模型介绍：LightGBM(Light Gradient Boosting Machine)

GBDT (Gradient Boosting Decision Tree) 是机器学习中一个长盛不衰的模型，其主要思想是利用弱分类器（决策树）迭代训练以得到最优模型，该模型具有训练效果好、不易过拟合等优点。GBDT 在工业界应用广泛，通常被用于点击率预测，搜索排序等任务。GBDT 也是各种数据挖掘竞赛的致命武器，据统计 Kaggle 上的比赛有一半以上的冠军方案都是基于 GBDT。

LightGBM （Light Gradient Boosting Machine）是一个实现 GBDT 算法的框架，支持高效率的并行训练，并且具有以下优点：

1. 更快的训练速度
2. 更低的内存消耗
3. 更好的准确率
4. 分布式支持，可以快速处理海量数据

LightGBM 提出的主要原因就是为了解决 GBDT 在海量数据遇到的问题，让 GBDT 可以更好更快地用于工业实践。

模型原理

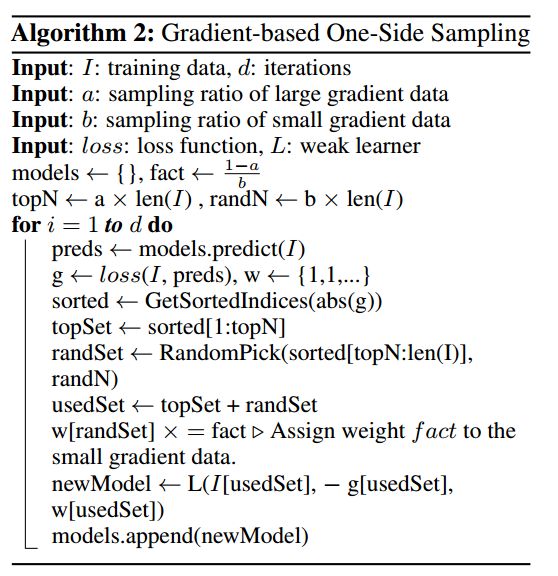
由于GDBT计算信息增益的过程中对每一个特征都需要扫描所有的样本，在面对大量数据或者特征维度很高时，效率和扩展性很难使人满意，LightGBM通过以下两个算法很好的解决这些问题：

1. GOSS(从样本角度）：排除大部分小梯度的样本，仅用剩余的样本计算信息增益。
2. EFB(从减少特征维度）：捆绑互斥特征，也就是他们很少同时取非零值（也就是用一个合成特征代替）

***GOSS***

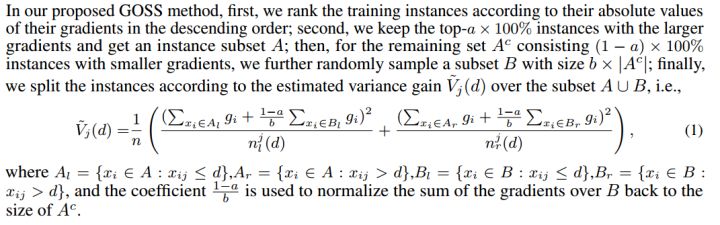
在AdaBoost中采用权重很好诠释了样本的重要性，GBDT没有这种权重，但是我们注意到每个数据样本的梯度可以用来做采样的信息。也就是，如果一个样本的梯度小，那么表明这个样本已经训练好了，它的训练误差很小了，我们可以丢弃这些数据。当然，改变数据分布会造成模型的精度损失。GOSS则通过保存大梯度样本，随机选取小梯度样本，并为其弥补上一个常数权重，这样，GOSS更关注训练不足的样本，同时也不会改变原始数据太多。

