万象洞察 | 10分钟了解GBDT+LR模型的来龙去脉

原创 刘君媛 中诚信征信 2018-01-26

小象说：

如果基础模型的效果差强人意，适当的改进往往可以提升模型学习能力，而基础模型的组合就是一种简单有效的常用方式。GBDT+LR模型作为一种混合模型，既带有GBDT树模型的天然特征处理属性，又不失LR广义线性模型方便易用的特点，犹如男女搭配，各显其长。

看一看

1、算法背景

2014年Facebook发表了一篇介绍将GBDT+LR模型用于其广告推荐系统的论文，之后，无论是Kaggle竞赛还是淘宝商品推荐，都有借鉴该论文中的GBDT+LR模型组合思想，即通过GBDT来发掘有区分度的特征和组合特征，来代替人工组合特征。

对于支撑互联网半壁江山的广告收入，推荐系统和CTR预估于其技术框架中占据重要地位，而LR模型则是其中最为常用的模型。

LR模型有以下特点：

- 计算复杂度低；
- 易于并行化处理；
- 易于得到离散化目标值0或1，利用sigmoid函数将传统线性模型的输出值映射到(0,1)区间；
- 学习能力限于线性特征，需要提前进行大量的特征工程得到有效的特征及特征组合。

输入LR模型的特征很重要，但是特征组合不能直接通过特征笛卡尔积获取，只能依靠人工经验。故而如何自动化进行特征工程，规范化LR模型使用流程是一个值得研究的问题。

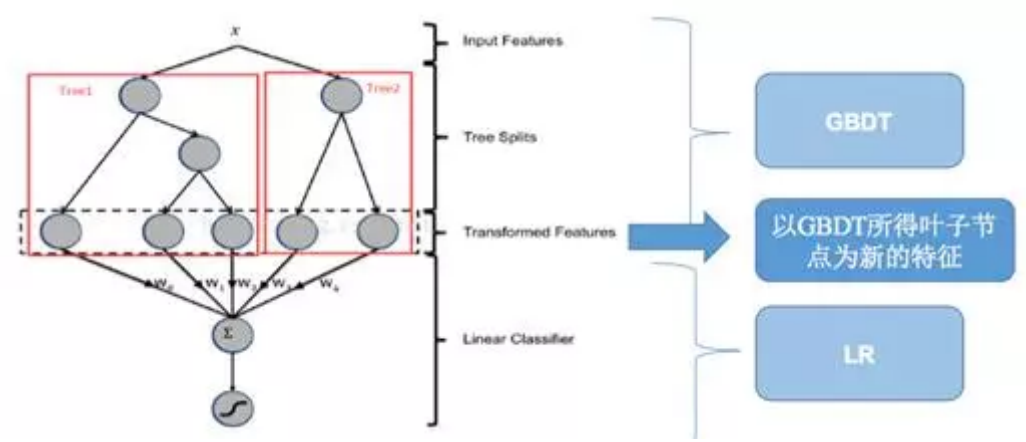
GBDT作为一种常用的树模型，可天然地对原始特征进行特征划分、特征组合和特征选择，并得到高阶特征属性和非线性映射。从而可将GBDT模型抽象为一个特征处理器，通过GBDT分析原始特征获取到更利于LR分析的新特征。这也正是GBDT+LR模型的核心思想——利用GBDT构造的新特征来训练LR模型。

2、算法原理及实现

前面简单介绍了GBDT+LR模型的产生背景和核心思想，接下来将会更为详细地描述GBDT+LR模型的算法组合思想和简单实现流程。

2.1、算法组合——stacking

stacking方法有些类似于农业中的嫁接，通过stacking方法组合的模型亦类似于嫁接植物，例如，解决了人类吃饭问题的杂交水稻。



如上图所示，GBDT算法的图示部分形如一棵倒过来的树，其根部即代表训练GBDT算法的原始数据集，经过树算法对原始数据的切分，可得到代表不同新特征的叶子节点。

再将GBDT所得的叶子节点输入LR算法，经过线性分析和sigmoid映射，即可得到模型分类结果。

以上的模型组合方式就是stacking方法，即将学习层模型对原始数据所得的预测结果作为新的特征集，并输入给输出层模型得到分类结果。Facebook论文中的GBDT+LR模型就采用了GBDT算法作为学习层，以LR算法为输出层。

2.2、算法流程& 代码简单实现

在这一部分中，GBDT+LR算法的代码实现语言为python，使用了sklearn包中的GradientBoostingClassifier和LogisticRegression函数作为GBDT模型和LR模型。

将训练集记为（X,y），其中X为原始特征，y为目标变量。

- 数据预处理

  对变量取值中的中英文字符、缺失值和正负无穷值进行处理。

- 数据集划分

  为了降低过拟合的风险，将训练集中的数据划分为两部分，

  一部分数据用于训练GBDT模型，另一部分数据通过训练好

  的GBDT模型得到新特征以训练LR模型。

  ```
  fromsklearn.model import train_test_split
  X_gbdt,X_lr,y_gbdt,y_lr= train_test_split(X,y,test_size=0.5)
  ```

