**STRATA演讲| 点击率(CTR)预估系统的进化：从LR到DNN**

**1. 引言**

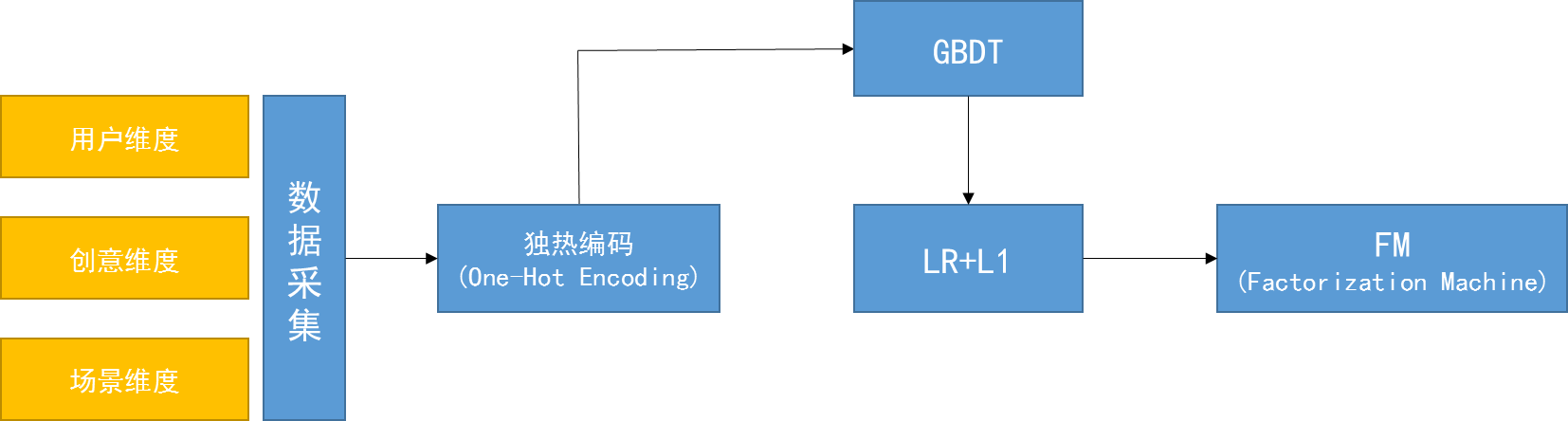
**什么是CTR，为什么要进行CTR预估？**

CTR（Click-Through-Rate）即点击通过率，指网络广告的点击到达率，即该广告的实际点击次数除以广告的展现量。

现代计算广告要求广告平台在正确的时间，正确的地点，给正确的人提供正确的创意。这意味着同一个广告位，对于不同的人来说，在不同的场景下看到的内容都不一致。而同一个广告主的广告一天可以出现在上万个广告位上。

