

**2020年春季学期  
计算学部《机器学习》课程**

**Lab 2实验报告**

|  |  |
| --- | --- |
| 姓名 | 王丁子睿 |
| 学号 | 1183710211 |
| 班号 | 1803104 |
| 电子邮件 | 1183710211@stu.hit.edu.cn |
| 手机号码 | 19845178018 |

**目录**

[1 问题概述 3](#_Toc53752014)

[2 数据生成 3](#_Toc53752015)

[3 问题求解 3](#_Toc53752016)

[3.1 无正则项 3](#_Toc53752017)

[3.2 含正则项 4](#_Toc53752018)

[4 应用 5](#_Toc53752019)

# 问题概述

给定一系列点集，每个点有一个分类，试求解一个分类模型，来根据每个点的坐标/属性，对每个点类别进行区分。

本实验主要解决二分类问题，但可以扩展到任意数量的分类。

# 数据生成

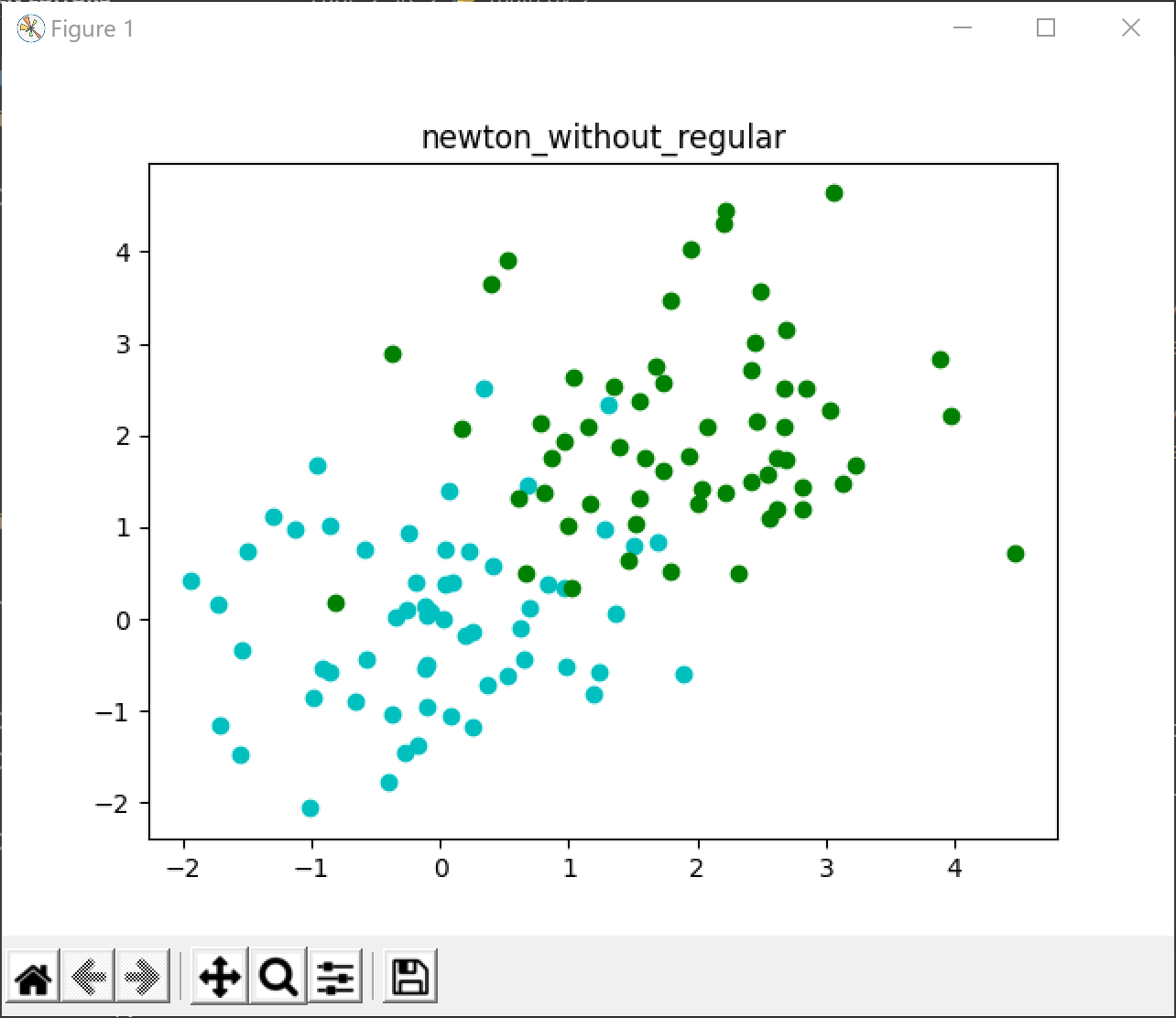
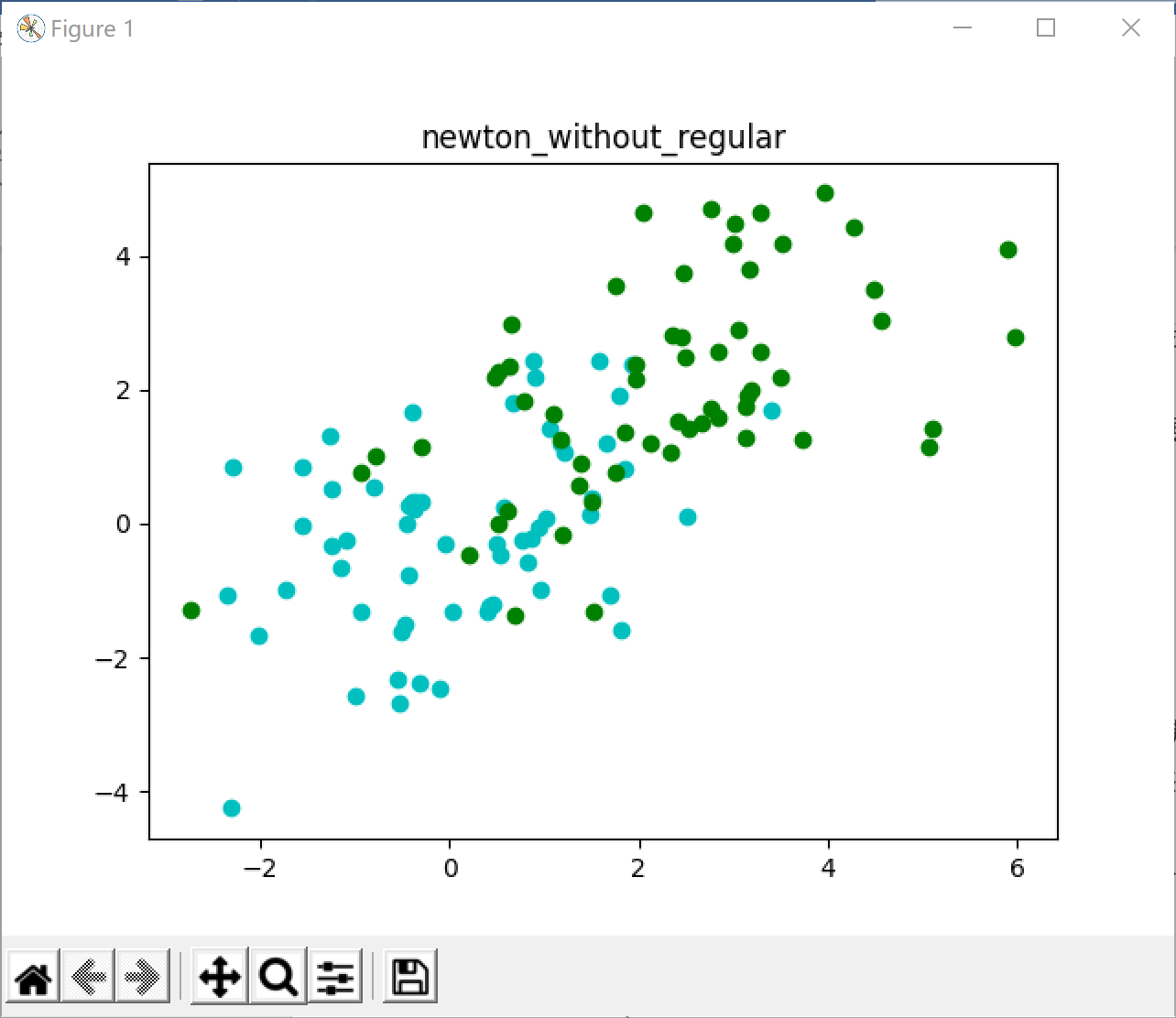
（代码见src/generate.py）

