

**模式识别大作业**

题 目 泰坦尼克号

学 院 信息科学与工程

专 业 控制科学与工程

组 员 肖若诚

指导教师 赵海涛

**完成日期： 2018 年 10 月23日**

**模式识别作业报告——泰坦尼克号**

**肖若诚 Y30180683**

**1 题目背景**

泰坦尼克号(RMS Titanic)是英国白星航运公司下辖的一艘奥林匹克邮轮，在其处女航行中，因与一座冰山相撞而至沉船。在这次事故中，有约2/3的人丧生。现在给定泰坦尼克号船上的乘客信息，需要通过一个模型来判断一名乘客在沉船灾难中能否最终存活下来。

题目数据包含一个训练集（train.csv）、一个测试集（test.csv）以及预测结果(gender\_submission）每名乘客有如下信息：

1. PassengerId : 乘客的唯一ID

2. Survived : 乘客最终是否存活(0 = No, 1 = Yes, 仅train.csv中包含此信息)

3. Pclass : 乘客的船票的等级(1 = 1st, 2 = 2nd, 3 = 3rd)

4. Name : 乘客名字

5. Sex : 乘客性别(male, female)

6. Age : 乘客年龄(Year)

7. Sibsp ：船上兄弟姐妹/配偶的人数

8. Parch : 船上父母/儿女的人数

9. Ticket : 船票号码

10. Fare : 船票价格

11. Cabin : 船舱号

12. Embarked : 出发港口(C = Cherbourg, Q = Queenstown, S = Southampton)

**2 题目分析**

2.1可行性分析

本题是一个典型的二分类问题

先从train.csv中提取每个乘客的多项特征和存活信息，使用特征和特征进行模型训练。

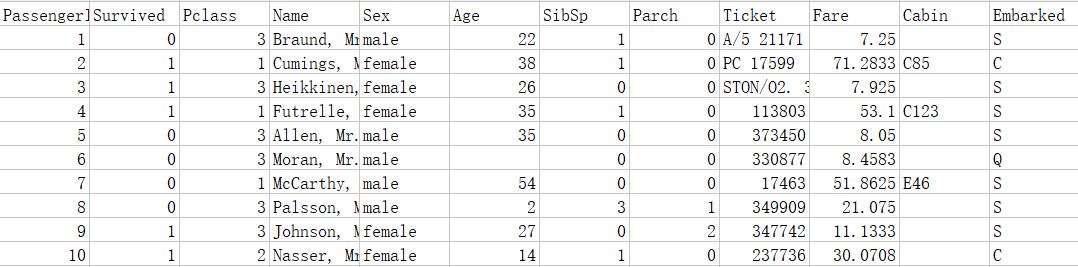
乘客的特征选取和处理比较重要。类别的特征，如Embarked、Pclass等，可以转换成one-hot encoding的表示形式。数值型的特征，如Age、Parch等可以进行范围标准化。一些很难处理和使用的特征，如Name等，则可以考虑直接忽略。

模型可以使用传统的分类模型，如SVM，Decision Tree，贝叶斯分类等。在训练过程中，要适当引入正则化项，防止模型过拟合，而致其泛化能力变差。

最后使用训练好的模型预测test.csv中的乘客最终能否存活。

本次试验我选用的是朴素贝叶斯。

2.2数据结构分析



**图2-1 train.csv数据格式**

如图2-1所示，共有十项特征信息，其中乘客姓名、船票价格对是否存活无明显的影响，船舱号由于数据缺失较多也一并舍弃。因此采用船票等级、性别、年龄、船上亲人人数、出发港口五个特征，其中亲人人数由SibSp和Parch两部分组成。

2.3朴素贝叶斯原理

贝叶斯分类模型如下：

 (2-1)

其中，X表示属性集，Y表示类变量，P(Y)为先验概率，P(X|Y)为类条件概率，P(X)为证据，P(Y|X)为后验概率。贝叶斯分类模型就是用先验概率P(Y)、类条件概率P(X|Y)和证据P(X)来表示后验概率。在比较Y的后验概率时，分母中的证据P(X)总是常数，因此可以忽略不计。先验概率P(Y)可以通过计算训练集中属于每个类的训练记录所占的比例很容易估计。对类条件概率P(X|Y)的估计，不同的实现决定不同的贝叶斯分类方法。

而在朴素贝叶斯分类中估计类条件概率P（X|Y）时假设属性之间条件独立，可表示为下式：

 (2-2)

其中，d为每条记录的属性个数

朴素贝叶斯分类模型如下：

 (2-3)

在该模型中，对于所有的Y，P(X)是固定的，所以上述模型等价于：

 (2-4)

