

**2020年春季学期  
计算学部《机器学习》课程**

**Lab 2实验报告  
logistic回归**

|  |  |
| --- | --- |
| 姓名 | 李卓君 |
| 学号 | 1180300210 |
| 班号 | 1803103 |
| 电子邮件 | zhuojunlee724@gmail.com |
| 手机号码 | 18845636557 |

**目录**

[1 实验目的与实验要求 2](#_Toc54030877)

[1.1 实验目的 2](#_Toc54030878)

[1.2 实验要求 2](#_Toc54030879)

[1.3 实验环境 2](#_Toc54030880)

[2 实验背景与原理 3](#_Toc54030881)

[2.1 实验背景 3](#_Toc54030882)

[2.2 实验原理 3](#_Toc54030883)

[3 4](#_Toc54030884)

建议写出：问题的描述，解决问题的思路，实验的做法，实验结果的分析，结论，自拟标题

# 实验目的与实验要求

## 实验目的

理解逻辑回归模型，掌握逻辑回归模型的参数估计算法。

## 实验要求

实现两种损失函数的参数估计（1，无惩罚项；2.加入对参数的惩罚），可以采用梯度下降、共轭梯度或者牛顿法等。

## 实验环境

Windows 10, Visual Studio Code, Python 3.8.4

# 实验背景与原理

## 实验背景

实验从分类问题出发，旨在用朴素贝叶斯方法作为理论指导解决二分类问题。以平面上的二维点集为例，对于横纵坐标均符合高斯分布的两类点，我们希望划分出一条边界可以尽可能的区分这两类点使得定义的代价函数最小，且当这两类点对应维度的方差相等时可以证明这条边界恰为低一维度的超平面。

## 实验原理

假定是一个维实值列向量，Y是一个满足伯努利分布的布尔值，令，目标函数。若对于给定的，各维度的变量相互独立，

满足高斯分布，根据高斯判别分析方法，我们根据先验概率和似然函数来计算后验分布。

以的计算为例，若满足朴素贝叶斯假设：

假设满足一维高斯分布，所以有其概率分布如下：

则有

如前所述，令对应维度的方差相等，则有

所以有

将常数项合并为，系数设为，于是有

满足多维高斯分布时可以进行推广，并得到以下结果

其中是共用的协方差矩阵，分别为两者的期望向量。

a是关于x是线性的，说明了在朴素贝叶斯假设的前提下进行高斯判别分析，对于n维的数据进行二分类，我们可以得到一条n-1维的超平面作为边界，例如，对于平面上的点集我们可以得到一条直线，对于空间中的点集我们则可以得到一个平面对其进行划分。

