

**模式识别作业报告**

题 目 朴素贝叶斯应用

学 院 信息科学与工程学院

专 业 控制科学与工程

姓 名 莫泽坤

学 号 Y30180666

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——朴素贝叶斯应用**

在本次模式识别课程中，我们已经学习了梯度下降法、最小二乘法、BP神经网络、奇异值分解法、主元分析法、朴素贝叶斯、logistics回归法等多种常用的模式识别算法。本作业报告主要使用朴素贝叶斯法，基于Matlab平台分别进行了广告点击概率预测与泰坦尼克号生还者预测的实验。

1. **算法原理**

贝叶斯分类器主要基于概率论中常用的贝叶斯公式：

式中，为条件为下，情况发生的概率。为情况，条件是的概率。为发生情况是的先验概率。则为发生条件是的先验概率，又称归一化因子，作用是使得结果的范围在0-100%。

对于条件为连续变量的情况，通常认为条件变量服从正态分布：

即：

对于条件为离散变量的情况，则统计各个变量值在整个样本中的出现概率：

式中，为样本总体，为条件为的样本集。

先验概率是不考虑各条件取值的情况下，对发生概率进行直接预测，其值的确定一般基于现实（如抛硬币正反面概率各约50%）或训练样本中的发生概率。

先验概率则是不考虑实际情况为何，只观察满足该条件的概率。使用全概率公式可展开为：

式中，表示为发生情况，即分别表示未发生情况下条件为的概率与为发生情况的概率。

以上是只有一种条件需要考虑的情况，实际上，对于分类或预测应用中，往往使用多个条件共同参与分类或预测，通过增加对数据的观察角度的方式提高正确率。由于不知道各个条件之间的关联性，就需要将每个条件的每种取值集合作为一个单独的整体条件。整体的可能条件集会非常庞大，而对于条件集的分布预测需要远多于条件个数的数据才能得到较为理想的分布参数。对于条件维数较多的情况，是难以取得理想的条件分布参数的。而朴素贝叶斯法就是从这里入手，通过认为各个条件独立，简化了对样本的需要量。在朴素贝叶斯法中有：

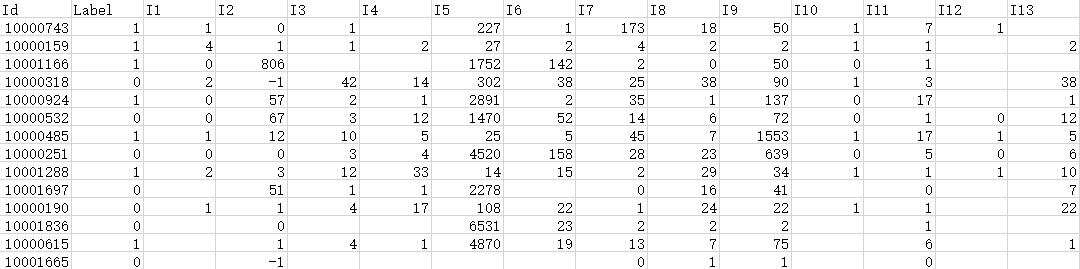
式中，D表示某种条件值的组合，该样本共有d种条件，即条件为d维（）。通过上式，对多维条件下的条件后验概率计算问题就被转化为了d个一维条件后验概率计算。则，朴素贝叶斯的计算公式为：

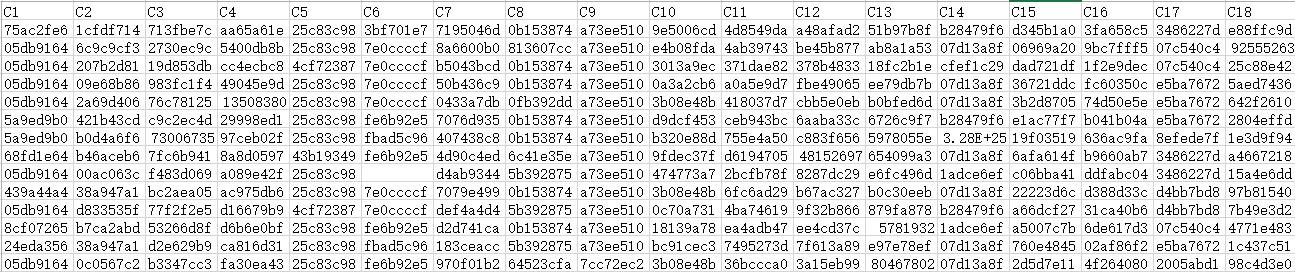
对某事件发生的概率计算，通过上式可以直接得到概率。而使用朴素贝叶斯分类器对样本进行分类时，由于不同事件下值一致，可以不计算，直接比较与的大小，就可得出样本应该分类为还是的预测结果。

