

**模式识别大作业**

题 目 logistic regression

学 院 信息科学与工程

专 业 控制科学与工程

组 员 张欣茹

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 regression**

组员：张欣茹 学号：Y30180697

一: logistic regression

1.1: logistic回归

回归，指研究一组随机变量(Y1 ，Y2 ，…，Yi)和另一组(X1，X2，…，Xk)[变量](https://baike.so.com/doc/2977483-3140692.html" \t "https://baike.so.com/doc/_blank)之间关系的统计分析方法，又称多重回归分析，常见的有利用最小二乘法进行线性回归等。普通线性回归主要用于连续变量的**预测**，即：线性回归的输出y的取值范围是整个实数区间。逻辑回归用于离散变量的**分类**，即它的输出y的取值范围是一个离散的集合，主要用于类的判别，其输出值y表示属于某一类的概率。经典的LR用于二分类问题，常用来预测。

1.2: logistic回归模型

以sigmoid作为回归模型，对于任意的x值，对应的y值都在区间(0,1)内，，具体公式如下：



令，则原式为



sigmoid函数图像如图1-1所示：

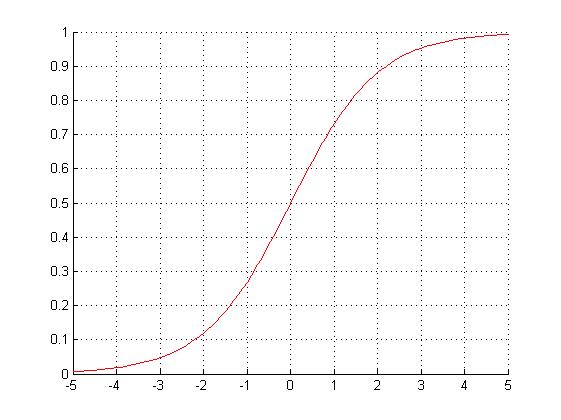


图1-1 sigmoid函数图像

因此logistic回归问题转化为通过训练数据估计参数W和b的问题。

1.3: 损失函数

假设有m个样本训练样本，表示样本i的输入和输出，以特征向量作为输入，输出为标签。通过训练样本，我们希望实现即：估计输出值与实际输出越接近越好。因此针对一个训练样本的损失函数为：



对于m个训练样本，定义成本函数代表全体训练集损失函数的平均值：



训练的目标就是找到w,b使得成本函数值最小。作业程序是将w，b两个参数合在一起为参数W：



其中n表示样本输入的行数即特征数。

二：梯度下降法

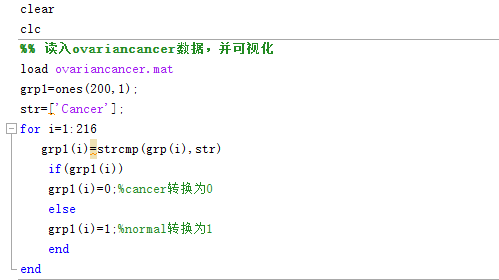
梯度下降法是一个[最优化](https://baike.so.com/doc/5868873-6081730.html" \t "https://baike.so.com/doc/_blank)[算法](https://baike.so.com/doc/2758411-2911336.html" \t "https://baike.so.com/doc/_blank)，通常也称为最速下降法，沿函数值下降最快的方向进行迭代。将梯度下降法应用到logistic回归，具体公式如下：



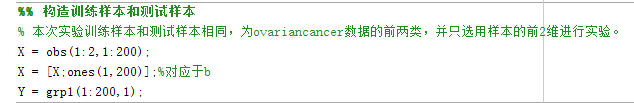
本文程序运用随机梯度下降法，其每次仅用一个样本点来更新参数，而且由于可以在新样本到来时对分类器进行增量式更新，因此随机梯度下降法是一个在线学习算法。程序里面的在每次更新参数的时候都会调整，每一次在数据集上更新时，是单调递减的，但是由于有一个常数项，永远不会减小到0，这样做可以保证多次迭代后新数据仍然有一定的影响但是影响又不会很大。

三：程序

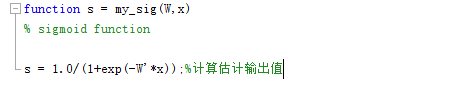
读入ovariancancer数据，其样本输出为字符串“cancer”、“normal”，所以首先对数据进行处理，转化为标签（0，1），具体程序如下：



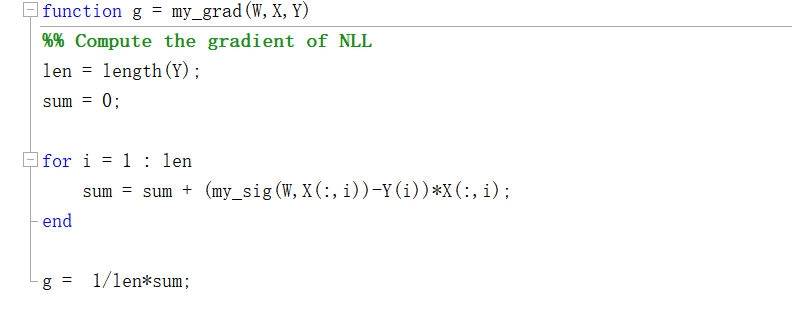
其次构建训练样本和测试样本，取100个样本中前两维进行训练，所以估计参数为。由于把w和b合在一起，所以在样本中加入元素为1的行向量，具体程序如下：



