《用户人口属性预测》华为极客比赛个人报告

姓名：李平

学号：16340121

项目分工：

LightGBM模型介绍

LightGBM模型原理

实验总结与展望

参考文献整理

个人工作介绍：

负责论文部分撰写工作，查找并参考了Ghost\_Hzp“LightGBM算法总结”的博文，借鉴了龙大鱼的知乎文章“LightGBM原理之论文详解”辅助理解了“LightGBM: A Highly Efficient Gradient Boosting Decision Tree”一文。

GBDT (Gradient Boosting Decision Tree)和XGBoost是机器学习中采用频率非常高的模型。但是当面对大规模数据高纬度特征时，时间和计算机资源的限制会变得愈发突出。由于两种算法对于计算特征的信息增益是，对于每个特征都要遍历整个数据集，时间复杂度为O(#DataSet\*#features)。如果将训练数据全部装入内存，则会限制训练数据的大小。如果不装入内存，频繁的读写操作又会消耗大量的时间，不能满足工业需求。

LightGBM通过down sample 数据集和特征来降低数据处理消耗。通过两个算法实现两个方法，GOSS和EFB.

GOSS算法对数据集大小进行欠采样。将数据集通过梯度对数据集进行裁剪，通过保留大梯度样本，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。

EFB算法对数据特征进行欠采样。通过绑定互斥特征（几乎不同时为0），替换为合成特征，避免了对0特征值的无用计算，更降低了样本维度，从而减少了计算成本。

关于算法更加详细的介绍，参见小队报告2.2LightGBM模型。

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. <https://blog.csdn.net/weixin_39807102/article/details/81912566>
3. <https://zhuanlan.zhihu.com/p/35155992>