山东大学 计算机科学与技术 学院

大数据分析实践 课程实验报告

学号：202200101012	姓名：康海洋	班级：数据 23

实验题目：

本次实验主要围绕宝可梦数据集进行分析，考察在拿到数据后如何对现有的数据进行预处理清洗操作，建立起对于脏数据、缺失数据等异常情况的一套完整流程的认识。

实验学时：2	实验日期：9.30

实验要求：

本次实验主要围绕宝可梦数据集进行分析，考察在拿到数据后如何对现有的数据进行预处理清洗操作，建立起对于脏数据、缺失数据等异常情况的一套完整流程的认识。

硬件环境：

计算机一台

软件环境：

Jupyter notebook

Python

实验步骤与内容：

1 库的导入与数据的读入

[5]:
```
import pandas as pd
from pandas import DataFrame
import numpy as np

primitive_data=pd.read_csv("data.csv",encoding='gbk')
primitive_data
```

[5]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
0	47	71	通辽	一般节点	1756	585	北京	网络核心	49636052613	1.000000e+11
1	47	74	通辽	一般节点	1756	776	北京	网络核心	50056871412	1.000000e+11
2	47	240	通辽	一般节点	1756	802	北京	网络核心	49453581081	1.000000e+11
3	47	241	通辽	一般节点	1997	464	天津	网络核心	49733361585	1.000000e+11
4	47	242	通辽	一般节点	474	672	哈尔滨	一般节点	50492573662	1.000000e+11
...	...	...	...	...	...	...	...	...	...	...
1113	1129	546	上海	网络核心	2050	502	石家庄	网络核心	48731433404	1.000000e+11
1114	1129	514	上海	网络核心	2473	946	吉林	一般节点	50060666120	1.000000e+11
1115	36036	499	长春	一般节点	1257	178	上海	网络核心	50545082113	1.000000e+11
1116	36422	346	天津	网络核心	1997	41	天津	网络核心	50628787089	1.000000e+11
1117	2701	619	大连	网络核心	2549	1070	沈阳	网络核心	48753971761	1.000000e+11

1118 rows × 10 columns

2. 2 删除多余的空行并进行过滤

采用 dropna 方法并指定参数为 any 删除多余的空行

[6]:
```
primitive_data_1=primitive_data.dropna(how='any')
primitive_data_1
```

[6]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
0	47	71	通辽	一般节点	1756	585	北京	网络核心	49636052613	1.000000e+11
1	47	74	通辽	一般节点	1756	776	北京	网络核心	50056871412	1.000000e+11
2	47	240	通辽	一般节点	1756	802	北京	网络核心	49453581081	1.000000e+11
3	47	241	通辽	一般节点	1997	464	天津	网络核心	49733361585	1.000000e+11
4	47	242	通辽	一般节点	474	672	哈尔滨	一般节点	50492573662	1.000000e+11
...	...	...	...	...	...	...	...	...	...	...
1113	1129	546	上海	网络核心	2050	502	石家庄	网络核心	48731433404	1.000000e+11
1114	1129	514	上海	网络核心	2473	946	吉林	一般节点	50060666120	1.000000e+11
1115	36036	499	长春	一般节点	1257	178	上海	网络核心	50545082113	1.000000e+11
1116	36422	346	天津	网络核心	1997	41	天津	网络核心	50628787089	1.000000e+11
1117	2701	619	大连	网络核心	2549	1070	沈阳	网络核心	48753971761	1.000000e+11

1118 rows × 10 columns

3 接下来过滤得到 traffic 不等于 0 且 from_level=一般节点的数据

[7]:
```
data_before_filter=primitive_data_1
data_after_filter_1=data_before_filter.loc[data_before_filter["traffic"]!=0]
data_after_filter_2=data_after_filter_1.loc[data_after_filter_1["from_level"]=='一般节点']
data_after_filter_2
```

[7]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
0	47	71	通辽	一般节点	1756	585	北京	网络核心	49636052613	1.000000e+11
1	47	74	通辽	一般节点	1756	776	北京	网络核心	50056871412	1.000000e+11
2	47	240	通辽	一般节点	1756	802	北京	网络核心	49453581081	1.000000e+11
3	47	241	通辽	一般节点	1997	464	天津	网络核心	49733361585	1.000000e+11
4	47	242	通辽	一般节点	474	672	哈尔滨	一般节点	50492573662	1.000000e+11
...	...	...	...	...	...	...	...	...	...	...
1097	2473	1460	吉林	一般节点	591	586	绥化	一般节点	48409925693	1.000000e+11
1103	36036	18	长春	一般节点	3443	650	青岛	网络核心	48663350759	1.000000e+11
1104	63	6	通辽	一般节点	36036	20	长春	一般节点	50355678076	1.000000e+11
1107	36036	52	长春	一般节点	1129	171	上海	网络核心	49345226162	1.000000e+11
1115	36036	499	长春	一般节点	1257	178	上海	网络核心	50545082113	1.000000e+11

550 rows × 10 columns

3. 4 对数据进行抽样

采取不同的采样方式采取 50 个样本并比较采样结果

- 加权采样：to_level 的值为一般节点与网络核心的权重之比为 1 : 5

```
weight_sample=data_before_sample.copy()
weight_sample['weight']=0
for i in weight_sample.index:
    if weight_sample.at[i,'to_level']=='一般节点':
        weight=1
    else:
        weight=5
    weight_sample.at[i,'weight']=weight

weight_sample_finish=weight_sample.sample(n=50,weights='weight')
#data_before_sample=data_before_sample[columns]
weight_sample_finish=weight_sample[columns]
weight_sample_finish
```

