# XGBoost在三分类任务上的应用

## 任务背景

课程预测三分类任务，整个训练集1253条，9列特征。测试集100条，需要对训练集进行建模后，预测测试集中每个样本的类别——三分类任务。

## 数据建模

XGBoost与 GBDT相同，但是在此基础上一些优化：使用二阶导数使损失函数更精准；加入正则项避免树过拟合；Block存储可以并行计算等。正是由于XGBoost的种种特质使得其在工业应用十分广泛，包括商店销售预测，高能物理事件分类，网站文本分类，顾客行为预测，动作检测，广告点击通过率预测，风险预测，大规模的在线课程辍学率预测。

XGBoost的目标函数由损失函数和正则化项两部分组成：

文本

描述已自动生成

对于XGBoost目标函数的优化主要包括三个步骤：第一步：二阶泰勒展开，去除常数项，优化损失函数项；第二步：正则化项展开，去除常数项，优化正则化项；第三步：合并一次项系数、二次项系数，得到最终目标函数。

XGBoost的开源库也比较多，使用频率最高的主要是LightGBM和XGBoost两个开源库，二者都是基于决策树提升(Tree Boosting)的工具，都拥有对输入要求不敏感、计算复杂度不高和效果好的特点，但LightGBM的特点主要在于：1. 减小数据对内存的使用，保证单个机器在不牺牲速度的情况下，尽可能地用上更多的数据；2.减小通信的代价，提升多机并行时的效率，实现在计算上的线性加速。更直观的说，XGBoost使用的是pre-sorted算法，能够更精确的找到数据分隔点；LightGBM使用的是histogram算法，占用的内存更低，数据分隔的复杂度更低。本文采用LightGBM库实现xgboost的建模。

## 模型训练

本任务提供了训练集，没有单独的验证集，因此利用sklearn中的train\_test\_split对数据进行切分，按照8:2的占比切分训练集和验证集，最终得到1002条训练集和251条验证集。模型参数设置如下：

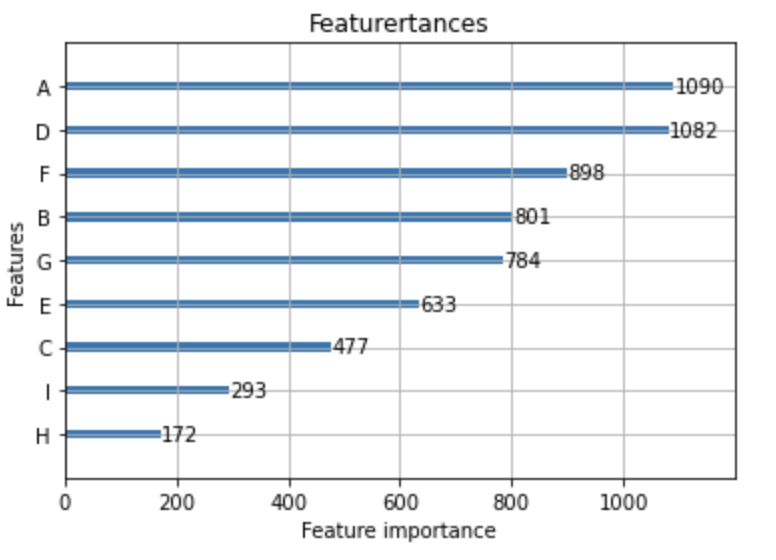
|  |  |
| --- | --- |
| parameter | definition |
| n\_estimators | 500 |
| objective | multiclass |
| boosting\_type | gbdt |
| metric | multi\_logloss |
| num\_class | 3 |
| num\_leaves | 8 |
| max\_depth | 8 |
| learning\_rate | 0.1 |
| lambda\_l1 | 0.1 |
| lambda\_l2 | 10 |

模型迭代500轮，其中在验证集上的多分类loss如图：

图形用户界面, 应用程序

描述已自动生成

可以看到模型在100轮左右基本已经收敛完毕，loss下降到0.2附近。模型在320轮达到最优，其验证集多分类loss=0.29。对于数据中的9列特征，模型对不同特征的重要程度分析如下：



可以看出最重要的两个特征为A列和D列，其中I列和H列的重要程度基本可以忽略。

此外模型在验证集上的效果如下：

表格

描述已自动生成

可以看出模型在-1和1类目上效果比较好，在0类目上效果不理想，这可能是类目不平均导致，本身训练集中0类的数据较少，因此模型对该类目的分布学习有限。而且0类目的效果拉低了宏平均的结果，而微平均的f1为0.88，与宏平均的f1值相差8个点。因此对于类目不平衡的数据主要还是要关注宏平均的结果。