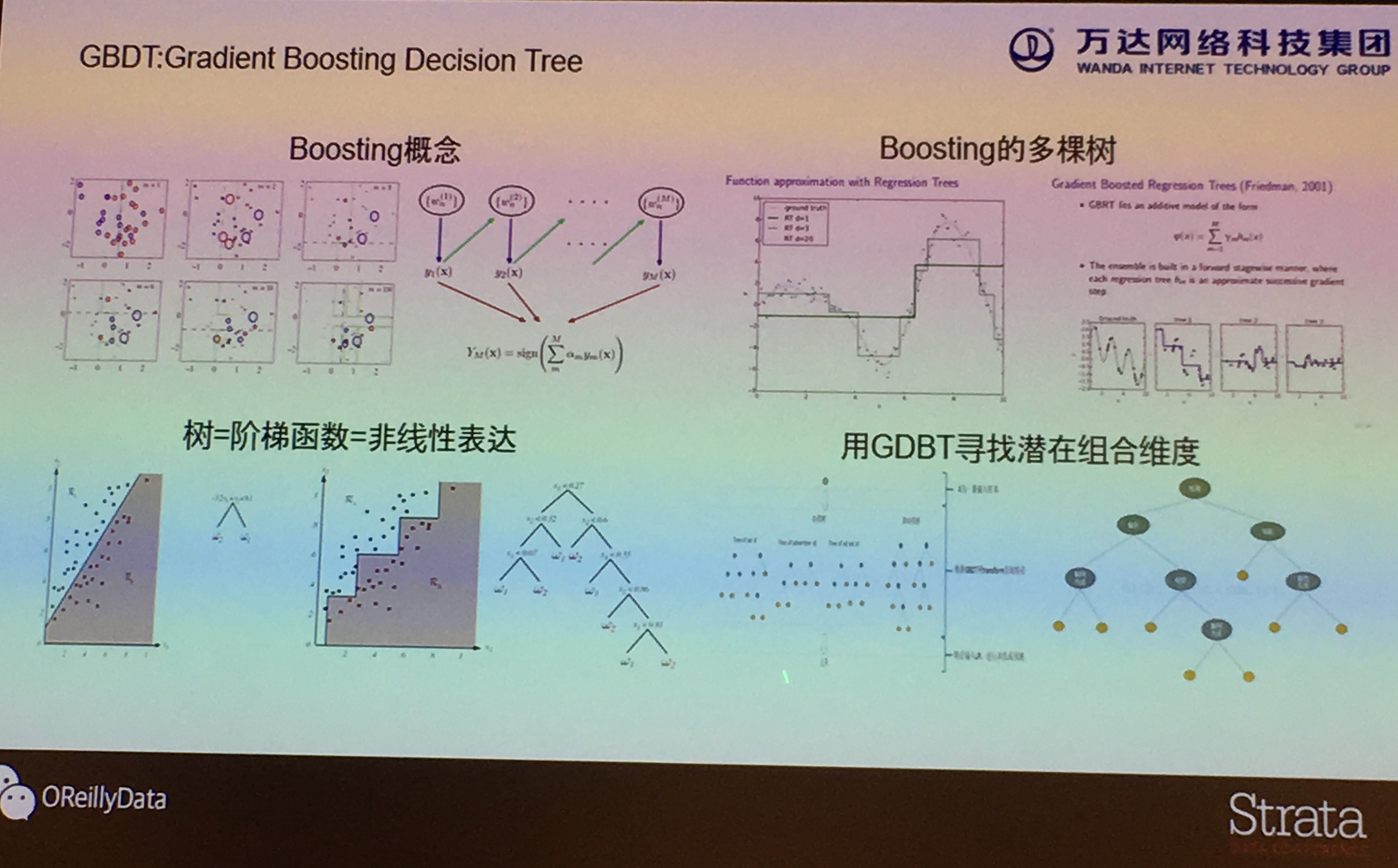
广告平台的最终目标是追求收益最大化，以CPC 广告（每次点击付费广告）为例，平台收益既与 CPC 单价有关，又与预测CTR 有关。良好的CTR预测算法可以使广告平台识别有效的人群和场景，提高平台的收益能力。互联网公司根据各自业务的特点，研发出了各种各样的CTR 预估模型及其变种，从较为简单的Logistic Regression发展到如今兴起的DNN算法。

本文主要结合STRATA演讲及网络相关博文，从算法和架构的角度为读者介绍CTR预估系统近几年的发展。

**2. 基于LR的经典的点击率预估流水线**

|  |  |  |
| --- | --- | --- |
| 数据采集 | 用户维度 | 用户的标签，包括网络行为、近期数据、长期数据和自身属性数据等 |
| 创意维度 | 广告素材自身的维度、色彩、文字、底标、类别等 |
| 场景维度 | 广告出现的位置等 |
| 独热编码 | | 连续值离散化 |
| GBDT | | 梯度提升决策树，二分类问题求解，产生树结构，将根节点和叶节点构成的潜在的组合维度加入到原来的维度池 |
| LR+L1 | | 产生稀疏解，新老维度交替 |
| FM（因子分解机） | | 解决稀疏数据下的特征组合问题 |

**2.1 GBDT(Gradient Boosting Decision Tree，梯度提升决策树)**



**（1）DT（Decision Tree 回归树）**

“树=阶梯函数=非线性表达”

在线性不可分的平面内，分割有两种方式。一种是利用非线性的函数进行分类，另一种用简单的线性函数分段分类，这种分段的线性可分等价于树形结构的表达形式。

GBDT中使用的是回归树而非决策树，通过最小化 log 损失函数找到最可能的分支，直到叶子节点上所有值唯一 (残差为0)，或者达到预设条件(树的深度)。若叶子节点上的值不唯一，则以该节点上的平均值作为预测值。在CTR的预测问题中，回归树模型主要用来生成可能的组合特征。

