比赛那些事儿

张海鹏

江南大学

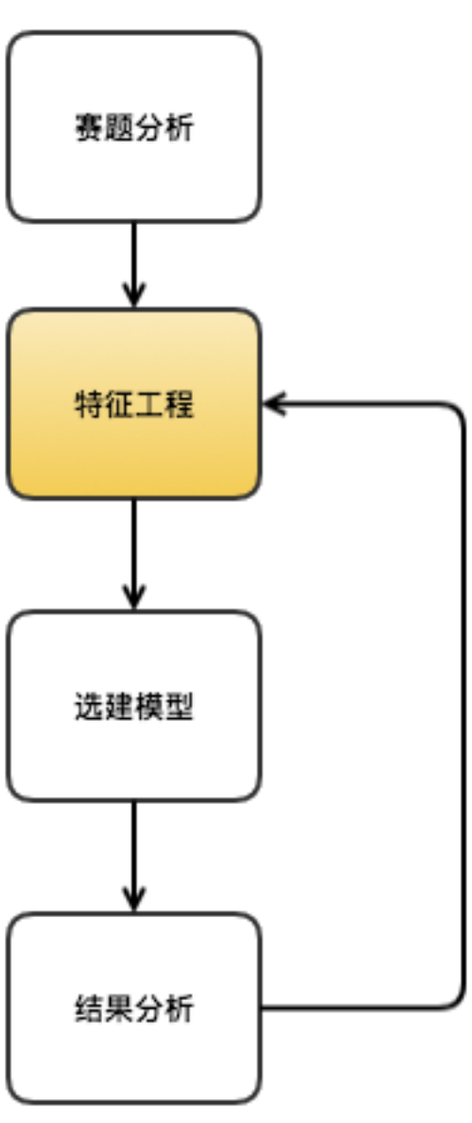
2017-11-18

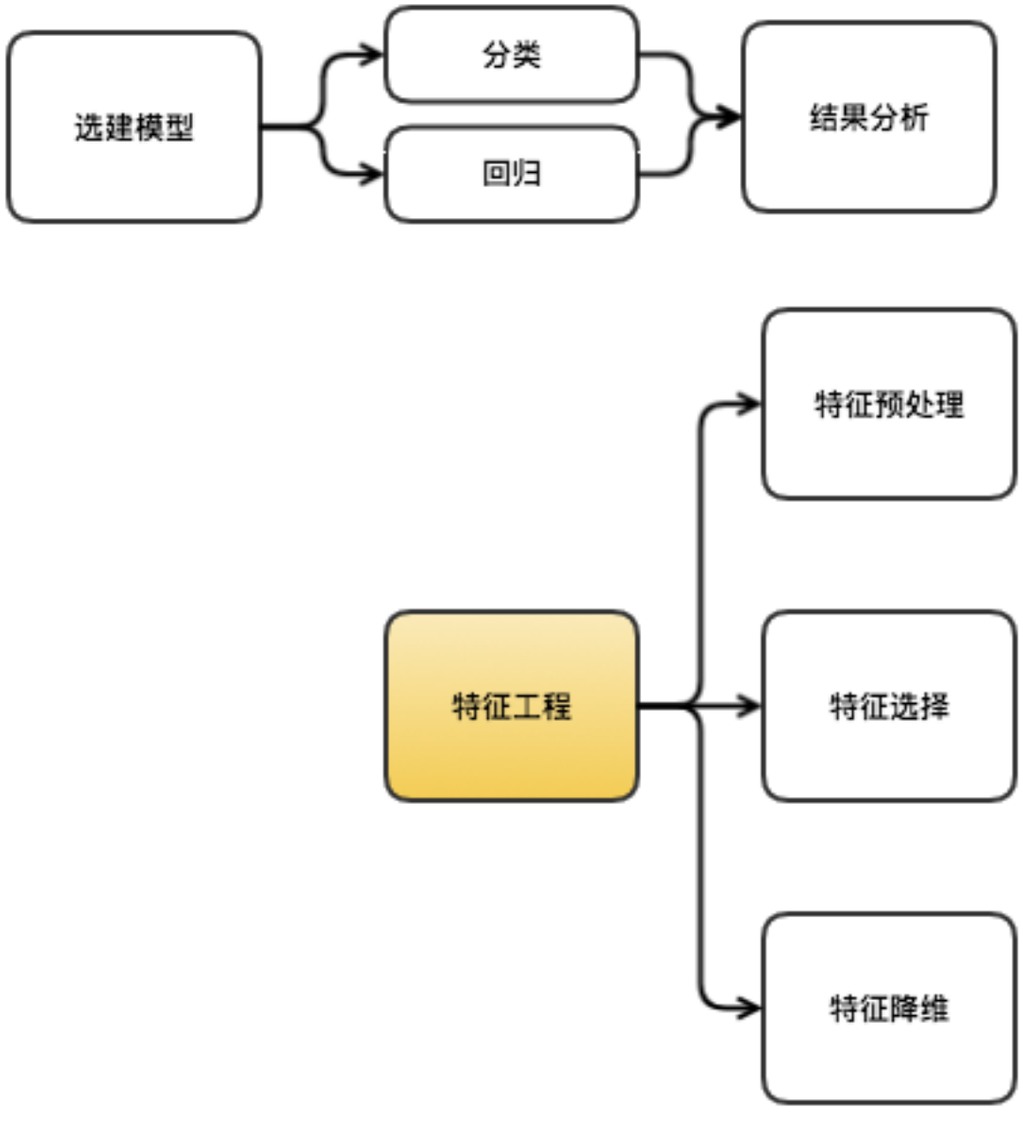
参赛经历

- “华为杯”第十四届中国研究生数学建模竞赛-基于监控视频的前景目标提取(国三)
- 全国并行应用挑战赛-基于NLP的金融营销活动情感分析(人工智能组银奖)
- 天池-蚂蚁金服: 商场中精确定位用户所在店铺(Rank:4% 119/2481-448/2905)
- 全球AI挑战赛-虚拟股票趋势预测
- 科赛网-携程: 出行产品未来14个月销量预测(Rank:15% 30/200)
- 天池-蚂蚁金服: IJCAI-2017口碑商家客流量预测(Rank:13% 541/4046)
- 科赛网-达观数据: 2017“达观杯”个性化推荐算法挑战赛

目录

工作流





代码示例

```
import numpy as np
import pandas as pd
import lightgbm as lgb
from sklearn.datasets import load_svmlight_file
from sklearn.metrics import accuracy_score
```

1

```
ltrain = lgb.Dataset(train_X, label=train_Y)
lval = lgb.Dataset(val_X, label=val_Y)
```

2

```
num_class = int( max(label) ) + 1
num_boost_round = 70
params = {  'task': ‘train',
        'boosting_type': 'gbdt',
        'objective': 'multiclass',
        ‘metric’: {'multi_error'},
        'num_class': num_class,
        'num_leaves': 31,
        'learning_rate': 0.02,
        'feature_fraction': 0.9,
        'bagging_fraction': 0.8,
