**逻辑回归的原理及应用**

逻辑回归是机器学习中的一种分类模型。虽然名字中带有回归，但是本质是一种分类算法。之所以没有同线性回归一样解决回归问题，是因为它在线性回归的结果上，加上了sigmoid激活函数。由于算法的简单高效，在实际中逻辑回归应用非常广泛。

逻辑回归就是解决二分类问题的利器。

逻辑回归的应用场景

* 天气晴雨预测
* 是否垃圾邮件
* 是否患病
* 是否金融诈骗
* 是否虚假账号

要掌握逻辑回归的原理，必须要解决这两个问题：

* 逻辑回归的输入值是什么
* 如何判断逻辑回归的输出

首先看第一个问题，逻辑回归的输入值是什么。

逻辑回归的输入，就是线性回归的输出，就是下面线性回归公式结果。

https://xiaoke-asset-prod.kaikeba.com/testing/pPK6Pv1m7j00DAyFOSz/testing/SOpHLQQm0SBf8uyttJH.png;base64

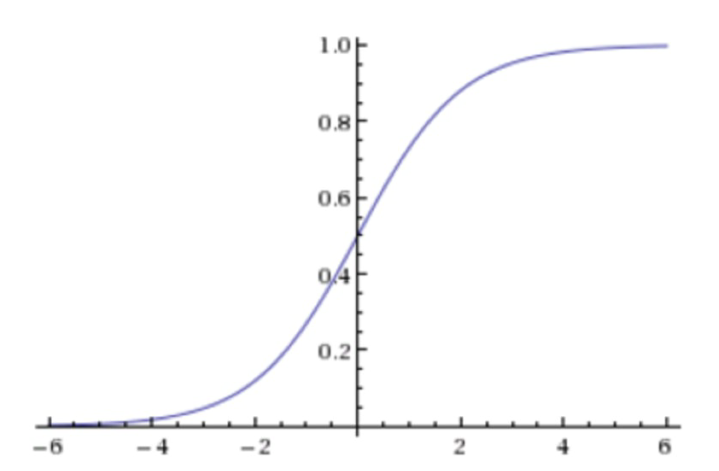
上面是逻辑回归的输入值，是逻辑回归与线性回归的联系之处，下面是逻辑回归区别于线性回归的地方，也就是我们要解决的第二个问题。我们先来了解一下sigmoid激活函数。sigmoid激活函数公式如下：

https://xiaoke-asset-prod.kaikeba.com/testing/eUjzVT3nkpc4qw33xQC/testing/xnrKzXO9cOEHGcRrJ6z.png;base64

**def logistic(x):**

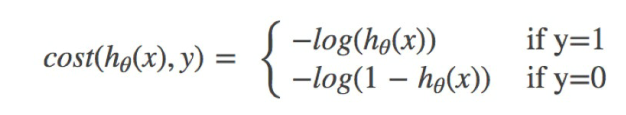
**return 1 / (1 + np.exp(-x))**

线性回归的结果输入到sigmoid激活函数当中，输出结果是 **[0, 1]** 区间中的一个概率值，默认0.5为阈值，逻辑回归最终的分类是通过属于某个类别的概率值来判断是否属于某个类别，并且这个类别默认标记为1（1表示正例），另外一个类别会标记为0（0表示反例）。



逻辑回归的损失以及优化

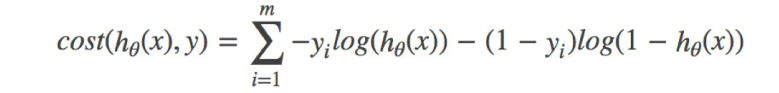
逻辑回归的损失，我们称之为对数似然损失，公式如下：

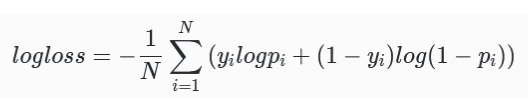


其中y为真实值，hΘ(x)为预测值。

无论何时，我们都希望损失函数值越小越好。分情况讨论，对应的损失函数值，当y=1时，我们希望hΘ(x)的值越大越好，当y=0时，我们希望hΘ(x)的值越小越好。

综合完整的损失函数：





分类评估方法

分类评估中的指标有：精确率、召回率、F1-score。

在分类任务下，预测结果与真实结果之间存在四种不同的组合，构成混淆矩阵。一个样本通过模型，得到预测结果，这个预测结果可以是正例，也可以是假例。样本本身有一个标签值，这个标签值可以是正例，也可以是假例。当真实结果和预测结果都为正例时，样本为真正例TP；当真实结果为正例，而预测结果为假例，那么样本为伪反例，顾名思义，假的反例；当真实结果为假例，预测结果却为正例，那么样本为伪正例；当真实结果和预测结果都为假例时，样本为真反例。而这些都是为了下面打基础。