给定一个协方差矩阵，据此生成一系列满足该协方差矩阵的高维正态分布的点集。

为了区分不同的类别，各类别满足的正态分布的平均数会有较大差别，从而有很明显的距离，在本实验中，取二维正态分布，两个分类满足的正态分布的平均数分别为和。

此外，若分布满足朴素贝叶斯矩阵，则分布的协方差矩阵为对角矩阵，本实验中取；否则，分布的协方差矩阵不为对角矩阵，本实验中取。

生成数据量满足训练集:验证集:测试集 = 60:20:20，测试集数据如下图所示：



具体数据见src/related\_normal\_distribution和src/unrelated\_normal\_distribution。

# 问题求解

## 无正则项

由于Logistic函数在分类问题中的良好的性质，本实验采用该函数来解决分类问题。

具体来说，设概率函数满足：

假设已知，则将代入之后，得到一个概率，若概率大于，则判定该点的分类为1；否则，则判定为0。

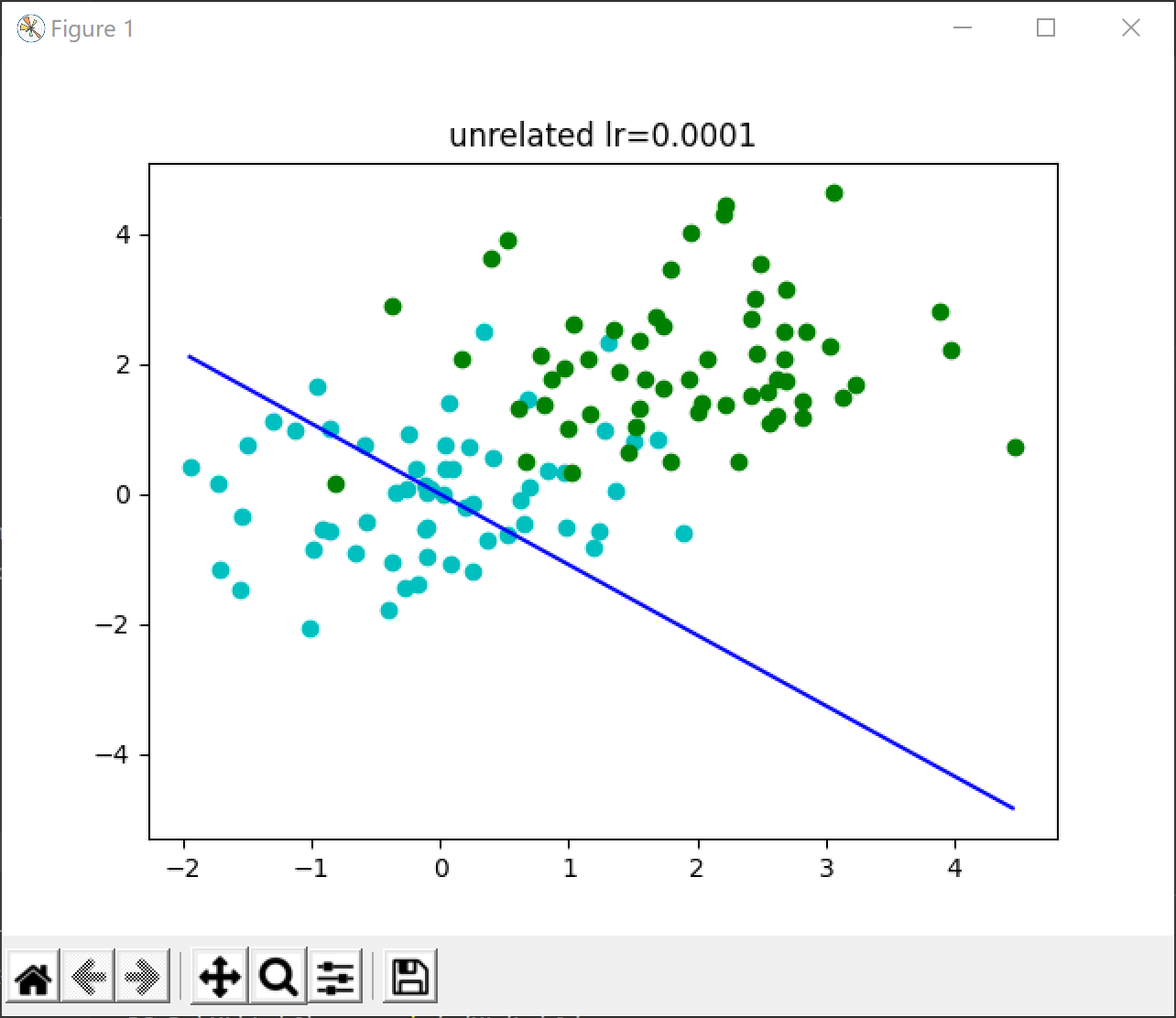
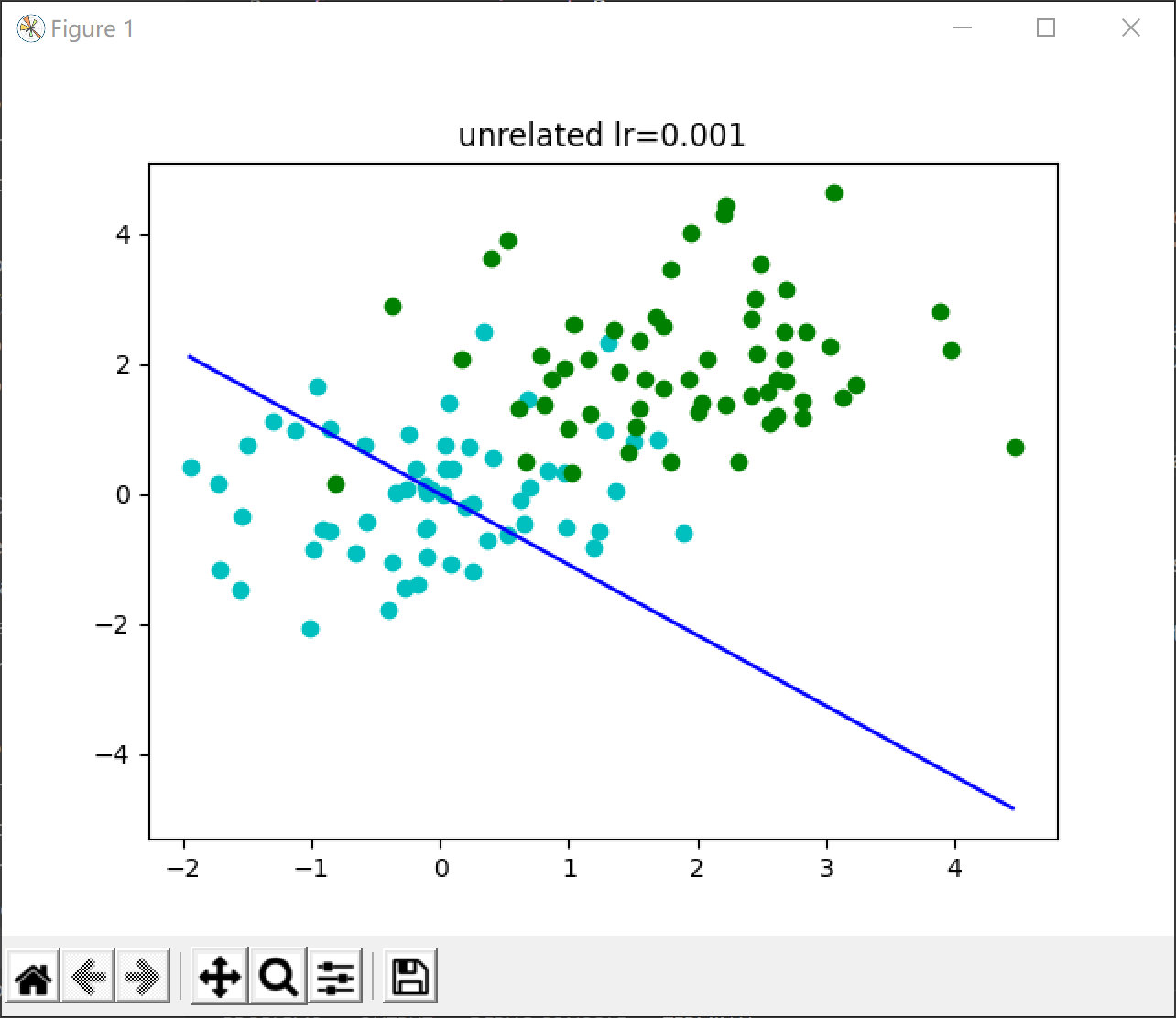
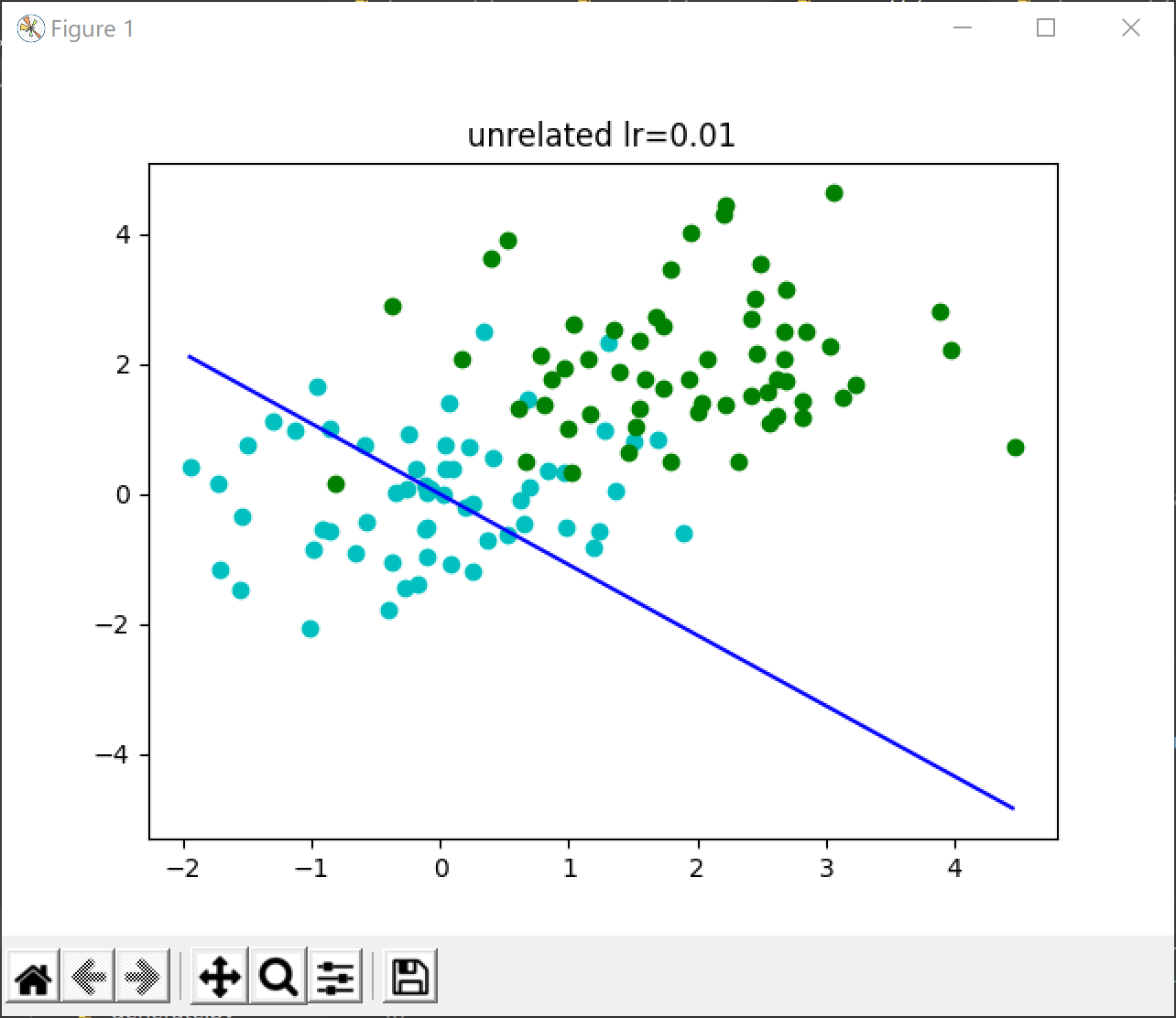
下面用似然函数法估计参数的值，易知似然函数为：