1. **广告点击预测实验**
2. **实验背景与数据介绍**

Criteo是一家第三方展示广告公司，与世界上超过4000家电子商务公司有合作关系。说到广告，关注的最多的就是点击率了。本题使用Criteo所共享的一周展示广告数据，数据中提炼了13个连续特征、26个离散特征和用户是否点击了该页面广告的标签。通过训练出合适的模型，预测用户在不同的特征下是否会点击广告。

本次实验的特征数据如图2-1所示。





**图2-1广告点击预测数据**

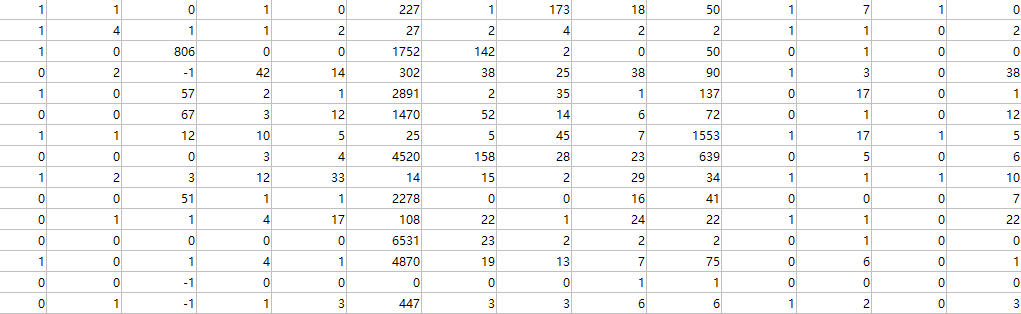
Id为用户标识码，Label为训练集中用户是否点击广告的标志，0表示用户未点击，1表示用户点击了广告，待测的样本集中不包含Label列。In代表着用户的连续特征，Cn代表用户的离散特征。在本次实验，我们只使用前13列连续特征值。

1. **实验过程**

本次实验共有一下几个步骤：首先，将数据导入至Matlab中并做简单的处理后分类为未点击情况与点击情况。分别计算两个情况下各条件的分布情况，训练得到条件的后验概率，并计算情况的先验概率。训练完毕后，应用朴素贝叶斯公式计算各样本的概率，并进行归一化后得到某用户点击广告的概率。

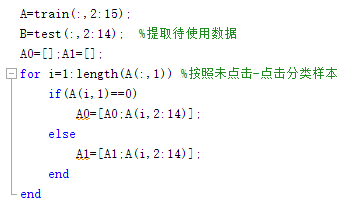
* 1. 数据处理与分类

由图2-1可知，有部分用户的部分特征是缺失数据的，这里简单处理为填充0值，得到了如图2-2的用户特征矩阵(A)。其中，第1列为Label，故共有14列。



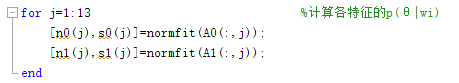
**图2-2处理后数据**

为了分别计算两种情况（未点击，点击）下各个特征的分布，通过Label对数据进行了分类，使得未点击下的特征值与点击下的特征值分至两个矩阵（A0,A1）中。代码如下：