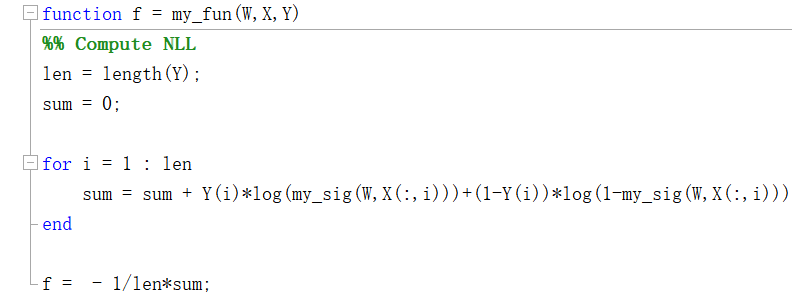
在用梯度下降法求解logistic回归前，计算估计输出值：



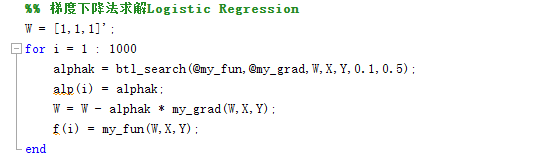
计算dw

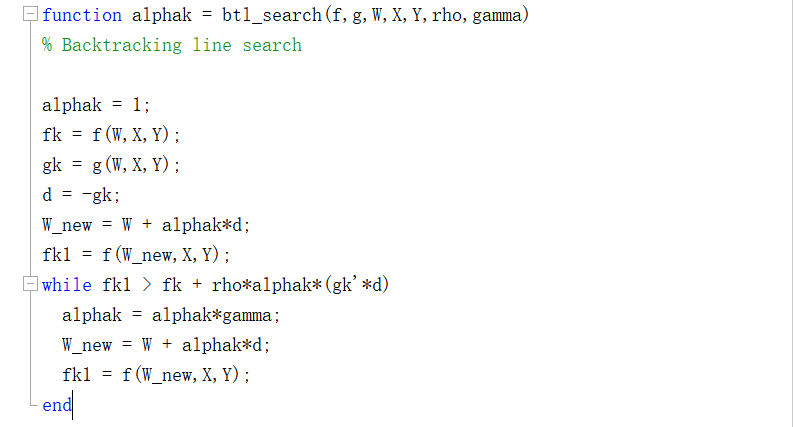


计算损失函数：

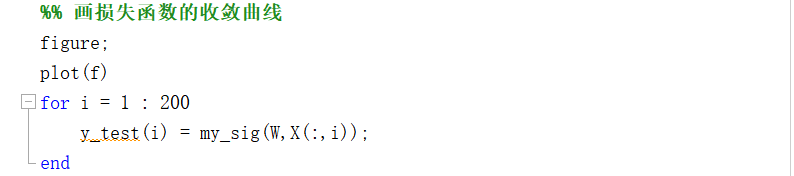


用梯度下降法求解logistic回归，迭代次数为1000：

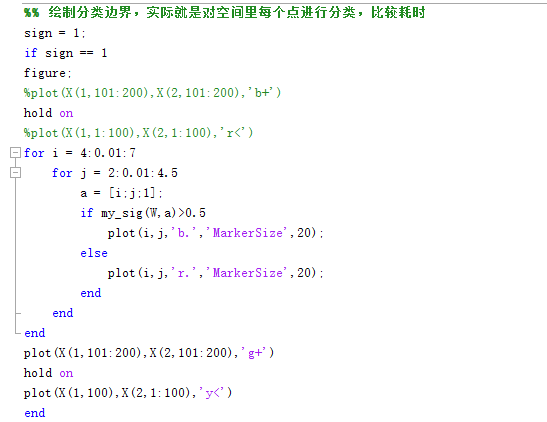




用测试样本进行测试，并绘制曲线：







四：实验结果



图4-1 损失函数收敛曲线



图4-2后验概率杆状图

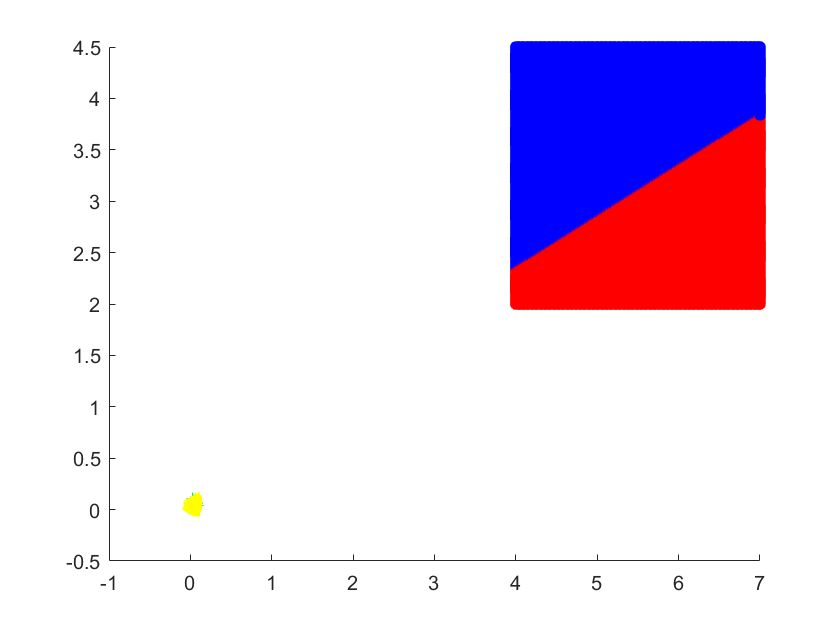


图4-3分类边界