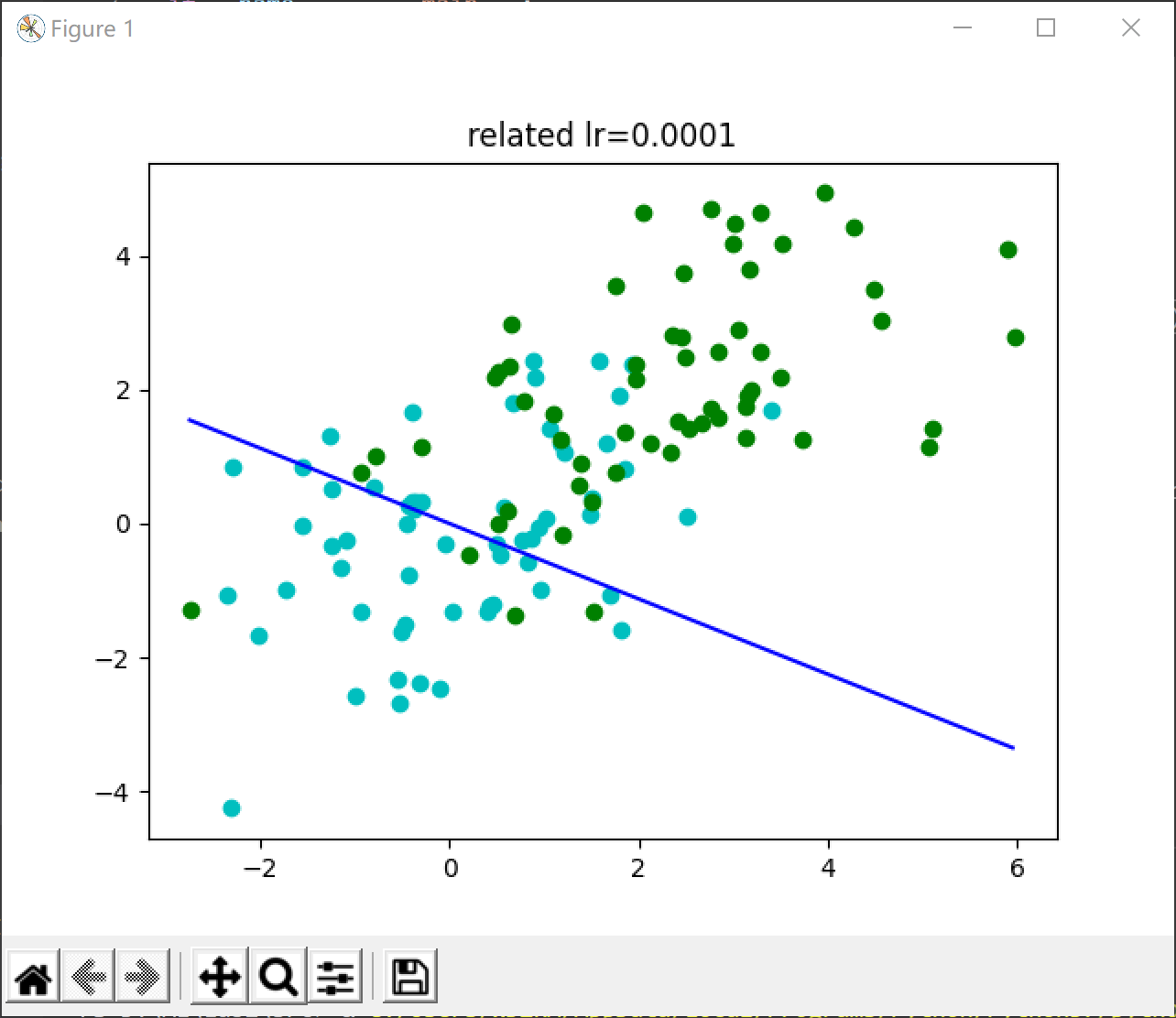
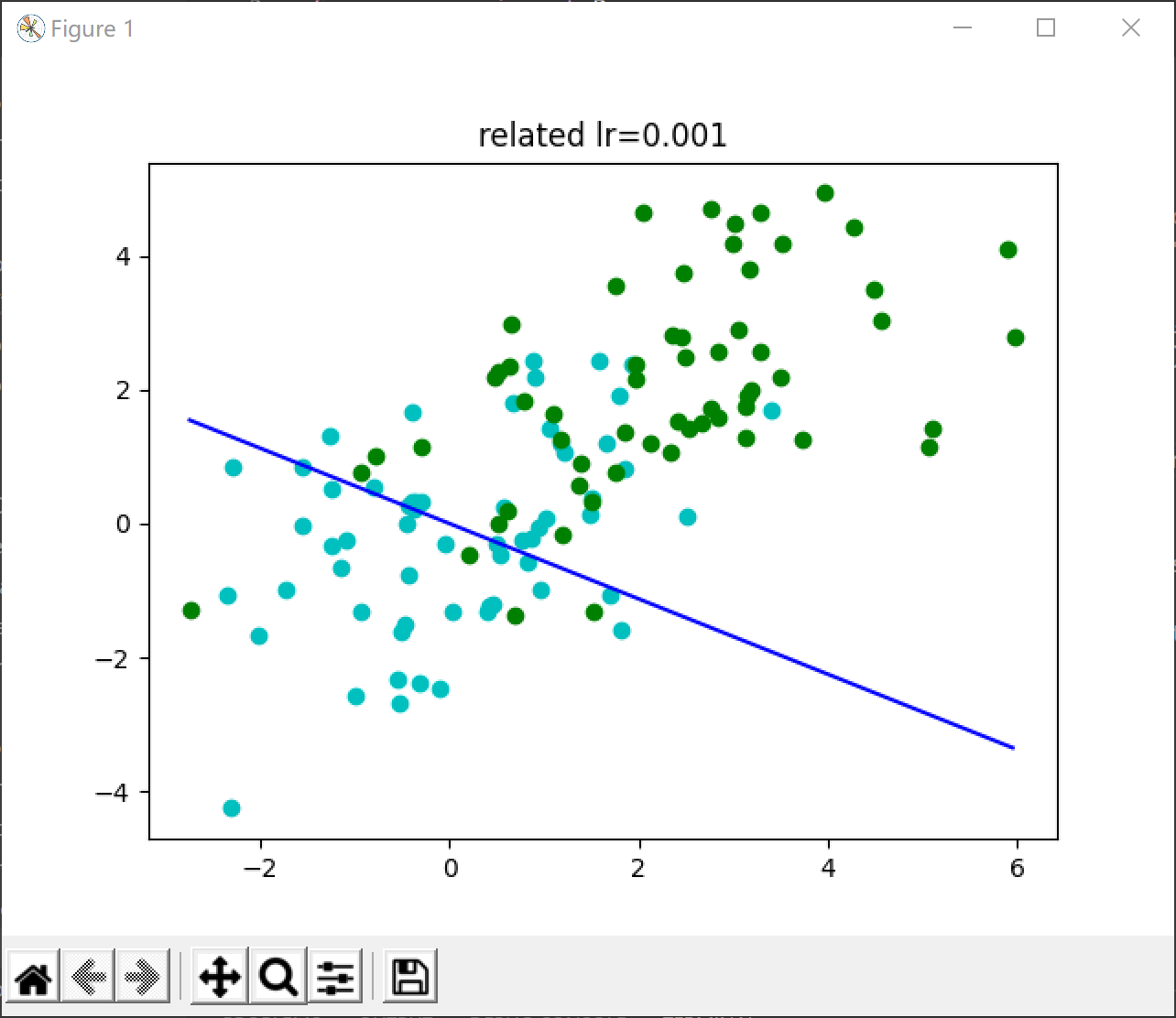
