**实验：逻辑回归**

张靖祥 计算机172 2017304010413

# 1 数据集介绍

参考：<https://blog.csdn.net/wangyuankl123/article/details/102777014>

癌症的30个特征，以及是否为癌症的标签。数据一共有569组30维。其中两个分类分别为良性：357

恶性：212

因为是30维度的数据，因此无法画图展示。

处理方法有3种：

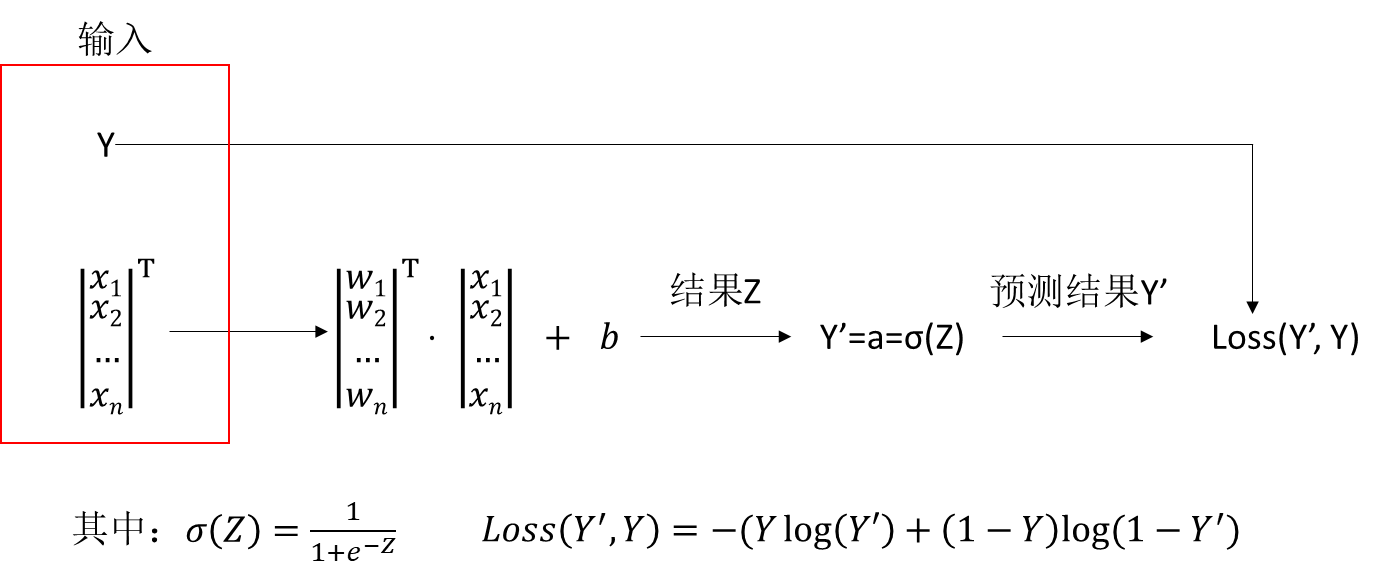
1. 仅仅选用前n个数据作为分类依据，但是必然存在大量数据丢失
2. 数据主成分分析，降为到更低的维度，然后进行逻辑回归，数据少量丢失
3. 直接对30个维度进行逻辑回归，速度可能较慢。

因此我将依次使用上述的**第一种和第三种**进行分析

# 2. 逻辑回归原理

**注：**此部分内容为我理论课的作业，因为解释的十分详细，所以在此拿来借鉴。

逻辑回归的计算流图如下



首先我们初始化n维向量x与常量b，可以初始化为0。因为逻辑回归是凸优化问题，所以无论从哪一个初始值出发，都会得到最优解，因此我们可以任意初始化x和b。

接下来将已知数据x1,x2 … xn输入到数据流图进行计算，得到的结果为Y’，与Y计算得到损失函数Loss。

通过计算得到的结果是Y’，而Y’>0.5的时候可以将其归为2类，Y’<0.5的时候可以将其归为1类。也就是说，样本得到2类的概率为Y’，而得到1类的概率为1-Y’。当真实结果Y=1的时候，我们希望Y’尽可能的大，Y=0的时候，我们希望1-Y’尽可能的大。此时定义了函数YY’(1-Y)(1-Y’)将而这统一起来，即：无论Y等于0还是1，我们都希望YY’(1-Y)(1-Y’)尽可能的大。而log是单调递增的函数，将YY’(1-Y)(1-Y’)取对数后得到的结果我们同样希望尽可能的大，即：取反后得到的损失函数尽可能的小。而log得到的函数容易计算导数，所以要做一个取对数的操作。

然后就是最关键的反向传播的过程：

由于，所以根据链式法则：

此时可以推导到w和b：

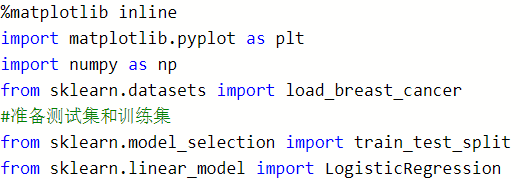
上式中，dwi和db即为要求得的导数值，X为输入向量，Y为输入的标签，a为计算得到的Y’。

