论文 | GBDT+LR作点击率预估的方法

原创 crackcell 有多少人工就有多少智能 2018-03-27

Table of Contents

1 前言

这篇文章是工业界点击率预估方向近几年比较重要的论文之一，提出了GBDT+LR的混合模型。用GBDT来进行特征值加工，对连续值特征有很不错的效果。

2 模型结构

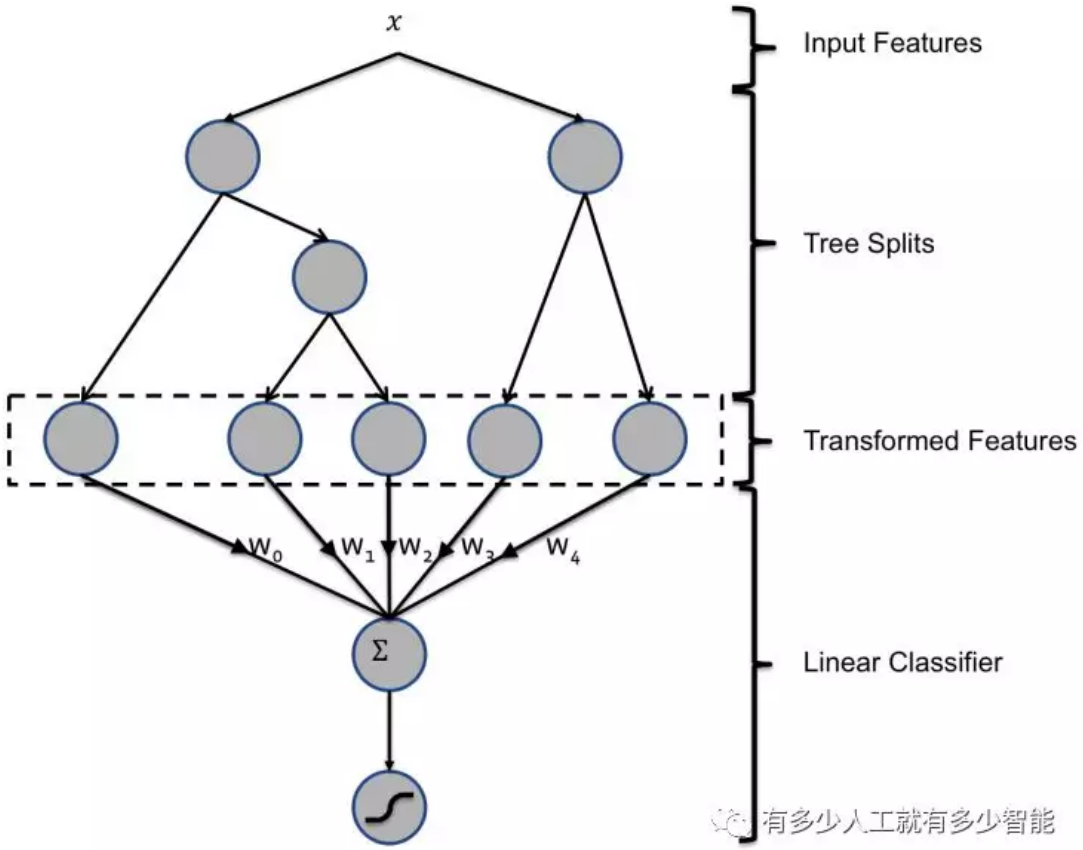

Figure 1: 混合模型结构

拿论文中的图举例。这个结构中有2棵树，树1有3个叶子节点，树2有2个叶子节点。特征x一个输入向量经过2棵树的处理，分别落到树1的第2个叶子节点和树2的第1个叶子节点，那么它的结果就是[0,1,0,1,0]。从这个角度讲，决策树起到的是一个编码器的作用，将连续值编码成一个one-hot向量。特征值在树中走过的每条路径就是一条特征处理规则。

最后将这个one-hot向量作为LR的输入。

2.1 特征编码器

先比较下GBDT和LR本身的一些特点：

1. 模型刻画能力：GBDT能更好刻画头部样本，LR能更好刻画长尾
2. Variance vs. Bias：GBDT偏variance，LR偏bias。所以需要注意控制GBDT的过拟合，而对于LR需要提升拟合能力

那么GBDT作为一个编码器，起了那些作用呢：

1. 对连续特征进行了离散化
2. 一棵树里面一条游走的路劲就是一个特征组合
3. GBDT树生长过程也是一个特征选择的过程

4. 多棵树表示多种特征组合的方法。早期树产生的特征组合侧重区分度，后期则侧重于关注分类效果不好的样本

3 模型更新

3.1 离线批量

传统的batch训练方式，不再赘述。

3.2 在线学习

文章中对LR的部分引入在线学习，来提高模型的时效性。原文中引入了基于SGD的在线学习方法，并对几种动态调节学习率的方法进行了比较。关于在线学习的部分，后续单独一个开一个帖子来讲。