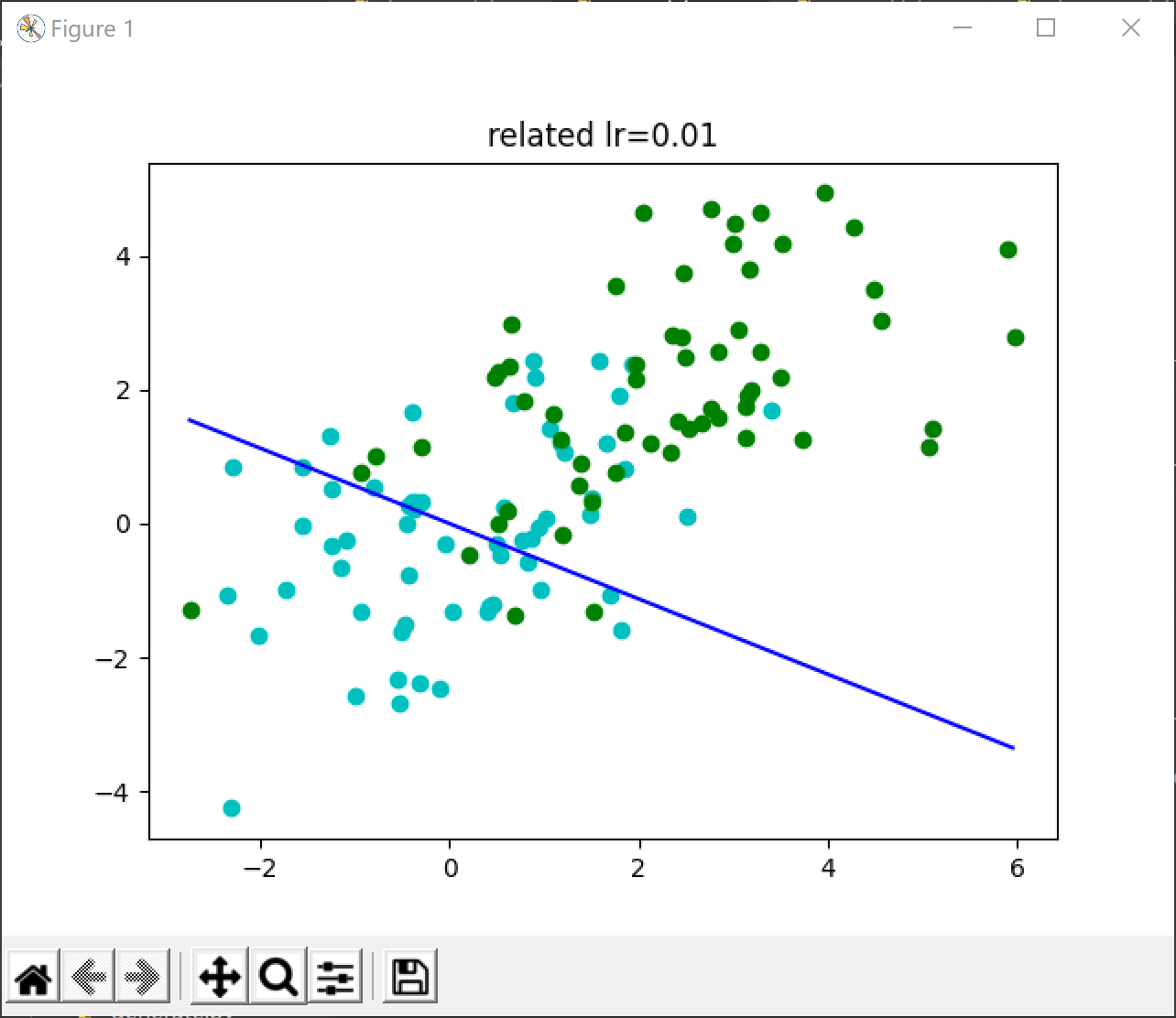
对其求对数，将得到的函数作为损失函数：