当属性是离散型时，类的先验概率可以通过训练集的各类样本出现的次数来估计，例如A类的先验概率=A类样本的数量/样本总数。类条件概率P(Xi=xi|Y=yj)可以根据类yj中属性值等于xi的训练实例的比例来估计。

当属性是连续型时，有两种方法来估计属性的类条件概率。第一种方法是把每个连续的属性离散化，然后用相应的离散区间替换连续属性值，但这种方法不好控制离散区间划分的粒度，如果粒度太细，就会因为每一个区间中训练记录太少而不能对P(X|Y)做出可靠的估计，如果粒度太粗，那么有些区间就会包含来自不同类的记录，因此失去了正确的决策边界。第二种方法是，可以假设连续变量服从某种概率分布，然后使用训练数据估计分布的参数，高斯分布通常被用来表示连续属性的类条件概率分布。

高斯分布有两个参数，均值μ和方差σ2。对每个类yi，属性Xi的类条件概率等于

 (2-5)

参数μij可以用类yj的所有训练记录关于Xi的样本均值来估计，同理σij2可以用这些训练记录的样本方差来估计。

**3 程序及测试**

3.1算法原理及程序实现

在五项特征当中，Pclass、Sex、SibSp and Parch以及Embarked均能看作是离散型变量，在数据导入matlab后通过一定的方式将其转换为数字并计算其出现的比例，即可作为类条件概率使用。在Age中无年龄者用-1填充，供进一步处理。代码如下：

clc

%-----------------------------数据转换------------------------

str=['male'];

for i=1:1309

Sex0(i,1)=strcmp(Sex(i,:),str) % 性别 男-0 女-1 转换

if(Sex0(i,:))

Sex1(i,1)=0;

else

Sex1(i,1)=1;

end

str2=char(Embarked(i,1)); %港口转换 C Q S- 1 2 3

switch(str2)

case'C'

Embarked1(i,1)=1;

case'Q'

Embarked1(i,1)=2;

case'S'

Embarked1(i,1)=3;

end

count(i,1)=SibSp(i,1)+Parch(i,1); %是否有亲友同行 有1 无0

if count(i,1)>0

count(i,1)=1;

end

end

% 无年龄者用-1填充

Age0=Age;

Age0(isnan(Age0(:,1)))=[-1];

在数据转换完成后，要将训练集中存活个体和死亡个体分开，为了方便进行测试，我将测试集(892~1309号)的数据续接在训练集之后（1~891号），测试集仅Survived列为空。并假设存活个体和死亡个体集合中年龄不为空群体的年龄为高斯分布，分别求取两组的均值及方差，程序如下：

%-------------------------------分类--------------------------

A=[ Pclass(1:891,1), Sex1(1:891,1), count(1:891,1), Embarked1(1:891,1), Age0(1:891,1)];

% 将不同类分开 A1 A2分别为死亡和存活群体 A3为测试群体，M1 M2分别为两群体中有年龄的个体

%假设服从正态分布 ，mu1 mu2 sigma1 sigma2分别为两类的均值和方差

j=1,k=1,m=1,n=1;

for i=1:891

if(Survived(i,1)==0)

A1(k,1:5)=A(i,1:5);

if(A1(k,5)>-1)

M1(m,1)=A1(k,5);

m=m+1;

end

k=k+1;

else

A2(j,1:5)=A(i,1:5);

if(A2(j,5)>-1)

M2(n,1)=A2(j,5);

n=n+1;

end

j=j+1;

end

end

A3=[ Pclass(892:1309,1), Sex1(892:1309,1), count(892:1309,1), Embarked1(892:1309,1), Age0(892:1309,1)];

[mu1,sigma1]=normfit(M1);

[mu2,sigma2]=normfit(M2);

之后分别计算测试集两类别中不同特征的类条件概率，采用tabulate函数可较容易得到不同数字索引以及类条件概率，在这里采用三维矩阵P1,P2记录前四项特征的类条件概率，程序如下：

%-------------------------计算各项离散概率值---------------------

P=length(A1)/length(A); %计算先验概率 P\_alive为存活概率

P\_alive=1-P;

P1=zeros(3,2,4); %求训练数据各变量概率 Pclass Sex1 count

P2=zeros(3,2,4);%Emparked1 ，P1 P2分别表示死亡和存活组概率

for i=1:4

B=tabulate(A1(:,i));

C=B(:,[1 3]);

d=max(size(C));

if(d==3)

P1(:,:,i)=C(:,:);

else

P1(1:2,:,i)=C(:,:);

P1(3,1,i)=2;

end

P1(:,2,i)= P1(:,2,i)/100;

end

for i=1:4

B=tabulate(A2(:,i));

C=B(:,[1 3]);

d=max(size(C));

if(d==3)

P2(:,:,i)=C(:,:);

else

P2(1:2,:,i)=C(:,:);