4 调参

4.1 树的数量

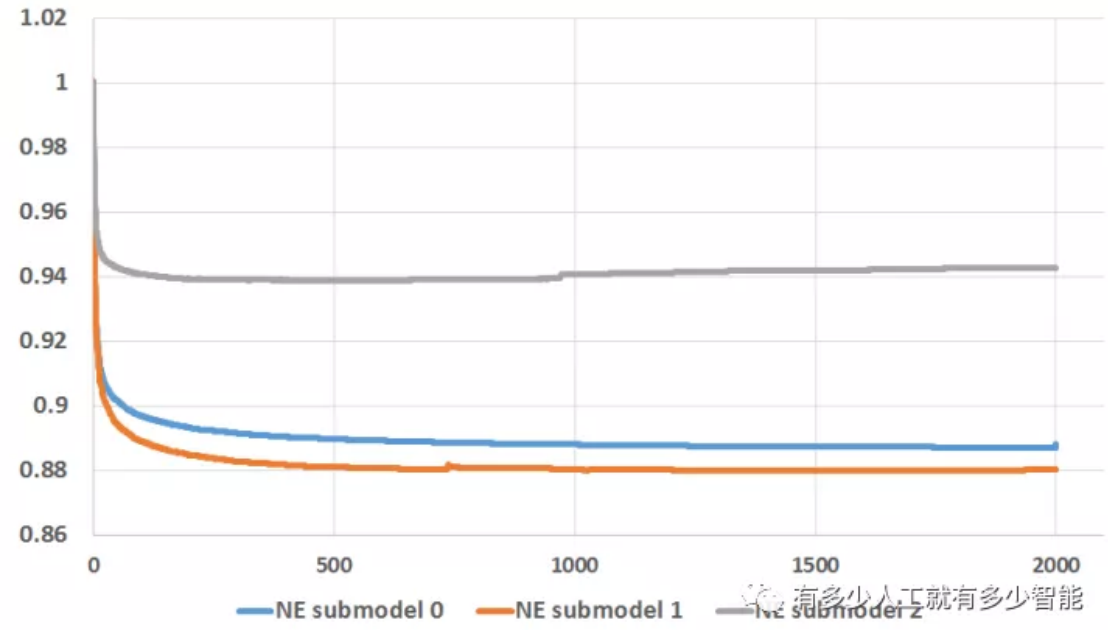

Figure 2: 模型树数量和NE的关系

文中尝试树的数量和NE的关系。统计熵下降的幅度累积和树数量的关系，大多数由500树贡献。同时NE submodel 2后续出现了效果衰退，这是由于过拟合。

4.2 特征数量

首先讨论下特征重要性的衡量。文中使用Boosting Feature Importance来度量特征重要性，它记录了在树分裂中用到的某个特征贡献的总的平方误差和的总和。

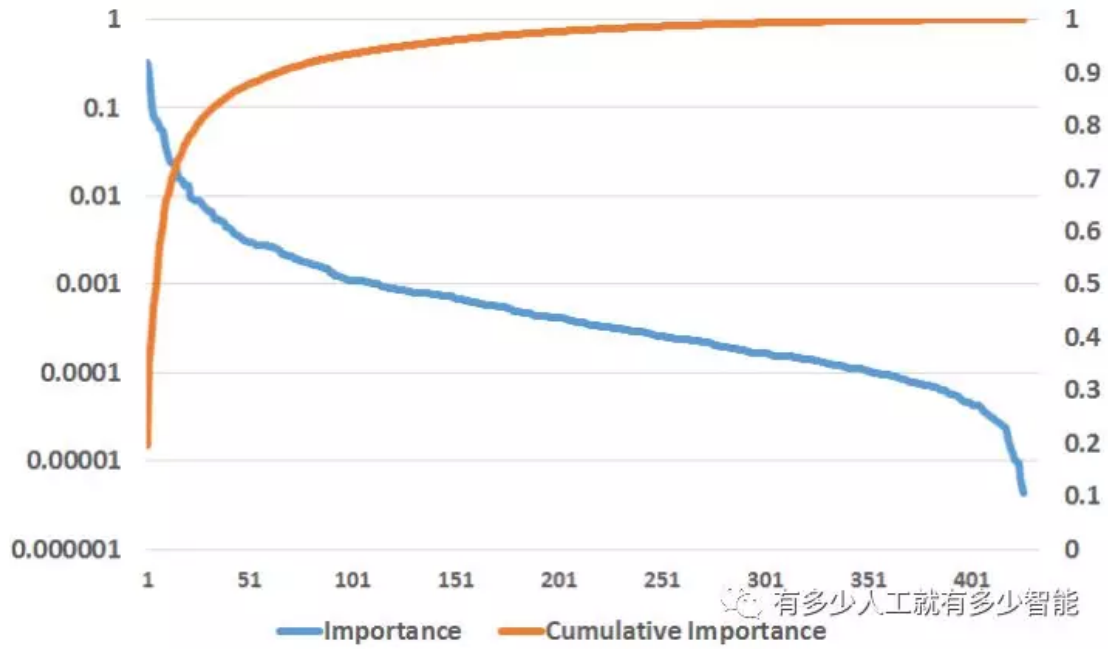

Figure 3: 特征数量和重要性的关系

从图中能知道，头部特征贡献了大多数的模型提升。结合这样的关系，可以指导特征的裁剪。

4.3 特征类型的选择

文中比较了两种特征：历史统计特征和场景特征。历史统计特征更加重要。但需要特别注意的是，场景特征对于新用户能启动有重要的意义。

4.4 减少训练数据量

文中对负样本作了降采样（Negtive down sampling）。尝试了不同的采样率，选择了一个提升比较大的。

4.5 线上模型较准

降采样能在线下带来训练速度的模型效果的提升，但会引入偏差，在线上预测的时候，需要进行校准（Calibration）。

具体计算方法是：

$$q = \frac{p}{p + \frac{1-p}{w}}$$

- p是线上预测时的正式预测值
- w是降采样率，也就是负样本和正样本采样率的比值

阅读原文