        'bagging_freq': 4,
        'verbose':0,
        #'device': 'gpu'
      }
```

3

```
num_boost_round = 70
model = lgb.train(params, ltrain, num_boost_round,
                  valid_sets=lval,
                  early_stopping_rounds=5)
```

4

```
y_pred = model.predict(test_X,
                  num_iteration=model.best_iteration)
```

5

```
acc = accuracy_score(test_Y, np.argmax(y_pred, axis=1))
print('Val acc is {}’.format(acc))
#with open('test.csv', 'a') as f:
#    f.write(str(mall_file.split(‘.’)[0])+
','+str(acc)+'\n')
```

6

```
model_path = model_dir+str(mall_file.split('.')[0] )
+’.model'
model.save_model(model_path,
 num_iteration=model.best_iteration)
```

7

经验总结

1.Gabage In, Gabage Out

2.(复杂特征+简单模型) / (简单特征+复杂模型)

3.结合具体问题，自定义metric/loss函数

4.模型融合：模型精度差别小+模型差异大

5.Try!!!特征工程没有系统化的理论

工具介绍

语言: Python3.X/Shell

IDE: VIM+PDB/Jupyter Notebook

常用库: scikit-learn/pandas/scipy/numpy/matplotlib/seaborn/

statsmodels/xgboost/lightGBM/Catboost

模型: SVM/RandomForest/GBDT

计算资源：深度学习机(2台)+计算集群(1台管理节点+4台计算节点)