接下来考虑关于参数的代价函数，因为a是线性的，所以可以用向量的乘积

来表示，后验分布函数就可以表示为一个关于矩阵和参数的函数：

由后验分布的定义知，这个函数实际上表示了满足了的特征下为正例的概率分布，由概率定义知，反例的概率分布为，由此，我们将二者统一化，有

对于m组满足条件独立分布的例子进行最大似然估计有

对取最小值时的即是我们所需的，考虑到这样计算不方便，我们对代价函数取一个负对数进行计算：

实际计算时往往还要对于样本数量取一个均值，所以有最终的代价函数如下

其梯度为

# logistic回归

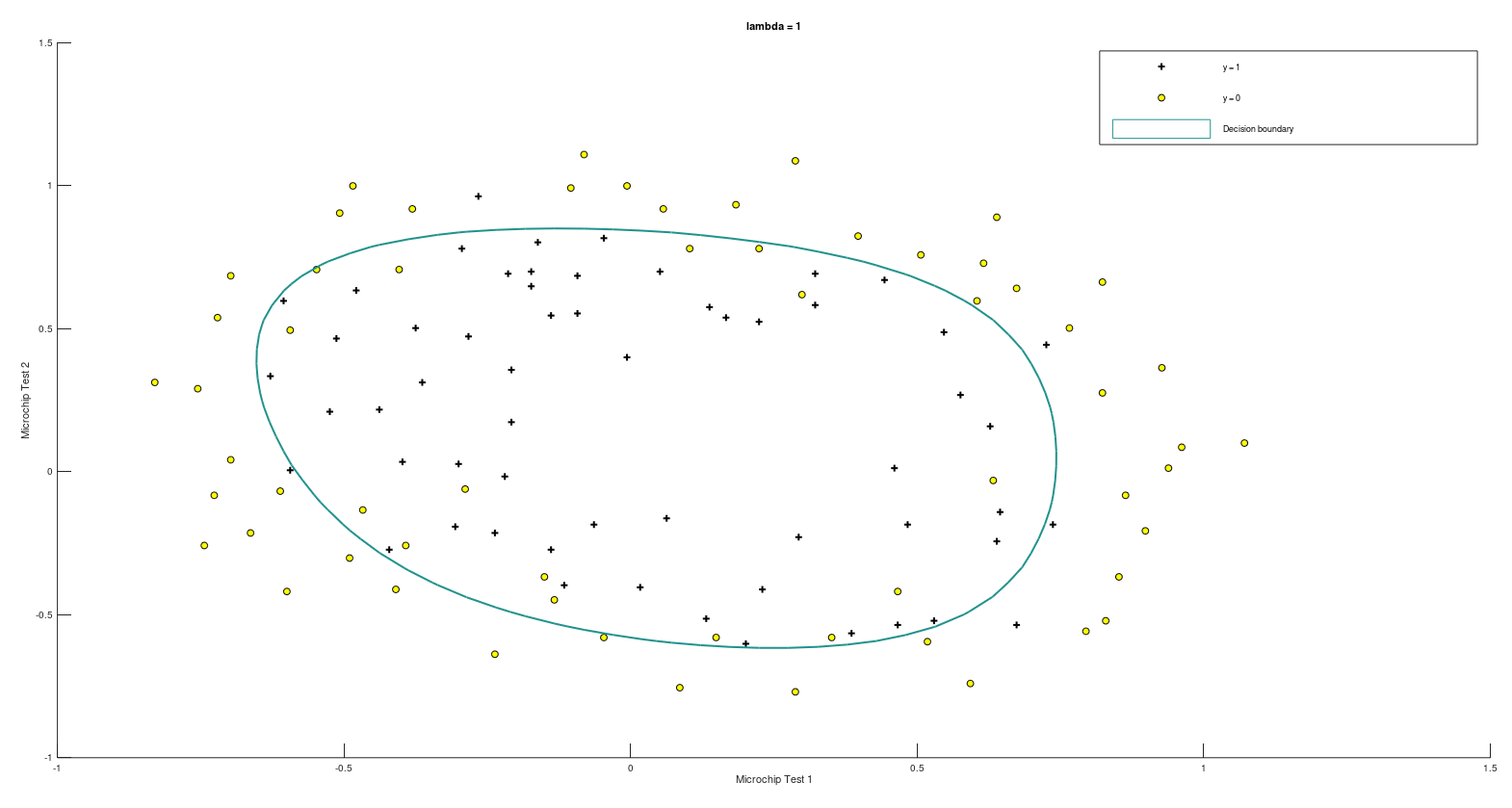
## 生成数据集

利用numpy库中的numpy.random.multivariate\_normal()函数来生成多维高斯分布的数据，为了便于以图表的形式表现，我在这里采用了二维高斯分布，设定期望向量和协方差矩阵等，进行传参会返回一个的向量，其中m表示生成样本个数。

生成协方差矩阵时需要注意，根据概率论的知识，当协方差矩阵非主对角线上的数为0时各维度满足独立分布条件，也就是满足贝叶斯假设，否则不满足。

## 代价函数

在上一小节中已经对于代价函数的方程进行了推导，增加惩罚项的方式与线性回归如出一辙，实际上，由于拟合出来的边界函数是一条直线(因为特征较少)，所以从直觉上判断，出现过拟合的可能几乎没有，而对于满足朴素贝叶斯的高斯分布的样本集，在上一节中也已经证明了边界函数为一条直线，但实际情况中数据不一定满足朴素贝叶斯假设和高斯分布，也就不能直接地用直线去拟合，这时用曲线拟合的情况下，加入惩罚项就十分有必要了，在吴恩达的课程实验中要求选做了这种情况下的数据的logistic回归。



## 参数计算

参数计算选用之前较为熟悉的梯度下降法，原理不再赘言。

## 实验结果

对于满足朴素贝叶斯假设的生成数据

