### 逻辑回归与线性回归的联系与区别

1. 模型假设不同。

自变量应变量是线性关系，

线性回归假设数据服从高斯分布，

对于所有的观测值，它们的误差项相互之间是独立的，并且服从正太分布，拥有同样的方差

逻辑回归的假设数据服从伯努利分布，非正及反，且假设样本为positive的概率用sigmoid函数表示

2. 寻找参数的方法不同。

线性回归使用的是最小二乘法来寻找参数，

逻辑回归使用的是最大似然估计的方法，运用梯度下降来求解参数

最小二乘法，使得误差最小的那个参数

最大似然估计是，使数据可能性最大的参数

3. 输出结果不同。

线性回归的输出结果是连续性的numerical数值型结果

逻辑回归则是处理分类问题，根据threshold进行分类，一般二分类问题就输出0或1

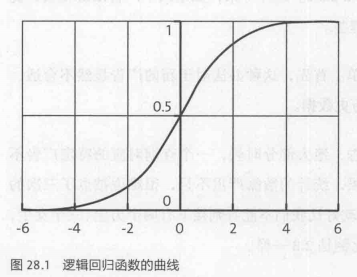
- Threshold：比如预测违约风险的时候，threshold越高，被认为违约的可能性就越低，以0.9为例的话，意味着违约可能性要达到0.9，才标记为违约

### 逻辑斯谛回归模型

逻辑斯谛回归模型的运用，首先在吴军的《数学之美》这本书中，运用广告的案例对为什么用使用逻辑斯谛回归模型进行了说明。以搜索广告为例子，之前的广告搜索在于预付广告金钱量的多少，一个广告的点击量取决于广告商投入费用的条件。但是在，面临完全新的广告（此时**没有历史数据量**作参考），旧的广告无人问津（**点击量少**），广告的点击量与**放置的位置**有关，一般越靠前，其点击量越多。因此，面对搜索广告的业务，如何**平衡这些因素，凸显广告的公平性，即满足顾客的搜索需求**，需要一个**好的模型来模拟这些影响因素的关系**。往往我们知道，可以引入**权重**对影响每一个因子在模型中的重要程度，因此基于多因素的条件下用统一的数学模型进行数学模型的建模，而**逻辑回归模型**可以满足以上的需求。从数学的角度来考虑，另外，考虑二分类的情况，即我们通常用0和1表示函数的预测类别，它们是一个离散的二分类情况，阶跃函数就是一种表示两种离散值的特殊函数。阶跃函数在0和1之间的瞬间跳跃的过程是非常难处理的，因此另一个函数叫做Sigmoid函数，具有类似的性质，且在数学上更容易处理！

（1）逻辑回归函数的模型如下：





可以看出，逻辑斯谛回归模型在-6-+6之外的区别，基本都贴近于0或者1，而它的函数曲线的范围却在0-1之间，是不是正好**符合概率的口味**呀。没错，逻辑斯谛回归模型就是利用它的**y的变化是在0-1之间的特性**，组合**自变量的值在负无穷到正无穷之**间的特点来进行分类概率预测的。

（2）进一步的，实现**多变量的融合**，图中的横坐标就是z，z可以用以下函数表示，引入各类权重来表示更影响参数的重要程度。





（3）以二项式逻辑斯谛回归模型为例子，





其中为了方便，是扩充的形式，表示与的内积（也就是上面的z），

（4）进一步的引入几率的概念，这是基于数学模型的特殊性提出的解决方式，对于一个事件它发生的概率为，那么它对应的，为什么要引入它，先别急，等到了计算损失函数的时候，就能发现它的妙用了。

几率为。

对数几率为



（5）重要的，我们需要对模型的参数（**也就是那些权重**）做出正确的估计，此时就需要用到最大似然函数的方法，通过最终的样本观察结果来反推这些参数最可能的值！

对于样本，y的值为0或者1

假设，

运用似然函数（获得所有观测样本的可能性乘积，使得它最大，基于这样条件下，求得的参数可能就是最优参数，这是极大似然函数的思想），其中最大似然函数是运用已经观测到的样本结果，来反推最有可能导致这样结果的参数值的一种方法。

