基于泰坦尼克号数据的分析与建模

（目标：基于泰坦尼克号数据进行数据探索和分析，并建立合理的分类模型）

摘要：本文对泰坦尼克号数据集进行探索性数据分析，以EDA、数值特征处理、类别特征处理等特征工程不同处理方式对不同的特征属性进行分析。针对预测乘船人员的生死分类问题，本文基于逻辑回归算法实现了简单高效的分类器，实验表明该分类器准确率为97%。

一、特征工程

1.1 数据介绍

泰坦尼克号有“永不沉没”的美誉，然而讽刺的是，在她的处女航中，泰坦尼克号便遭厄运——船上时间1912年4月14日23时40分左右，泰坦尼克号与一座冰山相撞，造成右舷船艏至船中部破裂，五座水密舱进水。次日凌晨2时20分左右，泰坦尼克船体断裂成两截后沉入大西洋底3700米处。2224名船员及乘客中，逾724人生还,1500人丧生,其中仅333具罹难者遗体被寻回。

图1.1 泰坦尼克号

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
0	1	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	0
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	1
2	3	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	1
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	1
4	5	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	0

图1.2 泰坦尼克号数据集

泰坦尼克号数据集包含12个字段，分别为：PassengerId代表乘客ID；Survived代表是否生存，0代表遇难，1代表还活着；Pclass代表船舱等级，1Upper，

2Middle，3Lower；Name代表姓名；Sex代表性别；Age代表年龄；SibSp代表兄弟姐妹及配偶个数；Parch代表父母或子女个数；Ticket代表乘客的船票号；Fare代表乘客的船票价；Cabin代表乘客所在的仓位（位置）；Embarked代表乘客登船口岸 。

1.2 数据集探索性分析

1.2.1 箱线图

箱形图（Box-plot）又称为盒须图、盒式图或箱线图，是一种用作显示一组数据分散情况资料的统计图，因形状如箱子而得名。在各种领域也经常被使用，常见于品质管理，它主要用于反映原始数据分布的特征，还可以进行多组数据分布特征的比较。箱线图的绘制方法是：先找出一组数据的最大值、最小值、中位数和两个四分位数；然后，连接两个四分位数画出箱子；再将最大值和最小值与箱子相连接，中位数在箱子中间。

图例显示了不同年龄情况下，遇难人员的存亡情况。由图可知，无论是对于生存还是死亡的人员来说，年龄的中位数都在28岁左右，并且20-40岁之间的人员占了绝大多数比例。此外，对于生还人员，年龄最大值在60岁左右，最小值不到一岁；对于遇难人员来说，年龄最大值在68岁左右，最小值同样不到一岁。相同的是两者都存在一些异常数据。

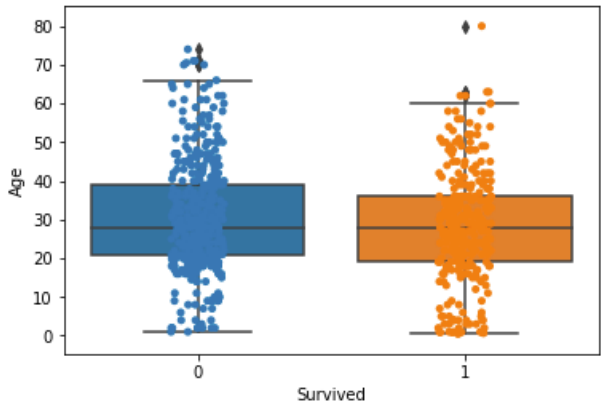

图1.3 箱线图

1.2.2

同样的方法去分析其他不同的图来分析不同的特征对于目标变量的关系以及不同的特征之间的关联，不同特征之间的融合对于目标变量之间的关系。

。 。 。 。 。 。

1.2.3

1.2.4

1.3 数值型数据处理

数值处理技巧包括：截断、二值化、分桶、缩放、缺失值处理、特征交叉、非线性编码、行统计量等常用技巧。

1.3.1 特征分桶处理

对于年龄特征值，样本数据分布大概为0-100之间的一个整数，为了将离散变量进行类别化，需要对特征值进行分桶处理。

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Fare	Embarked	Survived	Title
0	1	3	Braund, Mr. Owen Harris	0	22.0	1	0	7.2500	S	0	1.0
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	38.0	1	0	71.2833	C	1	3.0
2	3	3	Heikkinen, Miss. Laina	1	26.0	0	0	7.9250	S	1	2.0
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.0	1	0	53.1000	S	1	3.0
4	5	3	Allen, Mr. William Henry	0	35.0	0	0	8.0500	S	0	1.0

图1.4 原始年龄特征分布图

首先，将年龄特征分为5个区间，分别为小于16的区间，大于16并且小于32的区间，大于32并且小于48的区间，大于48并且小于64的区间，大于64的区间。

将年龄划分到不同的区间段后，得到下图。

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Fare	Embarked	Survived	Title	AgeBand
655	656	2	Hickman, Mr. Leonard Mark	0	1.0	2	0	73.5000	S	0	1.0	(16.336, 32.252]
70	71	2	Jenkin, Mr. Stephen Curnow	0	1.0	0	0	10.5000	S	0	1.0	(16.336, 32.252]
261	262	3	Asplund, Master. Edvin Rojj Felix	0	0.0	4	2	31.3875	S	1	0.0	(0.34, 16.336]
642	643	3	Skoog, Miss. Margit Elizabeth	1	0.0	3	2	27.9000	S	0	2.0	(0.34, 16.336]
42	43	3	Kraeff, Mr. Theodor	0	NaN	0	0	7.8958	C	0	1.0	NaN
533	534	3	Peter, Mrs. Catherine (Catherine Rizk)	1	NaN	0	2	22.3583	C	1	3.0	NaN
523	524	1	Hippach, Mrs. Louis Albert (Ida Sophia Fischer)	1	2.0	0	1	57.9792	C	1	3.0	(32.252, 48.168]