**（2）GB（Gradient Boosting 梯度提升）**

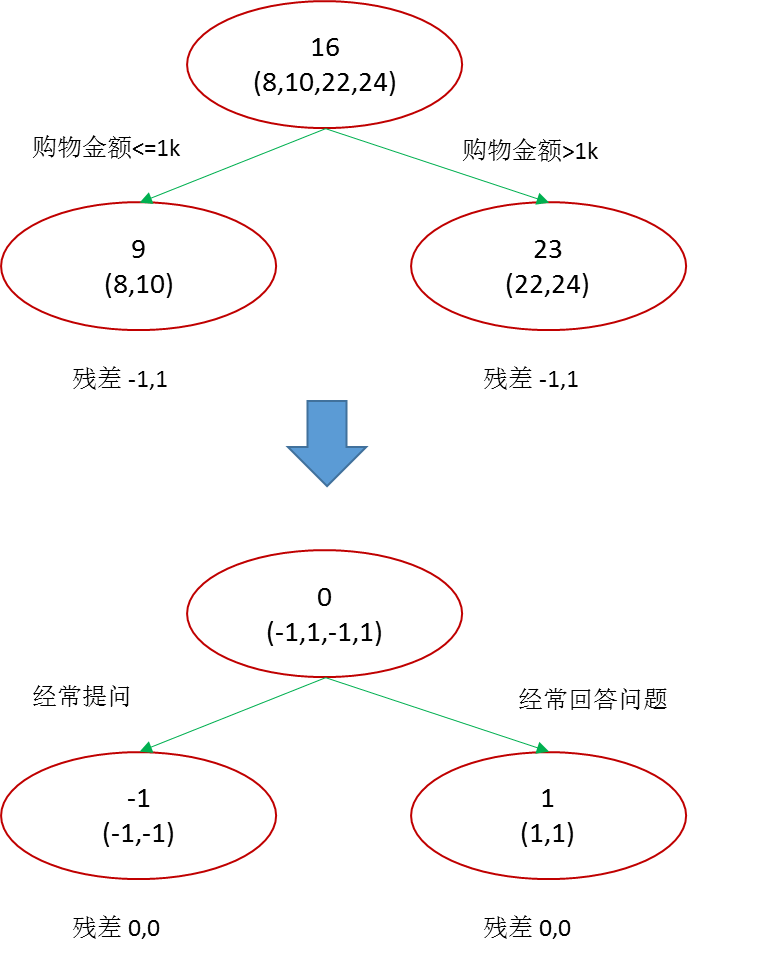
Boosting算法是求解分类问题中常用的算法，它通过对每次分类结果中的错分项增加权重，使得之后的分类器更加考虑到分错点的情况。

如上图所示，蓝色和红色点的分布结构复杂，不能够用一个分类器产生正确结果。在第一个相对简单的分类器的基础上，将每次模型错分点的权重提高，产生一个新的分类器。经过了M=6 M=30 M=130 的逐轮的迭代之后，形成了一个比较复杂的超平面，提高分类效果。具体在树型结构中，最直观的想法是每一棵树学习的是之前所有树结论的残差（对应错分点的权重），最终通过迭代生成的所有树的累加量判断分类结果。