* 1. 模型训练

由于特征为连续变量，因此我们认为他们服从正态分布。借助normfit函数分别计算了两种情况下13个特征各自的均值u与方差s。代码如下：



至此，得到了未点击广告下的各特征分布情况n0，s0与点击广告下的各特征分布情况n1，s1。它们都是13维向量。

此外，由于训练集与测试集来自同一广告公司同一广告的点击数据，所以假定测试集中用户点击的先验概率与训练集一致，得到：



未点击广告先验概率与点击广告先验概率。

* 1. 样本测试

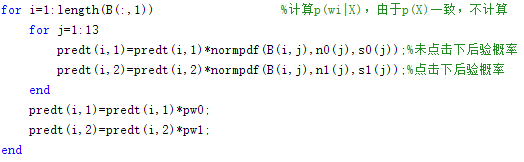
使用朴素贝叶斯公式对样本数据进行预测：

这里由于先不进行计算，归一化在稍后步骤中进行。

的计算实际就是在情况下，样本数据中第j个特征对训练集中的分布服从情况。因此使用normpdf函数计算各个数据中各特征的正态概率密度。代码如下：



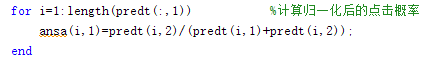
对13维特征的总后验概率计算，使用朴素贝叶斯思想，即：，并乘以情况的先验概率，得到：



可知，最后

由于该实验要求得出某用户点击广告的概率，故需要进行归一化处理。有以下公式：

代码实现为：



ansa变量按顺序存放着不同用户点击该广告的概率，如图2-3所示（转置显示）。



**图2-3广告点击预测概率部分结果**

**3 实验结果**

将以上结果与用户Id进行拼接并上传至Lintcode得到评估结果。Lintcode使用Logarithmic Loss对结果进行评估，经过以上实验过程，使用朴素贝叶斯得到的广告点击概率预测的结果为：损失值为1.25。

通过观察样本的数据可以发现，部分特征空缺较为严重，而之前的数据处理只是将空缺填补为0，这势必会降低预测的准确性。因此决定直接舍弃空缺较为严重的特征，只考虑特征I2-I9这八个特征进行点击的预测。得到结果Logarithmic Loss值为1.06，相对于使用全部特征，的确减少了误差度，这证明，在进行模式识别工作中，需要有选择地使用特征，对于不太理想的特征应予以抛弃。

该实验的完整代码如下：

A=train(:,2:15);

B=test(:,2:14); %train与test特征导入

A0=[];A1=[];

for i=1:length(A(:,1)) %对train中两种情况的特征分组

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

A0=[A0;A(i,2:14)];

else

A1=[A1;A(i,2:14)];

end

end

pw0=length(A0(:,1))/length(A(:,1)); %计算点击的先验

pw1=1-pw0;

for j=1:13 %得到特征的后验

[n0(j),s0(j)]=normfit(A0(:,j));

[n1(j),s1(j)]=normfit(A1(:,j));

end

predt=ones(length(B(:,1)),2);

for i=1:length(B(:,1)) %预测

for j=1:13

predt(i,1)=predt(i,1)\*normpdf(B(i,j),n0(j),s0(j));%未点击概率

predt(i,2)=predt(i,2)\*normpdf(B(i,j),n1(j),s1(j));%点击概率

end

predt(i,1)=predt(i,1)\*pw0;

predt(i,2)=predt(i,2)\*pw1;

end

ansa=[];

for i=1:length(predt(:,1)) %归一化

ansa(i,1)=predt(i,2)/(predt(i,1)+predt(i,2));

end

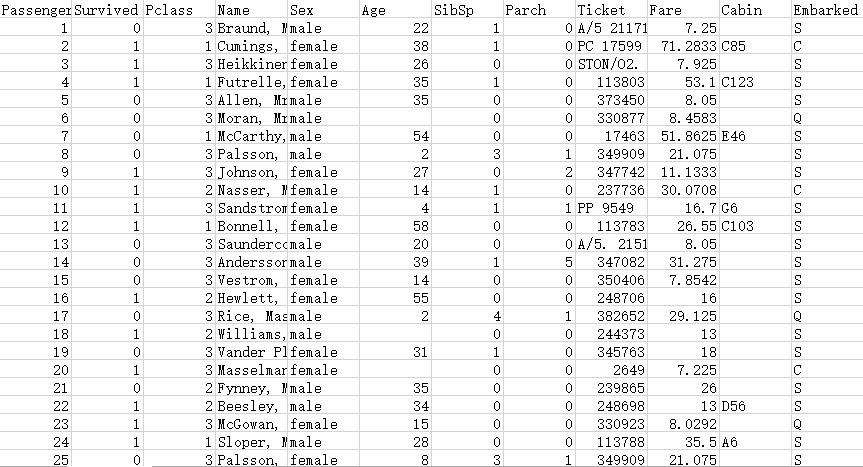
csvwrite('ans.csv',ansa,1);

