

1.数据要求

1.1对使用 excel 对原始数据经行处理数据

首先按 id 排序，将#NA 替换为 0，然后将 id 这一列再复制一列，并把两行中文去掉(很重要)，特征值顺序不变。选中全部数据，删除重复的数据；选中日期这一列，将其改为如下图格式。

	A	B	C	D	E	F	G	H	I	J	K
1				大屏互动	大屏互动	大屏互动	付费游戏	付费游	付费游	艺人互	艺人互动
2	酒吧ID	日期	开屏时长	用户数	gmV	霸屏次数	用户数	gmV	游戏场	用户数	gmV
3	68794268	2020/10/23	31950	4	40	4	0	0	0	0	0
4	38444999	2020/10/23	12732	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
5	26759610	2020/10/23	1608	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
6	30019168	2020/10/23	23490	6	58	10	0	0	0	0	0
7	51498539	2020/10/23	24066	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
8	40275565	2020/10/23	17844	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
9	36936599	2020/10/23	12474	12	67.699999	48	0	0	0	2	0
10	12338684	2020/10/23	23850	1	0	1	0	0	0	0	0
11	11654683	2020/10/23	17238	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
12	73638576	2020/10/23	17478	1	19	1	0	0	0	2	264
13	17095114	2020/10/23	84828	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
14	78896081	2020/10/23	13392	1	0	1	0	0	0	3	90
15	48637211	2020/10/23	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A

图 1 原始数据

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2																
3	10043210	10043210	2020-11-26	22260	6	350	21	17	1697	1	0	0				
4	10043210	10043210	2020-11-27	27906	2	0	6	4	105	1	0	0				
5	10043210	10043210	2020-11-28	14856	9	259	32	25	1708	1	0	0				
6	10043210	10043210	2020-11-29	11796	3	177	8	0	0	0	0	0				
7	10043210	10043210	2020-11-30	13554	3	0	5	9	445	1	0	0				
8	10043210	10043210	2020-12-01	14910	3	0	7	7	310	1	0	0				
9	10043210	10043210	2020-12-02	11166	0	0	0	17	270	1	0	0				
10	10043210	10043210	2020-12-03	0	0	0	0	0	0	0	0	0				
11	10043210	10043210	2020-12-25	12	0	0	0	0	0	0	0	0				
12	10043210	10043210	2021-01-04	0	0	0	0	0	0	0	0	0				
13	10064441	10064441	2020-12-29	4152	1	0	2	0	0	0	0	0				
14	10064441	10064441	2020-12-30	11100	1	0	12	0	0	0	0	0				
15	10064441	10064441	2020-12-31	4242	0	0	0	0	0	0	0	0				
16	10064441	10064441	2021-01-01	12450	1	74	3	0	0	0	0	0				
17	10120467	10120467	2020-11-25	16620	2	9	2	0	0	0	0	0				
18	10120467	10120467	2020-11-26	6072	0	0	0	0	0	0	0	0				
19	10120467	10120467	2020-11-27	16224	1	0	1	0	0	0	0	0				
20	10120467	10120467	2020-11-28	26016	0	0	0	0	0	0	0	0				
21	10120467	10120467	2020-11-29	27768	0	0	0	0	0	0	0	0				
22	10120467	10120467	2020-11-30	14988	0	0	0	0	0	0	0	0				

图 2 处理过后的数据

2.程序的使用

2.1config.py

该文件是对整个工程主要参数的配置，如数据的路径，训练的周期，学习率等等。

1.2.1 测试数据的路径

```
class DefaultConfig(object):  
    #测试集的路径  
    原始的测试数据路径='./原始data___true.csv'
```

图 3 测试的文件路径

在“原始测试数据路径=”后面写上已经被 excel 处理过的测试数据的路径。

1.2.2 是否使用 GPU

如果是将 config.py 文件中的 use_gpu 设为 True, 反之, 使用 False。

```
batch_size = 32 # batch size  
use_gpu = True # user GPU or not  
num_workers = 4 # how many workers for load  
print_freq = 20 # print info every N batch  
max_epoch = 200
```

图 4 参数的设置

3. 测试文件 test.py 的使用

直接运行, 程序运行结束后, 在“测试结果”文件夹下面找到一个 csv 文件, 里面每个 id 对应着一个标签, 1 对应流失, 0 对应没流失。

4. 训练网络

改变 config.py 文件里训练集的路径, root0 是 excel 处理后没流失的数据, root1 是流失数据的路径。

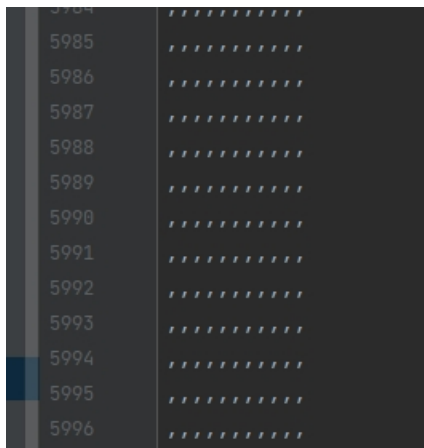
```
# 20数据  
train_data_root0_原始 = './原始data___False.csv'# 训练集存放路径  
train_data_root1_原始 = './原始data___true.csv'
```

图 5 训练数据的存放路径

然后运行 main.py 文件。训练网络这一步尽量避免, 最后训练的网络效果很不错, 训练之后这个网络就会被覆盖, 得到的结果可能没有原来的网络效果好。

5.注意事项

- 必须仔细阅读 README.md 文件
- 有时使用 Excel 处理表格数据，会使数据里出现很多“nan”，这时可能需要在 pycharm 里打开 csv 文件，处理数据。



5985	nan
5986	nan
5987	nan
5988	nan
5989	nan
5990	nan
5991	nan
5992	nan
5993	nan
5994	nan
5995	nan
5996	nan

图 6 需要除去的错误数据