P2(3,1,i)=2;

end

P2(:,2,i)= P2(:,2,i)/100;

end

在得到先验概率以及各特征的类条件概率后，可以进行后验概率的计算。由于各个变量是独立分布，因此按照式2-4直接将不同类特征的条件概率相乘，并乘以先验概率。对于有年龄数据的个体，额外乘上年龄关于是否存活的似然函数。比较每个测试个体在两类下的后验概率大小，取后验概率较大的一类为预测结果。程序如下：

post=ones(418,2);

for i=1:418

if(Age0(i+891)==-1)

for j=1:4

post(i,1)=post(i,1)\*(eq(A3(i,j),P1(1,1,j))\*P1(1,2,j)+eq(A3(i,j),P1(2,1,j))\*P1(2,2,j)+eq(A3(i,j),P1(3,1,j))\*P1(3,2,j));

post(i,2)=post(i,2)\*(eq(A3(i,j),P2(1,1,j))\*P2(1,2,j)+eq(A3(i,j),P2(2,1,j))\*P2(2,2,j)+eq(A3(i,j),P2(3,1,j))\*P1(3,2,j));

end

post(i,1)= post(i,1)\*P;

post(i,2)=post(i,2)\*(1-P);

else

prodt1(i)=normpdf(Age0(i+819),mu1,sigma1);

prodt2(i)=normpdf(Age0(i+819),mu2,sigma2);

for j=1:4

post(i,1)=post(i,1)\*(eq(A3(i,j),P1(1,1,j))\*P1(1,2,j)+eq(A3(i,j),P1(2,1,j))\*P1(2,2,j)+eq(A3(i,j),P1(3,1,j))\*P1(3,2,j));

post(i,2)=post(i,2)\*(eq(A3(i,j),P2(1,1,j))\*P2(1,2,j)+eq(A3(i,j),P2(2,1,j))\*P2(2,2,j)+eq(A3(i,j),P2(3,1,j))\*P1(3,2,j));

end

post(i,1)=prodt1(i)\*post(i,1)\*P;

post(i,2)=prodt2(i)\*post(i,2)\*(1-P);

end

end

for i=1:418

if(post(i,1)>post(i,2))

lable(i,1)=0;

else

lable(i,1)=1;

end

end

3.2测试结果及输出

将得到的预测结果lable与gender\_submission中的实际结果进行比较，计算准确率，并将lable以csv文件输出上传至Lintcode,程序如下：

%--------------------------计算正确率--------------------

% 测试结果数据导入名分别为Survived1和 PassengerID1

for r=1:418

Final(r,1)=xor(lable(r,1),Survived1(r,1));

end

accurary=1-sum(Final)/418

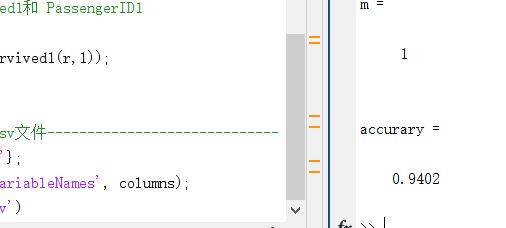
%--------------------------导出csv文件-------------------

columns={'PassengerId','Survived'};

data=table(PassengerId1,lable,'VariableNames', columns);

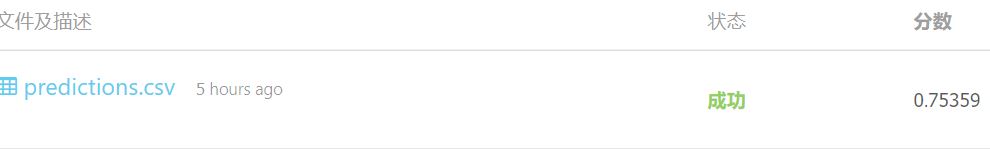
writetable(data, 'submission1.csv')

得到的准确率如图3-1所示，为0.9402。



**图3-1 准确率**

上传submission.csv至lintcode 得分如图3-2所示

**图3-2 lintcode得分**

**4 总结**

由于Python接触时间较短，语法知识掌握尚少，因此本次作业采用matlab编程。在初看题目之后，特征中离散型变量较多、各特征之间的关联度较低的特点让我将其与离散型朴素贝叶斯分类联系起来。在编程过程中，对年龄这一特征做过诸多尝试，例如用平均值填充无年龄者、划定年龄范围并人为设定条件概率等，最后均无明显效果，由于时间较短，就没再做过多尝试。

本次大作业让我对模式识别有了进一步认识，增加了我对模式识别课程的兴趣，也发现了做模式识别不仅需要良好的数学功底、编程能力，还有将生活实际与抽象思维相联系的思维逻辑。