具体算法如下：



先根据梯度对样本进行排序，选取a\*100%的top样本，再从剩余数据中随机选取b\*100%的样本，并乘以fact系数放大。

信息增益计算公式如下：



***EFB***

高维数据一般是稀疏的，可以设计一种损失最小的特征减少方法。并且，在稀疏特征空间中，许多特征都是互斥的，也就是它们几乎不同时取非0值。因此，我们可以安全的把这些互斥特征绑到一起形成一个特征，然后基于这些特征束构建直方图，这样又可以加速了。

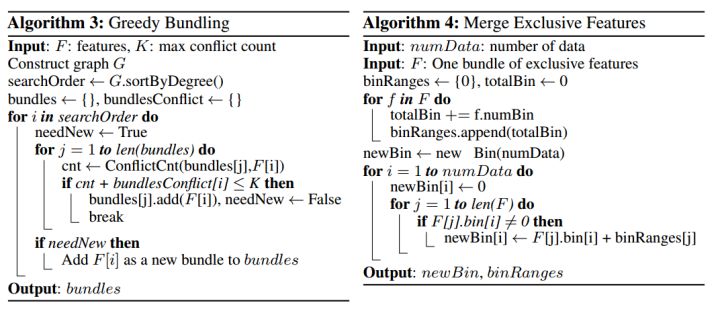
有两个问题待解决，如何判断哪些特征该绑到一起，如何构建绑定。

首先，转换到图着色问题。G=(V, E)，把关联矩阵G的每一行看成特征，从而得到|V|个特征，互斥束就图中颜色相同的顶点。图中点就是特征，边代表两个特征不互斥，也就是特征之间的冲突。如果算法允许小的冲突，可以得到更小的特征束数量，计算效率会更高。证明发现随机污染一小部分特征值，最多影响训练精度 ，是所有束中冲突最大的。通过选取合适的，我们可以很好的在效率和精度之间寻找平衡。

最后，排序就按照束的度来进行。当然，更一步优化是不够造图，直接根据非零值的数量排序，这个根据度排序很像，因为更多非0值意味着更高概率的冲突。更改了排序策略，可以避免重复。

第二个问题，合并特征，从而降低训练复杂度，关键是我

们可以确保原先特征值可以从特征束中识别出来。因为直方图存储的是特征的离散桶，而不是连续值，我们可以通过把互斥特征放到不同桶，从而构造一个特征束。这可以通过添加偏移实现。如，假设我们有2个特征在一个特征束中，原先特征A的范围为[0,10)，特征B的范围为[0,20)，我们给特征B加上一个偏移10，它就变成[10,30)，这样我们就可以执行安全的合并了，用特征束[0,30)代替特征A和B。具体算法如下。



总结与展望

手机用户的人口属性（如性别、年龄、常驻地等）数据一方面可以被用于个性化推荐服务，提升用户体验，另一方面可以用于手机用户群画像分析，帮助厂商了解产品的人群定位，优化产品设计。

本次实验我们使用了Light GBM模型训练了一个用户人口属性预测的模型，模型可以根据输入的用户性别和常住地信息，手机的硬件信息，软件信息，app使用历史记录等信息，输出预测的年龄结果。实验没有进行数据的特征提取工程，而是直接使用数据的原始信息进行训练，训练的模型结果经检验，预测准确率0.400900。

赛事网站排行榜准确率最高达到了0.662040，使用Keras搭建神经网络+XGboost做集成。后续我们的实验准确率可以通过特征提取、更换模型或者持续的参数调整来提升。

参考文献

1. Ke, Guolin and Meng, Qi and Finley, Thomas and Wang, Taifeng and Chen, Wei and Ma, Weidong and Ye, Qiwei and Liu, Tie-Yan. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30,2017:3146-3154.
2. GIGIX极客参赛数据：

[https://developer.huawei.com/consumer/cn/activity/devStarAI/algo/competition.html#/preliminary/info/digix-trail-02/data](https://developer.huawei.com/consumer/cn/activity/devStarAI/algo/competition.html" \l "/preliminary/info/digix-trail-02/data)