1. **泰坦尼克乘客幸存预测**

1 **实验背景与数据介绍**

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，需要通过算法来判断一名乘客在沉船灾难中能否最终存活下来。

该题目同样具有一个训练集用于训练模型，一个测试集用于测试分类器的分类性能。该题目提供的数据如图3-1：



**图3-1 泰坦尼克实验数据**

各列特征含义如下：

PassengerId : 乘客的唯一ID

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

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

Name : 乘客名字

Sex : 乘客性别(male, female)

Age : 乘客年龄(Year)

Sibsp ：船上兄弟姐妹/配偶的人数

Parch : 船上父母/儿女的人数

Ticket : 船票号码

Fare : 船票价格

Cabin : 船舱号

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

这些数据中，可以认为乘客舱室等级、性别、年龄、最为重要。因为按照历史实际，碰撞发生后，一等舱的乘客最先避难，因而其存活几率较另外两舱的乘客大。许多男性在避难中发扬绅士风度，将救生艇座位让给了妇女与老人儿童。因此，女性的存活率要高于男性，老人儿童的存活率也高于青壮年。此外，船上的兄弟姐妹、配偶数，父母、儿女数、可能影响乘客避难时的行为，也应考虑在内，船票价格可能可以反映乘客在船中的位置，也可能对避难产生影响，因此也作为分类特征。而乘客姓名、船票号码、船舱号、出发港口可以认为对避难没影响或影响较小，因而不作为分类特征。

2 **实验过程**

2.1 数据分析与处理

通过观察已有数据可以发现，该数据集部分乘客的年龄发生了缺失，上次实验是对缺失数据直接补0进行处理。此次考虑使用新的处理方式：通过其他部分特征，估测出该乘客的年龄。本次实验使用最小二乘法对缺失的年龄数据进行预测。这里对最小二乘法进行简单介绍。

对于一组散点数据，使用直线或曲线进行拟合。那么评估拟合好坏的指标就是在各个自变量一致的情况下，比对对应拟合值与实际值之间的误差大小。最小二乘法就是算出所有参数后，有：

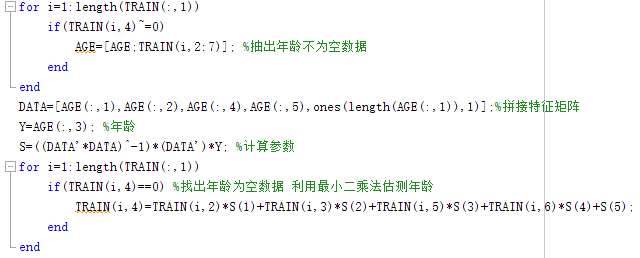
对于d维自变量，则待测参数也为d维，此时有。当X为n\*d维特征集，Y为n维列向量时，令，，则有：

则误差计算公式有：

其为最小值时则有：

则可估算出参数：

对于本次实验，则将年龄特征不空缺的乘客特征数据提取出来，并将其他特征（船舱、性别、携带家眷数、船票）拼接成特征组X，并补一列1，组成矩阵，Y则为这些乘客的年龄。通过上式得到参数B后，针对乘客年龄空缺的数据有。其中，，即为预测的年龄值。使用Matlab代码实现有：（以处理训练集train数据为例）



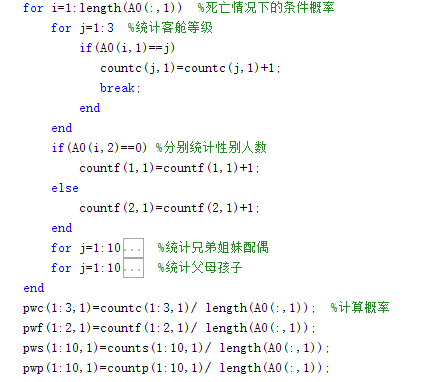
2.2 模型训练

对于连续变量的概率估计步骤与上一实验类似，不同的是，此次实验中，只有年龄和船票价格可以认为是连续变量，而船舱、性别、家眷数则为离散变量。对于他们的概率预测，需要使用下式进行：

