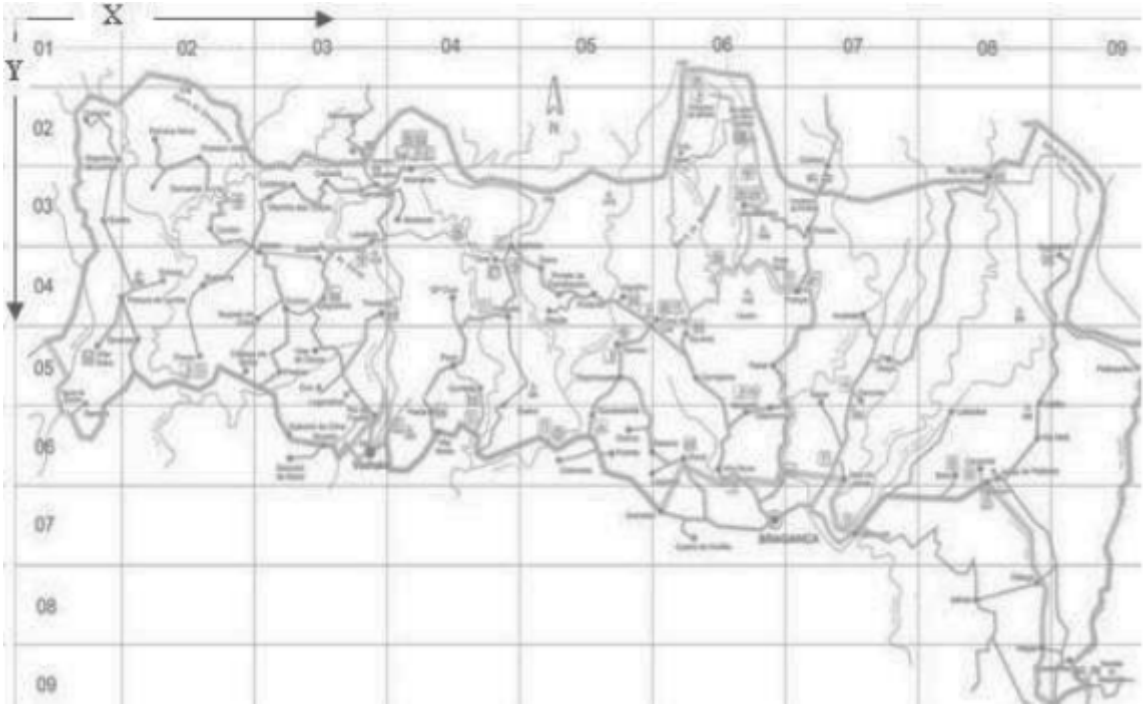
NNI进阶项目Feature Engineering文档

数据与特征是机器学习效果的关键要素；为了能够得到更好的预测模型，获取更好的训练数据正是特征工程所要实现的目的。



本实验所选择的数据集为森林火灾Forest Fires数据集。该数据集采用了葡萄牙东北部地区的森林火灾数据，考察了地域板块、月份(month)、星期(day)、燃油湿度(FFMC)、达夫水分度(DMC)、干旱度(DC)、初始扩展知数(ISI)、温度(temp)、相对湿度(RH)风力(wind)、雨水(rain)和地区大小(area)等十三个因素对于森林火灾规模的影响。

Attribute	Description	Value
X	x-axis spatial coordinate within the Montesinho park map	1 to 9
Y	y-axis spatial coordinate within the Montesinho park map	2 to 9
month	month of the year	"jan" to "dec"
day	day of the week	"mon" to "sun"
FFMC	FFMC index from the FWI system	18.7 to 96.20
DMC	DMC index from the FWI system	1.1 to 291.3
DC	DC index from the FWI system	7.9 to 860.6
ISI	ISI index from the FWI system	0.0 to 56.10
temp	temperature (in ℃)	2.2 to 33.30
RH	relative humidity (in %)	15.0 to 100
wind	wind speed (in km/h)	0.40 to 9.40
rain	outside rain in (mm/m^2)	0.0 to 6.4
area	the burned area of the forest (in *ha*)	0.00 to 1090.84

该数据集源自Paulo Cortez and Anibal Morais所著论文A Data Mining Approach to Predict Forest Fires using Meteorological Data：http://www3.dsi.uminho.pt/pcortez/fires.pdf。

