

**模式识别大作业**

题 目 Logistic回归

学 院 信息科学与工程

专 业 控制科学与控制工程

组 员 毛 盈

指导教师 赵海涛

**完成日期： 2018 年 10 月25日**

**模式识别作业报告——Logistic 回归**

姓名：毛盈 学号：Y30180664

本次作业我采用Logistic回归对泰坦尼克号上的存活者进行预测。Logistic回归是一种常用的机器学习方法，它常被用来估计某种事件的可能性或对事件进行分类。因为Logistic回归一般输出只有“0”和“1”两个类别，所以Logistic回归主要应用于二分类问题。

**一、Logistic回归原理**

在二分类问题中，对于给定数据集(x1,x2,…,xn), 其对应的输出一个 y∈{0,1},我们需要学习一个模型（预测函数）来预测输入数据的判断结果。所以首先要找到一个函数，这个函数的输出 y 只能有两种取值，即 0 或 1，分别来表示两种类别。最理想的函数是“单位阶跃”函数。但是阶跃函数并不连续，因此找到另一个Sigmoid函数：

 （1.1）

它的图像如图（1.1）所示，函数图像是一个取值在 0 和 1 之间的 S 型曲线。

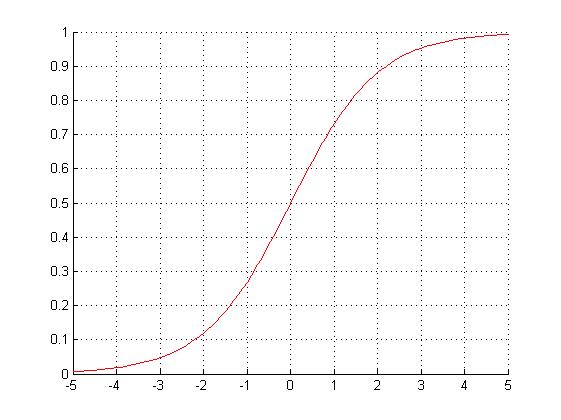


图1-1 Sigmoid函数图像

由图1-1可以看出，随着z的增大，y的值趋向于1，反之，趋向于0。当z取0时，函数值为0.5。因此可将使Sigmoid函数函数值大于0.5的数据归为一类，将使Sigmoid函数函数值小于0.5的数据归为另一类。

令，其中，w为回归系数，b为截距，因此可以得到Logistic回归的预测函数：

 （1.2）

（为预测值）的值表示结果取1的概率，因此对于输入x后，分类结果为类别1和类别0的概率如公式（1.3）所示。

 （1.3）

1. **Logistic回归的损失函数**

构造loss函数（error函数）如下：

 （2.1）

构造cost函数如下：

 （2.2）

1. **梯度下降法求的最小值**

根据梯度下降法得到的更新过程如公式（2.3），（2.4）所示：

 （2.3）

 （2.4）

其中，对J(w,b)求偏导如公式（2.5）所示：

 （2.5）

，

得到更新后的后，再次计算Sigmoid函数输出重复上述过程，直到准确率达到要求或训练次数达到设定值，即可得到分类需要的回归系数。

1. **MATLAB仿真试验**

4.1 数据导入

clear

clc

%% 读入fisheriris数据

load fisheriris.mat

str=['setosa'];

y=ones(100,1);

for i=1:100

y0(i)=strcmp(species(i),str)

if( y0(i))

y(i)=0;

else

y(i)=1;

end

end

从matlab数据库中导入fisheriris数据并进行处理。

4.2 构造样本数据

%% 构造训练样本和测试样本

% 选用样本的前2维进行实验。

X = meas(1:100,1:2)';

X = [X;ones(1,100)];

Y = y;

选择样本的前两维进行预测，需要注意的是，X需要构建三维，这是因为Y=WTX+b，三维分别为w1,w2和b。

4.3 下降梯度法更新回归系数

%% 梯度下降法求解Logistic Regression

W=rand(3,1);

alpha=0.001; %设置学习率为0.001

for i=1:5000

h=1./(1+exp(-W'\*X)); %构造预测函数

error=(Y-h'); %计算误差

W=W+alpha\*X\*error/100\*h\*(ones(1,100)-h)'; %列写损失函数

end

4.4 作图表示仿真结果

用下列程序画出后验概率杆状图，后验概率杆状图如图4-1所示。

用下列程序画出分类结果，分类结果如图4-2所示。

%% 绘制分类边界

figure();

a=min(X(1,:))-0.2;

b=max(X(1,:))+0.5;

c=min(X(2,:))-0.2;

d=max(X(2,:))+0.5;

for n=a:0.1:b

for m=c:0.1:d

if m>-W(1,1)/W(2,1)\*n-W(3,1)/W(2,1)

plot(n,m,'g.','MarkerSize',20);

else if m<-W(1,1)/W(2,1)\*n-W(3,1)/W(2,1)

plot(n,m,'y.','MarkerSize',20)

end

end

hold on;

end

end

axis([a b c d]);

