【任务2 - 逻辑回归算法梳理】时长：2天

1. **逻辑回归与线性回归的联系与区别**

1.联系：

线性回归+sigmoid函数=逻辑回归

2.区别：

（1）功能不同：

线性回归是做回归的，逻辑回归是做分类的。

（2）参数求解方法不同：

线性回归是用最小二乘法求解参数，逻辑回归是用梯度上升法求解参数

<https://blog.csdn.net/lx_ros/article/details/81263209>

逻辑回归与线性回归都属于广义线性回归模型,其区别与联系从以下几个方面比较：

分类与回归:回归模型就是预测一个连续变量(如降水量，价格等)。在分类问题中，预测属于某类的概率，可以看成回归问题。这可以说是使用回归算法的分类方法。

输出:直接使用线性回归的输出作为概率是有问题的，因为其值有可能小于0或者大于1,这是不符合实际情况的，逻辑回归的输出正是[0,1]区间。见下图，

参数估计方法：

线性回归中使用的是最小化平方误差损失函数，对偏离真实值越远的数据惩罚越严重。这样做会有什么问题呢？假如使用线性回归对{0,1}二分类问题做预测，则一个真值为1的样本，其预测值为50，那么将会对其产生很大的惩罚，这也和实际情况不符合，更大的预测值说明为1的可能性越大，而不应该惩罚的越严重。

逻辑回归使用对数似然函数进行参数估计，使用交叉熵作为损失函数，对预测错误的惩罚是随着输出的增大，逐渐逼近一个常数，这就不存在上述问题了1

也正是因为使用的参数估计的方法不同，线性回归模型更容易受到异常值(outlier)的影响，有可能需要不断变换阈值(threshold),线性回归分类的情况见下面两图:

无异常值的线性回归情况: 蓝线为求得的h(x)，上图中可选阈值为0.5作为判断肿瘤是否是良性。

有异常值的线性回归情况:

这个时候再想有好的预测效果需调整阈值为0.2，才能准确预测。

使用逻辑回归的方法进行分类，就明显对异常值有较好的稳定性。如下图:

参数解释:

线性回归中，独立变量的系数解释十分明了，就是保持其他变量不变时，改变单个变量因变量的改变量。

逻辑回归中，自变量系数的解释就要视情况而定了，要看选用的概率分布是什么，如二项式分布，泊松分布等

---------------------

版权声明：本文为CSDN博主「lx\_ros」的原创文章，遵循CC 4.0 by-sa版权协议，转载请附上原文出处链接及本声明。

原文链接：<https://blog.csdn.net/lx_ros/article/details/81263209>

1. 逻辑回归的原理

<https://blog.csdn.net/qq_41577045/article/details/80302991>

1. 逻辑回归损失函数推导及优化

<https://www.cnblogs.com/volcao/p/9314443.html>

1. 正则化与模型评估指标

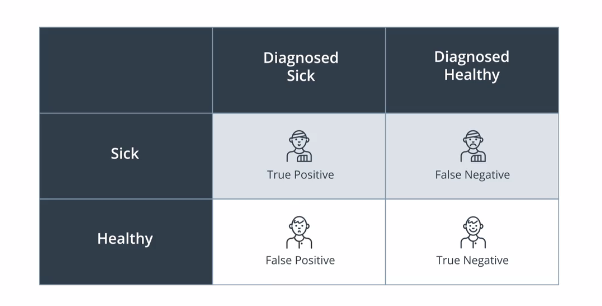
正则化https://www.cnblogs.com/90zeng/p/Regularization\_and\_model\_selection.html

模型评估指标

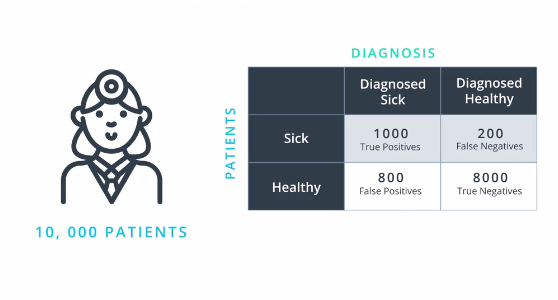
<https://cloud.tencent.com/developer/article/1407814>

