**Logistic回归**

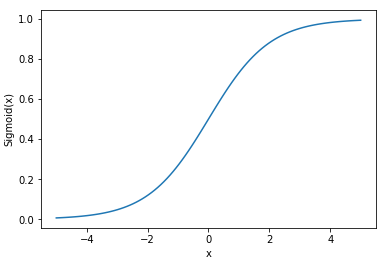
1. 基础知识
2. 利用Logistic回归进行分类的主要思想是：根据现有数据对分类边界线建立回归公式，以此来进行分类。
3. 训练分类器就是在寻找最佳拟合参数，使用最优化算法。
4. Logistic回归的特点：

优点：计算代价不高，易于理解和实现

缺点：容易欠拟合，分类精度可能不高

1. Sigmoid函数：

当x=0时，值为0.5；随着x的增大，值逼近于1；随着x的减小，值逼近于0。Sigmoid函数的值位于0~1之间，大于0.5的数据被分入1类，小于0.5的被归入0类。Sigmoid函数图像如下所示：



1. 基于最优化方法（梯度上升法）确定最佳回归系数

将Sigmoid函数的输出记为z，即

1. 梯度上升法的思想是：要找到某函数的最大值，最好的方法是沿着该函数的梯度方向搜索。将梯度记为▽，则函数f(x,y)的梯度表示为：

梯度以为着沿x的方向移动，沿y的方向移动。

1. 梯度上升算法到达每个点后都会重新估计移动的方向。迭代的过程中，梯度算子总是保证我们能选取到最佳的移动方向。梯度算子总是指向函数值增长最快的方向，这里说的是移动方向，将步长记为α，则梯度上升算法的迭代公式：

该算法一直被迭代执行，直至到达某个停止条件为止。

1. 梯度下降法

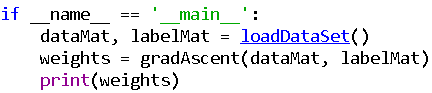
与梯度上升法一样，只是加法变为减法，对应的公式为：

梯度上升法用来求函数的最大值，而梯度下降法用来求函数的最小值。

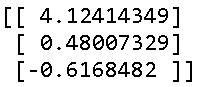
三、 使用梯度上升找到最佳参数





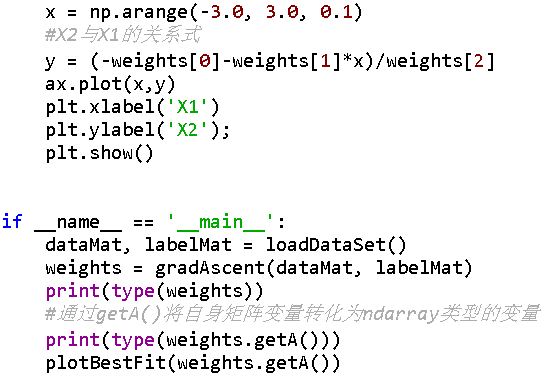


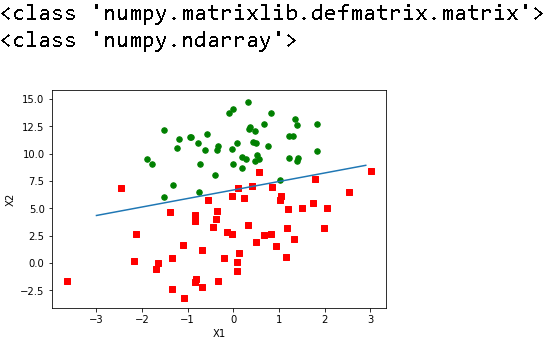
weights:



四、 分析数据：画出决策边界





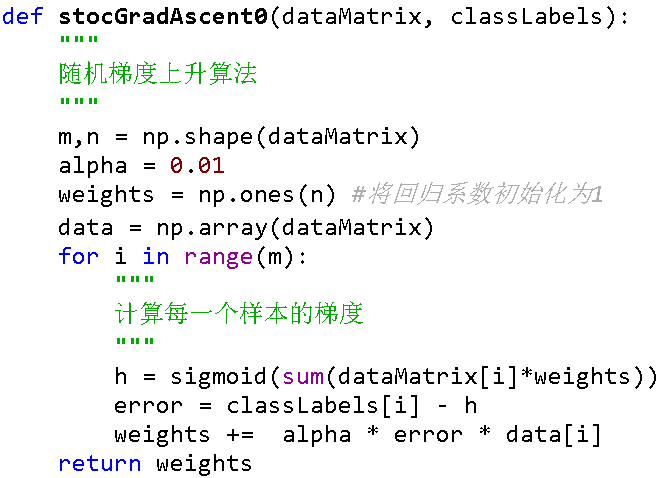


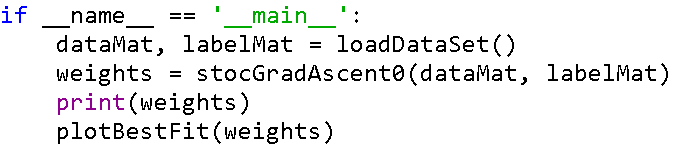
迭代了500次后，只错分了两个点，效果还不错。

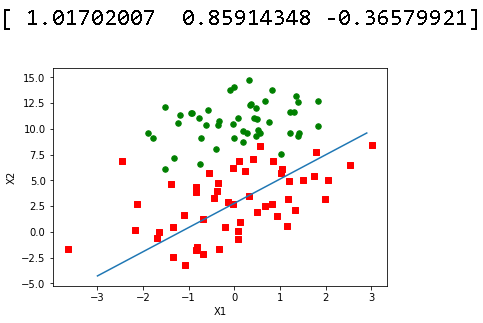
五、 训练算法：随机梯度上升

1. 由于上面的算法在数据集变大的时候，计算复杂度变高。

改进的算法：一次仅用一个样本点来更新回归系数，即随机梯度上升算法。







效果并不理想，错分了1/3的数据。

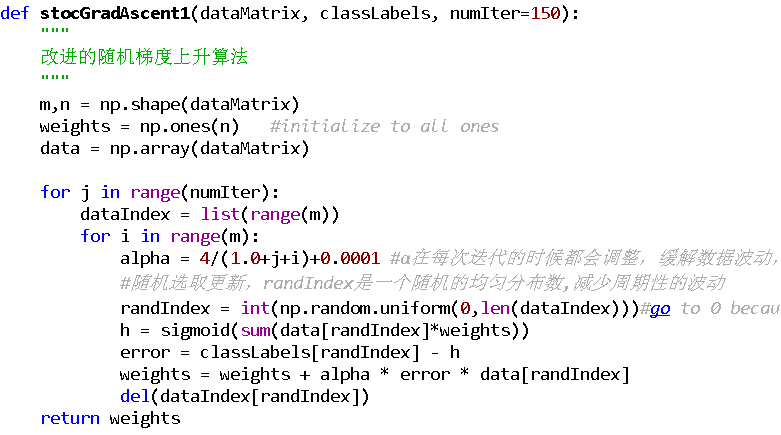
2. 一个判断优化算法优劣的可靠方法是看它是否收敛，也就是参数是否达到了稳定值。

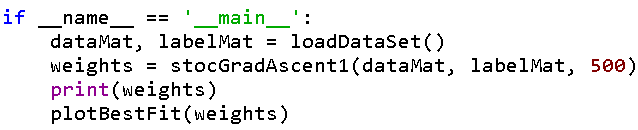
3. np.random.uniform(low,high,size)

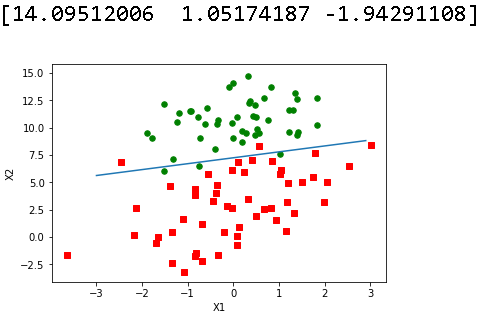
从一个均匀分布[low, high)中随机采样（左闭右开），输出为int型或者是一个数组。

1. 改进后：

随机选取样本来更新回归系数，减少了周期波动







六、 用疝气病症预测病马的死亡率

1. 收集数据：数据中有30%的值是缺失的

2. 用logistic算法进行回归分类

将测试集上每个特征向量乘以最优化方法得来的回归系数，再将该乘积结果求和，最后输入到Sigmoid函数中，如果对应的Sigmoid值大于0.5就预测类别为标签1，否则预测类别为标签0。