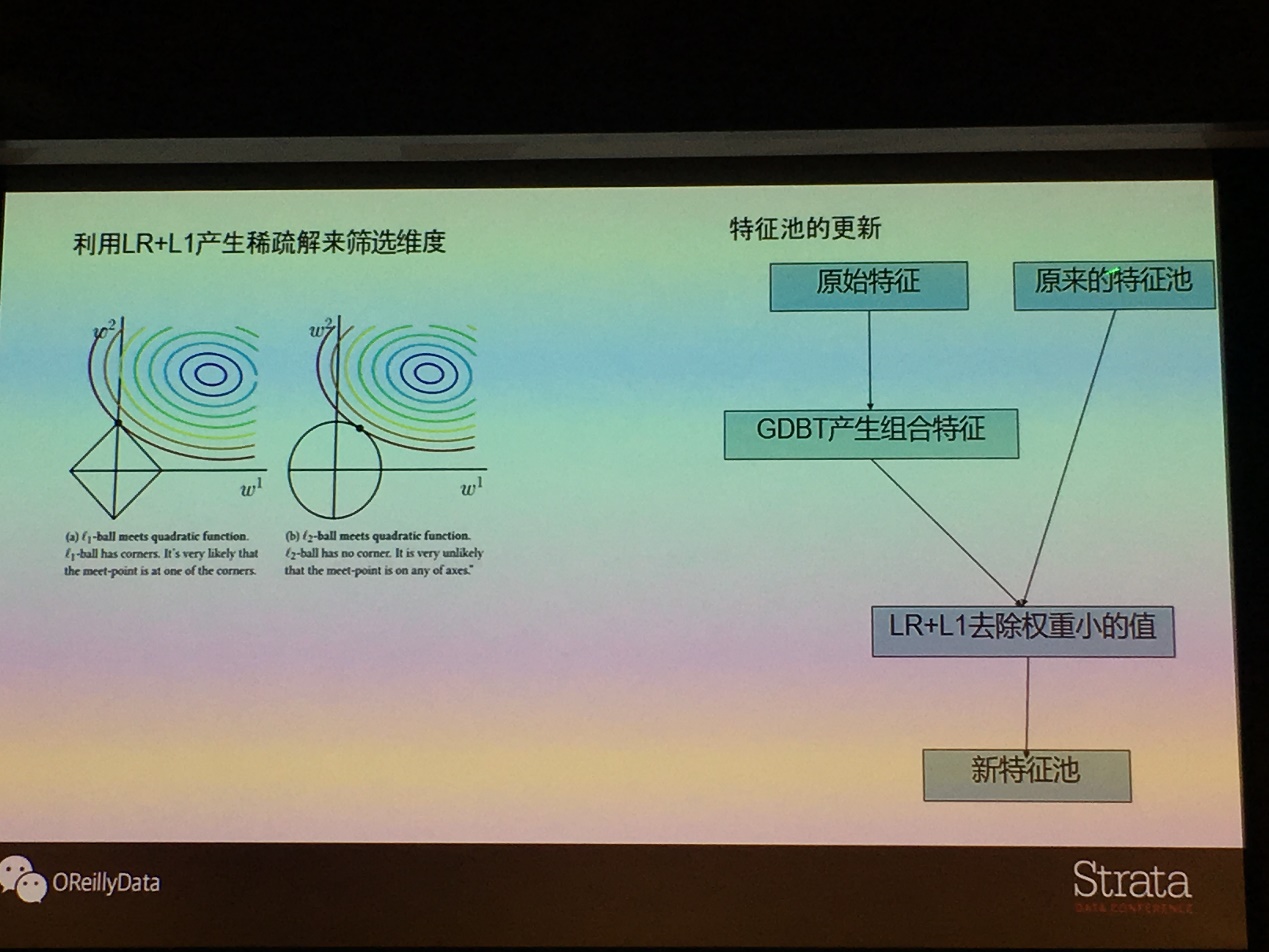
**（3）GBDT**

以年龄预测为例进行简单的工作过程说明。假设训练集只有4个样本ABCD，他们的年龄分别为8、10、22、24。其中A、B为小学二年级和四年级学生，C、D为应届毕业生和上班族。传统的回归树所需特征较多，很容易产生过拟合的问题。而在GBDT中，当我们限制树的深度和学习次数时，通过学习第一次分类器产生的残差，我们可以仅通过两次计算引入两个特征就能正确分类。这避免的过拟合的问题，同时在后续的学习中仅关注那些错分的权重。

在GBDT的预测过程中能够得到一系列的树形结构，具有重要的统计意义，代表潜在的特征组合。



**2.2. LR+L1产生稀疏解筛选、更新维度**



**（1）筛选维度，实现新老特征交替**

* 原特征池包括原始特征和之前处理好的组合特征原始特征（包括人工先验和机器学习）
* GBDT产生新的特征组合，加入到原特征池，产生一个更大的特征池（上亿维度）
* LR+L1算法除去权重小的特征，产生新特征池，从而实现了新老特征的交替。

**（2）为什么可以产生稀疏解？**

等价于拉格朗日中带约束的最优化问题，由上图所示，加上L1之后，在菱形的区间里，从等高线上来看，最优解更有可能向坐标轴偏。

**（3）为什么要进行特征更新？**

a. 如果不对新特征加以筛选，随着时间的推移，模型的训练的速度会越来越慢，对架构的压力也越大。

b. 从特征的实际含义来看，产品、潮流、偏好的更替也决定了更新特征池的必要性。以电商为例，旧产品下架，与这个商品有关的特征，就失去了意义；而新的时尚风格的出现，新的产品的上线，包含的新特征都具有新的学习价值。

