**Homework3 Logistic Regression 实验报告**

**实验要求：**

有2类二维空间点，A类和B类。

A类点以（0，0）为中心、（1，0；0，1）为协方差矩阵的二维高斯分布；

B类点以（1，2）为中心、（1，0；0，2）为协方差矩阵的二维高斯分布；

随机生成300个A类点，200个B类点，并用Logistic Regression的方法进行分类（GD和SGD）。画出不同epoch等参数下的结果。

**实验思路：**

1. 生成数据集：首先生成这500个点，以列表的形式表示位置坐标，再把所有的坐标放入一个大的列表中。再生成一个label列表，包含300个’A’和200个’B’。
2. 为了更加清晰直观，把生成的300个A类点和200个B类点分别用两种颜色的符号绘制在一个二维坐标系内。
3. 引入逻辑回归算法，包括定义sigmoid函数、定义梯度上升算法函数来求出参数θ

（由于数据样本较小，为保证更高的准确性，采用GD而不是SGD）。

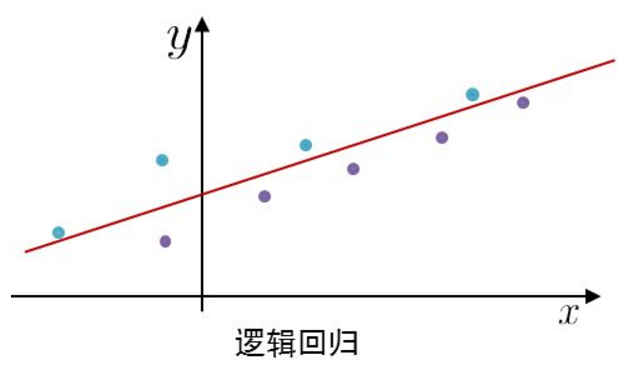
1. 增加附加功能：根据画出来的决策边界计算准确率，绘制epoch-accuracy曲线，求得最高正确率以及最高正确率下的最小的迭代次数。
2. 利用求出来的参数θ以及最佳且最小迭代次数画出决策边界。

**实验原理：**

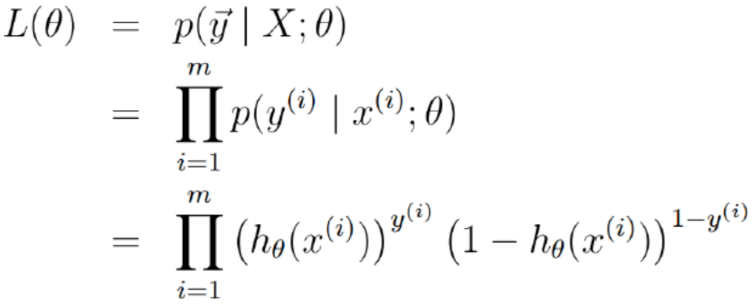
回归与分类：连续变量预测的定量输出称为回归；离散变量预测的定性输出称为分类。

线性回归完成的是回归拟合任务，而对于分类任务，我们同样需要一条线，但不是去拟合每个数据点，而是把不同类别的样本区分开来，即线性分类器。

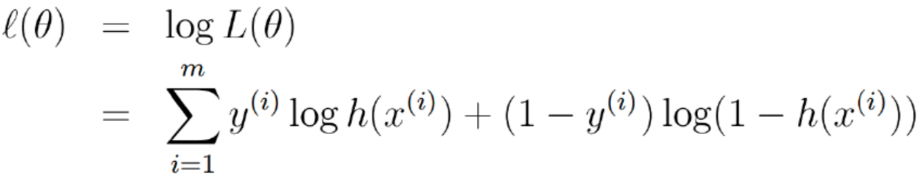
逻辑回归：是解决分类问题的一个算法，逻辑回归进行分类的主要思想是：根据现有数据对分类边界线建立回归公式，以此进行分类。



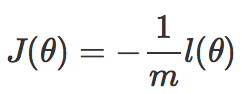
逻辑回归的似然函数：



对其似然函数求对数：

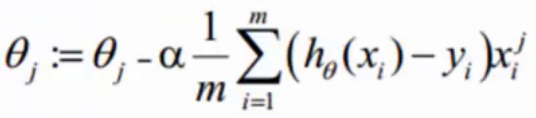


当似然函数为最大值时，得到的θ即可认为是模型的参数。求似然函数的最大值，可以使用一种方法，梯度上升，但可以对似然函数稍作处理，使之变为梯度下降，然后使用梯度下降的思想来求解此问题，变换的表达式如下：