- GBDT特征转化

首先，通过sklearn中的GradientBoostingClassifier得到GBDT模型，然后使用GBDT模型的fit方法训练模型，最后使用GBDT模型的apply方法得到新特征。

```
fromsklearn.ensemble import GradientBoostingClassifier
        gbdt = GradientBoostingClassifier()
        gbdt.fit(X_gbdt,y_gbdt)
        leaves = gbdt.apply(X_lr)[:,:,0]
```

- 特征独热化

使用sklearn.preprocessing中的OneHotEncoder将GBDT所得特征独热化。

```
fromsklearn.preprocessing import OneHotEncoder
        featutes_trans =OneHotEncoder.fit_transform(leaves)
```

- LR进行分类

用经过离散化处理的新特征训练LR模型并得到预测结果。

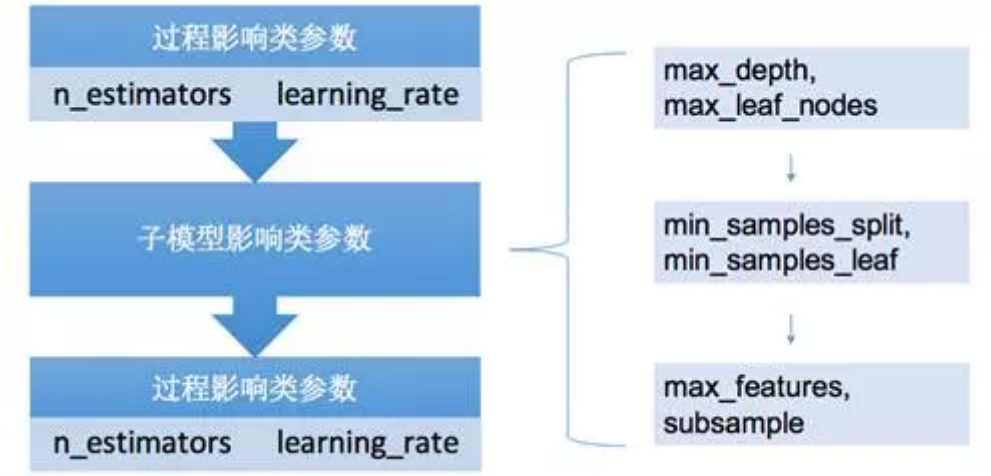
```
fromsklearn.linear_model import LogisticRegression
lr= LogisticRegression()
      lr.fit(features_trans,y_lr)
      lr.predict(features_trans)
      lr.predict_proba(features_trans)[:,1]
```

2.3、调参方法简述

构建了模型框架后，模型中的函数参数调整也是必不可少的。对模型参数的适当调整，往往可以有效提升模型的效果。

由于GBDT+LR模型无法整体使用GridSearchCV函数，所以调参时使用sklearn.cross_validation中的StratifiedKFold方法，将数据集进行k折交叉切分，然后以auc值为模型评估指标，对混合模型进行调参。

调参时的重点为GradientBoostingClassifier函数，可用如下图所示的调参顺序进行调参。



其中，n_estimators和learning_rate应该联合调参。

2.4、模型效果展示

在介绍了GBDT+LR模型的原理和实现流程之后，我们以一个1.5万条的数据样本为例，来比较直观地认识一下模型效果。

我们分别使用LR模型和GBDT+LR模型对样本数据集进行学习，通过模型所得的auc值和ks值，来评估和比较模型的效果。

模型	AUC_train	AUC_test	delta_AUC	ks_train	ks_test	delat_ks
LR	0.712	0.702	1.0%	0.331	0.318	1.3%
GBDT+LR	0.894	0.873	1.9%	0.618	0.592	2.6%

如上图所示，可知GBDT+LR模型的效果要更好一些，即GBDT所得的新特征的确更适合LR模型的分析。

3、算法引申

前面的内容描述了Facebook论文中GBDT+LR混合模型的算法原理并附有简单实现代码。然而，模型并不可孤立地比较好坏，模型的应用也要和应用场景及数据质量互相照应。

这一部分将会简单提供一些GBDT+LR混合模型的引申思路，希望对大家实际使用时有所裨益。

- 用FFM模型替代LR模型：
  直接将GBDT所得特征输入FFM模型；
- 用XGBoost模型替代GBDT模型；
- 将stacking模型学习层中的GBDT交叉检验；
- GBDT和LR模型使用model fusion，而不是stacking
- ……

—— THE END ——

THANKS

文 | 中诚信征信 追AI骑士 刘君媛

联系合作：ccx@ccx.cn

点击阅读原文申请产品试用

中诚信征信简介