将似然函数用对数似然函数表示为：

那么**对数几率**正好存在于这个等式中，

因此：对数似然函数进一步的转换为

（6）上一步骤最后推导出的公式正是逻辑斯谛回归模型的损失函数，因此我们最终的目的旨在获得最优参数，不要忘记我们的初衷，是利用极大似然函数对模型进行参数的估计。因此我们旨在获得对数似然函数的极大值，接下来运用梯度下降算法或者牛顿梯度下降算法进行参数的估计。



（7）权重的更新如下：，这里需要注意的是，因为是求解极大值，所以权重的更新应为减号。

（8）在迭代次数范围内，对权重不断地进行更新，最终获得最优的参数。

进一步的将二项式逻辑斯谛回归模型拓展到多项式逻辑斯谛回归模型中，多项式的逻辑斯谛回归模型





这一案例与就像是支持向量机中的一对多的模型，本来支持向量机是针对二分类模型的，那么对多分类模型而言，每次将一个标签当做一类对待，剩余标签当做另一类对待，不断地训练模型，就像是一棵单边的决策树，不断进行分类。

（9）重要的，在将来的样本预测阶段，如果概率值大于0.5，就认为这一类的标签为1，否则就是另一类为0.

以上就是逻辑斯谛回归模型的思路与算法。它的优点在于直接对分类可能性进行建模，无需假设数据的分布，避免了假设分布不准确所带来的问题，从而避免了假设分布不准确所带来的问题；不仅能预测出类别，而且还能近似概率预测，这对许多需利用概率辅助决策的任务很有用。对率函数是任意阶可导的凸函数，有很好的数学性质，现有的许多数值优化算法可直接用于求取最优解！

正则化与模型评估指标：



λ是正则项系数：  
• 若λ很大，说明对模型的复杂度惩罚大，对拟合数据的损失惩罚小，这样它就不会过分拟合数据，在训练数据上的偏差较大，在未知数据上的方差较小，但是可能出现欠拟合的现象；  
• 若λ很小，说明比较注重对训练数据的拟合，在训练数据上的偏差会小，但是可能会导致过拟合。

评估指标：

精确率，召回率，F1 score，ROC曲线，准确率

逻辑回归的优缺点：

优点：  
直接对分类的可能性建模，避免假设分布的不准确带来的影响；  
可得到近似概率预测，对利用概率辅助决策的事件有帮助；  
可直接用于求取最优解  
缺点：  
当特征空间很大时，逻辑回归性能不佳；  
容易欠拟合，一般准确度不太高；  
不能很好地处理大多数特征和变量

样本不均衡问题解决办法

一、样本的过采样和欠采样。

1.过采样：复制稀有类别的样本，通过增加此稀有类样本的数量来平衡数据集。该方法适用于数据量较小的情况。

2.欠抽样：从丰富类别的样本中随机选取和稀有类别相同数目的样本，通过减少丰富类的样本量来平衡数据集。该方法适用于数据量较大的情况。

sklearn参数

sklearn.linear\_model.LogisticRegression()

正则化选择参数（惩罚项的种类）：penalty: 可选‘l1’or ‘l2’,默认: ‘l2’

正则化系数λ的倒数：C: float,默认: 1.0

是否存在截距：fit\_intercept: bool,默认: True

solver{‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’},默认: ‘liblinear’

solver参数决定了我们对逻辑回归损失函数的优化方法，有四种算法可以选择，分别是：

a) liblinear：使用了开源的liblinear库实现，内部使用了坐标轴下降法来迭代优化损失函数，适用小样本。

b) lbfgs：拟牛顿法的一种，利用损失函数二阶导数矩阵即海森矩阵来迭代优化损失函数。

c) newton-cg：也是牛顿法家族的一种，利用损失函数二阶导数矩阵即海森矩阵来迭代优化损失函数。

d) sag：即随机平均梯度下降，是梯度下降法的变种，和普通梯度下降法的区别是每次迭代仅仅用一部分的样本来计算梯度，适合于大样本数据。

newton-cg, lbfgs和sag这三种优化算法时都需要损失函数的一阶或者二阶连续导数，只能用于L2正则化。而liblinear可以在L1正则化和L2正则化时使用。

Sklearn 中的参数

https://blog.csdn.net/loveliuzz/article/details/78708359