%% 画后验概率杆状图

for i = 1 : 100

x=X(:,i);

y\_test(i) = 1.0/(1+exp(-W'\*x));

end

stem(y\_test)

for j=1:100

x1=X(1,j);

y1=X(2,j);

if j<=50

plot(x1,y1,'r<');

else if j>50

plot(x1,y1,'k+');

end

end

hold on;

end



图4-1 后验概率杆状图

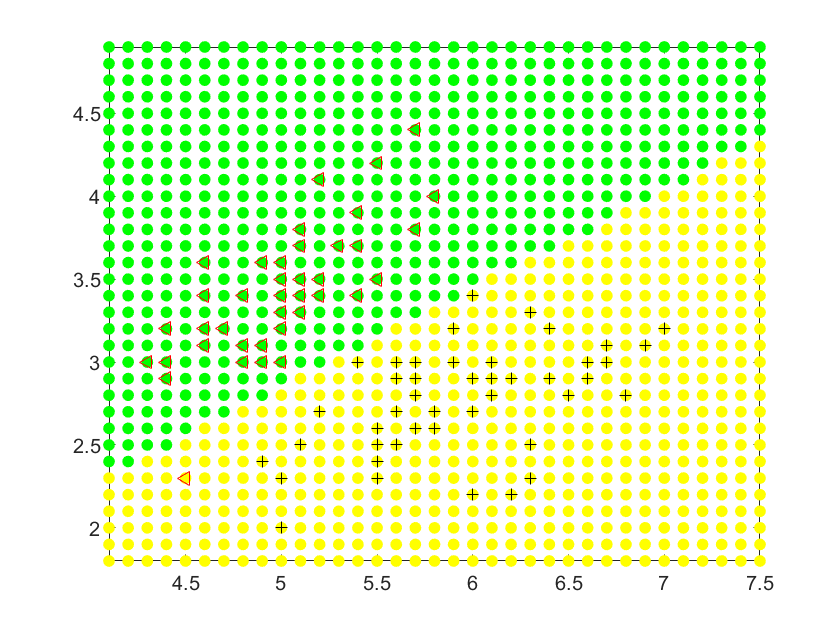


图4-2 分类结果图

1. **总结**

不知不觉一个月的《模式识别》课程已经接近尾声，在刚刚接收到编程大作业时，真是一筹莫展，完全不知道从哪里做起，即使是把上课老师讲的和吴恩达老师的视频理解了，也还是离目标很远。

在这次大作业上，遇到了很多的难题，即便是到交作业的前一刻这些问题也还是没有能够解决。首先是在做广告点击率的题目时，面对13个连续特征和26个类目特征束手无策，即便是通过查找资料了解了离散化的几种方法原理和one hot encoding的原理，也还是没有办法将其运用到编程上来。

其次，对于有多个特征的数据库没有办法准确地进行分类，因此在数据库的选择上我还是选择的最简单的iris数据库。

所以，这次的大作业我在对老师给的例子程序以及对logistic回归理论的理解上，完成了一个简单的小程序。

这次的《模式识别》大作业是进入研究生以来的第一次作业，而在这次手忙脚乱的作业中，我也深刻认识到了一点，大学以课堂为主，以基础理论为主的话，那么研究生阶段就不应该仅仅局限于课堂了，而实际编程一直以来都是我最为薄弱的地方，而这个决不能依靠老师们的课堂，而更应该注重课后的练习，所以说，赵老师所说的“课堂俩小时，课后20小时”也绝对不是开玩笑的。

最后，衷心感谢认真负责的赵老师，在赵老师的课上学会了很多实用的练习方法（lintcode刷题，网上视频教程等等），等结束了《模式识别》课程我还是会继续学习这门有用的学科，毕竟，选择了机器学习就是选择了终生学习！