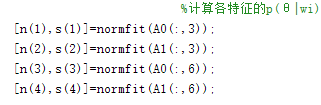
式中，为样本总体，为条件为的样本集。

对于后验概率则需要统计情况下，参数值为的个数为，则为情况下的样本总数。例如计算，则有：

其他特征类似，具体的代码实现如下：



由于连续变量的分布假设为正态分布。对于连续变量年龄、船票价格的概率计算代码实现如下：

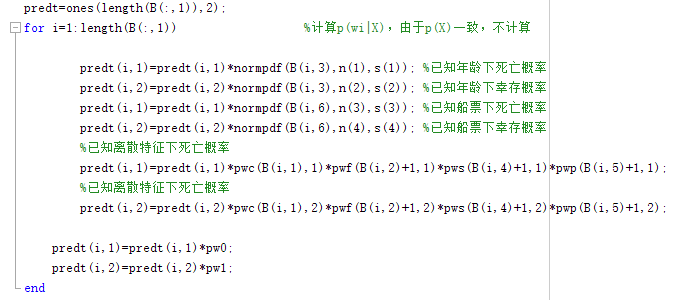


此外，对于幸存的先验概率计算，题干中已经提到约2/3死亡，而通过统计训练集中的死亡：幸存比约为：0.62:0.38。这里任选一个数据即可。

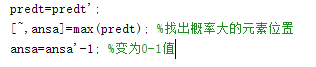
2.3 样本计算

和上一实验类似，使用朴素贝叶斯公式：

计算存活概率，代码为：



这里由于一致，仍然不进行计算。由于该实验的结果要求是直接对乘客的存活做出预测而非做出存活概率。因此，只要比对死亡概率与生存概率的大小，选择概率较大的情况即可输出结果。代码如下：



**3 实验结果**

将结果上传至Lintcode评估算法分类好坏。这里为了比较对数据的处理，也就是对空缺年龄的估测补全可否改善分类结果，分别对补0数据和估测数据进行了基于朴素贝叶斯的分类实验。如果仅对空缺的年龄数据做补0处理，则最终的幸存预测准确率为0.756.而如果对空缺的年龄数据使用最小二乘法进行估测补充处理，则实验最终的幸存预测准确率为0.763。可以看出，对缺失数据做预处理的确可以一定程度上提高最终的预测准确率。

该实验的完整代码如下：

%使用朴素贝叶斯估计泰坦尼克幸存者

%TRAIN,TEST是已经提取好的特征数组

A0=[];A1=[];AGE=[];

%最小二乘法

for i=1:length(TRAIN(:,1))

if(TRAIN(i,4)~=0)

AGE=[AGE;TRAIN(i,2:7)]; %抽出年龄不为空的数据

end

end

DATA=[AGE(:,1),AGE(:,2),AGE(:,4),AGE(:,5),ones(length(AGE(:,1)),1)];%拼接特征矩阵

Y=AGE(:,3); %年龄作为输出

S=((DATA'\*DATA)^-1)\*(DATA')\*Y; %计算参数

for i=1:length(TRAIN(:,1))

if(TRAIN(i,4)==0) %找出年龄为空的数据，利用最小二乘估测

TRAIN(i,4)=TRAIN(i,2)\*S(1)+TRAIN(i,3)\*S(2)+TRAIN(i,5)\*S(3)+TRAIN(i,6)\*S(4)+S(5);

end

end

for i=1:length(TEST(:,1))

if(TEST(i,3)==0)

TEST(i,3)=TEST(i,1)\*S(1)+TEST(i,2)\*S(2)+TEST(i,4)\*S(3)+TEST(i,5)\*S(4)+S(5);

end

end

for i=1:length(TRAIN(:,1)) %按照死亡-存活分类样本

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

A0=[A0;TRAIN(i,2:7)];

else

A1=[A1;TRAIN(i,2:7)];

end

end

pw0=length(A0(:,1))/length(TRAIN(:,1));

pw1=1-pw0;

%计算各特征概率

[n(1),s(1)]=normfit(A0(:,3));

[n(2),s(2)]=normfit(A1(:,3));

[n(3),s(3)]=normfit(A0(:,6));