[10]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
0	47	71	通辽	一般节点	1756	585	北京	网络核心	49636052613	1.000000e+11
1	47	74	通辽	一般节点	1756	776	北京	网络核心	50056871412	1.000000e+11
2	47	240	通辽	一般节点	1756	802	北京	网络核心	49453581081	1.000000e+11
3	47	241	通辽	一般节点	1997	464	天津	网络核心	49733361585	1.000000e+11
4	47	242	通辽	一般节点	474	672	哈尔滨	一般节点	50492573662	1.000000e+11
...	...	...	...	...	...	...	...	...	...	...
1097	2473	1460	吉林	一般节点	591	586	绥化	一般节点	48409925693	1.000000e+11
1103	36036	18	长春	一般节点	3443	650	青岛	网络核心	48663350759	1.000000e+11
1104	63	6	通辽	一般节点	36036	20	长春	一般节点	50355678076	1.000000e+11
1107	36036	52	长春	一般节点	1129	171	上海	网络核心	49345226162	1.000000e+11
1115	36036	499	长春	一般节点	1257	178	上海	网络核心	50545082113	1.000000e+11

550 rows × 10 columns

- 5 随机抽样

[11]:
```
random_sample=data_before_sample
random_sample_finish=random_sample.sample(n=50)
random_sample_finish=random_sample_finish[columns]
random_sample_finish
```

[11]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
170	787	60	玉溪	一般节点	4561	1025	成都	网络核心	49992676292	1.000000e+11
681	36036	20	长春	一般节点	1536	681	广州	网络核心	49317137743	1.000000e+11
75	180	84	呼和浩特	一般节点	1536	86	鄂尔多斯	网络核心	49100967003	1.000000e+11
128	474	1409	哈尔滨	一般节点	1756	1067	北京	网络核心	49473981680	1.000000e+11
616	36036	54	长春	一般节点	36272	105	太原	网络核心	50464551947	1.000000e+11
599	474	672	哈尔滨	一般节点	2050	336	石家庄	网络核心	51340689424	1.000000e+11
135	591	17	绥化	一般节点	3443	186	青岛	网络核心	49474305249	1.000000e+11
1059	47	252	通辽	一般节点	1997	250	天津	网络核心	50358481161	1.000000e+11
402	474	1399	哈尔滨	一般节点	180	252	呼和浩特	一般节点	49271182579	1.000000e+11
362	180	252	呼和浩特	一般节点	1997	724	天津	网络核心	49033191620	1.000000e+11
431	591	1104	绥化	一般节点	2549	852	沈阳	网络核心	49411244329	1.000000e+11
344	180	34	呼和浩特	一般节点	2050	295	石家庄	网络核心	50352242512	1.000000e+11
535	47	259	通辽	一般节点	2841	341	郑州	网络核心	51012708275	1.000000e+11
799	180	52	呼和浩特	一般节点	474	460	哈尔滨	一般节点	49553070694	1.000000e+11
134	591	15	绥化	一般节点	1385	1490	广州	网络核心	49228307349	1.000000e+11

- 6 分层抽样：根据 to_level 的值进行分层采样

根据比例一般节点抽 17 个，网络核心抽 33 个

[12]:
```
ybjd=data_before_sample.loc[data_before_sample['to_level']=='一般节点']
wlhx=data_before_sample.loc[data_before_sample['to_level']=='网络核心']
after_sample=pd.concat([ybjd.sample(17),wlhx.sample(33)])
after_sample
```

[12]:

	from_dev	from_port	from_city	from_level	to_dev	to_port	to_city	to_level	traffic	bandwidth
111	474	673	哈尔滨	一般节点	2473	799	吉林	一般节点	48852033101	1.000000e+11
86	180	226	呼和浩特	一般节点	36036	20	长春	一般节点	49248544673	1.000000e+11
21	63	58	通辽	一般节点	36036	54	长春	一般节点	48363382095	1.000000e+11
1063	47	314	通辽	一般节点	47	252	通辽	一般节点	49900452417	1.000000e+11
773	2473	762	吉林	一般节点	180	84	呼和浩特	一般节点	49702910101	1.000000e+11
33	63	286	通辽	一般节点	180	52	呼和浩特	一般节点	49725190236	1.000000e+11
732	96	141	呼和浩特	一般节点	36036	499	长春	一般节点	47474335913	1.000000e+11
979	2473	1043	吉林	一般节点	63	282	通辽	一般节点	49176857434	1.000000e+11
447	787	63	玉溪	一般节点	36036	54	长春	一般节点	49557001334	1.000000e+11
806	180	20	呼和浩特	一般节点	474	670	哈尔滨	一般节点	50581993828	1.000000e+11
705	47	242	通辽	一般节点	63	286	通辽	一般节点	49144860439	1.000000e+11
559	96	102	呼和浩特	一般节点	36036	52	长春	一般节点	49483965391	1.000000e+11
850	474	422	哈尔滨	一般节点	591	638	绥化	一般节点	51214123797	1.000000e+11
766	5058	144	南宁	一般节点	180	30	呼和浩特	一般节点	50481413185	1.000000e+11
622	180	20	呼和浩特	一般节点	36036	499	长春	一般节点	49636788433	1.000000e+11