对其求导得：

应用梯度下降法，每次令减去，控制梯度绝对值和小于时停止迭代，即可得到拟合函数。

取定，取，数据是否满足朴素贝叶斯（unrelated、related），得到的结果如下图所示：





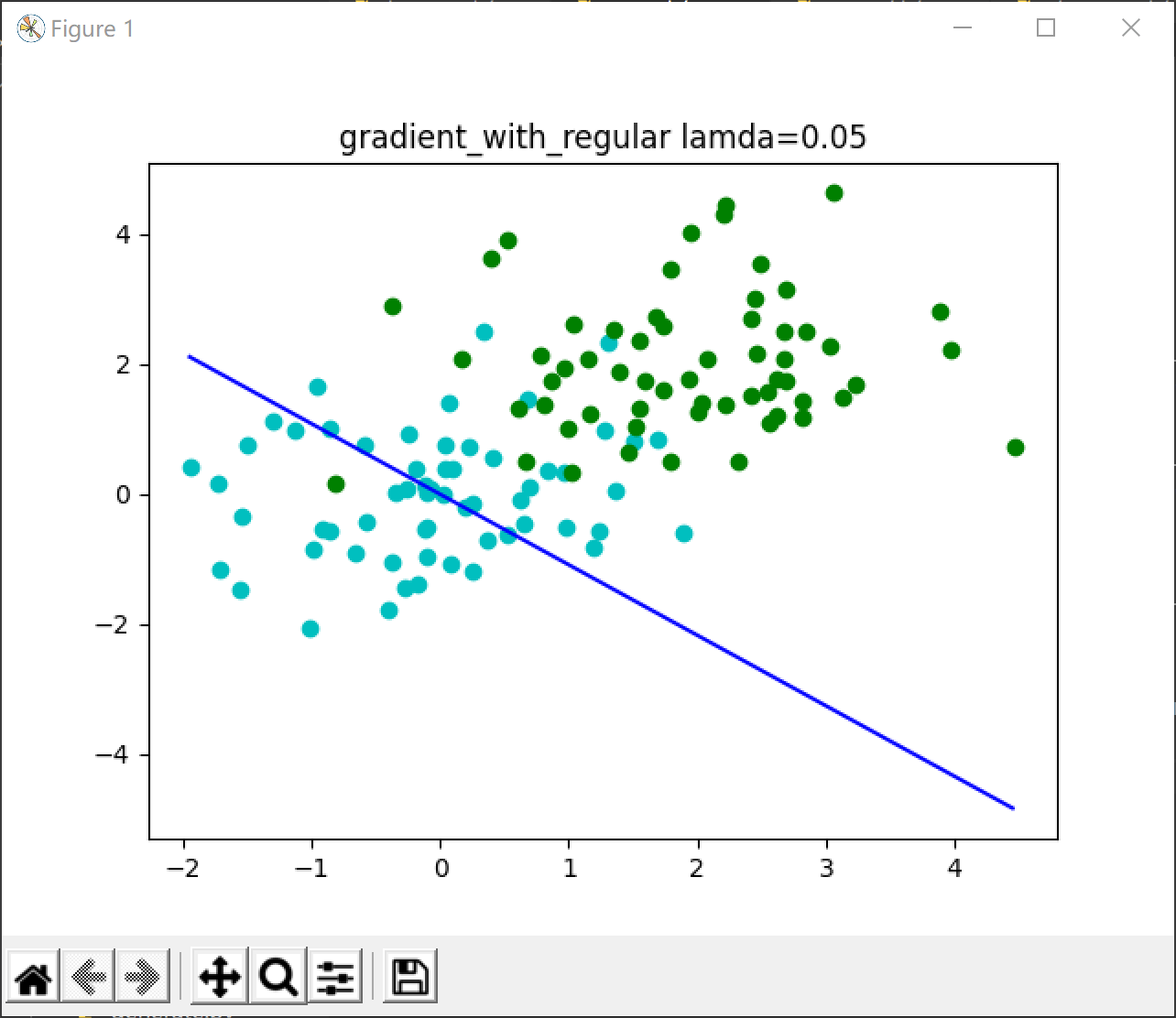
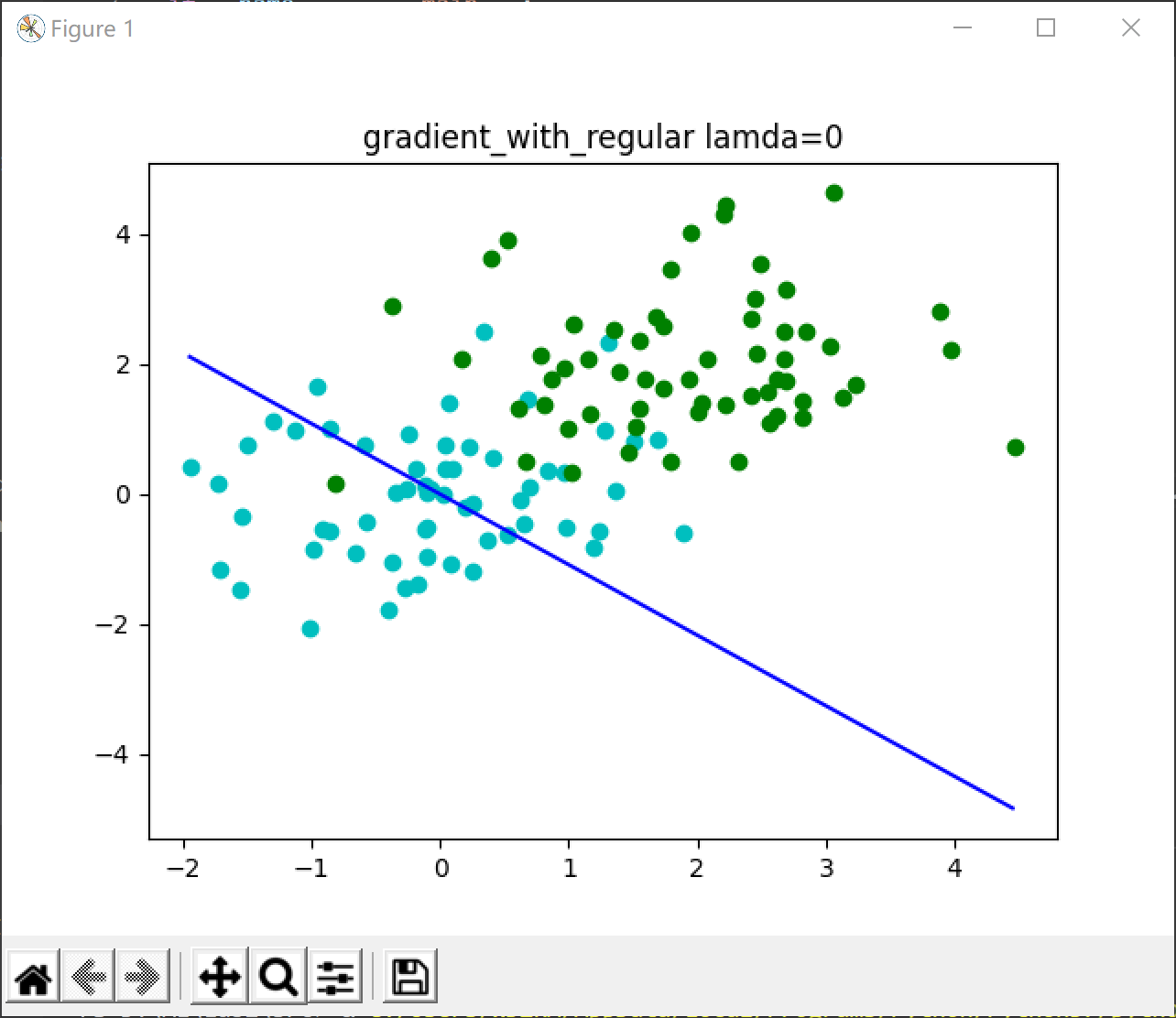
## 含正则项

在损失函数中引入正则项，即：

对其求导得：

依然采用梯度下降法进行求解。

取定，取，数据满足朴素贝叶斯，对比引入正则项前后的结果如图：

**

可以发现，引入正则项前后拟合直线的区别并不大，这是因为，该方法仅考虑了各点的一次项，因此并没有明显的过拟合现象。

# 应用

取UCI的数据集[MONK's Problems Data Set](http://archive.ics.uci.edu/ml/datasets/MONK%27s+Problems)进行测试，其为一个多元二分类问题的数据集，每个参数都是取一定范围内的一个整数值。

取超参数，得到的结果为：



由于matplotlib只能同时显示两个维度，而多元逻辑回归对任意两维的区分效果并不一定特别好，因此图像的参考价值有限，故不在此展示。