我们希望Loss尽可能的小，而求得的导数值正是直接对Loss产生影响的自变量的值的导数，所以要用该自变量X和b减去相应的导数值。

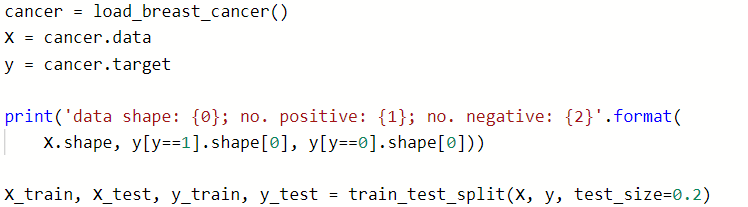
设定学习率为α，于是有：

# 3. 实验内容

调用的库

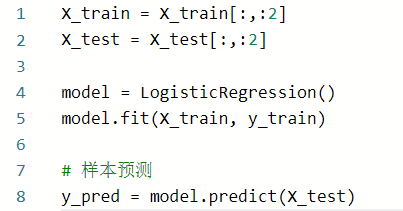


读取数据

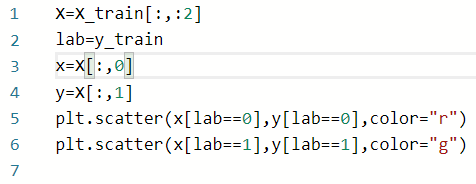


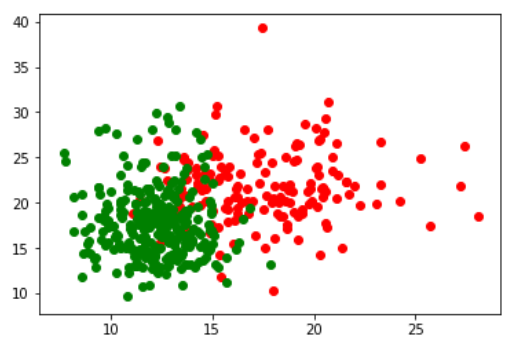
## 2.1选取部分参数作为分类依据

这里我以两个参数为例

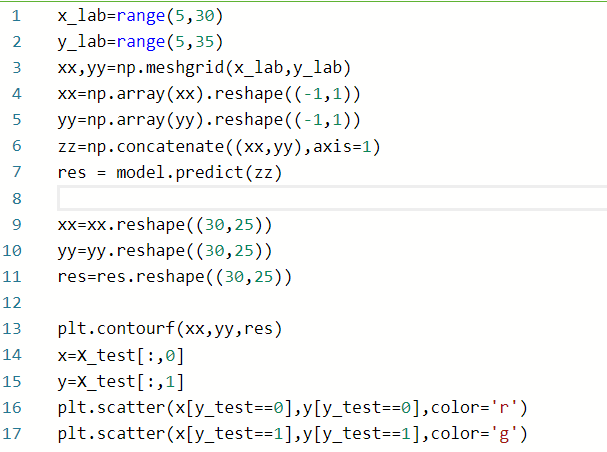


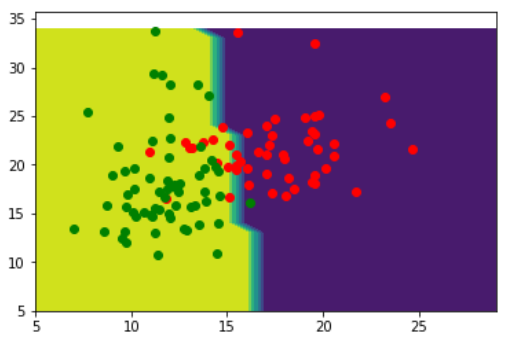
绘制该区域





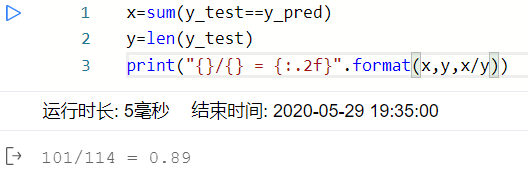
对预测结果进行画图





上图中，黄色区域为预测为非癌症的区域，蓝色区域为预测结果为癌症的区域，绿色的点表示真实值是非癌症，红色的点表示真实值是癌症，可见，逻辑回归只能对线性可分的问题进行区分，无法拟合复杂边界。

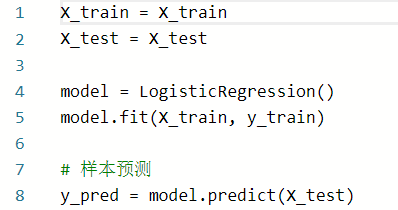
最终结果：



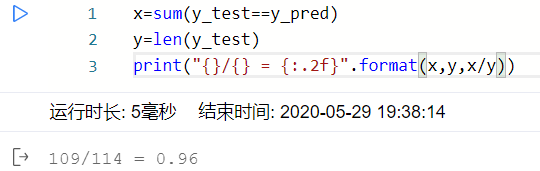
准确率百分之89

## 2.2直接进行逻辑回归

所有维度一起进行线性回归，因为维度大于2，因此无法绘图。



最后结果



准确率百分之96，比双变量的效果好