**N**

**P**

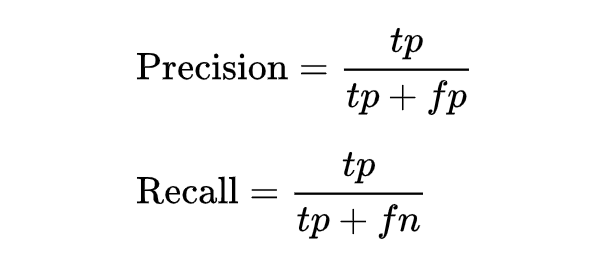
下面就是硬菜。

首先是**精确率(Precision)**，精确率是预测结果为正例样本中真实为正例的比例（预测出positive中，其中对了多少个），衡量模型查得准不准，计算公式是：精确率 = TP / (TP + FP)

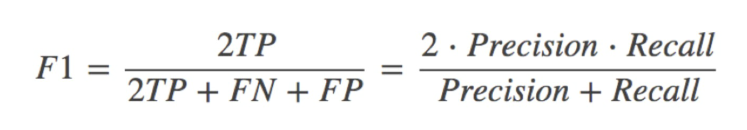


然后是**召回率(Recall)**，召回率是真实为正例的样本中预测结果为正例的比例（实际有多少个positive，预测对了多少个，即找到（recall）多少个），衡量模型查得全不全，评估模型对正样本的区分能力，计算公式为：召回率 = TP / (TP + FN)



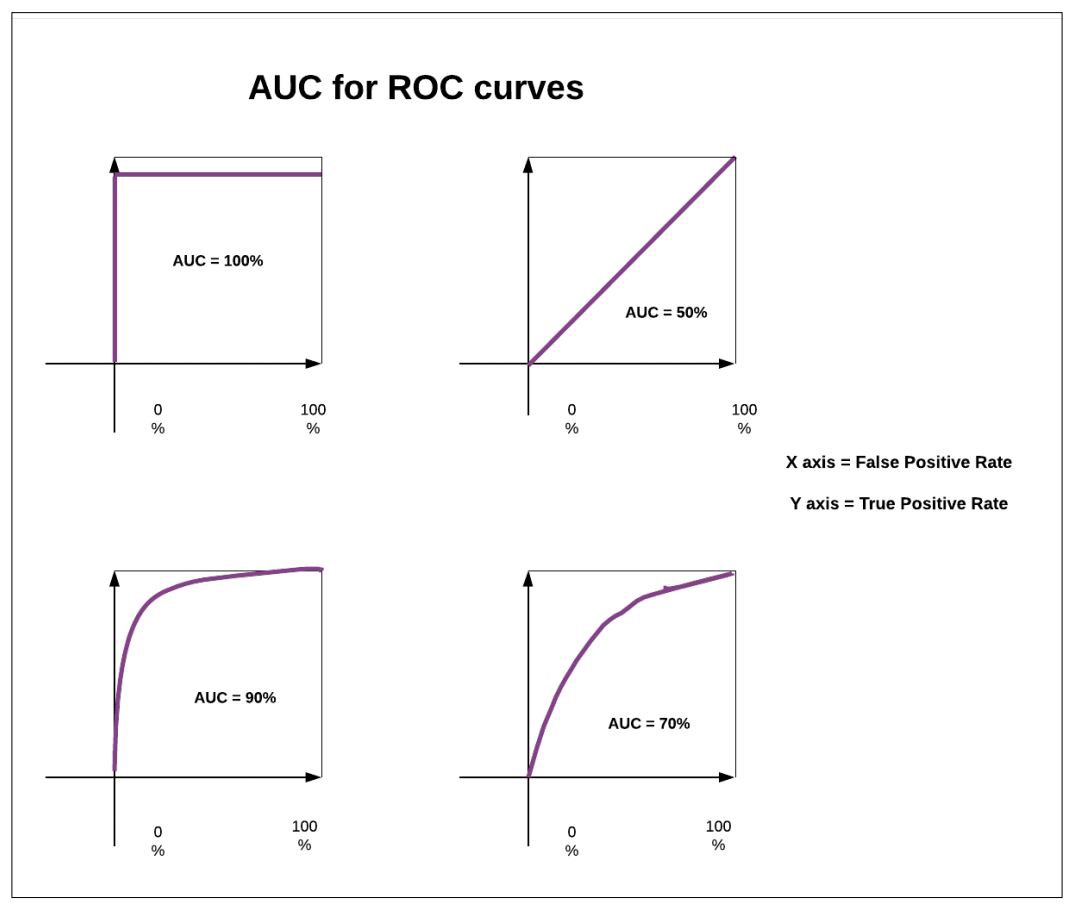


最后是**F1-score**，它综合反映了模型的稳健性，计算公式如下：



**AUC**：Area Under Curve）被定义为ROC曲线下与坐标轴围成的面积，显然这个面积的数值不会大于1。又由于ROC曲线一般都处于y=x这条直线的上方，所以AUC的取值范围在0.5和1之间。AUC越接近1.0，检测方法真实性越高；等于0.5时，则真实性最低，无应用价值。

**ROC（receiver operating characteristic curve）**接收者操作特征曲线。



ROC曲线是一个以FPR为横坐标，TPR为纵坐标的曲线。

TPR：真阳性率 ( true positive rate) ，又称敏感度(sensitivity)，TPR = TP / P = TP / (TP+FN)。

FPR：伪阳性率(false positive rate) ，FPR = FP / N = FP / (FP + TN)，其值为1-specificity。