一、运行环境

系统：windows10

环境：Anaconda / Visual Studio Code

PyTorch

导入包：运用pip工具安装numpy，pandas等关键包。

```
import nni
import logging
import numpy as np
import pandas as pd
import json
from fe_util import *
from model import *
```

二、文件说明

为方便标注数据，将数据集中的影响因素除月份日期外标注为I1至I11。并单独标注出ID列与target列分别表示数据条数，以及生成结果。

主函数文件

- 获取参数

```
RECEIVED_PARAMS = nni.get_next_parameter()
logger.info("Received params:\n", RECEIVED_PARAMS)
```

- 提取特征

```
nni.report_final_result({
    "default":val_score,
    "feature_importance":feature_imp
})
```

yml配置文件

```
authorName: None #作者名
experimentName: SanDro #项目名
trialConcurrency: 1 #同时运行的最大尝试数
maxExecDuration: 10h #最长持续时间
maxTrialNum: 2000 #最大尝试次数
trainingServicePlatform: local #本地训练，可选local, remote, pai
searchSpacePath: search_space.json #搜索空间文件
useAnnotation: false #是否允许注释方式配置搜索空间，可选true，false
tuner:  #调节器选项
  codeDir: .
  classFileName: autofe_tuner.py
  className: AutoFETuner
  classArgs:  #调节器算法参数
    optimize_mode: maximize
trial: #尝试选项
  command: python main.py
  codeDir: .
  gpuNum: 0
```

json搜索空间文件

```
{
    "count":[
        "I1","I2","I3","I4","I5","I6","I7","I8","I9","I10","I11"
    ],
    "crosscount":[
        [
            "I1","I2","I3","I4","I5","I6","I7","I8","I9","I10","I11"
        ],
        [
            "I1","I2","I3","I4","I5","I6","I7","I8","I9","I10","I11"
        ]
    ]
}
```