学习资源

1.打比赛，水群，交流，和大佬吹B

2.科赛网，天池的赛后总结文章，答辩视频

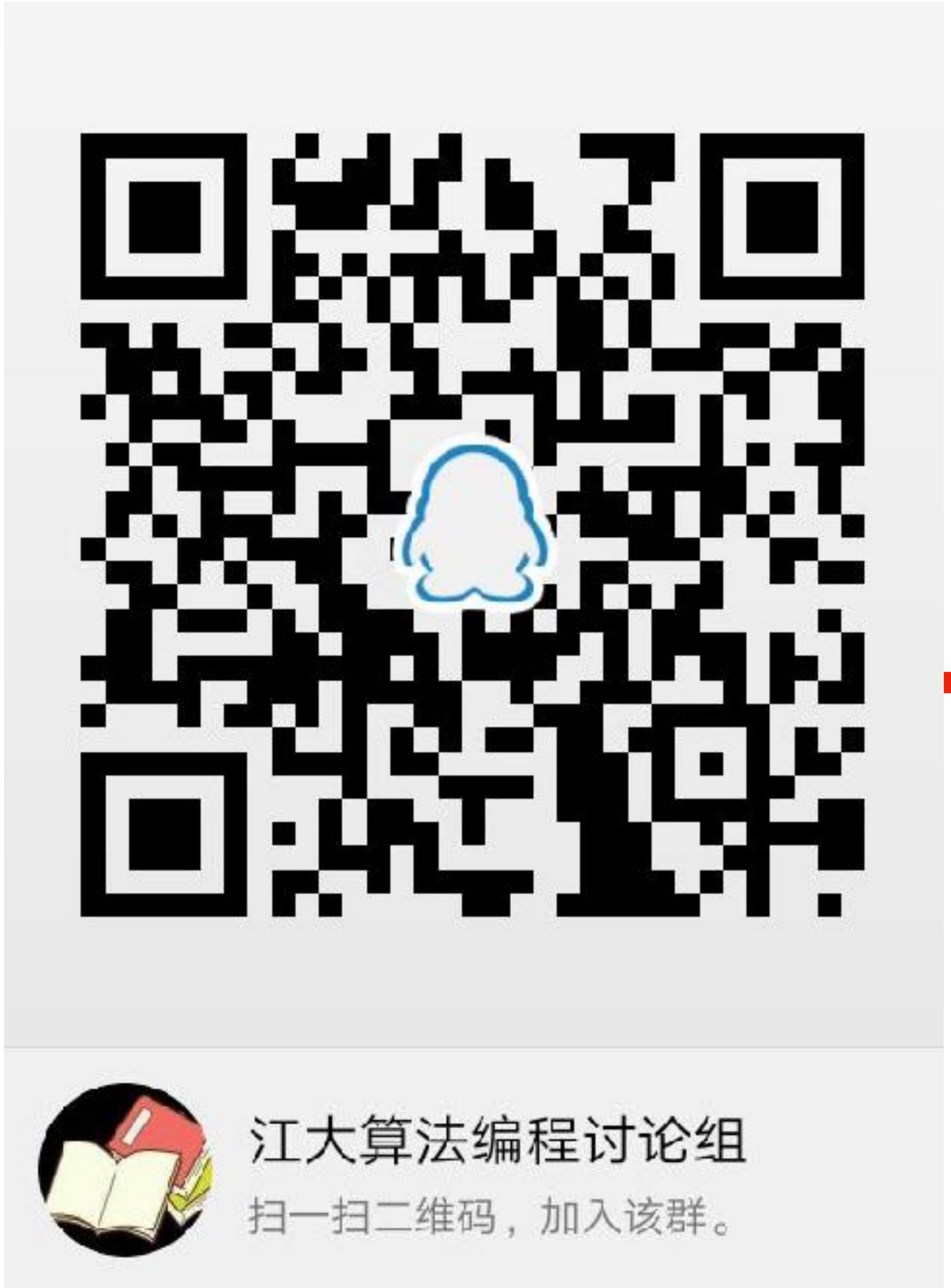
3.Kaggle的Kernel

4.XGBoost/LightGBM的原始paper，官方文档(sphinx)

5.台大的机器学习技法(KDD竞赛技巧)

6.七月在线

7.我的博客(https://zhpmatrix.github.io/)



TKS

大家有啥要交流的吗?

欢迎加入江大算法编程讨论组，
群号码：583307097