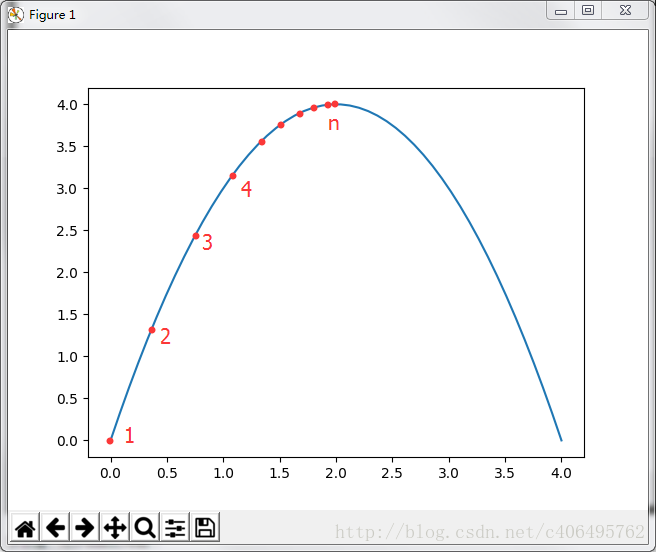
（由于乘了一个负的系数，所以梯度上升变梯度下降。）

梯度下降算法可使用下面的式子不断迭代更新得到最终结果：



α为步长，也就是学习速率，控制更新的幅度。

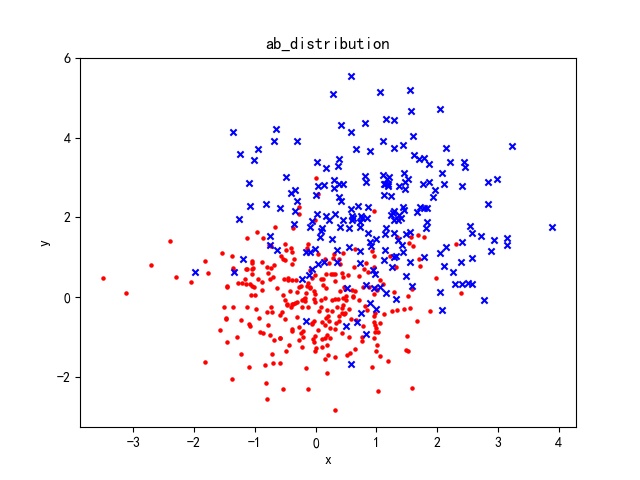
梯度上升：



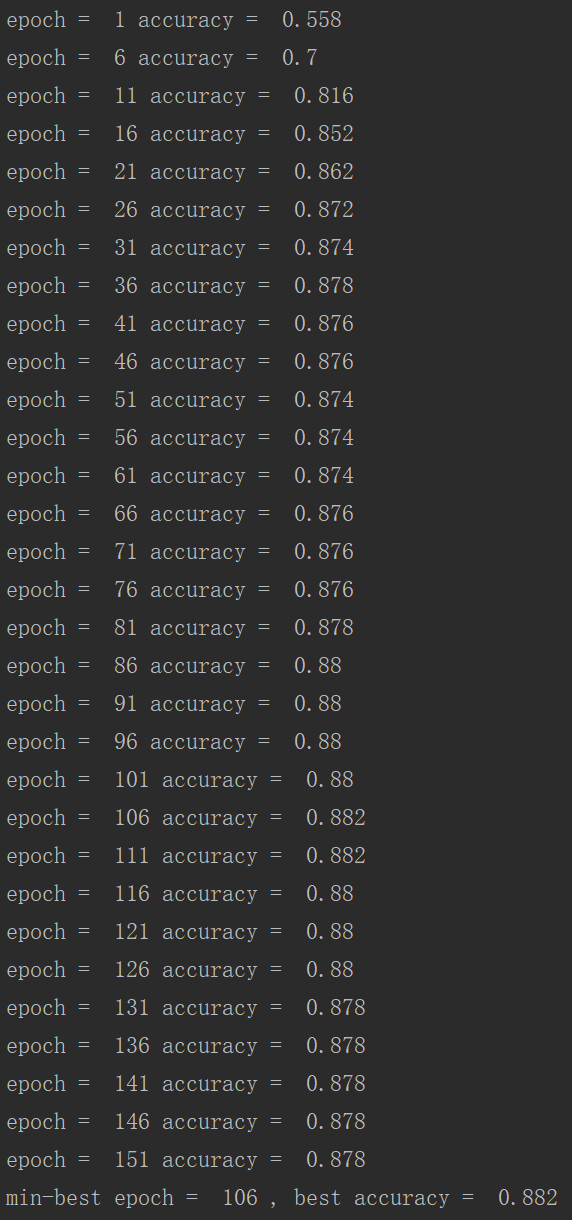
在本实验中，生成的500个点形成数据集，每一个点的横、纵坐标都是特征值，加上它们的类标签A、B（分别对应0、1），构建数组，生成特征向量，每一个点是一个训练样本，共3类变量、500个训练样本。把它们放在一个矩阵中，共500行，每一行是一个训练样本，3列，每一列代表一类变量。用逻辑回归算法求出参数θ，然后就能绘制出决策边界、计算正确率等。

**实验结果：**

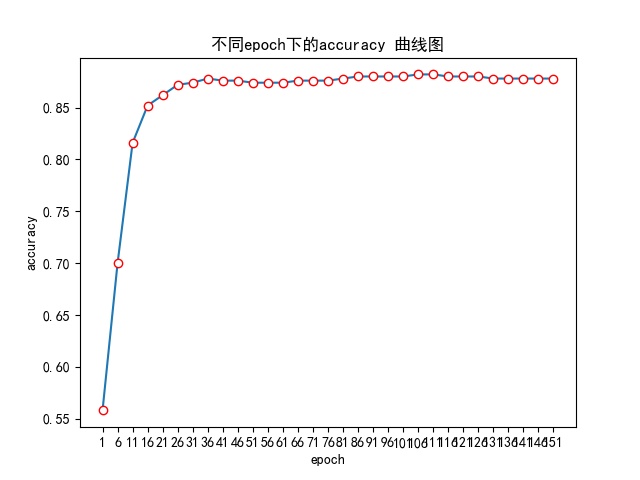
AB点分布图（A是红色点、B是蓝色叉）：



不同迭代次数epoch下的准确率：



迭代次数epoch与准确率曲线：



可以看到，在迭代次数很小时，准确率随着迭代次数的增加有明显的提升。而当迭代次数达到30次左右时，准确率达到最高值并基本维持不变。最高准确率是88.2%，达到最高准确率的最小迭代次数是106次。

用最佳且最小迭代次数绘制的最佳拟合决策分界线：

