数据挖掘作业 3　　分类

姓名：董永银

学号：2120171007

日期：2018.4.23

titanic 数据的分类分析报告

一.数据源

选取 https://www.kaggle.com/c/titanic/data 数据进行分类分析，共三个文件：gender_submission.csv、test.csv、train.csv，分别是提交格式、测试集和训练集。

二.数据分析

1、数据初探（一）数据概况

train.csv 文件里的内容，主要包含这么几列，可以简单地先判断一下那些数据比较有用：

PassengerId：只是个乘客序号；

Survived：最终是否存活；

Pclass：舱位，1 是头等舱，3 是最低等，从电影里看，这个影响还是挺大的；

Name：乘客姓名，除非是要算命，不然应该没啥影响；

Sex：性别，应该影响很大；

Age：年龄，有一部分数据缺失；

SibSp：一同上船的兄弟姐妹或配偶；

Parch：一同上船的父母或子女，目测这两项应该没啥影响吧，除非是要是一起死的那种；

Ticket：船票信息，比较乱，完全看不出有任何用处；

Fare：乘客票价，这个数据应该和 Pclass 有一定对应关系；

Cabin：客舱编号，应该不同的编号对应不同的位置，对逃生还是有一定影响的，然而这项数据缺失很多（204/891），所以我选择暂时忽略；

Embarked：上船地点，主要是 S（南安普顿）、C（瑟堡）、Q（皇后镇），这个应该也没啥影响，但不妨一试。

2、读取数据

```
dt_train=pd.read_csv('train.csv')
dt_test=pd.read_csv('test.csv')
```

	PassengerId	Survived	Pclass
0	1	0	3
1	2	1	1
2	3	1	3
3	4	1	1
4	5	0	3
5	6	0	3

	Name	Sex	Age	SibSp
0	Braund, Mr. Owen Harris	male	22.0	1
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1
2	Heikkinen, Miss. Laina	female	26.0	0
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1
4	Allen, Mr. William Henry	male	35.0	0
5	Moran, Mr. James	male	NaN	0

3、补充缺失值

表格中，年龄 Age 和舱房 Cabin 存在空值。用“0”补充上，表示“未知”含义

```
titanic_fill = titanic_df.fillna(value=0)
```

	Name	Sex	Age	SibSp
0	Braund, Mr. Owen Harris	male	22.0	1
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1
2	Heikkinen, Miss. Laina	female	26.0	0
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1
4	Allen, Mr. William Henry	male	35.0	0
5	Moran, Mr. James	male	0.0	0
6	McCarthy, Mr. Timothy J	male	54.0	0
7	Palsson, Master. Gosta Leonard	male	2.0	3
8	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0
9	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1
10	Sandstrom, Miss. Marguerite Rut	female	4.0	1
11	Bonnell, Miss. Elizabeth	female	58.0	0
12	Saundercock, Mr. William Henry	male	20.0	0
13	Andersson, Mr. Anders Johan	male	39.0	1
14	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0
15	Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0
16	Rice, Master. Eugene	male	2.0	4
17	Williams, Mr. Charles Eugene	male	0.0	0
18	Vander Planke, Mrs. Julius (Emelia Maria Vande...	female	31.0	1
19	Masselmani, Mrs. Fatima	female	0.0	0

4、在大致观察过这些数据后，不难发现，直观上乘客存活的几率与年龄、性别、舱位、登船号有很大关联（年龄、登船口等数据有缺失部分待处理），而与其他数据如姓名、家人数量、票价等关联度不大，或者说没有明显的关系，那我们接下来做一个简单的处理：

```
#去除乘客姓名、船票信息和客舱编号三个不打算使用的列
dt_train_p=dt_train.drop(['Name','Ticket','Cabin'],axis=1)
dt_test_p=dt_test.drop(['Name','Ticket','Cabin'],axis=1)
```

