**特征工程代码使用及说明**

**一、代码说明：**

1. 将正负样本按1：r的比例进行下采样，并可存入本地以便调用，
2. 增加用户识别特征：user\_id
3. 原19维特征+hour展开3个特征的特征工程

分批次从本地读取已下采样的数据进行特征工程，处理后存入本体。（预计获取400w数据需要1小时，若用linux可以快很多）

4、将训练集和测试集分开，将前9天的数据作为训练集，最后一天的作为测试集

1. 进行新增的7个特征工程处理

数据处理比较慢，10w数据大约需要20min

1. 文件说明：

5.1、w000\_w2\_FE\_30：特征工程所用代码

5.2、FE\_19.txt：除'id','device\_id','device\_ip','hour'外19维特征，特征工程所用到的前99%维度字典用数字进行编码，后1%用数字999表示。

Tip：FE\_19-取不同百分比数据作为other的字典，若需要可以采用其对应字典，将名字改为FE\_19.txt即可。文件说明：‘’FE\_19\_0.5other‘’为将后0.5%作为other。详细说明见下文。

5.3、FE\_hour.txt：生成hour的3个新特征的字典

5.4、w000\_w2\_1%dict\_obtion生成FE\_19.txt、FE\_hour.txt所用代码

5.5、utils.py统计数据所用函数

5.6、FE\_utils.py特征工程所用函数

**二、代码使用说明**

1、代码中共4块需要执行，分别对应上部分说明

2、需设置参数的地方标有**$$$，**注意参数说明并设置参数。

3、注意仅在每个$$$下方的**一个代码块**中需设置参数

4、除'id','device\_id','device\_ip','hour'外，FE\_19.txt使用不同比例特征编码所产生的特征维数说明。

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | other占比 | 处理后特征维数 | FE\_19对应字典 |
| 1 | 原始 | 32000 | - |
| 2 | 0.01% | 2666 | FE\_19\_0.01other |
| 3 | 0.10% | 2289 | FE\_19\_0.1other |
| 4 | 0.50% | 1801 | FE\_19\_0.5other |
| 5 | 1% | 1517 | FE\_19\_1other |
| 6 | 2% | 1220 | FE\_19\_2other |
| 7 | 3% | 1043 | FE\_19\_3other |
| 8 | 5% | 836 | - |
| 9 | 10% | 561 | - |