[n(4),s(4)]=normfit(A1(:,6));

countc=zeros(3,2);

countf=zeros(2,2);

counts=zeros(10,2);

countp=zeros(10,2);

for i=1:length(A0(:,1)) %死亡概率下的条件概率

for j=1:3 %统计客舱等级

if(A0(i,1)==j)

countc(j,1)=countc(j,1)+1;

break;

end

end

if(A0(i,2)==0) %·统计性别

countf(1,1)=countf(1,1)+1;

else

countf(2,1)=countf(2,1)+1;

end

for j=1:10 %统计兄弟姐妹配偶

if(A0(i,4)==j-1)

counts(j,1)=counts(j,1)+1;

break;

end

end

for j=1:10 %父母孩子

if(A0(i,5)==j-1)

countp(j,1)=countp(j,1)+1;

break;

end

end

end

pwc(1:3,1)=countc(1:3,1)/ length(A0(:,1)); %计算连续变量概率

pwf(1:2,1)=countf(1:2,1)/ length(A0(:,1));

pws(1:10,1)=counts(1:10,1)/ length(A0(:,1));

pwp(1:10,1)=countp(1:10,1)/ length(A0(:,1));

for i=1:length(A1(:,1))

for j=1:3

if(A1(i,1)==j)

countc(j,2)=countc(j,2)+1;

break;

end

end

if(A1(i,2)==0)

countf(1,2)=countf(1,2)+1;

else

countf(2,2)=countf(2,2)+1;

end

for j=1:10

if(A1(i,4)==j-1)

counts(j,2)=counts(j,2)+1;

break;

end

end

for j=1:10

if(A1(i,5)==j-1)

countp(j,2)=countp(j,2)+1;

break;

end

end

end

pwc(1:3,2)=countc(1:3,2)/ length(A1(:,1));

pwf(1:2,2)=countf(1:2,2)/ len gth(A1(:,1));

pws(1:10,2)=counts(1:10,2)/ length(A1(:,1));

pwp(1:10,2)=countp(1:10,2)/ length(A1(:,1));

B=TEST(:,1:6);

predt=ones(length(B(:,1)),2);

for i=1:length(B(:,1)) %计算后验概率

predt(i,1)=predt(i,1)\*normpdf(B(i,3),n(1),s(1)); %已知年龄死亡概率

predt(i,2)=predt(i,2)\*normpdf(B(i,3),n(2),s(2)); %已知年龄存活概率

predt(i,1)=predt(i,1)\*normpdf(B(i,6),n(3),s(3)); %已知船票死亡概率

predt(i,2)=predt(i,2)\*normpdf(B(i,6),n(4),s(4)); %已知船票存活概率

%已知离散特征死亡概率 predt(i,1)=predt(i,1)\*pwc(B(i,1),1)\*pwf(B(i,2)+1,1)\*pws(B(i,4)+1,1)\*pwp(B(i,5)+1,1);

%已知离散特征存活概率predt(i,2)=predt(i,2)\*pwc(B(i,1),2)\*pwf(B(i,2)+1,2)\*pws(B(i,4)+1,2)\*pwp(B(i,5)+1,2);

predt(i,1)=predt(i,1)\*pw0;

predt(i,2)=predt(i,2)\*pw1;

end

predt=predt';

[~,ansa]=max(predt); %找出概率大的元素的位置

ansa=ansa'-1; %变为0-1值

csvwrite('ans.csv',ansa,1);

**四 总结**

经过这次的实验，让我对朴素贝叶斯的用法有了更进一步的了解。当使用分类器对一堆没有什么头绪的数据进行分类并得出结果时，让我感受到了模式识别算法的强大，以及它广阔的应用领域。之后，很多时候拿到一堆数据，不再需要人为得借助数据绘出各种图像进行分类等复杂的操作，只需要交给计算机，尽管计算机并不知道数据背后具有的含义，但只要借助模式识别算法，仍然可以快速得到分类结果。本次实验用到的朴素贝叶斯法和最小二乘法只是模式识别中最简单的算法，相比如果对数据进行合理的预处理，并使用其他更加“高级”的算法，应该可以取得更高的识别正确率。

**附：文件说明**

本次附件一共有：

1作业报告

2 Bayes\_AD文件夹内为广告点击预测的相关文件，包括：

Bayes.m 程序文件

Bayes\_AD.mat 工作区变量

ans.csv Matlab输出的结果

submission.csv 拼接好的用于评估的结果

3 Titanic\_B文件夹内为泰坦尼克的相关文件，包括：

TITANIC\_B.m 程序文件

TITANIC\_B.mat 工作区变量

ans.csv Matlab输出的结果

submission.csv 拼接好的用于评估的结果

4 其他文件夹存放使用梯度下降、最小二乘、SVD的其他例子