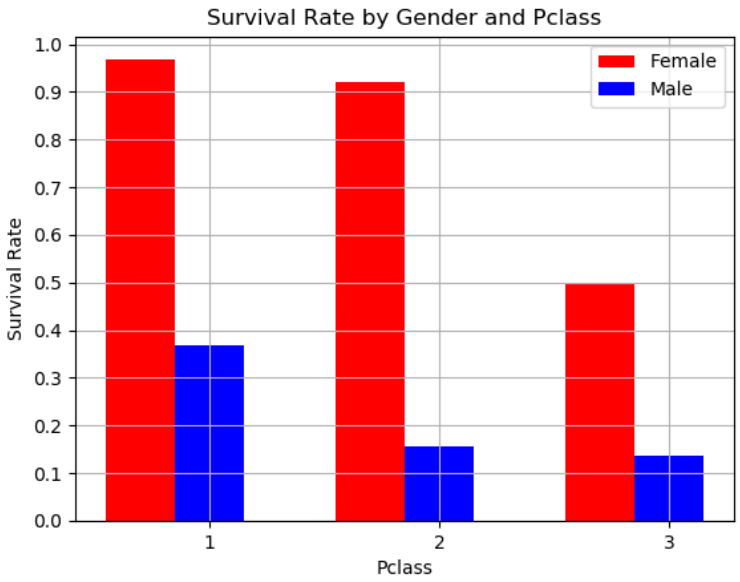
三.分类分析

1、舱位和性别，两者结合分析，

```
#按照性别和舱位分组聚合
Pclass_Gender_grouped=dt_train_p.groupby(['Sex','Pclass'])
 #计算存活率
PG_Survival_Rate=(Pclass_Gender_grouped.sum()/Pclass_Gender_grouped.count())['Survived']

x=np.array([1,2,3])
width=0.3
plt.bar(x-width,PG_Survival_Rate.female,width,color='r')
plt.bar(x,PG_Survival_Rate.male,width,color='b')
plt.title('Survival Rate by Gender and Pclass')
plt.xlabel('Pclass')
plt.ylabel('Survival Rate')
plt.xticks([1,2,3])
plt.yticks(np.arange(0.0, 1.1, 0.1))
plt.grid(True,linestyle='-',color='0.7')
plt.legend(['Female','Male'])
plt.show()  #画图
```

结论如图所示：



女性的存活率明显高于男性，即使是末等仓（Pclass=3），存活率也达到了 50.0%，高于男性中最高的头等舱高富帅（36.9%）。

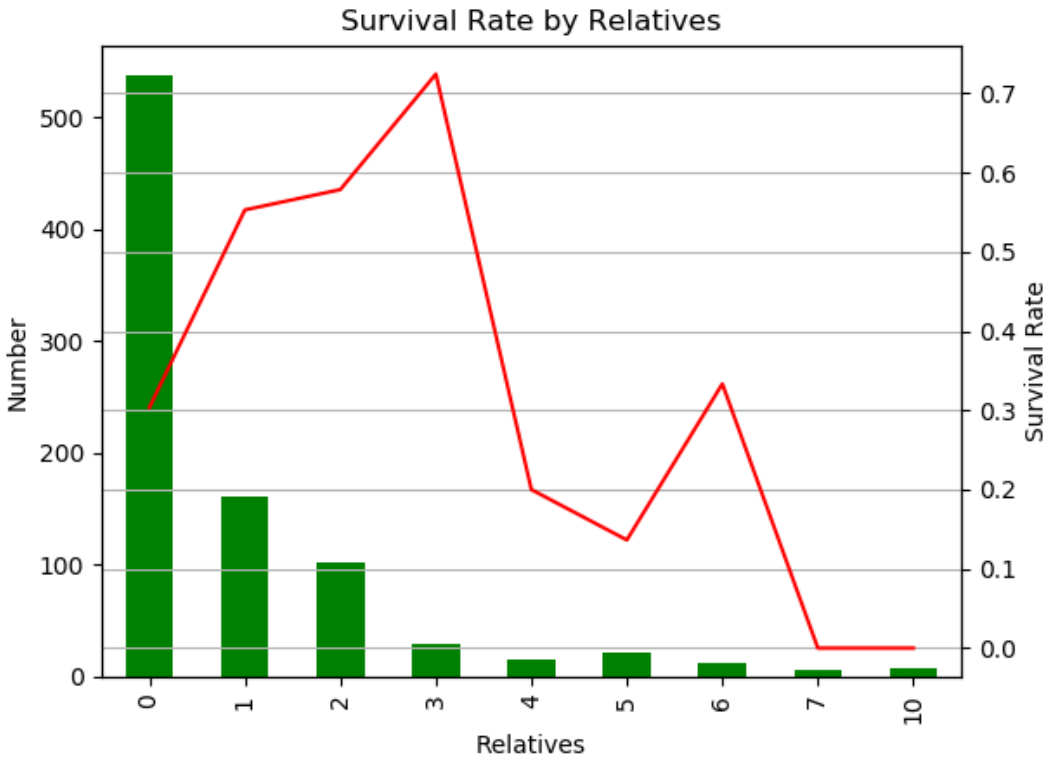
2、兄弟姐妹配偶（SibSP）以及父母子女（ParCh）结合分析

在此将两者数字加和，分组聚合，然后画图：

```
dt_train_p['Relatives']=dt_train_p['SibSp']+dt_train_p['Parch']
Rela_grouped=dt_train_p.groupby(['Relatives'])
Rela_Survival_Rate=(Rela_grouped.sum()/Rela_grouped.count())['Survived']
Rela_count=Rela_grouped.count()['Survived']

ax1=Rela_count.plot(kind='bar',color='g')
ax2=ax1.twinx()
ax2.plot(Rela_Survival_Rate.values,color='r')
ax1.set_xlabel('Relatives')
ax1.set_ylabel('Number')
ax2.set_ylabel('Survival Rate')
plt.title('Survival Rate by Relatives')
plt.grid(True,linestyle='-',color='0.7')
plt.show()
```

结论如图所示：



如果乘客有 1 个或 2 个或 3 个的亲属有助于提高存活率。

四.结论

（1）女性的存活率普遍高于男性；

（2）有 1~3 个亲属的人存活率要普遍高于其他区段；