520Systems.pdf)

Deep & wide

* [FNN:](https://link.zhihu.com/?target=https%3A//arxiv.org/pdf/1601.02376.pdf)
* [PNN:](https://link.zhihu.com/?target=https%3A//arxiv.org/pdf/1611.00144v1.pdf)
* [DeepFM: A Factorization-Machine based Neural Network for CTR Prediction:](https://link.zhihu.com/?target=https%3A//arxiv.org/abs/1703.04247)

这里有一个完整的CTR的解决方法，[http://blog.csdn.net/chengcheng1394/article/details/78940565](https://link.zhihu.com/?target=http%3A//blog.csdn.net/chengcheng1394/article/details/78940565" \t "_blank)，kaggle的数据集，比较适合学习，讲解地非常详细。