三、运行结果

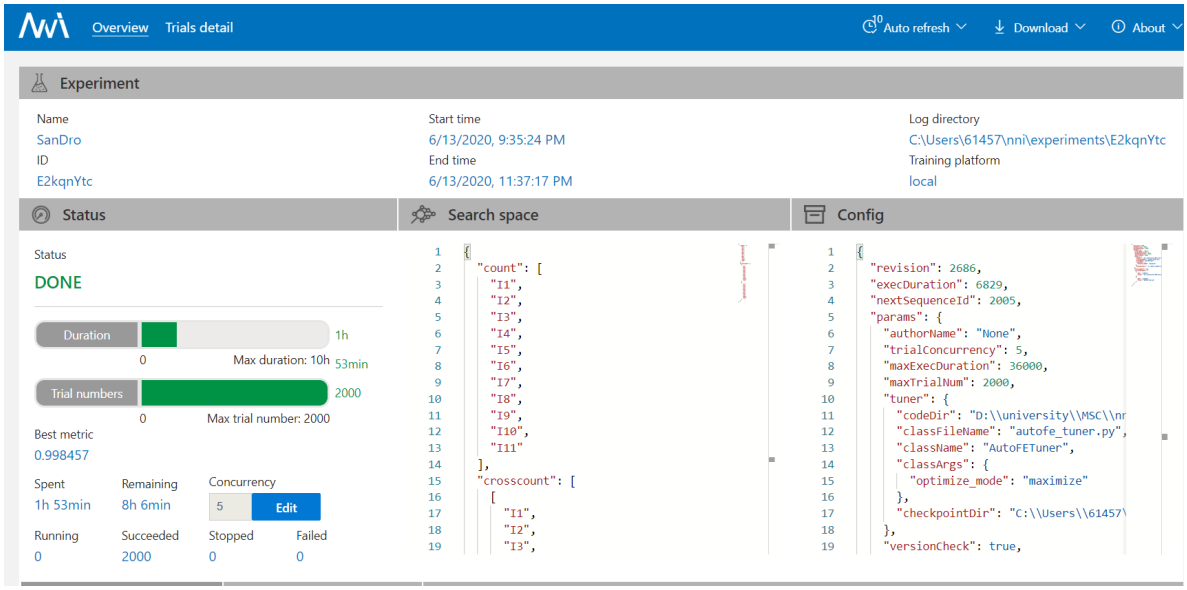
未添加NNI元素

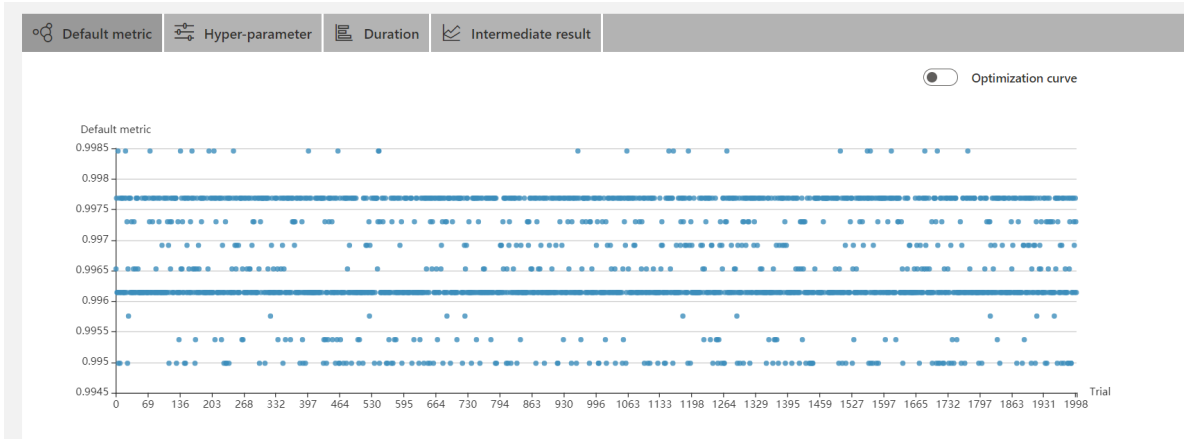
Baseline auc

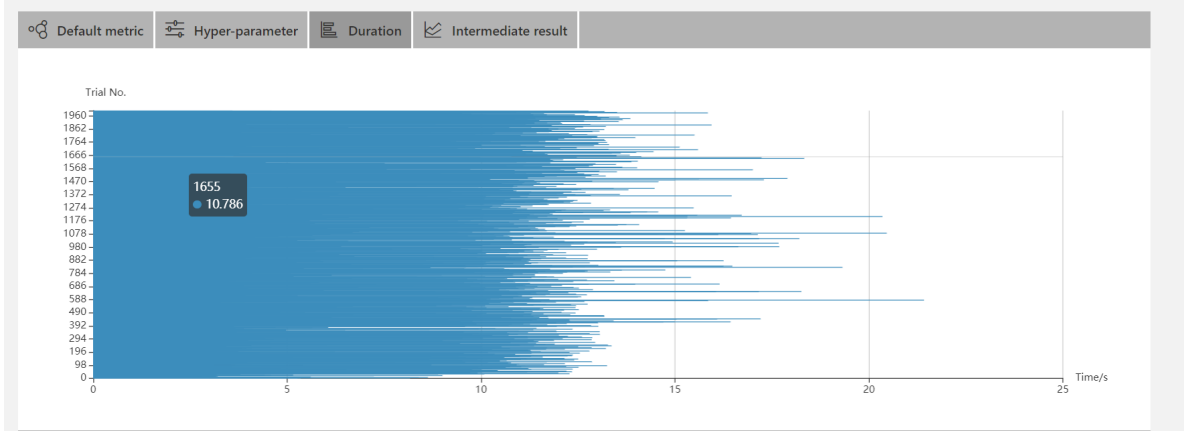
```
(nni_project) D:\university\MSC\nni\1.3\tabular_automl_NNI>python main.py
[06/13/2020, 09:30:48 PM] WARNING (nni) Requesting parameter without NNI framework, returning empty dict
Training until validation scores don't improve for 100 rounds
[50]    eval's auc: 0.994213
[100]   eval's auc: 0.995756
[150]   eval's auc: 0.996528
[200]   eval's auc: 0.996528
Early stopping, best iteration is:
[144]   eval's auc: 0.996528
D:\anaconda\anaconda\envs\nni_project\lib\site-packages\json_tricks\encoders.py:367: UserWarning: json-tricks: numpy sca
lar serialization is experimental and may work differently in future versions
```

添加NNI元素

AutoML auc







四、结论

Database	Baseline auc	AutoML auc
Forestfires	0.9965	0.9985