图1.5 年龄特征分区间分布图

然后，对于年龄的缺失值采用插补的方法，这里选用的是用中位数进行填补。接着，进行分桶处理，把在区间为0-16进行编号为桶0，把在区间为16-32进行编号为桶1，把在区间为32-48进行编号为桶2，把在区间为48-64进行编号为桶3，把在区间大于64进行编号为桶4，这样所有的年龄值都被分在了0-4五个桶里。

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Fare	Embarked	Survived	Title
0	1	3	Braund, Mr. Owen Harris	0	1.0	1	0	7.2500	S	0	1.0
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	2.0	1	0	71.2833	C	1	3.0
2	3	3	Heikkinen, Miss. Laina	1	1.0	0	0	7.9250	S	1	2.0
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	2.0	1	0	53.1000	S	1	3.0
4	5	3	Allen, Mr. William Henry	0	2.0	0	0	8.0500	S	0	1.0

图1.6 年龄特征分桶图

1.3.2。。。

1.3.3。。。

同样的方法去处理其余的数值特征。

1.4 类别型数据处理

。。。。。。。此处省略若干字

一步步处理完之后，在最后展示数据处理部分的部分结果截图。如下图所示。

	PassengerId	Pclass	Sex	Age	Fare	Embarked	Title	IsAlone	Age*Pclass
0	892	3	0	2.0	0	2	1.0	1	6.0
1	893	3	1	2.0	0	0	3.0	0	6.0
2	894	2	0	3.0	1	2	1.0	1	6.0
3	895	3	0	1.0	1	0	1.0	1	3.0
4	896	3	1	1.0	1	0	3.0	0	3.0

图1.7 数据处理效果图

二、数据建模

(在数据建模阶段，你可以利用数据探索的有用信息，基于某些算法设计合理的模型，这一部分需要写明你的设计细节以及创新点。针对你选定的具体数据集和不同应用场景（比如生物数据、交易数据、交通数据等等），设计细节可以是你具体的算法流程和模型结构 。实验结果、实验分析等内容放到单独的一节)

本文需要针对泰坦尼克号数据建立一个有效的分类模型，经过数据探索分析可知，泰坦尼克号数据数据是一个二分类问题。因此，本文基于逻辑回归模型进行数据建模，构建泰坦尼克号数据种类判别器。

下面可以介绍逻辑回归算法的基本原理。logistic回归是一种广义线性回归（generalized linear model），因此与多重线性回归分析有很多相同之处。它们的模型形式基本上相同，都具有 w‘x+b，其中w和b是待求参数，其区别在于他们的因变量不同，多重线性回归直接将w‘x+b作为因变量，即y =w‘x+b，而logistic回归则通过函数L将w‘x+b对应一个隐状态p，p =L(w‘x+b),然后根据p 与1-p的大小决定因变量的值。如果L是logistic函数，就是logistic回归，如果L是多项式函数就是多项式回归。

三、实验结果和分析

（展示你所做实验的内容，包括训练过程、分类测试结果，泰坦尼克号数据是一个比较简单的例子，在比较复杂的数据分析建模问题中，你应该对实验结果做出具体的解释。比如，你可以在建模阶段选用不同的方法，例如用支持向量机或者聚类方法，在这里展示不同方法的实验结果，并分析和对比实验结果，说明哪些地方好，哪些地方不好，以及展示）

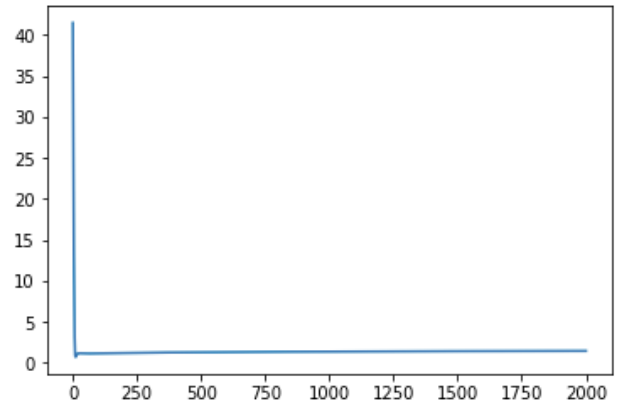

图1.8 数据实验效果图

关键代码分析

此处可对算法的核心代码进行简单的分析。

```
def gradAscent(dataMat,labelMat,alpha=0.0001,maxCycles=2000):
    """
    使用梯度上升算法，通过不断迭代改变参数的值来优化目标函数
```

```
"""
m,n=shape(dataMat)
weights=np.ones((n,1))
errors=[]
#循环迭代次数
for k in range(maxCycles):
    #求当前的sigmoid函数预测概率
    h=sigmoid(dataMat*weights)
    #**********************************************
    #此处计算真实类别和预测类别的差值
    #对logistic回归函数的对数释然函数的参数项求偏导
    error=(labelMat.T-h)
    #更新权值参数
    weights=weights+alpha*dataMat.transpose()*error
    #**********************************************
    err = error_function(dataMat,weights,labelMat)
    #print("err",err)
    errors.append(float(err))
return weights,error
```