GBDT + LR (Facebook, 2014)

paper: https://research.fb.com/wp-content/uploads/2016/11/practical-lessons-from-predicting-clicks-on-ads-at-facebook.pdf

1. 简介

GBDT+LR是早期使用在CTR预估上的一种方法。在GBDT+LR之前，使用最多的方法就是逻辑回归(LR)，LR使用了Sigmoid变换将函数值映射到0~1区间，映射后的函数值就是CTR的预估值。（从这里可以看出，LR很适合稀疏特征的学习） LR属于**线性模型**，学习能力十分有限，需要大量的**特征工程**来增加模型的学习能力。但大量的特征工程耗时耗力同时并不一定会带来效果提升。因此，如何自动发现有效的特征、**特征组合**，弥补人工经验不足，缩短LR特征实验周期，是亟需解决的问题。

FM模型通过隐变量的方式，自动计算**二阶**交叉特征的权重，但这种特征组合仅限于两两特征之间。后来都普遍用深度神经网络去挖掘更深层的特征组合关系，不过这便是后话了。其实，在全面使用神经网络之前，GBDT也是一种经常用来**发现特征组合**的有效思路。

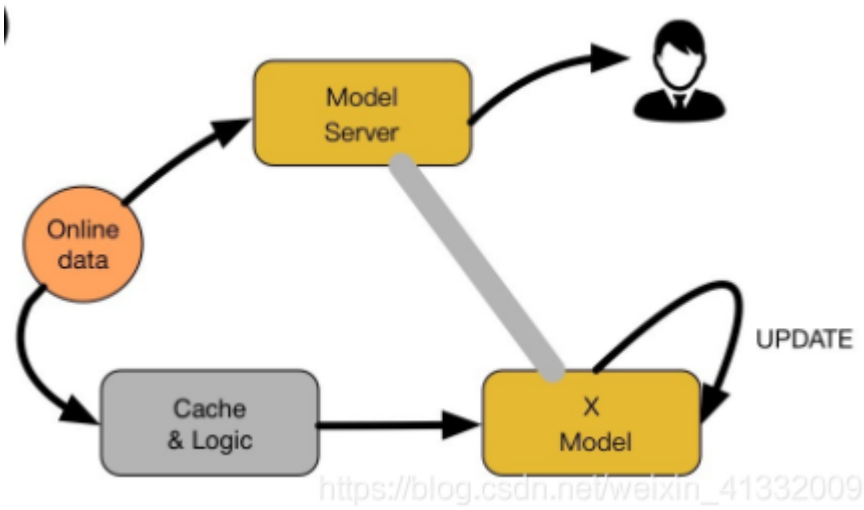
2. CTR预估模型的离线和在线训练

CTR预估的模型训练分为离线训练(offline)、在线训练(online)，其中离线部分目标主要是训练出可用模型，而在线部分则考虑模型上线后，性能可能随时间而出现下降，若出现这种情况，可选择使用**Online-Learning**来**在线更新**模型。

离线部分：

- 数据**收集**：收集和业务相关的数据，如在app位置进行埋点
- 预处理
- 构造数据集：切训练、测试、验证集
- 特征工程：对原始数据进行基本的特征处理，包括去除相关性大的特征，离散变量one-hot，连续特征离散化
- 模型选择：选择合理的机器学习模型来完成相应工作，原则是先从简入深，先找到baseline，然后逐步优化；
- 超参选择：利用gridsearch、randomsearch来进行超参选择，选择在离线数据集中性能最好的超参组合；
- **在线**A/B Test：选择优化过后的模型和原先模型（如baseline）进行A/B Test，若性能有提升则替换原先模型

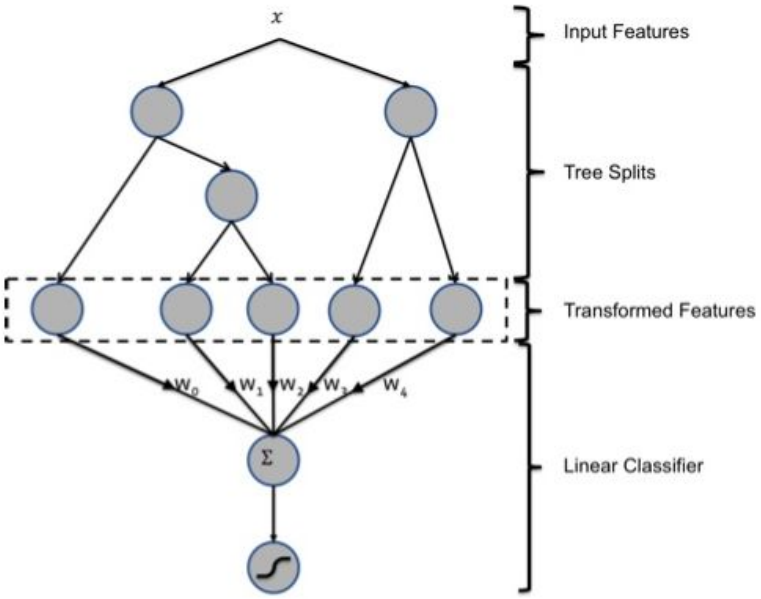
在线部分： 由于模型上线一段时间之后，会有新的用户、有新的商品出现，用户的兴趣也会发生变化，所以模型必须具有实时更新的能力。这个“实时”可以是分钟级、小时级、天级的更新。



- Model server部分：用户上线或者用户来了一个query，就从线上的模型取到结果，返回给用户。同时，将用户的行为记录下来，存入cache & logic模块
- Cache & Logic：设定简单过滤规则，过滤异常数据；收集足够多的新数据
- 模型更新：当Cache & Logic 收集到合适大小数据时，从Model Server中fetch一个最新的模型，然后对模型进行增量训练finetuning。之后在验证集上测试，如果效果比原始模型好，则更新model server的模型参数。这样，就完成了一次模型更新。

3. GBDT + LR 的结构

正如它的名字一样，GBDT+LR 由两部分组成，其中GBDT用来对训练集**提取特征**作为新的训练输入数据，LR作为新的训练输入数据的分类器。



图中共有两棵树，这两棵树是Gradient Boosting Tree，第二棵树学习的是第一棵树的残差。x为一条输入样本，遍历两棵树后，x样本分别落到两颗树的叶子节点上，每个叶子节点对应LR的一个特征。构造的新特征向量(Transformed Features)是取值0/1的。

举例来说：上图有两棵树(即两个弱分类器)，左树有三个叶子节点，右树有两个叶子节点，最终的特征即为五维的向量。对于输入x，假设他落在左树第一个节点，编码[1,0,0]，落在右树第二个节点则编码[0,1]，所以整体的编码为[1,0,0,0,1]，这类编码作为input，输入到LR中进行分类。在对原始数据进行GBDT提取为新的数据这一操作之后，数据不仅变得**稀疏**，而且可能会导致新的训练**数据特征维度过大**的问题（维度灾难，虽然在高维空间更加线性可分，但是在高维中训练得到的分类器其实相当于低维空间上的一个复杂非线性分类器）。因此，在Logistic Regression这一层中，可使用**正则化**来减少【过拟合】的风险，在Facebook的论文中采用的是L1正则化。

4. Q/A

- 为什么建树采用GBDT而非RF?

RF也是多棵树，但从效果上有实践证明不如GBDT。且GBDT前面的树，特征分裂主要体现对多数样本有区分度的特征；后面的树，主要体现的是经过前N颗树，残差仍然较大的少数样本。优先选用在整体上有区分度的特征，再选用针对少数样本有区分度的特征，思路更加合理，这应该也是用GBDT的原因。