**2.3. FM 因子分解机**

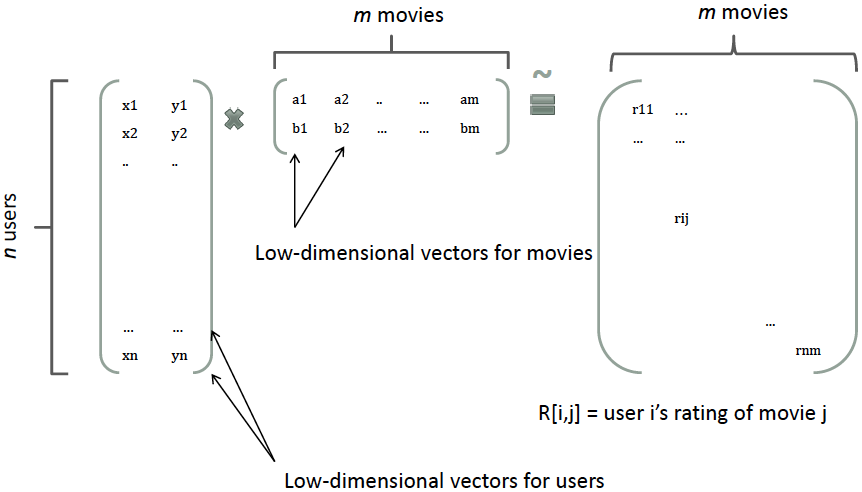
GBDT是在全量数据上进行数据的选择，难以应用到较稀疏的数据上。常见的场景包括：

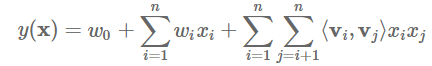
* 新用户上线，曝光数量不够，点击数据少，训练样本少
* 新产品上线，样本量不够（蓝色凉鞋卖得好，红色凉鞋很有可能也卖得好）

FM（Factorization Machine）是由Konstanz大学Steffen Rendle于2010年最早提出的，旨在解决稀疏数据下的特征组合问题。对于n个特征样本数量的二阶多项式模型来说，其组合特征的参数一共有n(n-1)/2 个，且两两独立。



然而，在数据稀疏性普遍存在的实际应用场景中，由于非零样本缺少，二次项参数的训练是很困难的。通过因式分解，将二次项参数矩阵分解为其隐向量矩阵的乘积，二次项的参数数量减少为 kn个，远少于多项式模型的参数数量，可以在样本稀疏的情况下相对合理地估计FM的二次项参数，这就是FM模型的核心思想。



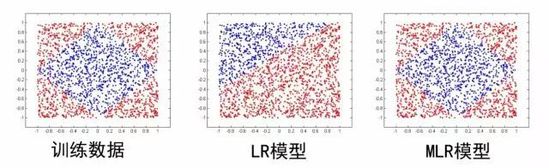


FM算法的好处是可以根据原有的数据的组合来判断没有出现过，或者出现数量不够的特征组合。

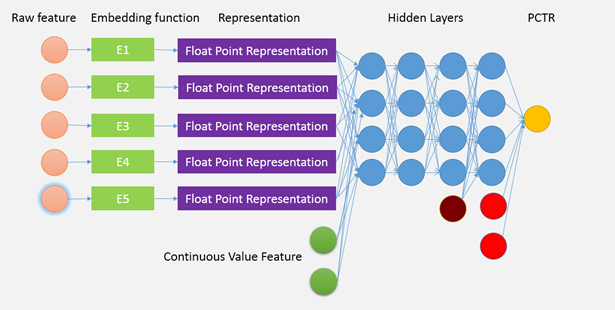
**3. MLR (mixed logistic regression, 混合逻辑回归)**

MLR可以看做是对LR的一个自然推广，它采用分而治之的思路，用分片线性的模式来拟合高维空间的非线性分类面。模型中的超参数分片数m可以较好地平衡模型的拟合与推广能力。当m=1时MLR就退化为普通的LR，m越大模型的拟合能力越强，但是模型参数规模随m线性增长，相应所需的训练样本也随之增长。因此实际应用中m需要根据实际情况进行选择。

MLR算法适合于工业级的大规模稀疏数据场景问题。



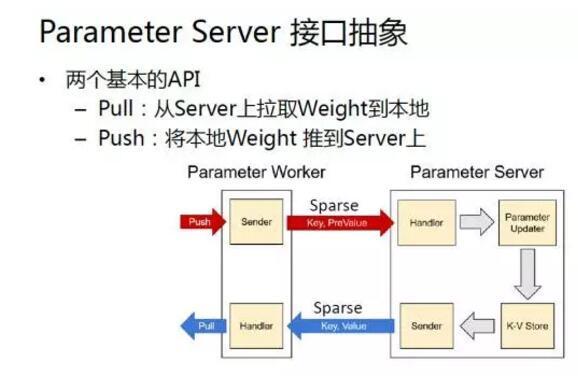
**4. DNN算法**



深度学习在CTR问题中的应用成为近年来比较火的尝试。由于广告系统包含海量离散特征，高达千万级别，如果全用 one-hot 表示，会产生维度爆炸，且十分稀疏。DNN 不支持这么多维的特征，而用人工构造高阶特征是不现实的。目前工业界方案是 FNN，即用 FM 做 Embedding，DNN 做训练，通过这种编码方式，进行比较大范围的模型压缩。