### **a.使用混淆矩阵评估模型**

如下图所示，我们以去医院就诊为例（生病为阳性，健康为阴性）。当一个病人被确诊为生病时，我们称之为True Positive（真阳性）;当一个健康的人被认为健康时，我们称之为True Negative（真阴性）；当一个病人被误诊为健康时，我们称之为False Negative（假阴性）；当一个健康的人被误诊为生病时，我们称之为False Positive（假阳性）。这就是一个典型的****混淆矩阵****，用来描述模型性能的一张表。

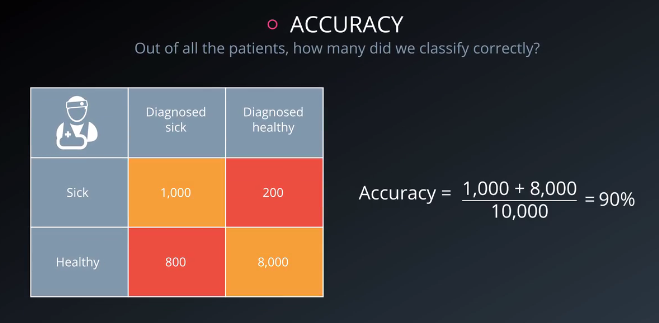


假设在一个模型中，我们共有10000名患者，其中有1000名患者被确诊，我们标进True Positive这一格；有200名患者被误诊为健康，我们标进False Negative这一格；有800名健康的人被误诊为患病，我们标进False Positive；有1000名健康的人被认为健康，我们标进True Negative这一格。



#### **b.使用准确率，查准率和查全率来评估模型**

* 准确率Accuracy 准确率(Accuracy)也是评估模型性能的一个指标。继续以上面10000名患者诊断为例。我们可以计算出准确率为：(1000+8000)/10000\*100%=90%，计算公式为：(True Positive样本数+True Negative样本数)/总样本数



* 查准率Precision 查准率指的是计算阳性样本中真阳性的比率。上面例子的查准率为：1000/(1000+800)\*100%=55.56%，计算公式为：True Positive样本数/总阳性样本数。
* 查全率Recall 查全率指的是真正患病的样本数在患病总样本数中的比率，即1000/(1000+200)=83.33%,计算公式为：True Positive样本数/(True Positive样本数+True Negative样本数)
* 总结 Accuracy = (预测正确的样本数)/(总样本数)=(TP+TN)/(TP+TN+FP+FN) Precision = (预测为阳性且正确预测的样本数)/(所有预测为阳性的样本数) = TP/(TP+FP) Recall = (预测为阳性且正确预测的样本数)/(所有真实情况为阳性的样本数) = TP/(TP+FN)

1. **逻辑回归的优缺点**

优点：

1）预测结果是界于0和1之间的概率；

2）可以适用于连续性和类别性自变量；

3）容易使用和解释；

缺点：

1）对模型中自变量多重共线性较为敏感，例如两个高度相关自变量同时放入模型，可能导致较弱的一个自变量回归符号不符合预期，符号被扭转。​需要利用因子分析或者变量聚类分析等手段来选择代表性的自变量，以减少候选变量之间的相关性；

2）预测结果呈“S”型，因此从log(odds)向概率转化的过程是非线性的，在两端随着​log(odds)值的变化，概率变化很小，边际值太小，slope太小，而中间概率的变化很大，很敏感。 导致很多区间的变量变化对目标概率的影响没有区分度，无法确定阀值。

1. **样本不均衡问题解决办法**

<https://blog.csdn.net/zhongjunlang/article/details/79568601>

1. sklearn参数

<https://blog.csdn.net/ustbclearwang/article/details/81235892>

Pip install playML