


# Kaggle知识点：树模型特征Embedding

尤而小屋 2022-07-18 00:00 发表于北京

以下文章来源于Coggle数据科学，作者Coggle



**Coggle数据科学**  
Coggle全称Communication For Kaggle，专注数据科学领域竞赛相关资讯分享。

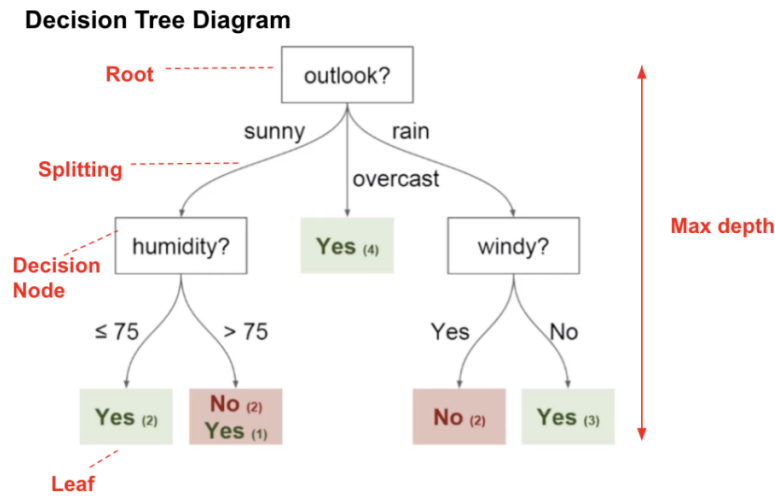
在对数据进行编码的过程中，经常会遇到一些非结构化的字段（如列表、文本），或者高维稀疏的字段。

在使用树模型的过程中，上述字段对树模型很不友好，会增加树模型的训练时间，一般情况需要通过人工特征提取，然后进行。

有没有一种可以适合树模型编码的操作呢？在树模型中可以通过叶子节点的次序作为进行编码，在Kaggle中称为Tree Categorical Embedding。

## Tree Categorical Embedding

在训练完树模型之后，可以通过对模型进行预测，通过节点逻辑的判断从根节点到叶子节点。



此时叶子节点中包含的样本类别（或标签均值）为最终的预测结果。这里想要具体的index，也就是样本预测到第几个叶子节点中。

在XGBoost中，拥有多棵树。则一个样本将会被编码为多个index，最终可以将index作为额外的类别特征再加入到模型训练。

## 具体API

### XGBoost

使用Learning API，设置pred\_leaf参数

```
import xgboost as xgb
from sklearn.datasets import make_classification

X, Y = make_classification(1000, 20)
dtrain = xgb.DMatrix(X, Y)
```

```
dtest = xgb.DMatrix(X)

param = {'max_depth':10, 'min_child_weight':1, 'learning_rate':0.1}
num_round = 200
bst = xgb.train(param, dtrain, num_round)
bst.predict(dtest, pred_leaf=True)
```

### LightGBM

使用sklearn API或者Learning API, 设置pred\_leaf参数

```
import lightgbm as lgb
from sklearn.datasets import make_classification

X, Y = make_classification(1000, 20)
dtrain = lgb.Dataset(X, Y)
dtest = lgb.Dataset(X)

param = {'max_depth':10, 'min_child_weight':1, 'learning_rate':0.1}
num_round = 200
bst = lgb.train(param, dtrain, num_round)
bst.predict(X, pred_leaf=True)
```

### CatBoost

使用calc\_leaf\_indexes函数

```
import catboost as cab
from sklearn.datasets import make_classification

X, Y = make_classification(1000, 20)
clf = cab.CatBoostClassifier(iterations=200)
clf.fit(X, Y)
clf.calc_leaf_indexes(X)
```

使用细节

1. **leaf index** 预测维度与具体树个数相关, 也就是与具体的round相关。
2. **leaf index** 的预测结果为类别类型。
3. **leaf index** 建议交叉验证编码, 避免自己训练并编码自己。

交叉验证实现: <https://www.kaggle.com/mmueller/categorical-embedding-with-xgb/script>





- MySQL必须掌握4种语言！
  - 三大树模型实战乳腺癌预测分类
  - Plotly+Pandas+Sklearn：实现用户聚类分群！
  - 用户群组分析，Python实现！
  - Kaggle可视化：黑色星期五画像分析
  - kaggle实战：可视化深度探索苹果AppStores
  - kaggle实战：6大回归模型预测航班票价
- 尤而小屋，一个温馨的小屋。小屋主人，一手代码谋求生存，一手掌勺享受生活，欢迎你的光临



**尤而小屋**  
尤而小屋，一个温馨且有爱的小屋🏠 小屋主人，一手代码谋求生存，一手掌勺享受生...  
261篇原创内容

公众号

喜欢此内容的人还喜欢

文本生成系列之retrieval augmentation（基础篇）  
NLP日志

Arcgis DEM融合与掩膜提取  
祥帅的小屋

一个有趣的小问题，threejs+chrome加载3d模型  
弱学狗的中间地带