**5. 参数服务器**

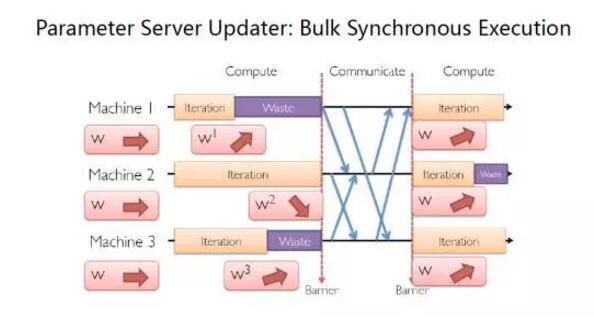
上文介绍了CTR在逻辑上的计算过程，然而针对工业级的数据量，在几亿的数据量和几十亿的维度上算法过程，不是单个电脑能够完成的，甚至也不是Spark、Haloop等分布式平台能够完成的。传统的计算构架很难满足企业的大数据分析需要，因此在架构上使用参数服务器，用它来解决大规模训练的问题。



框架由Server和Worker两个基本角色构成，通过Push和pull两个动作进行通讯。每一个Worker记载对应维度上的值，发送给Server，由Server审核后对每个权重进行更新，再重新发送给Worker，实现模型的迭代更新。

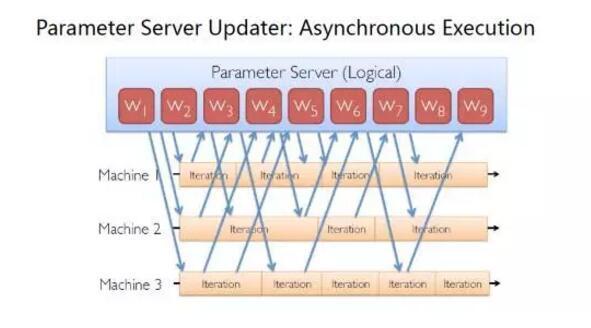
针对计算同步问题，参数服务器提供了三种等待的方式：

**（1）全等待（完全同步）**



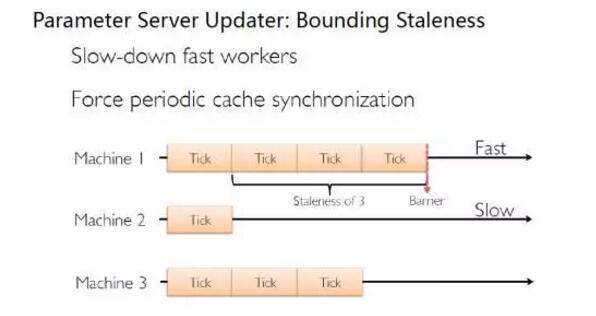
每一轮迭代都是所有的机器计算完成后再进行下一轮迭代，不会出现乱序执行导致的振荡问题，但是速度由最慢的那台机器决定，设备利用率较低。

**（2）不等待（完全异步）**



所有的机器自己执行完了进行下一轮，机器之间互相不可见，唯一的交互点是Parameter Server，优点是利用率高，没有等待的过程，缺点是容易产生震荡的情况，导致总体训练的时间延长。

**（3）弹性等待（有界异步）**



介于上述两种方案之间，称之为有界异步。最快的机器不能比最慢的机器快太多，两者之间有一定的界限。倍数（步长）可以自己设定，通过这种机制稳定快速地训练模型。

**参考（推荐）文章：**

CTR 预估模型的进化之路 <https://www.qcloud.com/community/article/701728>

GBDT基本概念和原理 <http://blog.csdn.net/u010159842/article/details/47276479>

深入FFM原理与实践 <https://tech.meituan.com/deep-understanding-of-ffm-principles-and-practices.html>

阿里妈妈首次公开自主研发MLR算法 <http://www.ebrun.com/20170615/235481.shtml>

淘宝参数服务器实践之路 <http://www.jianshu.com/p/496a23fdcf97>

闲聊DNN CTR预估模型 <http://www.52cs.org/?p=1046>